Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Μάθηση Πολλαπλοτήτων και Μη-Γραμμικές Δυναμικές Επαναληψιμότητας για Αναγνώριση Συναισθήματος από Φωνή σε Ποικίλες Χρονικές Κλίμακες

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΥΘΥΜΙΟΣ ΤΖΙΝΗΣ

Επιβλέπων : Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2018

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Μάθηση Πολλαπλοτήτων και Μη-Γραμμικές Δυναμικές Επαναληψιμότητας για Αναγνώριση Συναισθήματος από Φωνή σε Ποικίλες Χρονικές Κλίμακες

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΥΘΥΜΙΟΣ ΤΖΙΝΗΣ

Επιβλέπων : Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 21η Ιουνίου 2018.

Αλέξανδρος Ποταμιάνος Αναπληρωτής Καθηγητής Ε.Μ.Π.	Πέτρος Μαραγκός Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2018

Ευθύμιος Τζίνης

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Στον καθηγητή μαθηματικών μου στο Λύκειο, Γιώργο Φραγκουλόπουλο,
ο οποίος ήταν ο πρώτος που πίστεψε στις δυνατότητές μου και με βοήθησε να τις αναδείξω,
παρά τη μη συμβατική και ιδιαίτερα χαοτική φύση της ύπαρξής μου.

Περίληψη

Στην εργασία αυτή [1] διερευνούμε την αναγνώριση συναισθημάτων ομιλίας (SER) ακολουθώντας τρεις διαφορετικές προσεγγίσεις που περιγράφονται παρακάτω. Για την αξιολόγηση κάθε προσέγγισης, χρησιμοποιούμε πολλαπλά σύνολα δεδομένων και πειραματικές μεθόδους που ακολουθούνται και από τη βιβλιογραφία. Επιπλέον, ακολουθούνται τόσο οι μέθοδοι ταξινόμησης που βασίζονται σε ολόκληρες προτάσεις όσο και σε τμήματα ομιλίας, όπου κάθε συναισθηματική έκφραση αντιπροσωπεύεται από ένα διάνυσμα στοιχείων και από έναν κατάλογο διανυσμάτων, αντίστοιχα.

Πρώτον, διερευνάμε την αποτελεσματικότητα των διαφόρων χρονικών κλιμάκων (παραθύρου, φωνήματος, λέξης ή πρότασης) για να αποφασίσουμε το συναισθηματικό περιεχόμενο μιας φράσης ομιλίας τόσο για Χαμηλού Επιπέδου Περιγραφητές (LLD) (τοπικά χαρακτηριστικά) όσο και για στατιστικά χαρακτηριστικά (υψηλού επιπέδου περιγραφητές). Συνδυάζοντας τα ανατροφοδοτούμενα νευρωνικά δίκτυα (RNN) και τις στατιστικά χαρακτηριστικά πάνω σε τμήματα που αντιστοιχούν περίπου στη διάρκεια μερικών λέξεων, αναφέρουμε τα καλύτερα αποτελέσματα στην βάση δεδομένων IEMOCAP. Προφανώς, η επιλογή της κατάλληλης χρονικής κλίμακας είναι μια πολύ σημαντική παράμετρος για να υλοποιήσουμε συστήματα SER υψηλής απόδοσης.

Επιπλέον, διερευνάται η απόδοση των χαρακτηριστικών που μπορούν να καταγράψουν τη δυναμική μη γραμμικής επαναληψιμότητας (μη-γραμμικές υποτροπιάζουσες δυναμικές συστημάτων) που ενσωματώνεται στο σήμα ομιλίας για το SER. Η ανακατασκευή του χώρου φάσης κάθε πλαισίου ομιλίας και ο υπολογισμός του αντίστοιχου γραφήματος επαναληψιμότητας (RP) αποκαλύπτει πολύπλοκες δομές που μπορούν να μετρηθούν με την εκτέλεση της ποσοτικής ανάλυσης επαναληψιμότητας (RQA). Αυτά τα μέτρα συγκεντρώνονται υπολογίζοντας τιμές στατιστικών συναρτήσεων ανά συγκεκριμένες χρονικές κλίμακες τμημάτων συναισθηματικής ομιλίας η ακόμη και ολόκληρης της έκφρασης. Αναφέρουμε τα αποτελέσματα SER για την προτεινόμενη προσέγγιση σε τρεις βάσεις δεδομένων χρησιμοποιώντας διαφορετικές μεθόδους ταξινόμησης. Όταν συνδυάζουμε τα προτεινόμενα χαρακτηριστικά με τα παραδοσιακά σύνολα χαρακτηριστικών, παρατηρούνμε μια βελτίωση της μη σταθμισμένης ακρίβειας μέχρι 5.7 % και 10.7 % για τα πειράματα SER Εξαρτημένου-Ομιλητή (SD) και Ανεξαρτήτως-Ομιλητή (SI), αντίστοιχα. Ακολουθώντας μια προσέγγιση που βασίζεται σε τμήματα, επιδεικνύουμε τις καλύτερες επιδόσεις στην βάση δεδομένων IEMOCAP χρησιμοποιώντας ένα αμφίδρομο RNN με βάση την προσοχή.

Τέλος, μειώνουμε τη διάσταση των ακουστικών χαρακτηριστικών που χρησιμοποιούνται για SER, χρησιμοποιώντας αλγόριθμους μάθησης πολλαπλότητας. Στην ουσία, παρουσιάζουμε έναν νέο αλγόριθμο για τη μη γραμμική μάθηση πολλαπλοτήτων για ποικίλες εφαρμογές, χρησιμοποιώντας τεχνικές βελτιστοποίησης χωρίς τον υπολογισμό της παραγώγου. Χρησιμοποιώντας την ενοποιημένη φόρμουλα των αλγορίθμων Γενικής Αναζήτησης Μοτίβων (General Pattern Search) (GPS) είμαστε σε θέση να παρέχουμε εγγυήσεις θεωρητικής σύγκλισης μέχρι τα στάσιμα σημεία πρώτης τάξης για τον προτεινόμενο αλγόριθμο. Επιπλέον, επιδεικνύουμε πρακτικές βελτιώσεις στον προτεινόμενο αλγόριθμο όσον αφορά την υπολογιστική αποδοτικότητα, το ρυθμό σύγκλισης και την ακρίβεια της λύσης σε διάφορες πειραματικές ρυθμίσεις. Τα αποτελέσματά μας υποδεικνύουν ότι ο αλγόριθμος μας είναι σε θέση να βρει λύσεις στο γενικό πρόβλημα της πολυδιάστατης κλιμάκωσης (MDS) κάτω από πολλαπλές ρυθμίσεις. Σύμφωνα με την εστίασή μας στην αναγνώριση συναισθήματος από φωνή, αξιολογούμε το Pattern Search MDS όπως περιγράφεται παρακάτω. Κάθε συναισθηματική φράση αντιπροσωπεύεται από ένα διάνυσμα χαρακτηριστικών που βρίσκεται σε ένα ευκλείδιο χώρο μεγά-

[1] Κατα την διάρκεια εκπόνησης αυτής της διπλωματικής συνεγράφησαν τα κατόπιν άρθρα: [1], [2], [3], [4].

λης διάστασης. Προκειμένου να μειωθεί η διάσταση αυτών των συναισθηματικών χαρακτηριστικών, προσπαθούμε να προσεγγίσουμε μια εμβυθισμένη χαμηλής διαστάσεως πολλαπλότητα στην οποία διατηρούνται επίσης οι αρχικές αποστάσεις ανά ζευγάρι διανυσμάτων. Δείχνουμε ότι μπορεί να επιτευχθεί σημαντική μείωση όσον αφορά τη διαστασιμότητα των δεδομένων εισόδου και τον χρόνο εκπαίδευσης, διατηρώντας ταυτόχρονα την ακρίβεια του SER σε ανταγωνιστικό επίπεδο.

Λέξεις κλειδιά

Μη γραμμικές υποτροπιάζουσες δυναμικές, δυναμική επαναληψιμότητας, βαθιά μάθηση, μηχανική μάθηση, αναγνώριση συναισθημάτων από φωνή, γράφημα επαναληψιμότητας, χώρος φάσης, πολυδιάστατη κλιμάκωση, μάθηση πολλαπλοτήτων

Abstract

In this work [2] we investigate Speech Emotion Recognition (SER) by following three different approaches which are outlined below. For the evaluation of each approach, we use multiple datasets and experimental setups which are also followed by the literature. Moreover, both utterance-based and segment-based classification methods are followed where each emotional utterance is represented by one feature vector and a list of vectors, respectively.

First, we explore the efficacy of various time-scales (frame, phoneme, word or utterance) for deciding the emotional content of a speech utterance for both Low Level Descriptors (LLDs) (local features) and statistical functionals (global features). By combining Recurrent Neural Networks (RNNs) and statistical functionals over segments that roughly correspond to the duration of a couple of words, we report state-of-the-art results on IEMOCAP. Purportedly, choosing the appropriate time-scale is key for high performing SER systems.

In addition, we investigate the performance of features that can capture nonlinear recurrence dynamics embedded in the speech signal for SER. Reconstruction of the phase space of each speech frame and the computation of its respective Recurrence Plot (RP) reveals complex structures which can be measured by performing Recurrence Quantification Analysis (RQA). These measures are aggregated by using statistical functionals over segment and utterance periods. We report SER results for the proposed feature set on three databases using different classification methods. When fusing the proposed features with traditional feature sets, we show an improvement in unweighted accuracy of up to 5.7% and 10.7% on Speaker-Dependent (SD) and Speaker-Independent (SI) SER tasks, respectively, over the baseline feature set. Following a segment-based approach we demonstrate state-of-the-art performance on IEMOCAP using an attention-based Bidirectional RNN.

Finally, we reduce the dimensionality of acoustic features used for SER by using manifold learning algorithms. In essence, we present a novel algorithm for nonlinear manifold learning using derivative-free optimization techniques, namely, Pattern Search MDS. By using General Pattern Search (GPS) formulation we are able to provide theoretical convergence guarantees up to first order stationary points for the proposed algorithm. Moreover, we demonstrate practical improvements of the proposed algorithm in terms of computational efficiency, convergence rate and solution accuracy on various experimental setups. Our results suggest that our algorithm is capable of finding solutions to the general problem of multidimensional scaling (MDS) under multiple setups. In accordance with our focus on SER, we evaluate Pattern Search MDS as briefly discussed next. Each emotional utterance is represented by a feature vector lying in a high dimensional space. In order to reduce the dimensionality of these emotional feature vectors, we try to approximate an underlying low-dimensional manifold in which the initial pairwise distances are also preserved. We show that a significant reduction in terms of input dimensionality and training time can be achieved by simultaneously maintaining SER accuracy at a competitive level.

Key words

Nonlinear recurrence dynamics, deep learning, machine learning, speech emotion recognition, recurrence plot, phase space, multidimensional scaling, manifold learning

[2] Papers: [1], [2], [3], [4] have been conducted under the development of this thesis.

Ευχαριστίες

Πρώτα απ όλα, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στον επιβλέποντα καθηγητή Αλέξανδρο Ποταμιάνο, όχι μόνο για την ανεκτίμητη καθοδήγησή του κατά τη διεξαγωγή της διπλωματικής μου, αλλά και για τα κίνητρά που μου έδωσε κατά την διάρκεια εκπόνησής της με τέτοιο τρόπο που συνεχώς διέγειρε τις ερευνητικές μου ιδέες και με ωθούσε προς τα εμπρός ούτως ώστε να με κάνει να δουλεύω στο μέγιστο των δυνατοτήτων μου.

Επιπροσθέτως, θα ήθελα να ευχαριστήσω τους καθηγητές Πέτρο Μαραγκό και Γιώργο Σταμού που συμμετέχουν στην επιτροπή της διπλωματικής μου καθώς και τα μαθήματα τους στο Εθνικό Μετσόβιο Πολυτεχνείο (ΕΜΠ) που με ενέπνευσαν να εμβαθύνω στη μελέτη των δυναμικών συστημάτων και των νευρωνικών δικτύων.

Ευχαριστώ τον συνάδελφο και στενό μου φίλο Γιώργο Παρασκευόπουλο για την υψηλή ποιότητα έρευνας που δημιουργήσαμε, τις διορατικές συζητήσεις που ανταλλάξαμε αλλά και τις άγρυπνες νύχτες που συμβιώσαμε. Αλλά πάνω από όλα, θα ήθελα να τον ευχαριστήσω καθώς και όλους τους άλλους ανθρώπους της ομάδας, για τον ευχάριστο χρόνο που περάσαμε μαζί κατά το παρελθόν έτος.

Η αιώνια εκτίμησή μου πηγαίνει στους γονείς μου για την ανιδιοτελή αγάπη και υποστήριξη τους σε όλα αυτά τα χρόνια της ακαδημαϊκής μου ανάπτυξης. Συγκεκριμένα, θα ήθελα να τους ευχαριστήσω διότι μου μετέδωσαν τις ανεκτίμητες αρετές της επιμονής και της υπευθυνότητας που θα με συνοδεύουν για πάντα.

Τέλος, θα ήθελα να ευχαριστήσω όλους τους φίλους μου και όλους τους ανθρώπους που υπομονετικά έμειναν μαζί μου στις πιο δύσκολες μέρες της ζωής μου. Δεν θα ξεχάσω ποτέ την απεριόριστη ενθάρρυνσή σας όταν μοιραζόμουν μαζί μου τα όνειρα και τις φιλοδοξίες μου.

Ευθύμιος Τζίνης,

Αθήνα, 21η Ιουνίου 2018

Περιεχόμενα

Κατάλογος πινάκων

Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

1.1 Η αντίληψη των συναισθημάτων

Τα συναισθήματα διαδραματίζουν καθοριστικό ρόλο στην αλληλεπίδραση επικοινωνίας μεταξύ των ανθρώπων. Τα συναισθήματα αντικατοπτρίζουν την εσωτερική συναισθηματική κατάσταση του ομιλητή στην τρέχουσα στιγμή και με την ικανότητα να αντιληφθεί αυτές τις πληροφορίες, ο αντίστοιχος ομιλητής που συμμετέχει στον διάλογο είναι σε θέση να παρέχει περισσότερες εμπαθείς απαντήσεις [8]. Τα συναισθήματα εκφράζονται χρησιμοποιώντας πρωτίστως εκφράσεις του προσώπου, κινήσεις του σώματος και μέσω της ομιλίας. Για παράδειγμα, όταν κάοιος συμμετέχει σε ένα διάλογο με ένα άτομο που κλαίει, τότε ερμηνεύοντας το κοινωνικό μήνυμα ενός λυπημένου προσώπου με δάκρυα και την τρεμάμενη φωνή, μπορεί κανείς να ακολουθήσει μια καταπραϋντική προσέγγιση για να επικοινωνήσει το μήνυμά του.

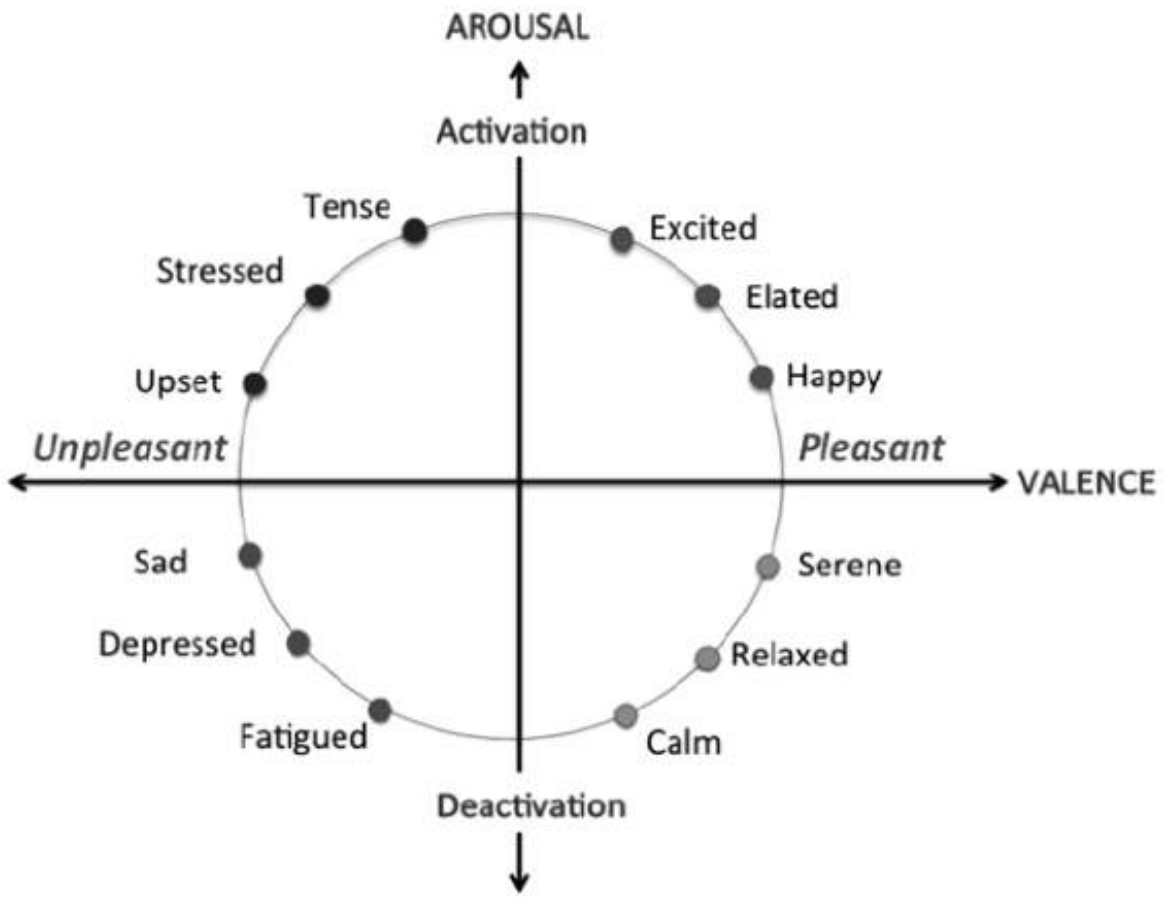
Μερικά από τα πιο σημαντικά συναισθήματα είναι ο θυμός, ο φόβος, η απέχθεια, η ευτυχία, η θλίψη και η έκπληξη που περιγράφουν εύστοχα το μοντέλο συναισθημάτων “Big Six” που εισήγαγε ο Ekmann [9]. η διάκριση των συναισθημάτων αναλύεται σε δύο άξονες του σθένους (valence) και της ενεργοποίησης (activation) [10] Σε αυτό το μοντέλο, κάθε συναίσθημα χαρτογραφείται σε μια συγκεκριμένη περιοχή του δισδιάστατου χώρου που παράγεται από τους δύο προαναφερθέντες άξονες όπως μπορεί το ίδιο ισχύει και για άλλα μοντέλα που δημιουργούν έναν τρισδιάστατο χώρο, χρησιμοποιώντας επίσης έναν άξονα απόκρισης για την κατάλληλη κωδικοποίηση του συναισθηματικού χάρτη [11].

Η πλειοψηφία αυτών των μοντέλων προέρχεται από βιολογική ή ψυχολογική άποψη για να περιγράψει τα συναισθήματα και τις ανθρώπινες συμπεριφορές. Μια ερώτηση που προκαλεί το ερώτημα είναι πώς μπορούμε πραγματικά να επικυρώσουμε την αποτελεσματικότητα αυτών των μοντέλων ή ακόμη και να τα αξιοποιήσουμε σε πραγματικές περιπτώσεις και πώς αυτά τα μοντέλα επεκτείνονται υπό διάφορες μορφές πληροφορίας (εικόνα, ήχος, κείμενο κλπ.).

1.2 Συναίσθημα στα σήματα ομιλίας

Η ομιλία είναι ένα από τα πιο κοινά κανάλια έκφρασης συναισθημάτων και μερικές φορές είναι το πιο βολικό. Για παράδειγμα, σε μια συνομιλία μέσω τηλεφώνου ή με άλλα μέσα, όπου δεν μπορεί κανείς να δει την έκφραση του προσώπου του συνομιλητή ή να καταλάβει το συναίσθημα μέσω της χροιάς της φωνής. Τα τελευταία χρόνια, οι ερευνητές προσπάθησαν να βρουν συσχετισμούς μεταξύ των χαρακτηριστικών των φωνητικών σημάτων και των συναισθημάτων που μεταφέρονται. Σύμφωνα με αυτή τη μέθοδο επεξεργασίας συναισθηματικής ομιλίας έχουν προταθεί ποικίλες διαταγμένες σχέσεις μεταξύ ακουστικών χαρακτηριστικών και συναισθημάτων. Στην ουσία, αυτές οι σχέσεις είναι ποιοτικοί δείκτες του τρόπου με τον οποίο κάθε συναισθηματικό στοιχείο χαρτογραφείται στο εύρος των τιμών που παρέχεται από ένα χαρακτηριστικό ομιλίας.

Ορισμένες σχέσεις είναι αρκετά διαισθητικές, όπως μια θυμωμένη φράση πιθανότατα θα αποδώσει υψηλότερο ενεργειακό περιεχόμενο από μια λυπημένη φράση. Φυσικά υπάρχει μια αναμφισβήτητη διαφοροποίηση μεταξύ διαφορετικών ομιλητών αλλά χωρίς απώλεια στη γενίκευση μπορούμε να καταλήξουμε στο συμπέρασμα ότι γενικά τα συναισθήματα που κατατάσσονται υψηλότερα στον



Σχήμα 1.1: Διδιάστατος χάρτης συναισθημάτων με ενεργοποίηση και σθένος [5]

άξονα σθένους / ενεργοποίησης θα κατατάσσονται επίσης υψηλότερα σε επίπεδο ενεργειακού περιεχομένου [12]. Σε αυτό το πλαίσιο, το SER μπορεί να θεωρηθεί ως ένα πιο δύσκολο έργο και αντ ’αυτού μια πιο εύκολη περίπτωση προσπάθησε να επιλυθεί, η οποία είναι συνεχής αναγνώριση συναισθημάτων. Βάσει των προηγούμενων δηλώσεων σχετικά με τους άξονες της διέγερσης και του σθένος μπορούμε να υποθέσουμε ότι κάθε συναίσθημα θα μπορούσε να αντιπροσωπεύεται εύστοχα στο 2D επίπεδο αυτών των συνεχών αξόνων. Έτσι, θα μπορούσαμε να προβλέψουμε για κάθε συναισθηματική φράση το αντίστοιχο σκορ σε κάθε έναν από αυτούς τους άξονες και μετά θα υποθέσουμε ότι θα μπορούσαμε να αναθέσουμε το αντίστοιχο διακριτό συναίσθημα. Επιπλέον, η θεμελιώδης συχνότητα της φωνητικής οδού ή του pitch όπως λέγεται είναι επίσης ένας καλός διαχωρισμός μεταξύ ορισμένων βασικών συναισθημάτων. Ουσιαστικά, το pitch αντιστοιχεί στη συχνότητα των ταλαντώσεων που παράγει η φωνητική πτυχή κάτω από την παραγωγή ομιλίας και έτσι ένας υψηλότερος τόνος στη φωνή σχετίζεται άμεσα με την τιμή στιγμιαίου pitch. Τα σήματα ομιλίας που περιέχουν φόβο ή θυμό τείνουν να παρουσιάζουν πολύ υψηλές τιμές ανύψωσης, ενώ τα σήματα που περιέχουν τη θλίψη ή την αηδία χαρτογραφούνται σε χαμηλότερες τιμές της θεμελιώδους συχνότητας [13]. Αν και αυτές οι αντιστοιχίσεις δεν μπορούν να αποτελέσουν ένα αντικειμενικό κριτήριο για τη διάκριση συναισθημάτων, ήταν τα πρώτα φώτα στο ταξίδι της αναγνώρισης των συναισθημάτων λόγω της εύκολης ερμηνείας τους.

Ωστόσο, τα συναισθήματα συσχετίζονται επίσης με συγκεκριμένες περιοχές τιμών από ακουστικά χαρακτηριστικά που δεν είναι τόσο κοντά στην διαίσθησή μας όπως αυτά που περιγράφηκαν προηγουμένως. Για παράδειγμα, στο [14] διεξάγεται διεξοδική και εξαντλητική μελέτη για την εύρεση συσχετισμών με μια ποικιλία ακουστικών χαρακτηριστικών και βασικών συναισθημάτων. Για παράδειγμα, η υψηλότερη συχνότητα ομιλίας μπορεί να υποδηλώνει μια έκφραση συναισθημάτων θυμού, φόβου ή αηδιασμού ενώ αντίθετα οι χαμηλότερες τιμές μπορεί να αντιστοιχούν στη θλίψη ή

την ευτυχία. Οι απότομες αλλαγές στις τιμές του βήματος με την πάροδο του χρόνου μπορεί να αντιπροσωπεύουν άγχος ή οργή σε σύγκριση με τις ομαλές μεταβάσεις που είναι πιο πιθανό να βρεθούν σε εκφράσεις ευτυχίας ή φόβου.

Για να αναλύσουμε αντικειμενικά το συναισθηματικό περιεχόμενο των σημάτων ομιλίας δεν μπορούμε να χρησιμοποιήσουμε αυτές τις καλά μελετημένες αλλά πάνω από όλα ποιοτικές σχέσεις μεταξύ ακουστικών χαρακτηριστικών και υποκείμενων συναισθημάτων. Σε μια τέτοια αόριστα καθορισμένη περιοχή όπως η αντίληψη των συναισθημάτων, οι ποιοτικοί διακριτικοί παράγοντες συχνά υποβάλλονται σε συζήτηση και δεν μπορούν να προσφέρουν λύση σε πραγματικά σενάρια όπου θα πρέπει να συναχθεί η συναισθηματική κατάσταση ενός ομιλητή. Πρέπει να είμαστε σε θέση να κωδικοποιούμε τις συναισθηματικές πληροφορίες από την ομιλία με τέτοιο τρόπο ώστε να διευκολύνουμε την αναγνώριση συναισθημάτων σε πραγματικό χρόνο και να γενικεύουμε σε πολλαπλούς ομιλητές. Σε αυτό το πλαίσιο, η ποσοτική ανάλυση των συναισθημάτων στα σήματα ομιλίας φαίνεται να είναι περισσότερο αναγκαιότητα παρά αντικείμενο μίας φιλοσοφικής διαμάχης.

1.3 Αυτόματη αναγνώριση συναίσθηματων από την ομιλία

Σε αυτή την ενότητα παρέχουμε ένα αρχικό κίνητρο από τα θεωρητικά μοντέλα της συναισθηματικής αντίληψης σε ποσοτικοποιημένες μεθόδους συναισθηματικής υπολογιστικής και κάποιες εφαρμογές της τελευταίας στην ενότητα 1.3.1. Επιπλέον, συζητάμε μερικά από τα πιο ευρέως χρησιμοποιούμενα σύνολα ακουστικών χαρακτηριστικών και μεθόδων ταξινόμησης που προτείνονται στη βιβλιογραφία στις ενότητες 1.3.2 και 1.3.3, αντίστοιχα. Με τον όρο μέθοδο εδώ αναφέρουμε τον συνδυασμό της στρατηγικής εξαγωγής χαρακτηριστικών παράλληλα με τον συνδυασμό της με τη χρήση ενός ταξινομητή που χρησιμοποιείται για την πρόβλεψη της αντίστοιχης συναισθηματικής ετικέτας για τη διεξαγωγή της δοκιμασίας. Τέλος, στην ενότητα 1.3.4, περιγράφουμε σύντομα μερικές από τις τελευταίες τεχνολογικές προσεγγίσεις για την αναγνώριση συναισθήματος ομιλίας (SER) που παρουσιάζονται στη βιβλιογραφία. Από εδώ και πέρα, όταν χρησιμοποιούμε τον όρο SER, αναφερόμαστε στην αυτόματη αναγνώριση συναισθημάτων ομιλίας που εκτελείται σε ένα υπολογιστικό σύστημα αντί να χρησιμοποιεί ένα θεωρητικό πλαίσιο για την ερμηνεία του συναισθηματικού περιεχομένου.

1.3.1 Κίνητρο

Η εκτεταμένη ανάλυση της ανθρώπινης αντίληψης και της δημιουργίας συναισθημάτων έδωσε ουσιαστική διαίσθηση για την εφαρμογή τεχνητών μοντέλων που είναι σε θέση να συλλάβουν τα σήματα συμπεριφοράς [15]. Η οικοδόμηση συναισθηματικά συνειδητοποιημένων διεπαφών ανθρώπου-μηχανής (HMI) τελικά βασίζεται στην κατανόηση της υποκείμενης δυναμικής των συναισθηματικών συνθηκών και στην ενσωμάτωση της γνώσης σε ένα HMI. Συνεπώς, ο στόχος της αυτόνομης αναγνώρισης συναίσθημα ομιλίας (SER) είναι να οικοδομήσουμε μια μηχανή ικανή να αλληλεπιδρά με τους ανθρώπους και να ερμηνεύει και να χρησιμοποιεί με προσοχή τα συναισθηματικά σήματα από την ομιλία. Όλα αυτά τα μοντέλα ανήκουν στην ευρύτερη κατηγορία των μοντέλων τεχνητής νοημοσύνης (AI). Το τελευταίο προέρχεται από ένα πολυτομεακό πλαίσιο όπως η νευροεπιστήμη, η φυσική, η βιολογία, η ψυχολογία και, τέλος, η επιστήμη των υπολογιστών. Σε αυτή τη μελέτη ακολουθούμε μια υπολογιστική προσέγγιση, αλλά δεν πρέπει να αγνοούμε τη σημαντική συμβολή άλλων κλάδων για να μπορέσουμε να φτιάξουμε μοντέλα SER.

Με βάση τη μετάβαση από την αφηρημένη φύση των συναισθημάτων σε μια ποσοτική προσέγγιση όπου όλες οι πληροφορίες θα πρέπει να κωδικοποιούνται σε ένα διάνυσμα μηδενικών και άσων, πρέπει να θυμηθούμε κάποιες από τις πιθανές χρήσεις ενός συστήματος SER. Για παράδειγμα, το SER είναι το κλειδί για την κατασκευή ευφυών προσαρμοστικών HMI στην συναισθηματική κατάσταση του χρήστη, ειδικά σε περιπτώσεις όπως τηλεφωνικά κέντρα όπου δεν υπάρχει άλλη πληροφορία. [16]. Σε γενικές γραμμές, τα συστήματα αυτόματου διαλόγου, τα οποία προσαρμόζονται στην εσωτερική κατάσταση του χρήστη, είναι ζωτικής σημασίας στοιχεία για ένα τεράστιο όγκο εφαρμογών, συμπεριλαμβανομένων των παραγόντων επιτήρησης και διερμηνείας [17]. Προχωρώντας ένα βήμα

παραπέρα, η καταγραφή της συναισθηματικής κατάστασης ενός ομιλητή θα μπορούσε να χρησιμοποιηθεί για προβλέψεις ακόμη υψηλότερου επιπέδου, όπως η εμπλοκή ενός χρήστη σε μια συγκεκριμένη αλληλεπίδραση διαλόγου.

Συγκεκριμένα, στο [4] μελετήσαμε έναν αυτόματο ταξινομητή για να αποφασίσουμε αν ένα παιδί πάσχει από διαταραχή του φάσματος του αυτισμού (ASD). Έχουμε ενσωματώσει ένα καινοτόμο ψυχολογικό μοντέλο για τον προσδιορισμό του πραγματικού επιπέδου εμπλοκής για κάθε έκφραση της αλληλεπίδρασης γονέα-παιδιού. Πραγματοποιήσαμε ένα δυαδικό πείραμα ταξινόμησης χρησιμοποιώντας έναν ταξινομητή με διανύσματα υποστήριξης (SVM) εκπαιδευμένο σε ακουστικά, γλωσσικά και διαλογικά χαρακτηριστικά που προέρχονται από το σύνολο δεδομένων που έχουμε εισαγάγει με εγγραφές βίντεο. Δείξαμε ότι τα παιδιά με τυπική ανάπτυξη (TD) παρουσίασαν ένα πιο ντετερμινιστικό μοντέλο συμπεριφοράς απ' ότι τα παιδιά της ASD, ενώ τα χαρακτηριστικά που σχετίζονται με το βίντεο των γονέων βρέθηκαν ότι είναι τα καλύτερα προγνωστικά για το επίπεδο εμπλοκής του παιδιού. Το άρθρο που γράψαμε παρέχει βαθύτερη γνώση στις κοινωνικές και γνωστικές δομές των παιδιών με ASD. Σε γενικές γραμμές, οι προαναφερθείσες μελέτες εμφανίζουν μόνο ένα μικρό κλάσμα από αυτό που η συναισθηματική πληροφορική θα μπορούσε να δώσει πίσω στην ανθρωπότητα [18].

1.3.2 Ακουστικά χαρακτηριστικά

Ένα από τα πιο απαιτητικά προβλήματα του SER είναι η εύρεση ενός αντιπροσωπευτικού συνόλου χαρακτηριστικών συναισθημάτων και η βέλτιστη χρονική κλίμακα για την εξάσκηση συναισθηματικών πλαισίων. Πληροφορία προσοδική, φασματική και ποιότητα φωνής συντελούν Χαρακτηριστικά Χαμηλού Επιπέδου (LLD), που εξάγονται από τα πλαίσια ομιλίας και έχουν χρησιμοποιηθεί εκτενώς για το SER [19]. Ορισμένες από τις πιο ευρέως χρησιμοποιούμενες LLD ή τοπικά χαρακτηριστικά είναι οι συντελεστές συχνότητας Cepstral (MFCC), η θεμελιώδης συχνότητα, η βραχυχρόνια ενέργεια από τα πλαίσια, η ταχύτητα μηδενικής διέλευσης (ZCR) και η αναλογία αρμονικών προς θόρυβο (HNR). Όλα αυτά τα LLD εξάγονται απευθείας από τα πλαίσια ομιλίας που αντιστοιχούν σε παράθυρα $20 - 100$ ms. Έτσι, για να μπορέσουμε να παρέχουμε μια χρονική συνάθροιση των συναισθηματικών πληροφοριών για μεγαλύτερες περιόδους ομιλίας από τα πλαίσια, χρησιμοποιούνται παγκόσμια χαρακτηριστικά. Το τελευταίο σύνολο χαρακτηριστικών υπολογίζεται ως στατιστικά σε σχέση με τα τοπικά χαρακτηριστικά για τη δεδομένη ομιλία ή τμήματα αυτής [20].

Κατά τα τελευταία χρόνια, μια ποικιλία LLDs, που εξάγονται σε επίπεδο πλαισίου και αργότερα συγκεντρώνονται χρησιμοποιώντας στατιστικές ή έχουν μοντελοποιηθεί απευθείας ως ακολουθία πλαισίων, έχει προταθεί στη βιβλιογραφία ως ικανή να καταγράψει το συναισθηματικό περιεχόμενο [19]. Ο κύριος όγκος της εργασίας στην εξαγωγή ακουστικών χαρακτηριστικών περιστρέφεται γύρω από την παραγλωσσιακή ανάλυση των σημάτων ομιλίας που προέρχονται πρωτίστως από τις προσοδικές και φασματικές αναπαραστάσεις [21]. Ορισμένες άλλες λειτουργίες όπως το jitter και το shimmer έχουν επίσης αποδειχθεί ότι παρέχουν πρόσθετες πληροφορίες για την ταξινόμηση των μορφών ανθρώπινων ομιλητών και των επιπέδων διέγερσης [22] και χρησιμοποιούνται αργότερα για το SER [23]. Όλα αυτά τα χειροποίητα χαρακτηριστικά έχουν ελεγχθεί διεξοδικά για τις παραγλωσσιακές τους ιδιότητες και την αποτελεσματικότητά τους στην καταγραφή συναισθηματικού περιεχομένου από την ομιλία. Μία υποδειγματική βάση των πιο ουσιαστικών προσεγγίσεων που υπάρχουν στον τομέα της παραγλωσσιακής ανάλυσης δίνεται από τους Schuller και Batliner στο [24].

Ορισμένα σύνολα χαρακτηριστικών, που περιέχουν τους περισσότερους από τους προαναφερθέντες περιγραφικούς δείκτες καθώς και ένα σύνολο στατιστικών συναρτήσεων που εφαρμόζονται πάνω σε αυτά, έχουν αποδώσει εντυπωσιακές ακριβείς ταξινομήσεις για διάφορα πειράματα SER. Ένα από τα πρώτα συμπαγή σύνολα χαρακτηριστικών που παρουσιάστηκαν ήταν ένα διάνυσμα ακουστικών χαρακτηριστικών μήκους 384 στο Paralinguistic Challenge του Interspeech 2009 [25] (σύνολο χαρακτηριστικών IS09). Αυτό το σύνολο αποτελείται από τα LLD που αναφέρθηκαν προηγουμένως σε αυτήν την ενότητα και είχαν ως αποτέλεσμα την ακρίβεια έως και 70.1% και 65.1% για πειράματα με 2 και 5 κλάσεις για την ταξινόμηση συναισθημάτων στο Fau-Aibo σύνολο δεδομένων [26]. Ένα πολύ μεγαλύτερο σύνολο χαρακτηριστικών μήκους 1582 έχει εισαχθεί στο Interspeech2010 (σύνολο χαρα-

κτηριστικών IS10) από τον Schuller[23], όπου δοκιμάστηκε για αναγνώριση ηλικίας, φύλου και το επίπεδο ενδιαφέροντος. Σε αυτό το σύνολο, μερικοί επιπλέον περιγραφείς όπως το jitter, το shimmer, η πιθανότητα να υπάρχει φωνή, τα φασματικά ζεύγη γραμμών και οι λογαριθμικές τράπεζες Συχνοτήτων Μελ έχουν προστεθεί στα LLDs της προηγούμενης εργασίας. Παραδόξως, αυτό το σετ εξακολουθεί να χρησιμοποιείται και αποδίδει κορυφαία αποτελέσματα ακόμη και στις σημερινές προσεγγίσεις [2], [27],[28]. Για το σκοπό αυτό, παρουσιάστηκε ένα πολύ ευρύτερο σύνολο χαρακτηριστικών ομιλίας σε μια επακόλουθη παράλληλη πρόκληση του Interspeech2013 [29], παρουσιάζοντας ένα διάνυσμα χαρακτηριστικών (IS13 Feature Set) μήκους 6373 για κάθε ομιλία. Το προτεινόμενο σύνολο χαρακτηριστικών έχει αξιολογηθεί με την αναγνώριση των συναισθημάτων, του αυτισμού και των κοινωνικών σημάτων από τα φωνητικά συνθήματα.

Επιπλέον, μερικά έργα συνδυάζουν διαφορετικά σύνολα χαρακτηριστικών για την ταξινόμηση των συναισθημάτων συνδυάζοντας διαφορετικούς τρόπους ή σύνολα χαρακτηριστικών που εξάγονται από διαφορετικό τομέα. Για παράδειγμα, στο [30] χρησιμοποιήθηκαν τρεις διαφορετικοί τρόποι ενημέρωσης για την εξαγωγή χαρακτηριστικών, δηλαδή για κείμενο, ήχο και εικόνα. Η τελική αναπαράσταση συναισθηματικής ομιλίας συνίσταται στη συνένωση των τριών διανυσμάτων χαρακτηριστικών που χρησιμεύει ως είσοδος για τον ταξινομητή. Ειδικότερα, η πολυτροπική προσέγγιση της επεξεργασίας εικόνας δεν σχετίζεται άμεσα με το SER, καθώς περιλαμβάνει ένα κανάλι πληροφοριών το οποίο δεν μπορεί να ληφθεί αποκλειστικά από τον ήχο, αλλά εμπίπτει στο γενικό πρόβλημα της αυτόματης αναγνώρισης συναισθημάτων. Επιπλέον, οι γλωσσικές πληροφορίες θα μπορούσαν να ληφθούν από τον ήχο με την οικοδόμηση του ακουστικού μοντέλου και των γλωσσικών μοντέλων για την εκτέλεση της αυτόματης αναγνώρισης ομιλίας (ASR) [6]. Σε αυτό το πλαίσιο, τα ακουστικά χαρακτηριστικά έχουν συνδυαστεί με γλωσσικά χαρακτηριστικά για τη βελτίωση της απόδοσης των συστημάτων SER [31]. Η συσχέτιση αυτών των διανυσμάτων χαρακτηριστικών ονομάζεται συχνά ”πρώιμη σύντηξη” ή μόνο ”σύντηξη”.

Άλλα έργα χρησιμοποιούν πολύ πιο συμπαγή σύνολα χαρακτηριστικών για SER. Στο [32] χρησιμοποιείται ένα υποσύνολο του προαναφερθέντος συνόλου χαρακτηριστικών IS10 [23] για την προσαρμογή τομέα στις εργασίες SER. Επιπλέον, λόγω της τεράστιας διαστασιμότητας των προτεινόμενων συνόλων χαρακτηριστικών, πολλές προσεγγίσεις περιλαμβάνουν μια τεχνική μείωσης των διαστάσεων προκειμένου να ελαττωθεί το υπολογιστικό φορτίο στη διαδικασία εκπαίδευσης των μοντέλων. Συγκεκριμένα, υπενθυμίζουμε ότι σε μερικά έργα χρησιμοποιούνται ορισμένες τεχνικές γραμμικής μείωσης διαστάσεων όπως η Ανάλυση σε κυρίαρχε Συνιστώσες (PCA) [33]. Στην μεταγενέστερη εργασία, ο PCA εκτελείται πάνω σε ένα διάνυσμα χαρακτηριστικών με βάση την θεμελιώδη συχνότητα και την ενέργεια, προκειμένου να ελαχιστοποιηθεί η διαστασιοποίηση των διανυσμάτων εισόδου, αλλά και να ενισχυθεί η τελική ακρίβεια ταξινόμησης. Επιπλέον, διάφορες τεχνικές επιλογής χαρακτηριστικών γνώρισαν επίσης άνθηση υπό την τεράστια διαστασιμότητα των προτεινόμενων συνόλων χαρακτηριστικών για το SER. Για παράδειγμα, στο [34] Ensemble Random Forest to Trees (ERFTrees) έχουν εισαχθεί για την επιλογή των πιο αντιπροσωπευτικών χαρακτηριστικών καθώς και τεχνικές μη γραμμικής μείωσης διαστάσεων για τη μείωση του χώρου εισόδου των διανυσμάτων που χρησιμοποιούνται για SER.

Στο ίδιο πλαίσιο της ελάχιστης επεξεργασίας του λόγου και της εξαγωγής χαρακτηριστικών, υπάρχει ανάγκη να μειωθεί ο συνολικός χρόνος της εξαγωγής χαρακτηριστικών και της προεπεξεργασίας δεδομένων γενικά. Για το σκοπό αυτό, στο [35], κάθε συναισθηματική φράση αντιπροσωπεύεται μόνο από το φασματογράφημά της και έτσι απαιτείται ελάχιστη επεξεργασία λόγου. Στην ουσία, ένας μετασχηματισμός Fourier εφαρμόζεται για όλη την διαδικασία της προετοιμασίας των δεδομένων και της εξαγωγής χαρακτηριστικών πριν έρθει μπροστά η τελική διαδικασία εκπαίδευσης και πρόβλεψης. Οι αναπαραστάσεις του φασματικού σπεκτρογράμματος έχουν επίσης χρησιμοποιηθεί από σήματα της γλωττίδας για την τελική ταξινόμηση των συναισθημάτων σε μια γενικότερη προσέγγιση μαθησιακής εκπροσώπησης [36]. Μια άλλη κοινή τεχνική εξαγωγής χαρακτηριστικών είναι η χρήση ενός αυτόματου κωδικοποιητή για την αυτόματη εξαγωγή χαρακτηριστικών γνωρισμάτων από φασματικές αναπαραστάσεις, όπως Mel Τράπεζες Φίλτρων (MFB) [27], φασματογραφήματα από ακατέργαστο σήμα [37].

Άλλες λιγότερο δημοφιλείς μέθοδοι εξαγωγής χαρακτηριστικών βασίζονται στο μετασχηματισμό του σήματος εισόδου με διαφορετική ανάλυση από το Fourier ή το Cepstrum. Για παράδειγμα, στο [38] χρησιμοποιείται μετασχηματισμός πακέτων wavelet για μεταγενέστερη εξαγωγή χαρακτηριστικών. Τα κύματα καθορίζονται από συντελεστές οι οποίοι είναι παράμετροι που πρέπει να ρυθμιστούν κατά τη διάρκεια της φάσης εκπαίδευσης για τη βελτιστοποίηση μιας συνάρτησης κόστους όπως η εντροπία μεταξύ συναισθηματικών κλάσεων. Οι συντελεστές wavelet είναι ικανοί να καταγράψουν τις πληροφορίες χαμηλών συχνοτήτων κοντά στις συχνότητες γενιάς ανθρώπινης ομιλίας [39]. Επιπλέον, στο [40] Teager Energy Operator χρησιμοποιείται για το SER που είναι ένας μη γραμμικός τελεστής που χρησιμοποιείται συχνά για την εκτίμηση της πραγματικής ενέργειας του φίλτρου πηγής στη διαδικασία διαμόρφωσης AM-FM [41]. Ορισμένα άλλα μη γραμμικά χαρακτηριστικά έχουν επίσης δοκιμαστεί σε προβλήματα SER που προέρχονται από το φάσμα των σημάτων ομιλίας [42] και την στιγμιαία φάση και εύρος του πλαισίου ομιλίας [43]. Τα μη γραμμικά χαρακτηριστικά για το SER είναι στην πραγματικότητα η κύρια έννοια του Κεφαλαίου 4. Συγκεκριμένα, στην ενότητα 4.2, συζητούμε λεπτομερώς άλλες προσεγγίσεις οι οποίες χρησιμοποιούν επίσης μη γραμμική ανάλυση σημάτων ομιλίας για το SER.

Πρόσφατα, εισήχθησαν σημαντικά λιγότερες χρονοβόρες προσεγγίσεις προεπεξεργασίας, όπου τα συμβατικά χαρακτηριστικά ομιλίας συμπληρώνονται από τις αναπαραστάσεις ανεπεξέργαστων σημάτων. Παραδόξως, στην [44] χρησιμοποιείται μόνο η ακατανόητη αναπαράσταση σήματος κάθε συναισθηματικής έκφρασης ως διανύσματος χαρακτηριστικών εισόδου και ταυτόχρονα έχουν αναφερθεί ανταγωνιστικά αποτελέσματα στη συνεχή αναγνώριση συναισθημάτων.

1.3.3 Μέθοδοι ταξινόμησης

Από τη μία πλευρά, η εξαγωγή χαρακτηριστικών ικανών να καταγράψουν τη συναισθηματική κατάσταση του ομιλητή είναι μια δύσκολη εργασία για τον SER, αλλά αυτό είναι μόνο ένα μέρος των παραμέτρων που πρέπει να προσδιοριστούν προκειμένου να καθοριστεί η ολιστική προσέγγιση που θα ακολουθηθεί για την τελική ταξινόμηση. Η επιλογή του κατάλληλου ταξινομητή, ο τρόπος με τον οποίο θα εκπαιδευτεί, η μεταφορά μαθησιακών και άλλων τεχνικών μεταξύ διαφορετικών μοντέλων και άλλων παραμέτρων είναι επίσης βασικά στοιχεία για την εκτέλεση του SER. Έτσι, ο χώρος των παραμέτρων που κάποιος πρέπει να ψάξει προτού καταλήξει στο συμπέρασμα για την προσέγγισή του γίνεται τεράστιος εάν αναγνωρίσουμε το γεγονός ότι νέα μοντέλα ικανά να μοντελοποιούν πλαίσια ομιλίας καθώς και άλλα μοντέλα τα οποία είναι έντονα συνδεδεμένα με μοντέλα από άλλους τομείς όπως το κείμενο και η εικόνα. Για παράδειγμα, τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs) πρωτοεμφανίστηκαν για την αναγνώριση εικόνας των χειρόγραφων ψηφίων [45] αλλά έχουν χρησιμοποιηθεί εκτενώς για SER χρησιμοποιώντας ως είσοδο το φασματογράφημα των συναισθηματικών προτάσεων [35]. Σε αυτή την ενότητα δεν επιδιώκουμε να παρουσιάσουμε εξαντλητικά όλες τις ιστορικές συνεισφορές άλλων ερευνητών στο SER, αλλά εστιάζουμε στην εμφάνιση μερικών από τις πιο σημαντικές από την άποψη της καινοτομίας. Στην επόμενη ενότητα (Τμήμα 1.3.4) θα καλύψουμε παρόμοιες πτυχές, αλλά από την άποψη της αποτελεσματικότητας και εστιάζοντας κυρίως στα ευρήματα που αναφέρουν τα καλύτερα ποσοστά επιτυχίας σε SER.

Οι προτεινόμενες προσεγγίσεις SER διαφέρουν κυρίως από τη συσσωμάτωση και τη χρονική μοντελοποίηση της ακολουθίας εισόδου των LLD ή με άλλα λόγια των τοπικών χαρακτηριστικών που αντιστοιχούν στις αναπαραστάσεις πλαισίου. Μπορούμε εύκολα να διαιρέσουμε τις προσεγγίσεις SER σε τρεις κύριες κατηγορίες. Συγκεκριμένα, 1) με βάση όλη την πρόταση ή ομιλία 2) με βάση το τμήμα ομιλίας και 3) άμεσες προσεγγίσεις που εξηγούνται κατωτέρω με περισσότερες λεπτομέρειες. Για κάθε κατηγορία παρουσιάζουμε μερικές από τις πιο καλές προσεγγίσεις ως προς την απόδοση στις εργασίες SER.

Προσεγγίσεις βασισμένες σε όλη την ομιλία: Σε αυτές τις προσεγγίσεις, οι συναισθηματικές δηλώσεις κατανέμονται αρχικά σε αλληλεπικαλυπτόμενα πλαίσια που είναι μια κοινή τεχνική επεξεργασίας ομιλίας για την εξαγωγή βραχυπρόθεσμων χαρακτηριστικών [6]. Για κάθε πλαίσιο εξάγεται ένα σύνολο LLD και στη συνέχεια συγκεντρώνονται σε όλο το μήκος έκφρασης εφαρμόζοντας στατιστικές λειτουργίες και κατά συνέπεια δημιουργώντας μια παγκόσμια στατιστική αναπαράσταση

για κάθε συναισθηματική φράση. Αυτή η τελική αναπαράσταση χρησιμοποιείται ως είσοδος για έναν ταξινομητή που εκπαιδεύεται σε ένα επιλεγμένο κλάσμα του συνόλου δεδομένων και ελέγχεται στις υπόλοιπες διαθέσιμες δηλώσεις. Οι προσεγγίσεις που βασίζονται στην κατανόηση είναι οι πιο απλές για εφαρμογή, αλλά σε συνδυασμό με τους ισχυρούς ταξινομητές παρουσιάζουν καλά αποτελέσματα για διάφορες εργασίες SER [20]. Για παράδειγμα, η χρήση μόνο στατιστικών σε επίπεδο γνώσεων από LLD που σχετίζονται με την θεμελιώδη συχνότητα και μοντέλα Μείγματος Γκαουσιανών (GMMs) για την τελική ταξινόμηση αποδίδει ακρίβεια αναγνώρισης έως και 77% [46]. Μια παρόμοια προσέγγιση είναι να χρησιμοποιήθει ένας ταξινομητής SVM με στατιστικά στοιχεία [47]. Το σετ χαρακτηριστικών IS09 δοκιμάστηκε σε μια ποικιλία από συναισθηματικές βάσεις δεδομένων ομιλίας με πολυεπίπεδο SVM για την τελική πρόβλεψη κάθε συναισθήματος [25]. Επιπλέον, το σύνολο χαρακτηριστικών IS10 έχει αξιολογηθεί με SER χρησιμοποιώντας πολλαπλά ζευγαρωτά SVMs για διακριτική ταξινόμηση συναισθημάτων [23]. Το ίδιο σύνολο χαρακτηριστικών έχει επίσης χρησιμοποιηθεί για την ενσωμάτωση των πληροφοριών φύλου στο προαναφερόμενο σύνολο χαρακτηριστικών IS10 χρησιμοποιώντας Αυτόματους Κωδικοποιητές με ακύρωση ήχου (DAEs) και στη συνέχεια ένα SVM εφαρμόζεται για το SER [48]. Επιπλέον, χρησιμοποιήθηκε ένα μοντέλο ταξινόμησης Ensemble Softmax Regression (ESR) με στατιστικά στοιχεία που βασίζονται σε ακρόαση παρόμοια με το σύνολο χαρακτηριστικών IS10 [49]. Τέλος, μια προσέγγιση διδασκαλίας πολλαπλών εργασιών, όπου ένα δίκτυο Νευρωνικών Βαθιάς Πίστης (DBN) εκπαιδεύεται ταυτόχρονα σε διακρητή SER και συνεχή αναγνώριση συναισθημάτων ως παλινδρόμηση στην διέγερση και το σθένος, χρησιμοποιώντας σαν είσοδο το ίδιο σύνολο χαρακτηριστικών IS10 [28].

Προσεγγίσεις βασισμένες σε τμήματα ομιλίας: Μια προσέγγιση αυτής της κατηγορίας είναι αρκετά παρόμοια με μια προσέγγιση βασισμένη σε όλη την ομιλία, παρά το γεγονός ότι τα στατιστικά στοιχεία δεν εφαρμόζονται άμεσα σε ολόκληρο το μήκος έκφρασης. Αντίθετα, τα πλαίσια της όλης ομιλίας ομαδοποιούνται σε τμήματα ομιλίας τα οποία είναι γενικά βραχύτερα σε διάρκεια από ότι η ίδια η ίδια η έκφραση. Σε αυτό το πλαίσιο, κάθε τμήμα ομιλίας αντιπροσωπεύεται από ένα σύνολο στατιστικών χαρακτηριστικών οι οποίες εφαρμόζονται σε LLD που εξάγονται από τα πλαίσια που ανήκουν στη συγκεκριμένη χρονική διάρκεια. Οι προσεγγίσεις με βάση το τμήμα έχουν δείξει ότι ο υπολογισμός των στατιστικών συναρτήσεων σε LLDs σε κατάλληλο χρονικό διάστημα αποδίδει σημαντική βελτίωση της απόδοσης για τα συστήματα SER [50], [2]. Ωστόσο, σε αυτές τις προσεγγίσεις η είσοδος κάθε συναισθηματικής έκφρασης είναι μια ακολουθία διανυσμάτων και έτσι συμβαίνει μια νέα πρόκληση στη διαδικασία εκπαίδευσης των μοντέλων. Η μοντελοποίηση της ακολουθίας εισόδου των διανυσμάτων χαρακτηριστικών δεν είναι μια τετριμμένη εργασία στο SER. Χρησιμοποιείται μια τεχνική πλειοψηφικής ψηφοφορίας στο [51], όπου η απόφαση της τελικής συναισθηματικής ετικέτας βασίζεται στις posterior πιθανότητες ενός ταξινομητή GMM, ο οποίος εκπαιδεύεται και δοκιμάζεται σε ένα επίπεδο υποπρότασης. Αντίθετα, αυτή η υπόθεση δεν ισχύει εν γένει επειδή μια οργισμένη έκφραση συνδέεται στενότερα με μια συμβολή που οδηγείται από γεγονότα και όχι με την προσδοκία να παρατηρήσουμε θυμωμένο περιεχόμενο σε όλα τα πλαίσια μιας δεδομένης έκφρασης. Στο [52], μια υβριδική υλοποίηση που περιλαμβάνει ένα απλό Τεχνητό Νευρωνικό Δίκτυο (ANN) από στατιστικά στοιχεία που βασίζονται σε όλη την ομιλία, καθώς και ένα Hidden Markov Model (HMM) προκειμένου να αποκτήσει πληροφορίες από τμήματα ομιλίας. Στο [53], η Provost εισήγαγε μια μεταβλητή προσέγγιση μήκους παραθύρου για την καταγραφή της συναισθηματικής ομιλίας σε κλίμακα υπό-πρότασης. Για κάθε παράθυρο εκτιμήθηκε η παρουσία ενός συναισθηματικού προφίλ (Emotion Profile (EP)) και η τελική ροή συναισθημάτων διαμορφώθηκε με μια τροχιά που χρησίμευσε ως είσοδος για ένα HMM. Η σύντηξη των EP και LLD σε ένα μοντέλο ιεραρχικής συσχέτισης με πολλαπλά στρώματα που αντιστοιχούν σε κάθε τύπο χαρακτηριστικού που προέρχεται από πολλαπλές χρονικές κλίμακες, χρησιμοποιήθηκε στο [54]. Συγκεκριμένα, αναπαραστάσεις χαρακτηριστικών που αφορούσαν όλη την πρόταση, τμήματα αυτής και παράθυρά της εξήχθησαν για κάθε συναισθηματική πρόταση και χρησιμοποιήθηκαν σε διαφορετικά στρώματα των ιεραρχικών μοντέλων. Οι πιθανοτητες για την ταξινόμηση συναισθήματος της μονάδας κάθε στρώματος τροφοδοτήθηκαν σε έναν ταξινομητή SVM που πρότεινε έναν απλό τρόπο να εκτελέσει την εκμάθηση μεταφοράς [55] στο SER. Με άλλα λόγια οι πιθανότητες που αντιστοιχούν σε κάθε συναισθηματική τάξη μάθαιναν σε διαφορε-

τική χρονική κλίμακα και χρησιμοποιήθηκαν για να εκτελέσουν την τελική ταξινόμηση στο επίπεδο ολόκληρης της πρότασης.

Άμεσες Προσεγγίσεις: Η χρονική συσσώρευση χαρακτηριστικών σε επίπεδο πλαισίου που χρησιμοποιεί στατιστικά χαρακτηριστικά σε τμήματα ομιλίας ή ολόκληρη την ομιλία δεν είναι ένας βέλτιστος τρόπος διατήρησης του συναισθηματικού περιεχομένου που μεταφέρεται από μια φωνητική παραπομπή [27]. Έτσι, αυτή η κατηγορία προσεγγίσεων SER συνήθως αναφέρεται στην άμεση μοντελοποίηση των συναισθηματικών πληροφοριών που προέρχονται από πλαίσια-LLD ή ακόμα και το ίδιο το σήμα φωνής. Συνεπώς, η πολυπλοκότητα αυτού του προβλήματος σχετίζεται στενά με το τεράστιο μήκος της ακολουθίας που υποτίθεται ότι έχει διαμορφωθεί. Πολλές περιοχές μέσα σε μια ομιλία αντιστοιχούν στη σιωπή ή σε μια ουδέτερη κατάσταση της συναισθηματικής κατάστασης του ομιλητή που πρέπει να αγνοηθεί για την τελική ταξινόμηση των συναισθημάτων [56]. Πρώτον, χρησιμοποιήθηκαν τα HMM για να μοντελοποιηθεί η ακολουθία των LLDs σε επίπεδο πλαισίου [57]. Ωστόσο, με την έλευση της βαθιάς μάθησης, πολλά πράγματα στον τρόπο της μοντελοποίησης ακολουθιών των διανυσμάτων χαρακτηριστικών έχουν αλλάξει [58]. Τα πλέον σύγχρονα αποτελέσματα έχουν αναφερθεί σε πολλές εργασίες όπως η ταξινόμηση / κατάτμηση εικόνας, η ASR, η ανίχνευση αντικειμένων κλπ. Όλα αυτά τα μοντέλα έχουν επίσης χρησιμοποιηθεί για SER. Προοδευτικά, βαθύτερες αρχιτεκτονικές που έμαθαν ανεξάρτητα συναισθηματικό πλαίσιο από απλά τοπικά χαρακτηριστικά, εφαρμόστηκαν. Δηλαδή, τα Βαθειά Νευρωνικά Δίκτυα (DNN) [60], τα μηχανήματα ακραίας μάθησης [61], ,[62] εκπαιδεύτηκαν σε φορείς LLD που αντιστοιχούσαν σε 1 ή περισσότερα διαδοχικά πλαίσια. Ένα ειδικό μοντέλο RNN το οποίο είναι ικανό να συντηρεί μακροπρόθεσμες εξαρτήσεις των ακολουθιών εισόδου και ονομάζεται μονάδα LSTM [64] έχει χρησιμοποιηθεί εκτενώς για SER και έχει εκπαιδευτεί σε χαρακτηριστικά πλαισίου [65], [66]. Στις τελευταίες δύο προσεγγίσεις, ένας μηχανισμός προσοχής στην κορυφή του LSTM ενισχύει την ακρίβεια αναγνώρισης παρέχοντας έναν μηχανισμό προσοχής για την ανίχνευση φαινομένου στην τεράστια ακολουθία εισόδου [67]. Με τον τρόπο αυτό, χρησιμοποιούνται για το σκοπό αυτό τα βάρη (οι τιμές που μαθαίνονται κατά την εκπαίδευση) του στρώματος προσοχής. Έτσι, οι περιοχές της ακολουθίας εισόδου που αντιστοιχούν σε σιωπηλά πλαίσια ή σε μια ουδέτερη συναισθηματική κατάσταση, παίρνουν χαμηλότερο βάρος κατά την διαδικασία της συναισθηματικής ταξινόμησης. Οι LLD που χρησιμοποιήθηκαν στις προαναφερθείσες μελέτες είναι παρόμοιοι με εκείνες που συζητήθηκαν στην προηγούμενη ενότητα (Ενότητα 1.3.2) που αποτελείται από περιγραφείς πλαισίων που εμπίπτουν στις τρεις κύριες κατηγορίες ποιότητας φωνής, θεμελιώδους συχνοτητας και φάσματος. Συνεπώς, η χρονική κλίμακα απόφασης για κάθε συναισθηματική κατάσταση βασίστηκε σε επίπεδο πλαισίου (30ms) ή επίπεδο φωνημάτων (10-30 αλληλουχίες LLDs). Σε άλλες προσεγγίσεις, όπως το [44], ένα μονοδιάστατο CNN χρησιμοποιείται από το ακατέργαστο σήμα εισόδου για συνεχή αναγνώριση συναισθημάτων τόσο στους άξονες ενεργοποίησης όσο και στους άξονες σθένους.

1.3.4 Προσεγγίσεις τελευταίας τεχνολογίας

Πρόσφατα, τα DNN μαζί με τους ΕΛΜ χρησιμοποιήθηκαν [60]. LLDs εξήχθησαν από πλαίσια και συνενώθηκαν σε κομμάτια 265ms. Τα υψηλότερα ενεργειακά τμήματα τροφοδοτήθηκαν σε DNN και η τάξη των συναισθημάτων υπολογίστηκαν για καθεμία από αυτές τις πιθανότητες. Στατιστικά αυτών των πιθανοτήτων τροφοδοτήθηκαν στον ELM για την ταξινόμηση σε επίπεδο ολόκληρης της ομιλίας, ενώ το μοντέλο δοκιμάστηκε στη βάση δεδομένων Interactive Emotional Dyadic Motion Capture (IEMOCAP) [68]. Metallinou και άλλοι [63] κατέδειξαν τη σημασία της χρήσης διπλής κατεύθυνσης LSTM (BLSTMs) και γενικά RNNs για την ενσωμάτωση μακροπρόθεσμου χρονικού πλαισίου σε SER με πειραματισμό σε χώρο ενεργοποίησης-σθένους. Lee και άλλοι [61] πρότειναν ένα σχήμα CNN με Connectionist Temporal Classification (CTC), στο οποίο κάθε πλαίσιο μπορεί να αντιστοιχιστεί σε όλες τις συναισθηματικές κλάσεις και μια επιπλέον ετικέτα NULL για πλαίσια σιωπής. Εφαρμόζουν το ELM πάνω από το BLSTM και παρουσίασαν σημαντική βελτίωση στο IEMOCAP αυξάνοντας σχετικά την ακρίβεια γύρω στο 12

Το σύγχρονο έργο SER επικεντρώνεται επίσης στην ταξινόμηση του συναισθήματος χρησιμοποιώντας χαρακτηριστικά φασματογράφων με ελάχιστη επεξεργασία ομιλίας. Στο [36] ο Ghosh και

άλλοι εισήγαγαν καινοτόμους τρόπους για την κατάρτιση ενός BLSTM, όπως η εκπροσώπηση και η μεταφορά της μάθησης. Ο πρώτος βασίζεται στα σήματα ροής του γλωττίδα ενώ ο τελευταίος μαθαίνει την αναπαράσταση των συναισθημάτων από το χώρο ενεργοποίησης-σθένους και οι δύο μέθοδοι δοκιμάστηκαν σε αυτοσχέδιες και συγγραφικές φράσεις. Σε μια εγκατάσταση πολλαπλών εργασιών μάθησης Xia και άλλοι [28] χρησιμοποίησαν με επιτυχία πληροφορίες αλληλεπίδρασης ενεργοποίησης και ανέφεραν τα καλύτερα αποτελέσματα με τα DBN πάνω σε παραστάσεις βασισμένες σε ολόκληρη την πρόταση. Επιπλέον, η εκμάθηση πολλαπλών εργασιών έχει επίσης χρησιμοποιηθεί με DNNs σε πειραματικό σχήμα cross-corpus που δείχνει την μεγάλη δυνατότητα γενίκευσης αυτών των μεθόδων [69]. Τέλος, η εκμάθηση μεταφοράς έχει επίσης διερευνηθεί συνδυάζοντας προοδευτικά νευρωνικά δίκτυα για SER σε μία από τις τελευταίες μελέτες [70].

Με βάση την τεράστια εκφραστικότητα των RNN και την αναγκαιότητα να τονιστούν τα συναισθηματικά πυκνά πλαίσια ομιλίας, αναπτύχθηκαν μοντέλα προσοχής για το SER. Στο [66] πολύ καλά αποτελέσματα παρουσιάστηκαν στο IEMOCAP. Ένα BLSTM με βάση ένα σύστημα προσοχής χρησιμοποιήθηκε για μια καλύτερη εννοιολογική συναισθηματική υποδιαίρεση και απέδωσε απόλυτη βελτίωση $1,46\%$ στην ακρίβεια από το νευρωνικό που δεν χρησιμοποιούσε κανένα μοντέλο προσοχής. Τέλος, οι Misramadi και άλλοι [65] εξήγαγαν ακατέργαστα φασματικά χαρακτηριστικά και LLD για την εκπαίδευση ενός BLSTM. Έδειξαν ότι τα LLD υπερέβησαν τα ακατέργαστα φασματικά χαρακτηριστικά και χρησιμοποίησαν ένα σταθμισμένο στρώμα συγκέντρωσης με μαχανισμό προσοχής στην κορυφή, προκειμένου να ασχοληθεί με την ψηφοφορία των συναισθηματικά-άσχετων πλαισίων. Επιπλέον, συγκρίθηκε μια ποικιλία μεθόδων συγκέντρωσης, κατά την τελευταία κρυφή στρώση, για την τελική ταξινόμηση των ομιλιών.

Σε μια πρόσφατη μελέτη [35], συγκρίθηκαν διάφορες βαθιές αρχιτεκτονικές. Τα μοντέλα εκπαιδεύτηκαν σε χαρακτηριστικά από σπεκτρογράμματα και τα καλύτερα αποτελέσματα της IEMOCAP αναφέρθηκαν χρησιμοποιώντας το CNN. Σε αυτό το πλαίσιο, το CNN [27] έχει χρησιμοποιηθεί για το SER προς την κατεύθυνση της χρήσης ενός συνόλου ελάχιστων χαρακτηριστικών των MFB ικανών να συντηρούν σημαντικές αναπαραστάσεις από περιγραφείς σε επίπεδο πλαισίου. Επιπλέον, συνελικτικοί σποροτοί κωδικοποιητές (SAE) για την εκμάθηση χαρακτηριστικών χαρακτηριστικών από σπεκτρογράμματα συναισθηματικών ομιλιών έχουν επίσης μελετηθεί [37] κάτω από πολλαπλές πειραματικές ρυθμίσεις SER. Στην τελευταία προσέγγιση, οι αναπαραστάσεις που μαθαίνονται αντιστοιχούν στα βάρη των συνελικτικών φίλτρων μετά την εφαρμογή ενός πυρήνα ικανού να διατηρήσει τους τοπικούς αμετάβλητους φορείς χαρακτηριστικών. Αυτοί οι φορείς χρησιμεύουν ως είσοδος για μια εκπαίδευση και δοκιμή SVM.

Τα προαναφερθέντα σύνολα χαρακτηριστικών, οι μέθοδοι ταξινόμησης και οι πλέον σύγχρονες προσεγγίσεις αποτελούν μόνο ένα κλάσμα των ποικίλων μεθόδων που υπάρχουν στη βιβλιογραφία. Επειδή ο σκοπός αυτής της μελέτης δεν σχετίζεται αποκλειστικά με την εξαντλητική ανασκόπηση όλων των διαθέσιμων μεθόδων, προτείνουμε στον αναγνώστη να αναζητήσει πρόσθετες πληροφορίες σχετικά με τις σχετικές δημοσιεύσεις στον τομέα στις ακόλουθες αναθεωρήσεις [19], [20] και [21].

1.4 Προκλήσεις

Παρόλο που πολλές νέες προκλήσεις έχουν δημιουργηθεί συνεχώς, καθώς επινοήθηκαν νέα μοντέλα και προσεγγίσεις για την οικοδόμηση αποτελεσματικότερων μοντέλων SER, εστιάζουμε σε τρία βασικά προβλήματα τα οποία πάντα προκαλούσαν σύγχυση στην έρευνα SER εδώ και χρόνια. Αυτές οι προκλήσεις εξηγούνται συνοπτικά παρακάτω, καθώς και οι ερωτήσεις τους για την πολυπλοκότητα και την πιθανότητα να εξευρεθεί πραγματικά μια καλή λύση στο πρόβλημα.

1.4.1 Χρονικές Κλίμακες της Απόφασης για το Συναισθηματικό Περιεχόμενο μιας Ομιλίας

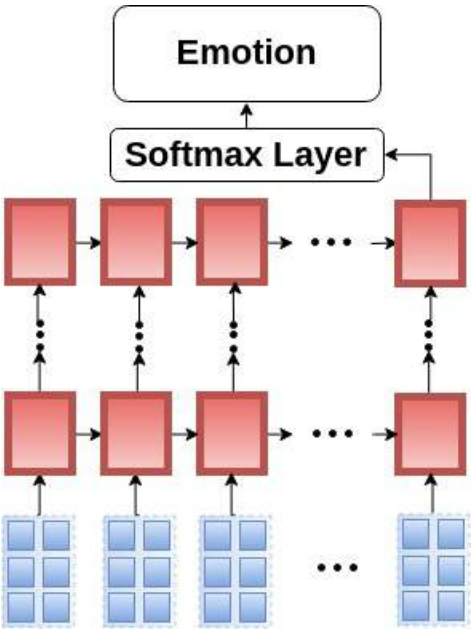
Σε γενικές γραμμές, το συναισθηματικό περιεχόμενο θα μπορούσε να ληφθεί σε διαφορετικές χρονικές κλίμακες. Η δομή του ανθρώπινου ακουστικού μοντέλου επεξεργασίας στον εγκέφαλο εξαρτά-

ται σε μεγάλο βαθμό από αυτό το είδος λειτουργικής συμπεριφοράς [71]. Η αντίληψη του ανθρώπινου ήχου θα μπορούσε να διαμορφωθεί επαρκώς με έναν πολυδιάστατο φασματο-χρονικό υποδοχέα εκπεμπόμενων ήχων [71]. Αυτό είναι ενδεικτικό της διαδικασίας που ο εγκέφαλος αναλαμβάνει για να αναγνωρίσει τα φωνητικά συνθήματα και κατά συνέπεια, να ερμηνεύσει τις πληροφορίες υψηλότερου επιπέδου που μεταφέρονται. Διαφορετικές χρονικές κλίμακες επικεντρώνονται σε διαφορετικές πτυχές αυτής της πληροφορίας που μεταβιβάζεται. Είναι σημαντικό να κατανοήσουμε τις ιδιοτροπίες κάθε χρονικής κλίμακας, προκειμένου να δημιουργήσουμε επιτυχημένα μοντέλα SER.

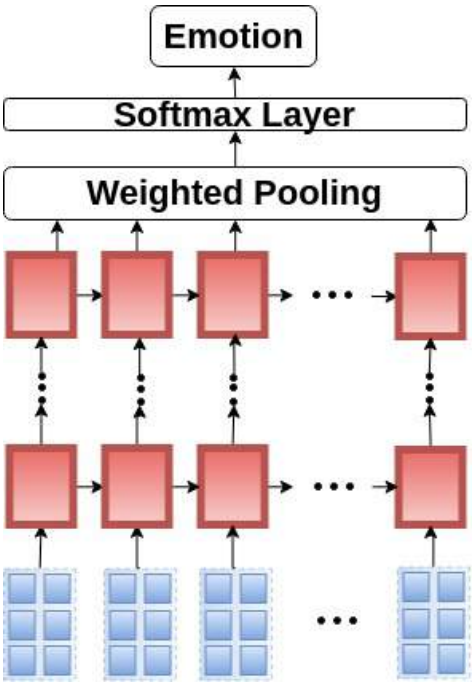
Όπως αναφέρθηκε στην προηγούμενη ενότητα, οι προσεγγίσεις με βάση τμήματα ομιλίας και οι άμεσες προσεγγίσεις αντιμετωπίζουν το πρόβλημα της μοντελοποίησης μιας σειράς χαρακτηριστικών προκειμένου να εκπαιδεύσουν ένα μοντέλο που μπορεί να προβλέψει το συναισθηματικό περιεχόμενο μιας έκφρασης. Κάθε χρονική βαθμίδα της ακολουθίας εισόδου του φορέα θα μπορούσε να αντιστοιχεί σε ένα χρονικό κλιμάκιο που χαρακτηρίζει το επίπεδο από το οποίο αντλείται η απόφαση της συναισθηματικής ετικέτας. Αυτό το χρονικό διάστημα, το οποίο επίσης αναφέρεται ως χρονικό επίπεδο [50], θα μπορούσε να αντιστοιχεί σε μια διάρκεια ομιλίας ενός πλαισίου, ενός φωνήματος, μιας συλλαβής, μιας λέξης, ενός τμήματος ομιλίας αυθαίρετου μήκους ή ακόμα και της ίδιας της ομιλίας. Αν η αρχιτεκτονική κάθε μοντέλου μεταβάλλει σημαντικά την κωδικοποίηση της ακολουθίας εισόδου των διανυσμάτων χαρακτηριστικών, η βάση της απόφασης έκφρασης διαμορφώνεται από την αντίστοιχη χρονική κλίμακα για κάθε διάνυσμα εισόδου.

Υπενθυμίζοντας τις προσεγγίσεις SER στη βιβλιογραφία, μπορούμε να δούμε ότι έχουν χρησιμοποιηθεί διαφορετικές χρονικές κλίμακες για την εξαγωγή ακουστικών χαρακτηριστικών. Ωστόσο, η επιλογή κάθε χρονικής κλίμακας για την εξαγωγή κάθε συνόλου χαρακτηριστικών (τοπικά ή παγκόσμια χαρακτηριστικά) δεν έχει ακόμη αναλυθεί προσεκτικά. Αυτό μπορεί να οδηγήσει σε μια ad-hoc επιλογή της αντίστοιχης χρονικής κλίμακας, καθώς επίσης και σε έναν τρόπο αντιμετώπισης του συναισθηματικού λόγου που είναι αρκετά μακριά από αυτόν που έχουμε να σκεφτόμαστε ως άνθρωποι. Ειδικά όταν η προσέγγιση βασίζεται σε πλαίσια ή τμήματα υπο-προτάσεων. Για παράδειγμα, ακολουθώντας μια προσέγγιση που βασίζεται σε τμήματα για την εξαγωγή χαρακτηριστικών και στη συνέχεια εφαρμόζοντας ένα SVM σε κάθε τμήμα προκειμένου να συναγάγουμε την αντίστοιχη ετικέτα συναισθημάτων για καθένα από αυτά δεν είναι αντιπροσωπευτική του ανθρώπινου συμπεράσματος των συναισθημάτων. Συγκεκριμένα, εάν μια έκφραση αποτελείται από τμήματα των 3 δευτερολέπτων και μόνο ένα πλαίσιο ομιλίας περιέχει μια εκδήλωση θυμού, τότε ολόκληρη η έκφραση θα πρέπει να χαρακτηριστεί ως θυμωμένη, παρά το γεγονός ότι όλα τα άλλα τμήματα που περιλαμβάνονται είναι ουδέτερα. Εάν η επιλεγμένη συνάρτηση συνάθροισης για να συναγάγουμε τη συναισθηματική ετικέτα είναι μια συνάρτηση πλειοψηφίας, τότε η ταξινόμηση θα θεωρηθεί ότι προβλέπει μια ουδέτερη διατύπωση αντί για μια θυμωμένη. Το ίδιο πρόβλημα υπάρχει και για τις άμεσες προσεγγίσεις όπου κάθε πλαίσιο 20 ms αναγκάζεται να ψηφίσει για το υπάρχον συναίσθημα σε αυτό το χρονικό διάστημα [61] εκτελώντας έναν αλγόριθμο Μεγιστοποίησης της Προσδοκίας (EM) για την ταξινόμηση όλων των πλαισίων. Πώς θα μπορούσαμε να είμαστε βέβαιοι ότι ένα πλαίσιο μερικών χιλιοστών του δευτερολέπτου θα μπορούσε να περικλείει διακριτικές πληροφορίες για μια ευτυχισμένη εκδήλωση;

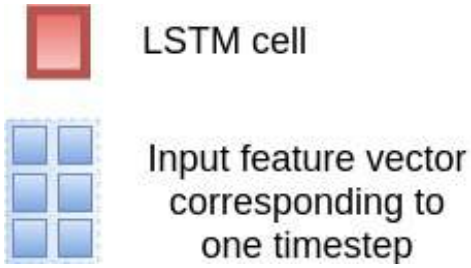
Τα πρόσφατα χρησιμοποιούμενα μοντέλα όπως: HMMs [52] και RNNs [65] μοντελοποιούν την ακολουθία εισόδου των διανυσμάτων χαρακτηριστικών χωρίς να απαιτούν ρητά το αποτέλεσμα ταξινόμησης από κάθε βήμα της ακολουθίας εισόδου. Αυτά τα μοντέλα κατά τη διάρκεια της διαδικασία εκπαίδευσης κωδικοποιούν τις συναισθηματικές πληροφορίες της ακολουθίας με την τιμωρία του σφάλματος πρόβλεψης που παράγεται από τις ενεργοποιήσεις του στρώματος εξόδου τους. Παρόλο που μια ποικιλία αρχιτεκτονικών RNN διαφέρουν σημαντικά μεταξύ τους από την άποψη της χρονικής συνάθροισης που εφαρμόζεται προκειμένου να συναχθεί το συναισθηματικό περιεχόμενο κάθε έκφρασης. Στο σχήμα 1.2 εμφανίζονται δύο από τους πιο γνωστούς τύπους αρχιτεκτονικών RNN. Στο σχήμα 1.2a, είναι προφανές ότι μόνο οι ενεργοποιήσεις του τελικού χρονικού βήματος συμβάλλουν στην απόφαση της πρόβλεψης συναισθήματος. Αντίθετα, στο Σχήμα 1.2b οι ενεργοποιήσεις από όλα τα βήματα χρόνου επηρεάζουν την τελική απόφαση ταξινόμησης. Αν και δεν πρέπει να αγνοούμε το γεγονός ότι οι αρχιτεκτονικές RNN, όπως οι LSTMs, θα μπορούσαν να κωδικοποιήσουν την πληροφορία της ακολουθίας εισόδου των χαρακτηριστικών στις ενεργοποιήσεις του τελικού χρονικού βήματος



(a) Ταξινόμηση RNN χρησιμοποιώντας μόνο ενεργοποιήσεις από το τελικό βήμα



(b) Ταξινόμηση RNN χρησιμοποιώντας τις ενεργοποιήσεις όλων των χρονικών βημάτων μετά από έναν αφινικό μετασχηματισμό που εφαρμόζεται με σταθμισμένη συγκέντρωση



(c) Χάρτης συμβόλων

Σχήμα 1.2: Συναισθηματική ταξινόμηση χρησιμοποιώντας διαφορετικές αρχιτεκτονικές RNN

και έτσι η αρχική αρχιτεκτονική χρησιμοποιείται ευρέως [64]. Ωστόσο, εάν η ακολουθία εισόδου είναι αρκετά μεγάλη τότε η ροή κωδικοποίησης πληροφοριών δεν είναι τόσο ομαλή όσο αναμένεται να είναι. Για παράδειγμα, αν υποθέσουμε ότι οι συνιστώσες εισόδου αντιστοιχούν σε LLD που εξάγονται από πλαίσια 20 ms με επικαλυπτικό λόγο 0.5, τότε μια δήλωση 4 δευτερολέπτων θα περιέχει συνολικό αριθμό 400 διανυσμάτων χαρακτηριστικών. Εάν στην αρχή της ακολουθίας βρίσκονται κάποια ενδεικτικά LLD επιπέδου πλαισίου μιας έκφρασης θυμού και μετά από αυτό όλα τα υπόλοιπα πλαίσια αντιπροσωπεύουν μια ουδέτερη συναισθηματική κατάσταση τότε στο τέλος της ακολουθίας όπου βασίζεται η συναισθηματική απόφαση η συμβολή των προηγούμενων πλαισίων να μειωθεί [72]. Αυτός είναι ο λόγος που χρησιμοποιούνται αμφίδρομα RNN [63] και οι μηχανισμοί προσοχής [66] για να ανακουφίσουν αυτό το πρόβλημα που σχετίζεται με την καταστροφή της πληροφορίας μέσω πολλαπλών επαναλήψεων των χρονικών βημάτων.

Στην ουσία, όλες οι προτεινόμενες αρχιτεκτονικές στη βιβλιογραφία θα μπορούσαν να μειωθούν σε αυτούς τους δύο τύπους αρχιτεκτονικών [65], που απεικονίζονται στα Figures 1.2a και 1.2b. Για παράδειγμα, ένα μοντέλο προσοχής θα ήταν απλώς μια άλλη σταθμισμένη στρατηγική συγκέντρωσης που χρησιμοποιείται για να προσδιοριστούν τα βάρη για κάθε χρονικό βήμα. Συνεπώς, οι αντίστοιχες ενεργοποιήσεις κάθε χρονικού βήματος θα συνέβαλαν στην τελική συναισθηματική απόφαση μετά

την εφαρμογή ενός στρώματος softmax. Το συναισθηματικό περιεχόμενο δεν προκύπτει μόνο από το ίδιο το μοντέλο αλλά από το σύνολο των αντιπροσωπευτικών χαρακτηριστικών που εξάγονται κάτω από ένα συγκεκριμένο χρονικό διάστημα. Ειδικότερα, οι χρονικές κλίμακες κάτω από τις οποίες αντλούμε την απόφαση για το συναισθηματικό περιεχόμενο συνδέεται στενά με τις χρονικές κλίμακες των χαρακτηριστικών εισόδου παρά το γεγονός ότι τα RNNs φαίνεται να συλλάβουν εξαρτήσεις σε πολυπαραμετρικές χρονικές κλίμακες. Η προαναφερθείσα αρχιτεκτονική RNN δεν αλλάζει αυτό το χρονικό διάστημα μεταξύ των επόμενων στρώσεων, όπως κάνουν τα CNNs με τα μέγιστα / μεσαία και άλλα στρώματα συγκέντρωσης μεταξύ των [73].

Λαμβάνοντας υπόψη τον συνδυασμό της ποικιλίας των ακουστικών χαρακτηριστικών (τοπικών περιγραφητών και γενικών στατιστικών μεγαλύτερης διάρκειας) που εξάγονται σε διαφορετικό χρονικό διάστημα προκύπτουν μερικές πολύ ενδιαφέρουσες ερωτήσεις και δεν έχουν ακόμη διερευνηθεί πλήρως.

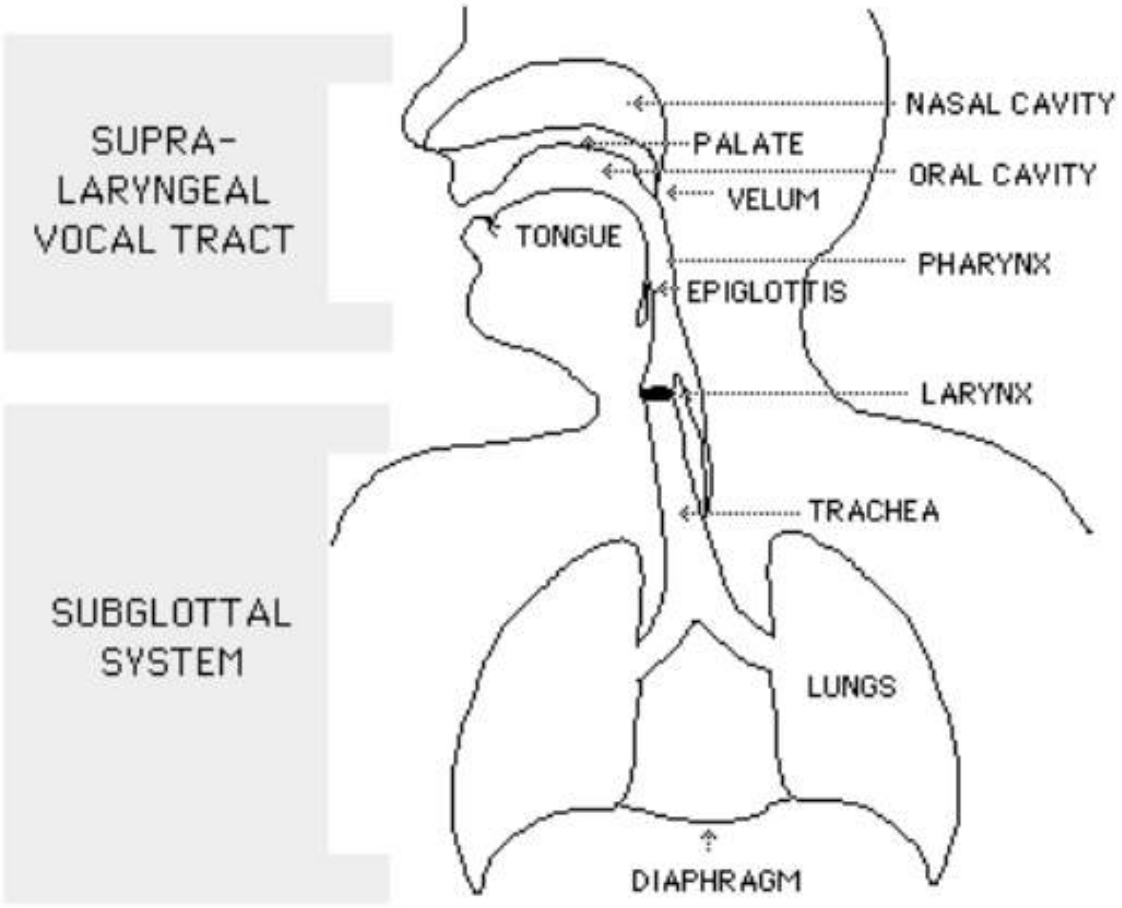
- Το χρονοδιάγραμμα για την εξαγωγή συναισθηματικού περιεχομένου είναι διαφορετικό για κάθε σύνολο χαρακτηριστικών;
- Πώς θα μπορούσαμε να καθορίσουμε μια ικανοποιητική χρονική κλίμακα για κάθε τύπο συνόλων χαρακτηριστικών που θα μπορούσε καταλλήλως να συλλάβει συναισθηματική πληροφορία;
- Αν ενώσουμε μερικά διανύσματα χαρακτηριστικών για κάθε τύπο χαρακτηριστικών και τροφοδοτήσουμε τον RNN με μια πολύ μικρότερη ακολουθία διανυσμάτων, μπορούμε να ενισχύσουμε την ακρίβεια αναγνώρισης;

1.4.2 Μη γραμμικά φαινόμενα στην παραγωγή ομιλίας

Το μοντέλο της παραγωγής ομιλίας έχει αναθεωρηθεί και αναλυθεί εκτενώς σε όλα αυτά τα χρόνια [74]. Στην Εικόνα 1.3 [1], το φυσιολογικό σύστημα παραγωγής ανθρώπινης ομιλίας εμφανίζεται. Στην ουσία, οι πνεύμονες αντιστοιχούν στην πηγή της ισχύος στο σύστημα αυτό που παράγει τη ροή του αέρα. Το αναπνευστικό σύστημα παράγει κραδασμούς των φωνητικών πτυχών στον λάρυγγα εάν η ροή του αέρα πιέζεται μέσω της γλωττίδας ή μια παροδική ώθηση για τη μη φωνήενη νεύρωση. Οι παραγόμενες αέριες δονήσεις φιλτράρονται αργότερα από την φωνητική οδό που αντιστοιχεί στον λάρυγγα, τον φάρυγγα και τη ρινική κοιλότητα. Η γλώσσα, οι γνάθοι, τα χείλη λειτουργούν ως αρθρωτές που μεταβάλλουν τα χαρακτηριστικά συντονισμού του άνω μέρους του συστήματος [74].

Από μια πιο τεχνική άποψη, μπορούμε να μοντελοποιήσουμε το προαναφερθέν σύστημα παραγωγής ομιλίας με ένα μοντέλο φίλτρου πηγής (source-filter model) όπως ορίζεται στο [6] το οποίο απεικονίζεται στο Σχήμα 1.4. Παρόλο που έχουν προταθεί ποικίλα μοντέλα για να αντικατοπτρίζουν σωστά την πραγματική υποκείμενη δυναμική του συστήματος παραγωγής ομιλίας, εστιάζουμε στην πιο συνηθισμένη που είναι το μοντέλο πηγαίου φίλτρου [75]. Το πιο δύσκολο συστατικό του προαναφερθέντος μοντέλου είναι η φωνητική πτυχή όσον αφορά την ανάλυση και τη διαμόρφωση. Ουσιαστικά, οι στροβιλισμοί, οι διεισδύσεις και οι διακυμάνσεις που προκύπτουν από βιοφυσικές διεργασίες οδηγούν σε ορισμένα χαοτικά φαινόμενα στο φωνητικό σύστημα παραγωγής [76]. Οι απεριοδικές δονήσεις των φωνητικών πτυχών ενδέχεται να είναι δελεαστικές για να τις χαρακτηριστούν ως τυχαίες διαταραχές αλλά, αντιθέτως, παρουσιάζουν έναν εξ ολοκλήρου αιτιοκρατικό χαρακτήρα που βρίσκεται σε χαοτικό καθεστώς. Στην πραγματικότητα, όλοι αυτοί οι ήχοι αντιστοιχούν σε μια λεκάνη έλξης στους αντίστοιχους χώρους φάσης τους (ο χώρος που περιλαμβάνει όλες τις πιθανές καταστάσεις ενός συστήματος) [78]. Ο τελευταίος χώρος διέπεται από περιοδικές δομές που παρουσιάζουν μοτίβα επαναληψιμότητας που είναι ενδεικτικά της πραγματικής φύσης του υπό ανάλυση συστήματος [79]. Η ενδογενής μη γραμμικότητα των φωνητικών πτυχών έχει επίσης μελετηθεί για παθολογικές καταστάσεις χρησιμοποιώντας σπεκτρογράμματα στενής ζώνης [80]. Τα τελευταία αποτελέσματα δείχνουν ότι η φύση της δυναμικής των φωνητικών πτυχών είναι πραγματικά μη γραμμική και δεν αποτελεί μόνο μέρος μιας γενικής υπόθεσης του συστήματος.

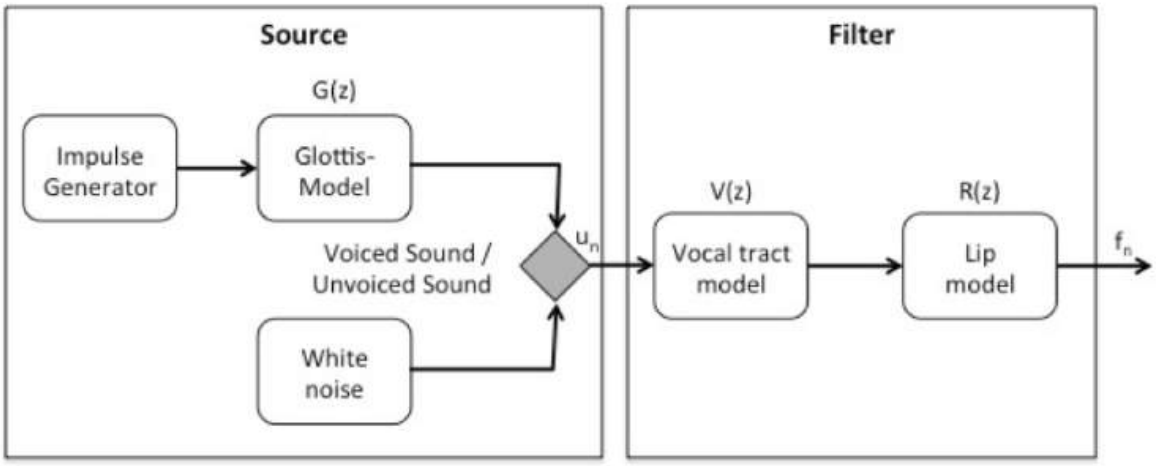
[1] Το Σχήμα 1.3 βρέθηκε στο: http://sail.usc.edu/~lgoldste/General_Phonetics/Source_Filter/test.html



Σχήμα 1.3: Ανθρώπινο σύστημα παραγωγής ομιλίας

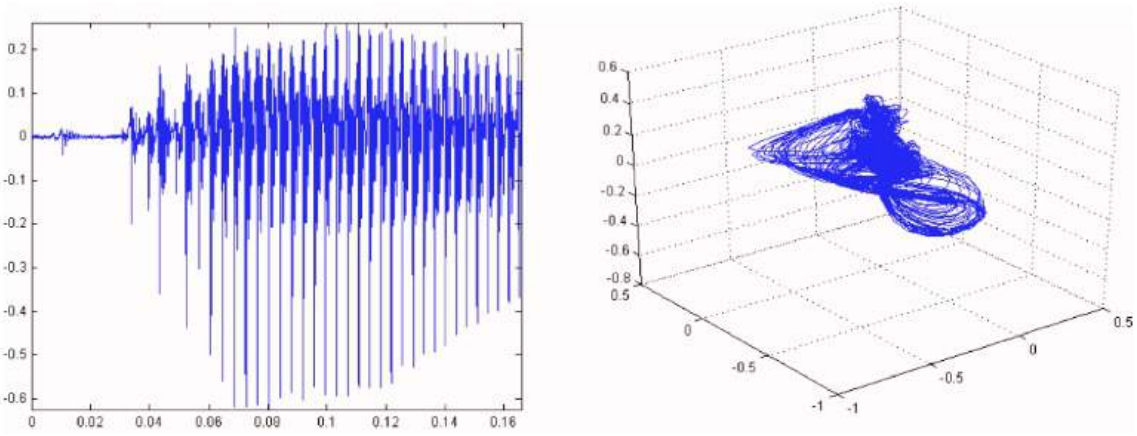
Τα περισσότερα από τα σύνολα χαρακτηριστικών που περιγράφονται στην Ενότητα 1.3.2 βασίζονται σε LLD που εξάγονται υπό την προϋπόθεση ότι ένα μοντέλο γραμμικού πηγαίου φίλτρου δημιουργεί ομιλία. Ωστόσο, οι ταλαντώσεις των φωνητικών χορδών και η δυναμική ρευστού του φωνητικού συστήματος εμφανίζουν συχνά άκρως μη γραμμικές δυναμικές ιδιότητες οι οποίες μπορεί να μην αναπαρίστανται σωστά από συμβατικά LLDs. Το γεγονός ότι είναι πολύ πιο εύκολο να εξαχθούν ακουστικά χαρακτηριστικά με την παραδοχή της γραμμικότητας και της ακινησίας του σήματος, όπως απαιτεί το μετασχηματισμό Fourier, κάνει τις μη γραμμικές παραστάσεις ομιλίας να παραμεληθούν εντελώς στα παραδοσιακά σύνολα χαρακτηριστικών. Επιπλέον, τα συνηθισμένα σύνολα χαρακτηριστικών περιλαμβάνουν τα χαρακτηριστικά ανύψωσης που σχετίζονται με την εκτίμηση της θεμελιώδους συχνότητας της φωνητικής οδού. Ωστόσο, στην παραγωγή ομιλίας μόνο τα φωνήεντα σχετίζονται με περιοχές περιοδικότητας από δυναμική άποψη συστήματος. Τα πλαίσια σιωπής μεταξύ δύο φωνημάτων που προφέρονται αντιστοιχούν σε τυχαίο θόρυβο, καθώς και σε τρεχούμενα συμφώνια, στα οποία η υποκείμενη δυναμική τους αντιστοιχεί σε χαοτικά πρότυπα [81]. Και τα δύο προαναφερθέντα φωνήματα προκαλούν θορυβώδη έξοδο σε χαρακτηριστικά συναρτήσει της θεμελιώδους συχνότητας, τα οποία υποθέτουν περιοδικότητα μέσα στο πλαίσιο εξαγωγής.

Θα μπορούσαμε να αναλύσουμε τις δυναμικές ιδιότητες του συστήματος παραγωγής ομιλίας όπως οποιοδήποτε μη γραμμικό δυναμικό σύστημα. Για να γίνει αυτό, μία από τις πιο σημαντικές μεθόδους είναι η ανακατασκευή του χώρου φάσης [82]. Ο τελευταίος χώρος είναι στην πραγματικότητα μια εμβύθινση του σήματος στον χρόνο σε ένα χώρο υψηλότερων διαστάσεων χρησιμοποιώντας καθυστερημένες εκδοχές του σήματος και ονομάζεται επίσης χώρος κατάστασης. Διαισθητικά, αυτή είναι μια πιο εκφραστική αναπαράσταση του ίδιου δυναμικού συστήματος, το οποίο είναι πολύ σύνθετο στον τομέα του χρόνου, αλλά μπορεί να ξεδιπλωθεί χρησιμοποιώντας περισσότερες διαστάσεις και κατάλ-



Σχήμα 1.4: Μοντέλο πηγής-φίλτρου παραγωγής ομιλίας [6]

ληλες χρονικές καθυστερήσεις για το συγκεκριμένο σύστημα [78]. Για παράδειγμα, στην απεικόνιση 1.5 η εμφάνιση του φωνήματος */a/* εμφανίζεται στο πεδίο χρόνου (αριστερά) και στην αντίστοιχη ανακατασκευή του χώρου (δεξιά). Από τον απεικονιζόμενο χώρο φάσης του Σχήματος 1.5, είναι προφανές ότι οι εμβυθισμένες τροχιές του χώρου φάσης εμφανίζουν συμπεριφορά επαναληψιμότητας ή υποτροπιάζουσα δυναμική ειδικά όταν το αντίστοιχο σήμα στο χρονικό πεδίο παρουσιάζει περιοδικότητα. Οι τροχιές που φαίνεται να είναι παρόμοιες στο χώρο φάσης ή παράλληλες σε τοπική κλίμακα αντικατοπτρίζουν τη συν-εξέλιξη καταστάσεων για το σύστημα που αναλύεται. Το τελευταίο εμφανίζεται τελικά όταν δημιουργούνται ελκυστές στον ανακατασκευασμένο χώρο φάσης του σήματος. Οι ελκυστές είναι σύνολα αριθμητικών τιμών προς τα οποία ένα σύστημα τείνει να εξελιχθεί, για μια ευρεία ποικιλία συνθηκών εκκίνησης του συστήματος [83]. Αυτή η ασυμπτωτική συμπεριφορά των τροχιών γύρω από τους περίεργους ελκυστήρες θα μπορούσε να είναι κατατοπιστική για τις ιδιότητες επαναληψιμότητας της υποκείμενης δυναμικής του συστήματος που είναι ένα πλαίσιο ομιλίας στην περίπτωσή μας. Ωστόσο, η επανάληψη δεν μπορεί να ληφθεί σωστά χρησιμοποιώντας μόνο εκτιμήσεις των πιο σημαντικών συχνοτήτων, αλλά αυτές οι απεικονίσεις στηρίζουν την εκμετάλλευση των σύνθετων δομών χρησιμοποιώντας εργαλεία ικανά να καταγράψουν αυτές τις πληροφορίες και να τα χρησιμοποιήσουν για επεξεργασία ομιλίας. Ειδικά στο SER, όπου οι στροβιλισμοί είναι πολύ πιο εμφανείς στα συναισθηματικά φωνητικά σήματα, θα πρέπει επίσης να προσπαθήσουμε να καταγράψουμε αυτές τις πληροφορίες για την ενίσχυση της απόδοσης των συστημάτων SER.



Σχήμα 1.5: Ένα φώνημα /α/ και ο αντίστοιχος χώρος φάσης που παρουσιάζει υποτροπιάζουσα δυναμική [7]

Λαμβάνοντας υπόψη τη μη γραμμική φύση της δυναμικής του συτήματος παραγωγής φωνής, μια πρακτική πρόκληση είναι να βρεθεί μια σειρά από μη γραμμικά χαρακτηριστικά που να μπορούν να συλλάβουν σωστά τις συναισθηματικές πληροφορίες λαμβάνοντας υπόψη την πραγματική ταυτότητα της υποκείμενης δυναμικής. Οι πληροφορίες επαναληψιμότητας του χώρου φάσης (PS) ή του χώρου κατάστασης των σημάτων ομιλίας είναι ιδιαίτερα ενδεικτικές της πραγματικής φύσης της υποκείμενης δυναμικής και μπορούν να καταγράψουν μη γραμμικές εξαρτήσεις των δεδομένων εισόδου [84]. Αν και έχει μελετηθεί μια ποικιλία μη γραμμικών χαρακτηριστικών για το SER όπως: Teager Energy Operator (TEO) [40], χαρακτηριστικά διαμόρφωσης από το στιγμιαίο πλάτος και φάση [43] καθώς και γεωμετρικά μέτρα από την τροχιά PS [42], οι πληροφορίες επαναληψιμότητας του PS δεν έχουν ακόμη διερευνηθεί. Ορισμένες αναδυόμενες ερωτήσεις και προκλήσεις είναι οι εξής:

- Είναι δυνατή η εξαγωγή πληροφοριών μη γραμμικής επαναληψιμότητας από σήματα ομιλίας προκειμένου να χρησιμοποιηθούν ως ακουστικές πληροφορίες;
- Πώς μπορούμε να ενσωματώσουμε πληροφορίες μη γραμμικών υποτροπιαζουσών δυναμικών των σημάτων φωνής για SER;
- Οι πληροφορίες επαναληψιμότητας και η μη γραμμική ανάλυση αυτών θα μπορούσαν να ενισχύσουν την ακρίβεια των συστημάτων SER;

1.4.3 Η κατάρα της διαστατικότητας και η προσπάθεια μείωσης των διαστάσεων εισόδου

Παρόλο που τα προαναφερθέντα χαρακτηριστικά που έχουν οριστεί στην Ενότητα 1.3.2 έχουν αποδειχθεί ότι αποτυπώνουν συστηματικά το συναισθηματικό περιεχόμενο από τα σήματα ομιλίας, ο χώρος των χαρακτηριστικών τους είναι αρκετά μεγάλος και αυξάνεται καθώς προστίθενται νέα χαρακτηριστικά στα προηγούμενα σύνολα. Παραδείγματος χάριν, τα έργα του Schuller στο [25], [23], [29] έχουν αυξήσει το μέγεθος των προτεινόμενων χαρακτηριστικών για παραγλωσσικά καθήκοντα από την προηγούμενη αναφορά. Δηλαδή, τα αρχικά 384 χαρακτηριστικά [25] έχουν επεκταθεί σε 1582 [23] και κατά συνέπεια έχουν ως αποτέλεσμα διανύσματα 6373 χαρακτηριστικών σε [29]. Παρόλο που τα προσφάτως προστεθέντα χαρακτηριστικά θα μπορούσαν να αποτυπώνουν ανεξάρτητα σημαντικές συναισθηματικές πληροφορίες, θα ήταν πολύ πιθανό ότι δεν προσφέρουν περαιτέρω διακριτική ικανότητα σε ολόκληρο το σύνολο των χαρακτηριστικών. Φυσικά δεν μπορούμε να είμαστε σίγουροι για τη συμβολή του κάθε χαρακτηριστικού στην ακρίβεια αναγνώρισης καθώς η τελευταία προσέγγιση απαιτεί μια εξαντλητική αναζήτηση στους συνδυασμούς που παράγονται από το σύνολο ισχύος του συνόλου χαρακτηριστικών.

Τα μοντέλα τα οποία αυτά τα χαρακτηριστικά εκπαιδεύονται υποφέρουν από αυτό που ονομάζεται "Κατάρα της Διαστασικότητας" ή "Curse of Dimensionality" [85]. Το τελευταίο πρόβλημα αναφέρεται σε φαινόμενα μείωσης των μετρήσεων απόδοσης, αριθμητικής αστάθειας, σπανιότητας και ανομοιότητας των δεδομένων εισόδου όταν το μοντέλο εκπαιδεύεται σε δεδομένα μεγάλης διαστάσεως. Ο τεράστιος αριθμός ιδιοτήτων διαστάσεων μπορεί να είναι καταστροφικός για τη διαδικασία εκπαίδευσης ενός μοντέλου λόγω των εκθετικά αυξημένων (από άποψη διαστάσεων εισόδου) χρόνου και δειγμάτων που απαιτούνται για τη διαδικασία.

Για το σκοπό αυτό, έχουν προταθεί πολλές τεχνικές επιλογής χαρακτηριστικών [34] καθώς και μια ποικιλία τεχνικών μείωσης διαστάσεων [86]. Στην πρώτη προσέγγιση, επιλέγεται ένα υποσύνολο ολόκληρου του συνόλου χαρακτηριστικών χρησιμοποιώντας μια ευρετική βάση με βάση την μέση αμοιβαία πληροφορία ή την εντροπία μεταξύ ορισμένων συνόλων χαρακτηριστικών. Αντίθετα, γενικά οι τεχνικές μείωσης των διαστάσεων επιδιώκουν να βρουν μια χαμηλότερη παράσταση διαστάσεων $\mathbf{x} \in \mathbf{R}^n$ των αρχικών ιδιοτήτων υψηλών διαστάσεων που βρίσκονται στο $\mathbf{x} \in \mathbf{R}^D$, όπου συνήθως $D >> n$ και διατηρώντας παράλληλα τις γεωμετρικές ιδιότητες του χώρου των υψηλν διαστάσεων. Ωστόσο, οι μειωμένες διαστάσεις του χώρου εισόδου χάνουν τη φυσική τους σημασία καθώς ο αλγόριθμος μείωσης των διαστάσεων παράγει την υψηλότερη διαστασιακή αναπαράσταση ανεξάρτητα από τις πραγματικές φυσικές ιδιότητες του χώρου εξόδου. Μερικοί αλγόριθμοι, όπως το

PCA [33], προσπαθούν να παράγουν τις παραστάσεις χαμηλών διαστάσεων χρησιμοποιώντας γραμμικούς συνδυασμούς των διαστάσεων των αρχικών χαρακτηριστικών ενώ άλλες μη γραμμικές μέθοδοι προτείνουν την ανασυγκρότηση μιας μη γραμμικής εκβυθισμένης πολλαπλότητας $\mathcal{M}$ που διατηρεί την εγγενή γεωμετρία του χώρου μεγάλης διαστάσεως. Η τελευταία προσέγγιση αποκαλείται συχνά "Μάθηση πολλαπλότητας" και έχει χρησιμοποιηθεί εκτενώς σε διάφορους τομείς όπου η διάσταση των δεδομένων εισόδου υπερβαίνει τα υπολογιστικά όρια που θέτουν οι περιορισμοί των αρχιτεκτονικών που χρησιμοποιειούνται για να προσεγγίστει αριθμητικά η πολλαπλότητα $\mathcal{M}$ [87].

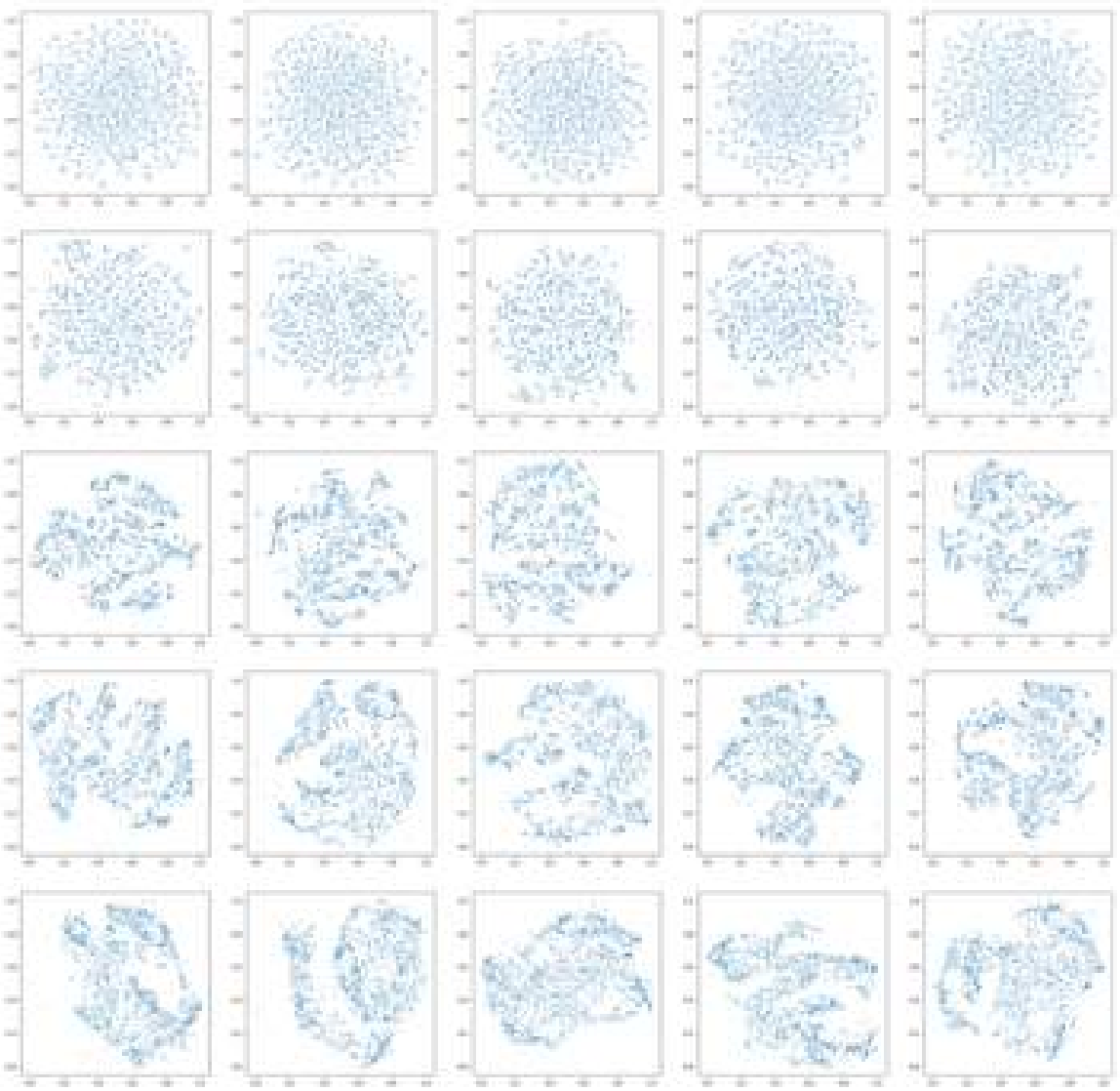
Εάν τα δεδομένα μεγάλης διαστάσεων εισόδου $\mathbf{R}^D$ μπορούν να περιγραφούν από μια μικρή διαστάσεων $\mathcal{M}$ πολλαπλότητα που είναι εκβυθισμένη σε ένα μικρότερο χώρο διαστάσεων του $\mathbf{R}^n$ τότε μπορεί να μειώσει σημαντικά την υπολογιστική πολυπλοκότητα και το χρόνο των μοντέλων μας με την ανακατασκευή αυτής της πολλαπλότητας. Αυτό ισχύει συχνά για τα δεδομένα ήχου, κειμένου, εικόνας και πολυμέσων όπου οι φορείς εισόδου περιγράφονται από φορείς διαστάσεων υψηλών διαστάσεων στους οποίους οι πραγματικές πληροφοριακές μονάδες θα μπορούσαν να είναι αραιές. Ως αποτέλεσμα, θα μπορούσαμε να προσεγγίσουμε την υποκείμενη κατανομή δεδομένων $p(\mathbf{X})$, όπου $\mathbf{X}$ είναι τα δεδομένα εισόδου, αξιοποιώντας τις διάφορες μεθόδους μάθησης πολλαπλότητα. Είναι ενδιαφέρον ότι οι περισσότερες από αυτές τις τεχνικές μάθησης πολλαπλότητων δεν βασίζονται στις ετικέτες των δεδομένων εισόδου για να συμπεράνουν την κατανομή εισόδου $p(\mathbf{X})$.

Σε σπάνιες περιπτώσεις όπου η κατανομή των δεδομένων εισόδου μπορεί να προσεγγιστεί κατάλληλα χρησιμοποιώντας διαστρωματικούς ή τρισδιάστατους χώρους στόχων, αποκτάμε μια πιο διαισθητική αναπαράσταση της διανομής. Σε γενικές γραμμές, αυτό δεν ισχύει σε χώρους με μεγάλες διαστάσεις όπου λείπει η απεικόνιση των πολλαπλών διανομέων και των χώρων που προκύπτουν. Χωρίς απώλεια της γενικότητας, εστιάζουμε στην οπτικοποίηση διαφόρων πηγών δεδομένων χρησιμοποιώντας αλγόριθμο μείωσης διαστατικότητας t-Distributed Stochastic Neighboring Embedding (t-SNE) [88].

Για παράδειγμα, μπορούμε να πάρουμε μια βαθύτερη εικόνα της υποκείμενης κατανομής πιθανοτήτων των MFCCs, τα οποία είναι ίσως τα πιο κοινά χαρακτηριστικά που χρησιμοποιούνται στην επεξεργασία ομιλίας. Στο σχήμα 1.6 [2] μια μαθηματική 2-διαστάσεων πολλαπλότητα από 512 αρχείων ήχου χρησιμοποιώντας το t-SNE για διάφορες διαμορφώσεις και ως στατιστικές λειτουργίες διανυσμάτων εισόδου που εφαρμόζονται πάνω σε παραστάσεις MFCC με βάση το πλαίσιο. Οι δύο παράμετροι του t-SNE που διασκορπίζονται σε κάθετο και οριζόντιο άξονα είναι η πολυπλοκότητα και ο αριθμός των επαναλήψεων που θα αναγκαστούν να τρέξουν αντίστοιχα. Τώρα, κάθε κυματομορφή ήχου εκφράζεται με ένα διάνυσμα 2-D αντί $13 \cdot N_{stats}$, όπου N_{stats} είναι ο αριθμός των στατιστικών συναρτήσεων που εφαρμόζονται σε όλα τα πλαίσια που αντιπροσωπεύονται από 13 MFCCs. Είναι προφανές ότι μπορούμε να χρησιμοποιήσουμε αυτήν την πολλαπλότητα χαμηλής διάστασης που διατηρεί τις γεωμετρικές ιδιότητες και τη δομή του χώρου μεγάλου διαστάσεων χαρακτηριστικών προκειμένου να εκτελέσουμε τη συσσωμάτωση ή την ταξινόμηση με ένα λιγότερο αραιό χώρο χαρακτηριστικών. Στο σχήμα 1.6 μπορούμε να δούμε ότι εμφανίζονται διαφορετικές περιοχές ενδιαφέροντος στο δισδιάστατο επίπεδο ενώ τα σημεία συγκεντρώνονται σε συγκεκριμένες περιοχές. Πιθανότατα, τα σημεία στα δισδιάστατα επίπεδα που φαίνονται να βρίσκονται αρκετά κοντά ή βρίσκονται σε γειτονικές περιοχές θα ακολουθούσαν μια παρόμοια συμπεριφορά στον χώρο μεγάλης διαστάσεως. Συνεπώς, θα μπορούσαμε να απεικονίσουμε στην πραγματικότητα τη δυνατότητα διακριτικής ικανότητας των MFCCs αν αυτά τα χαρακτηριστικά καθιστούν πράγματι τις αντίστοιχες τάξεις γραμμικά διαχωριζόμενες στον χώρο μεγάλης διαστάσεως.

Στηριζόμενοι σε αυτό, μπορούμε να επωφεληθούμε από τη μείωση του χώρου εισόδου και των διαστάσεών του προκειμένου να απεικονίσουμε τις περιοχές όπου βρίσκονται τα δείγματα κάθε κλάσης. Προφανώς, εάν όλα τα δείγματα κάθε κλάσης βρίσκονται σε περιοχές που διαχωρίζονται εύκολα, τότε η κατανομή εισόδου $p(\mathbf{X})$ των διανυσματικών μεγεθών μεγάλου διαστάσεων θα μπορούσε να προσεγγιστεί χρησιμοποιώντας μόνο μια 2-διαστάσεων πολλαπλότητα $\mathcal{M}$. Παρόλο που η εστίασή μας έγκειται στην επεξεργασία ομιλίας και ειδικά στο SER, παρουσιάζουμε τα αποτελέσματα των 2-διαστάσεων πολλαπλοτήτων χρησιμοποιώντας το t-SNE τόσο στην αναγνώριση εικόνας όσο και

[2] Το σχήμα βρέθηκε στο: http://alexanderfabisch.github.io/t-sne-in-scikit-learn.html



Σχήμα 1.6: Μάθηση πολλαπλοτήτων δύο διαστάσεων χρησιμοποιώντας t-SNE με διαφορετική παραμετροποίηση για μια ποικιλία ήχων

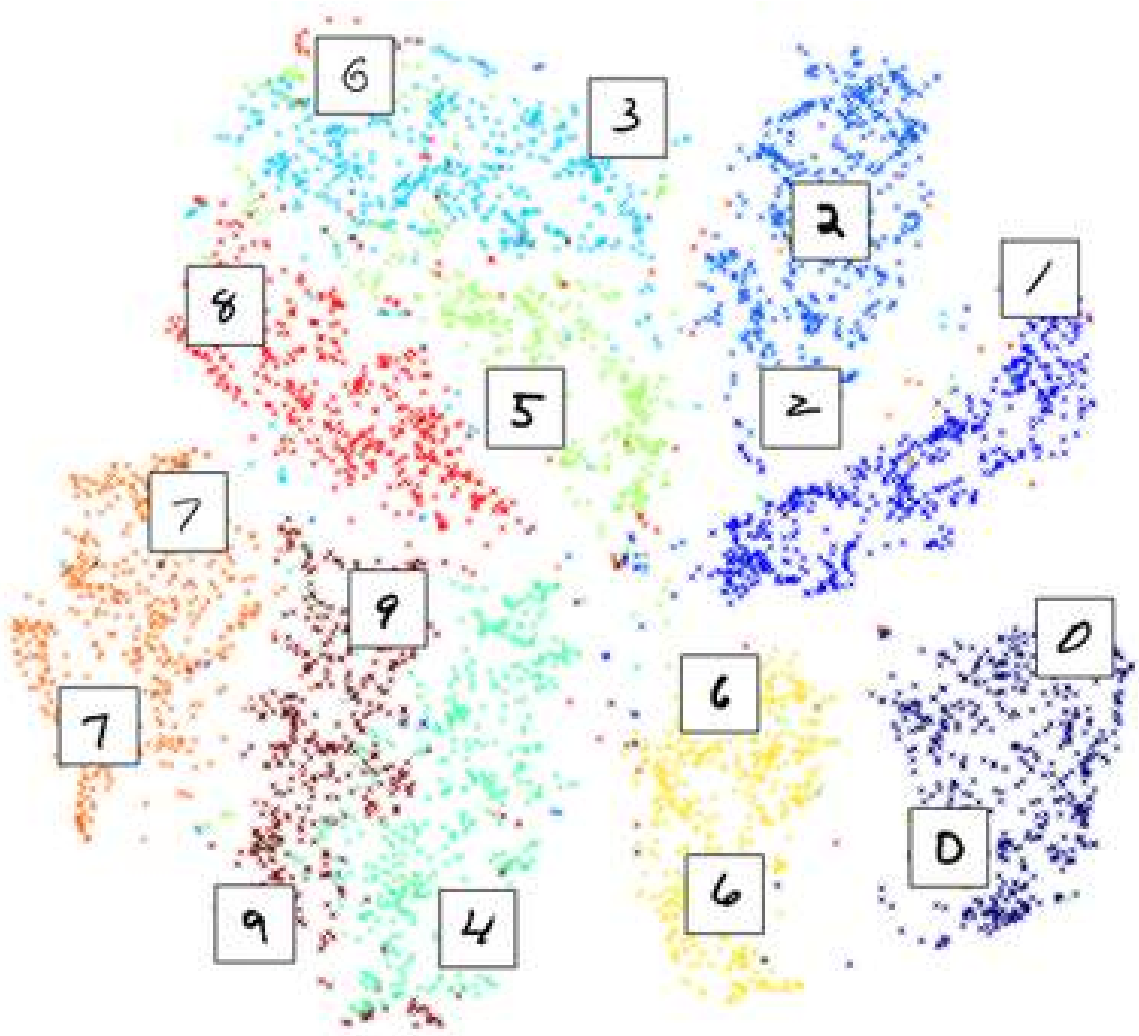
στην ταξινόμηση κειμένου στα Figures 1.7 [3] και 1.8 [4], αντίστοιχα.

Στο σχήμα 1.7, οι εικόνες του MNIST dataset [89], οι οποίες είναι 28×28 εικόνες χειρόγραφων ψηφίων χρησιμοποιούνται ως είσοδος για τον t-SNE. Το διάνυσμα εισαγωγής 784 πίξελς για κάθε εικόνα μειώνεται σε ένα διδιάστατο διάνυσμα με μια μη γραμμική διαδικασία εκμάθησης πολλαπλοτήτων. Όπως φαίνεται στο σχήμα, η μάθηση πολλαπλότητας παρέχει μια ισχυρή ένδειξη επί της όρασης ότι αυτή η τεχνική παράγει δισδιάστατες αναπαραστάσεις όπου οι κλάσεις όλων των δειγμάτων είναι γραμμικά διαχωρίσιμες μεταξύ τους. Προκειμένου να προσφέρουμε ένα πλαίσιο για την ταξινόμηση, πρέπει πρώτα να εκτελέσουμε τον αλγόριθμο μάθησης πολλαπλότητας στην υψηλή διάσταση εισόδου, να πάρουμε τις νέες διαστάσεις χαμηλής διάστασης και έπειτα να αξιολογήσουμε κάθε ταξινομητή στις παραστάσεις με μειωμένες διαστάσεις. Επιπλέον, το ίδιο ισχύει και σε άλλους τομείς, όπως η ταξινόμηση κειμένου, όπου οι συνιστώσες εισόδου αντιστοιχούν στις αναπαραστάσεις του φορέα των λέξεων. Στο σχήμα 1.8, οι δισδιάστατες αναπαραστάσεις που παίρνουμε μετά την εφαρμογή t-SNE κατά την αρχική ενσωμάτωση λέξεων εμφανίζονται για μια παρτίδα λέξεων που ανήκουν σε διάφορα περιβάλλοντα. Όπως είναι επίσης εμφανές με μια οπτική επιθεώρηση του σχήμα-

[3] Το σχήμα βρέθηκε στο: https://medium.com/@LeonFedden/comparative-audio-analysis

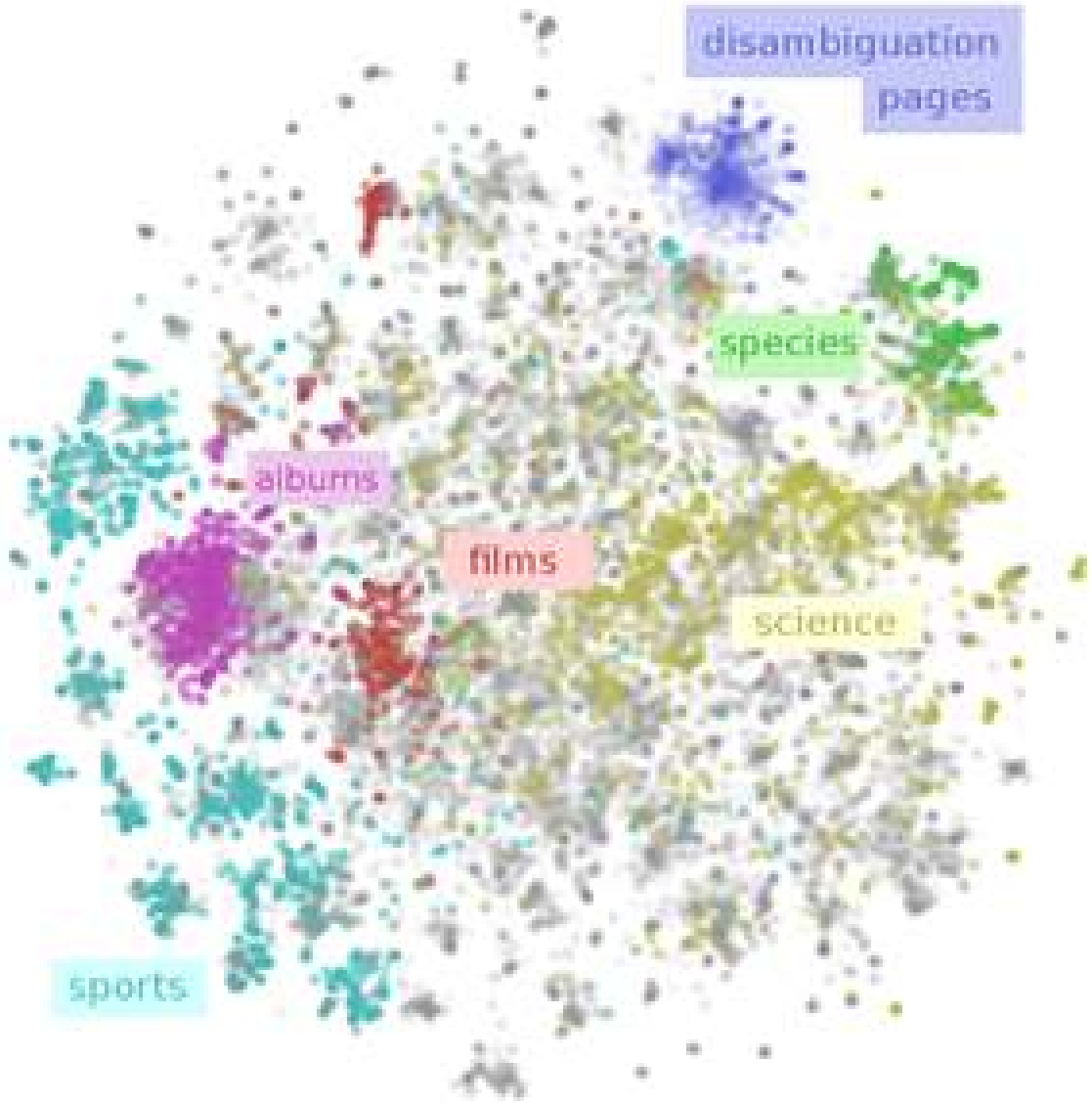
[4] Το σχήμα βρέθηκε στο: http://colah.github.io/posts/2015-01-Visualizing-Representations/

τος, οι λέξεις σε παρόμοια περιβάλλοντα εμφανίζονται παρόμοια στον δισδιάστατο χάρτη. Αυτή είναι μια έντονη ένδειξη ότι οι επαγόμενες λέξεις ενσωματώνουν διακριτικές πληροφορίες για το πλαίσιο στο οποίο κάθε λέξη ανήκει, αλλά κάτω από μια αραιή αναπαράσταση, όσον αφορά την πυκνότητα των περιεχομένων πληροφοριών. Συνολικά, μπορούμε να υποθέσουμε ότι μπορούμε να επεκτείνουμε αυτές τις μεθόδους για το SER και ενδεχομένως να μειώσουμε τη διαστατικότητα των χρησιμοποιούμενων ακουστικών χαρακτηριστικών δραστικά χωρίς να χάσουμε πολλά από την άποψη της απόδοσης αναγνώρισης.



Σχήμα 1.7: 2D πολλαπλότητα της βάσης δεδομένων MNIST που έμαθε μια εκτέλεση t-SNE

Είναι σημαντικό τα δεδομένα εισόδου να προσφέρουν μια επαρκή δειγματοληψία της υποκείμενης κατανομής προκειμένου να είναι σε θέση η $\mathcal{M}$ να ανασυγκροτήσει την προαναφερθείσα γεωμετρία του χώρου. Διαφορετικά, οι προσεγγίσεις μάθησης πολλαπλοτήτων όπως οποιαδήποτε άλλη τεχνική εκμάθησης μηχανών θα θεωρούνται ανεπιτυχείς. Παρόλο που χρησιμοποιήσαμε μόνο τον t-SNE για την προηγούμενη ανάλυση απεικόνισης, υπάρχει ένας τεράστιος αριθμός διαφορετικών αλγορίθμων μάθησης πολλαπλοτήτων παρουσιαζόμενων στη βιβλιογραφία και δεν γνωρίζουμε ποια θα αποφέρει την πιο καλή λύση για ακουστικά χαρακτηριστικά που χρησιμοποιούνται στο SER. Δεν υπάρχει καλύτερος αλγόριθμος για την μάθηση πολλαπλοτήτων, αλλά υπάρχει ένα ενδιαφέρον φαινόμενο αντιστρόφων συγκινονούντων δοχείων μεταξύ της εκφραστικότητας, της απαιτούμενης μνήμης και του χρόνου εκτέλεσης [86]. Η εύρεση μιας ισχυρής προσέγγισης μάθησης πολλαπλοτήτων είναι πιο κοντά σε ένα πιο παγκόσμιο πρόβλημα μηχανικής μάθησης λόγω της γενικότητας των παρεχόμενων λύσεων που μπορούν να χρησιμοποιηθούν επαρκώς σε διαφορετικά καθήκοντα (όπως είδαμε προηγουμένως



Σχήμα 1.8: 2D πολλαπλότητα των εκβυθισμένων διανυσμάτων λέξεων για μια ποικιλία συμφραζομένων

με τον t-SNE) χωρίς καμία παραδοχή σχετικά με την πραγματική ταυτότητα της πηγής δεδομένα που χρησιμοποιούνται ως είσοδοι.

Παρόλα αυτά, οι περισσότεροι από τους αλγορίθμους που προτείνονται στη βιβλιογραφία βασίζονται στον υπολογισμό παραγώγων τα οποία είναι υπολογιστικά αναποτελεσματικά και λιγότερο κλιμακούμενα σε προβλήματα όπου η ομαλότητα δεν μπορεί να θεωρηθεί. Συγκεκριμένα, η πλειονότητα αυτών των αλγορίθμων μπορεί να μειωθεί στο πρόβλημα της βελτιστοποίηση μιας ντετερμινιστικής συνάρτησης απώλειας f. Δεδομένου αυτού του στόχου ελαχιστοποίησης, συνήθως χρησιμοποιούν μεθόδους που βασίζονται στον υπολογισμό της παραγώγου αυτής της συνάρτησης-στόχου για να βρουν ένα ολικό ή τοπικό βέλτιστο. Σε πολλές περιπτώσεις, ωστόσο, η συνάρτηση απώλειας είναι μη παραγωγίσιμη ή η εκτίμηση της κλίσης της μπορεί να είναι υπολογιστικά δαπανηρή. Επιπλέον, οι αλγόριθμοι που βασίζονται σε κλίση συνήθως αποδίδουν μια αργή σύγκλιση, καθώς απαιτούνται πολλαπλές επαναλήψεις προκειμένου να ελαχιστοποιηθεί η συνάρτηση απώλειας ικανοποιητικά.

Όλες οι προαναφερθείσες αναλύσεις δημιουργούν ένα τεράστιο σύνολο προκλήσεων που πρέπει να αντιμετωπιστούν και ερωτήσεις που πρέπει να απαντηθούν.

- Θα μπορούσαμε να προσφέρουμε μια νέα, εκφραστική και ταυτόχρονα αποτελεσματική προ-

σέγγιση μάθησης πολλαπλοτήτων προκειμένου να πραγματοποιήσουμε τη μη γραμμική μείωση των διαστάσεων;

- Θα μπορούσαμε να αποφύγουμε τον υπολογισμό των παραγώγων χωρίς να χάσουμε την ικανότητα της προσέγγισης των πολλαπλοτήτων;
- Πώς μπορούμε να προσφέρουμε μια θεωρητική απόδειξη για τη σύγκλιση ενός τέτοιου αλγορίθμου;
- Αυτός ο αλγόριθμος παρέχει μια γενική λύση για τις αποκλίνουσες πηγές δεδομένων χωρίς να κάνει οποιαδήποτε παραδοχή σχετικά με την πηγή όπου εξάγονται τα διανύσματα χαρακτηριστικών;
- Ο πειραματισμός σε ακουστικά σύνολα χαρακτηριστικών για δεδομένα SER προσφέρει σημαντική βελτίωση;
- Μπορούμε να χρησιμοποιήσουμε τον προτεινόμενο αλγόριθμο πολλαπλότητας μάθησης προκειμένου να παρέχουμε απεικονίσεις των δισδιάστατων χαρτών διαφορετικών συναισθημάτων, που παράγονται από μια μη γραμμική μείωση των διαστάσεων των ακουστικών χαρακτηριστικών εισόδου;
- Οι συναισθηματικοί χάρτες από ακουστικά χαρακτηριστικά θα ήταν παρόμοιοι με αυτούς που παρουσιάζονται στα σχήματα 1.7 και 1.8 και θα παρέχουν διορατικές ποιοτικές πληροφορίες σχετικά με τις αναπαραστάσεις που δημιουργούνται;

1.5 Στόχοι και Συνεισφορές

Οι στόχοι αυτής της διπλωματικής καθοδηγούνται από τα ερωτήματα που τίθενται στην Ενότητα 1.4 για όλους τους υπο-τομείς ενδιαφέροντος. Με άλλα λόγια, η επιτυχία αυτής της εργασίας θα μπορούσε να αξιολογηθεί με βάση:

- Είτε οι προσεγγίσεις, οι παρατηρήσεις και τα αποτελέσματα μας δίνουν τις απαντήσεις στα ερωτήματα που τίθενται στην Ενότητα 1.4.
- Εάν αντιμετωπίσουμε τις προκλήσεις και αντιμετωπίσουμε το πλήρες φάσμα του ενδιαφέροντός μας, όπως περιγράφονται στο 1.4.

Αυτή η εργασία έχει τριπλή συμβολή, η οποία μπορεί να περιγραφεί συνοπτικά από τα ακόλουθα:

1. **Εξετάζουμε την αποτελεσματικότητα της επιλογής της κατάλληλης χρονικής κλίμακας όταν χρησιμοποιούμε RNNs για SER:** Δείχνουμε ότι η επιλογή της κατάλληλης χρονικής κλίμακας για LLD (τοπικά χαρακτηριστικά) και στατιστικά (γενικά χαρακτηριστικά) είναι το κλειδί για ένα σύστημα SER υψηλής απόδοσης. Αξιολογούμε την προσέγγισή μας χρησιμοποιώντας τοπικά και γενικά χαρακτηριστικά και αξιολογούμε την απόδοση σε διάφορες χρονικές κλίμακες (πλαίσιο, φωνή, λέξη ή ομιλία). Η ανάλυση σε διάφορες χρονικλες κλίμακες παρέχει τόσο ποιοτικές όσο και ποσοτικές πληροφορίες για το πώς τα διαφορετικά σύνολα χαρακτηριστικών συλλαμβάνουν τις πληροφορίες σε διαφορετικά χρονικες κλίμακες. Δείχνουμε ότι για τα μοντέλα RNN, η εξαγωγή στατιστικών χαρακτηριστικών πάνω σε τμήματα ομιλίας που αντιστοιχούν περίπου στη διάρκεια μερικών λέξεων παράγει την καλύτερη ακρίβεια. Αναφέρουμε την καλύτερη απόδοση SER στην βάση δεδομένων IEMOCAP με σημαντικά χαμηλότερο μοντέλο απο άποψη υπολογιστικής πολυπλοκότητας και μηχανισμών.
2. **Παρουσιάζουμε ένα νέο σύνολο μη γραμμικών χαρακτηριστικών που μπορούν να καταγράψουν τα πρότυπα της δυναμικής επαναληψιμότητας και υποτροπιαζουσών δυναμικών και τα αξιολογούμε κάτω από διαφορετικά μοντέλα και πειράματα SER:** Διερευνούμε την

απόδοση χαρακτηριστικών που μπορούν να καταγράψουν τη μη γραμμική δυναμική επαναληψιμότητας ενσωματωμένη στο σήμα ομιλίας για SER. Η ανακατασκευή του χώρου φάσης κάθε πλαισίου ομιλίας και ο υπολογισμός των αντίστοιχων RP αποκαλύπτει πολύπλοκες δομές που μπορούν να μετρηθούν με την εκτέλεση του RQA. Αυτά τα μέτρα συγκεντρώνονται χρησιμοποιώντας στατιστικές λειτουργίες σε περιόδους περιορισμού και έκφρασης. Αναφέρουμε τα αποτελέσματα SER για την προτεινόμενη λειτουργία σε τρεις βάσεις δεδομένων χρησιμοποιώντας διαφορετικές μεθόδους ταξινόμησης. Όταν συνδυάζουμε τα προτεινόμενα χαρακτηριστικά με τα παραδοσιακά σύνολα χαρακτηριστικών, π.χ. [23], εμφανίζουμε μια βελτίωση στην μη σταθμισμένη ακρίβεια έως και 5.7% και 10.7% για πειράματα SER Εξαρτημένου-Ομιλητή (SD) και Ανεξαρτήτως-Ομιλητή (SI), αντίστοιχα, πάνω από το ποσοστό που πετυχαίνουμε χρησιμοποιώντας μονο τα χαρακτηριστικα IS10 [23]. Μετά από μια προσέγγιση που βασίζεται σε τμήματα, επιδεικνύουμε τις καλύτερες επιδόσεις στο IEMOCAP χρησιμοποιώντας αμφίδρομο επαναλαμβανόμενο νευρικό δίκτυο με μηχανισμό προσοχής στην κορυφή.

3. **Προτείνουμε έναν νέο αλγόριθμο για μάθηση πολλαπλοτήτων που βασίζεται στη βελτιστοποίηση χωρίς παραγώγιση αντί για τις συμβατικές προσεγγίσεις που βασίζονται στον υπολογισμό της κλίσης σε κάθε εποχή, τον Pattern Search MDS:** Οι τελευταίες προσεγγίσεις απαιτούν τον υπολογισμό της κλίσης μίας συνάρτησης απώλειας προκειμένου να επιλεγεί η βέλτιστη κίνηση στον χώρο εξερεύνησης. Ωστόσο, ο αλγόριθμός μας δεν απαιτεί ομαλές ιδιότητες της συνάρτησης απώλειας που πρέπει να ελαχιστοποιηθεί, καθώς και οποιαδήποτε παραδοχή σχετικά με τη διαφοροποίησή του. Συγκεκριμένα, προτείνουμε μια επέκταση της κλασσικής μεθόδου MDS, όπου αντί να εκτελούμε τον αλγόριθμο της μείωσης της κλίσης, δοκιμάζουμε και αξιολογούμε πιθανές "κινήσεις" σε μια σφαίρα σταθερής ακτίνας για κάθε σημείο του ενσωματωμένου χώρου. Μια θεωρητική απόδειξη της σύγκλισης του αλγορίθμου σε σταθερό σημείο μπορεί να αποδειχθεί με τη διαμόρφωση του προτεινόμενου αλγορίθμου ως στιγμιότυπου της υπερκλάσσης των μεθόδων αναζήτησης γενικού προτύπου (GPS). Η αξιολόγηση τόσο σε καθαρά όσο και σε θορυβώδη συνθετικά σύνολα δεδομένων δείχνει ότι το Pattern Search MDS μπορεί να συμπεράνει με ακρίβεια την εγγενή γεωμετρία των πολλαπλών που είναι ενσωματωμένες σε χώρους μεγάλης διαστάσεως. Επιπλέον, τα πειράματα σε πραγματικά δεδομένα, ακόμη και υπό θορυβώδεις συνθήκες, αποδεικνύουν ότι η προτεινόμενη αναζήτηση μοτίβου MDS αποδίδει ανταγωνιστικά αποτελέσματα σε σύγκριση με την βιβλιογραφία σε μια ποικιλία εργασιών όπως η ταξινόμηση της εικόνας και η σημασιολογική ομοιότητα των ενσωματώσεων κειμένου. Για τα πειράματα SER χρησιμοποιούμε τα σύνολα χαρακτηριστικών που περιγράφονται στην ενότητα 4 καθώς και ο συνδυασμός τους και οραματιζόμαστε τους μαθημένους 2 -μεταβλητικούς συναισθηματικούς χάρτες αυτών των ακουστικών συνόλων.

1.6 Οργάνωση αυτής της Διπλωματικής Εργασίας

Το υπόλοιπο της παρούσας εργασίας οργανώνεται ως εξής. Διαχωρίζουμε τις προαναφερθείσες συνεισφορές (βλέπε Ενότητα 1.5) σε τρία κεφάλαια 3, 4, 5. Κάθε κεφάλαιο μπορεί να θεωρηθεί αυτοτελές από άποψη σημειογραφίας, σχετικής εργασίας γύρω από το θέμα που αναλύει, πειράματα και συμπεράσματα που αντλούνται από τα αποτελέσματα του πρώτου. Αυτό το μοντέλο οργάνωσης έχει επιλεγεί λόγω των διαφορετικών μεθόδων, μοντέλων και θεμάτων που παρουσιάζονται σε κάθε κεφάλαιο.

Πρώτον, στο Κεφάλαιο 2, εισάγουμε κάποια βασική μαθηματική σημειογραφία (Section 2.1), αναλύουμε μερικές μεθόδους ταξινόμησης και μάθηση με επίβλεψη κάτω από την οποία εκπαιδεύουμε τα μοντέλα μας (Ενότητα 2.2). Παρουσιάζουμε επίσης τη μαθηματική διατύπωση και τη χρήση των RNN καθώς και τον τρόπο με τον οποίο είναι δομημένα και συνδυασμένα με μηχανισμούς προσοχής και αμφίδρομες τοπολογίες (Τμήμα 2.3). Επιπλέον, παρουσιάζουμε την επισημοποίηση για την ανασυγκρότηση PS, την εξαγωγή των RP και τα αντίστοιχα μέτρα RQA στα τμήματα 2.4.1, 2.5 και 2.6. Στις ενότητες 2.7 και 2.8 παρέχουμε τη μαθηματική διατύπωση για μεθόδους πολυδιάστατης κλιμάκωσης (MDS) και γενικής αναζήτησης προτύπων (GPS).

Το πρώτο μέρος αυτής της εργασίας αποτελείται από την ανάλυση των συστημάτων SER υπό διαφορετικές χρονικές κλίμακες και ειδικά για τα συστήματα SER που βασίζονται σε RNN. Η προαναφερθείσα ανάλυση καλύπτεται στο κεφάλαιο 3 όπου αναλύουμε την αποτελεσματικότητα διαφορετικών χρονικών κλιμάκων όταν χρησιμοποιούμε RNNs για SER. Ορισμένα προηγούμενα κίνητρα για αυτό το μέρος της εργασίας δίνονται στην ενότητα 3.1 και μια εκτεταμένη ανασκόπηση καθώς και ποιοτική ανάλυση των μεθόδων για τεχνικές παγκόσμιας και τοπικής εξαγωγής ακουστικών χαρακτηριστικών παρουσιάζεται στην ενότητα 3.2. Επιπλέον, τα προτεινόμενα συστήματα SER συζητούνται στην ενότητα 3.3 ενώ η προετοιμασία για τη διεξαγωγή πειραμάτων με αυτές τις αρχιτεκτονικές παρέχεται στην ενότητα 3.4. Τα τελικά αποτελέσματα και τα ευρήματα σχετικά με τον τρόπο με τον οποίο οι χρονικές κλίμακες απόφασης επηρεάζουν την απόδοση συστημάτων SER καθώς και μια σύγκριση με τη βιβλιογραφία παρουσιάζεται στην ενότητα 3.5.

Το δεύτερο μέρος της παρούσας διπλωματικής εργασίας αποτελείται από το κεφάλαιο 4 στο οποίο διερευνάμε την ενσωμάτωση των μη γραμμικών ακουστικών χαρακτηριστικών που αντικατοπτρίζουν την περίπλοκη δυναμική επαναληψιμότητας του συναισθηματικού λόγου. Συγκεκριμένα, στην ενότητα 4.1 συζητάμε τις ιδέες που οδήγησαν αυτή την προσέγγιση ενώ στην ενότητα 4.2 παραθέτουμε ορισμένες από τις πιο σημαντικές προσεγγίσεις για τη χρήση πληροφοριών που προέρχονται από RPs και γενικά μη γραμμικά χαρακτηριστικά ομιλίας SER. Στα τμήματα 4.3 και 4.4, καθορίζουμε πώς υπολογίζονται τα PS και RP για κάθε πλαίσιο ομιλίας και παρέχουμε ποιοτικά ευρήματα για αυτές τις δομές για διαφορετικούς ομιλητές και συναισθηματικές εκδηλώσεις. Καθορίζουμε αυστηρά τα σύνολα ακουστικών χαρακτηριστικών που θα εξαχθούν για το SER και τις μεθόδους ταξινόμησης που θα χρησιμοποιήσουμε στα τμήματα 4.5 και 4.6, αντίστοιχα. Παρουσιάζουμε τα σύνολα δεδομένων που χρησιμοποιούμε για την αξιολόγηση και των τριών συνόλων χαρακτηριστικών σε όλα τα πειράματα SER, καθώς και τα τελικά αποτελέσματα σε σύγκριση με τις προσεγγίσεις της βιβλιογραφίας στην ενότητα 4.7.

Το τελευταίο μέρος αυτής της εργασίας αποτελείται από το κεφάλαιο 5 όπου προτείνουμε τον αλγόριθμο Pattern Search MDS για να επιτύχουμε μη γραμμική μείωση διαστασιμότητας χωρίς τον υπολογισμό της κλίσης σε κάθε εποχή. Στην ενότητα 5.1 συζητάμε τις ιδέες που οδήγησαν αυτή την προσέγγιση ενώ στην ενότητα 5.2 συζητούμε εν συντομία κάποιες από τις πιο σημαντικές προσεγγίσεις που βρέθηκαν στη βιβλιογραφία για να μειώσουμε τη διαστασιμότητα. Περιγράφουμε τον προτεινόμενο αλγόριθμο στην ενότητα 5.3 και προτείνουμε επίσης μερικές από τις προσεγγίσεις προκειμένου να τερματίσουμε τον αλγόριθμό μας γρηγορότερα. Επίσης συζητήσουμε μερικά trade-offs και επιταχύνσεις στο τμήμα 5.4. Μειώνουμε τον προτεινόμενο αλγόριθμο στη γενική τάξη μεθόδων GPS στην ενότητα 5.5 και αποδεικνύουμε τη σύγκλιση του θεωρητικά, χρησιμοποιώντας αποδεδειγμένα θεωρήματα για το ενοποιημένο πλαίσιο GPS στο τμήμα 5.6. Αξιολογούμε τον αλγόριθμό μας για διάφορες εργασίες αναγνώρισης και εργασίες διατήρησης γεωμετρίας για τη λύση που παρέχεται από τον αλγόριθμό μας στην Ενότητα 5.7. Συγκεκριμένα, δοκιμάζουμε τον αλγόριθμό μας για τον SER χρησιμοποιώντας πανομοιότυπες πειραματικές ρυθμίσεις και σύνολα ακουστικών χαρακτηριστικών που εισήχθησαν στο Κεφάλαιο 4 (βλέπε Ενότητα 5.7.7). Τέλος, παρέχουμε μια σύγκριση για τις παραγόμενες πολλαπλές χρήσεις από τις διάφορες μεθόδους μείωσης των διαστάσεων στην ενότητα 5.8 σε $\mathbb{R}^2$ και $\mathbb{R}^3$.

Τέλος, η περιγραφή αυτής της εργασίας είναι το Κεφάλαιο 6 όπου συνοψίζουμε τα συμπεράσματα που έχουμε αντλήσει από την ανάλυση όλων των προηγούμενων Κεφαλαίων (βλέπε Κεφάλαιο 6.1) καθώς επίσης επισημαίνουμε κάποιες πιθανές κατευθύνσεις που θα μπορούσαν να βασιστούν σε αυτό το έργο (βλ. Ενότητα 6.2).

Κεφάλαιο 2

Τεχνικό Υπόβαθρο

2.1 Σημειογραφία

Δηλώνουμε πραγματικούς, ακέραιους και φυσικούς αριθμούς ως $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{N}$ αντίστοιχα. Οι αριθμοί αντιπροσωπεύονται από γράμματα χωρίς έντονα γράμματα, τα διανύσματα με έντονα και οι μήτρες σημειώνονται με κεφαλαία γράμματα. Όλα τα διανύσματα θεωρούνται διανύσματα στήλης εκτός αν ορίζονται σαφώς ως διανύσματα σειράς. Για ένα διάνυσμα $\mathbf{z} \in \mathbb{R}^n$, $\|\mathbf{z}\|_1 = \sum_{i=1}^{n} |z_i|$ είναι η ℓ_1 νόρμα και $\|\mathbf{z}\|_2 = \sqrt{\sum_{i=1}^{n} z_i^2}$ είναι η ℓ_2 νόρμα, όπου z_i είναι το i-στοιχείο του $\mathbf{z}$. Επιπλέον, ο τελευταίος συμβολισμός θα μπορούσε να οριστεί είτε ως $z(i)$. Με το $\mathbf{A} \in \mathbb{R}^{n \times m}$ υποδηλώνουμε έναν πραγματικό πίνακα με n σειρές και m στήλες. Επιπρόσθετα, η στήλη j του matrix $\mathbf{A}$ και η καταχώρησή της στη γραμμή i και στήλη j αναφέρονται ως $\mathbf{a}_j$ και a_{ij}, αντίστοιχα. Μπορούμε επίσης να ορίσουμε την καταχώρηση μιας μήτρας $\mathbf{A}$ στη στήλη i a και j th ως $\mathbf{A}_{ij}$ αλλά αυτό θα οριστεί ειδικά για κάθε εξίσωση χωριστά. Το ίχνος του matrix $\mathbf{A}$ εμφανίζεται ως $tr(\mathbf{A})$ και η Frobenious νόρμα του ορίζεται ως $||\mathbf{A}||_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2}$. Ο τετραγωνικός πίνακας ταυτότητας με n rows σημειώνεται ως $\mathbf{I}_n \in \mathbb{R}^{n \times n}$. Για τους πίνακες $\mathbf{A} \in \mathbb{R}^{n \times m}$ και $\mathbf{B} \in \mathbb{R}^{n \times m}$ ορίζουμε το Hadamard προϊόν τους $\mathbf{A} \odot \mathbf{B}$. Το n-οστό καρτεσιανό προϊόν πάνω από n σύνολα $S_1, ..., S_n$ σημειώνεται με $\{(s_1, ..., s_n) : s_i \in S_i$. Ως $\mathbf{X}^{(k)}$ αναφέρεται η εκτίμηση μιας μεταβλητής $\mathbf{X}$ στην k-οστή επανάληψη ενός επαναληπτικού αλγόριθμου. Ορίζουμε την πιθανότητα του γεγονότος Ω δεδομένου ότι το συμβάν Ω' έχει ήδη συμβεί με: $p(\Omega|\Omega')$. Για να γίνει ευκολότερη η γενίκευση των συναρτήσεων $f : \mathbb{R} \to \mathbb{R}$ σε διανύσματα πραγματικών αριθμών $\mathbf{x} = [x_1, ..., x_n]$, υποθέτουμε ότι η άμεση εφαρμογή της συνάρτησης στο φορέα που χρησιμοποιούμε $f(\mathbf{x})$ ως διαισθητική επέκταση της συνάρτησης f στο $\mathbb{R}^n$ όπου $f(\mathbf{x}) = [f(x_1), ..., f(x_n)]$. Δηλώνουμε με $\mathbf{a}||\mathbf{b}$ τη σύζευξη των διανυσμάτων $\mathbf{a}$ και $\mathbf{b}$. Ορίζουμε τον πολλαπλασιασμό στοιχείων για δύο διανύσματα $\mathbf{a}$ και $\mathbf{b}$ με $\mathbf{a} \odot \mathbf{b}$ όμοια με πρίν.

2.2 Μοντέλα Ταξινόμησης

Σε αυτό το κεφάλαιο θα αναλύσουμε τα μοντέλα ταξινόμησης που χρησιμοποιούνται σε αυτή τη διπλωματική εργασία. Σε αυτή την εργασία, όταν αντιμετωπίζουμε μια εργασία ταξινόμησης, τότε θα αναλάβουμε μια εποπτευόμενη εργασία εκμάθησης μηχανής, όπου είναι γνωστή η ετικέτα $\mathcal{Y}_i$ κάθε δείγματος $\mathbf{x}_i$ του εκπαιδευτικού σετ. Σε γενικές γραμμές, υπάρχουν πολλές άλλες ρυθμίσεις όπου είναι γνωστές κάποιες ετικέτες (ημι-εποπτευόμενη μάθηση - semi-supervised learning) καθώς και ρυθμίσεις όπου δεν διατίθεται ετικέτα των δεδομένων εισόδου για τον χρόνο εκπαίδευσης (μη επιτηρούμενη μάθηση - unsupervised-learning) [58]. Αφού εκπαιδεύσουμε το μοντέλο μας, αξιολογούμε το πρότυπό μας χρησιμοποιώντας σε ένα ξεχωριστό υποσύνολο των διαθέσιμων δεδομένων που ονομάζεται “σύνολο δοκιμής” ή “test-set”. Τα σύνολα αυτά πρέπει να έχουν κενή τομή.

2.2.1 Συνάρτηση Απώλειας

Βελτιστοποιούμε τις παραμέτρους $\boldsymbol{\theta}$ του μοντέλου μας, ελαχιστοποιώντας μια αντικειμενική συνάρτηση $\mathcal{J}(\boldsymbol{\theta})$ (επίσης αποκαλούμενη συνάρτηση απώλειας). Στην εποπτευόμενη εκπαίδευση, η συνάρτηση απώλειας ορίζεται επίσης από την εκπαίδευση $\mathbf{X}_{\mathbf{train}}$ και τις αντίστοιχες ετικέτες $\mathcal{Y}_{train}$.

Σε SER και σε άλλα πειράματα έχουμε συνήθως πάνω από δύο διακριτές ετικέτες για τα δεδομένα μας και ως εκ τούτου χρησιμοποιούμε μια κατηγορική σταυροειδή εντροπία ως συνάρτηση στόχο που προσπαθούμε να ελαχιστοποιήσουμε. Για παράδειγμα, αν έχουμε δείγματα εκπαίδευσης N και $\mathcal{C}$ συναισθηματικές ετικέτες, τότε μπορούμε να μετατρέψουμε τις κατηγοριοποιημένες ετικέτες σε ένα διάνυσμα μήκους $\mathcal{C}$. Στην τελευταία παράσταση, κάθε διάνυσμα έχει μηδενικά παντού εκτός από τον δείκτη που αντιστοιχεί στην κατηγοριακή ετικέτα ενός δείγματος $\mathbf{x}_k$ όπου έχει την τιμή 1. Συγκεκριμένα, μετατρέπουμε την κατηγοριοποιημένη ετικέτα ως εξής:

$$\mathcal{Y}_k \rightarrow \quad \mathbf{y}_k = [0, ..., \underbrace{1}_{corresponding\ \ label\ \ index\ \ c}, ..., 0] \tag{2.1}$$

Τώρα ο δείκτης c σε αυτή την διάνυσμα αντιπροσωπεύει την ίδια κλάση με $\mathcal{Y}_k$ η οποία είναι η αληθινή σχολιασμένη ετικέτα που αντιστοιχεί στο δείγμα $\mathbf{x}_k$. Σε αυτή την εποπτευόμενη πειραματική διάταξη μάθησης, κάθε δείγμα αντιστοιχεί σε μια πλειάδα 1) μιας αναπαράστασης διάνυσματος δεδομένων $\mathbf{x}_k$ και 2) μιας διανυσματικής απεικόνισης $\mathbf{y}_k$ που αφορεί την αντίστοιχη ετικέτα του δείγματος αυτού.

Υποθέτοντας ότι το μοντέλο μας εκπαιδεύεται σε ένα σύνολο N δειγμάτων δεδομένων $\{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^{N}$ η πρόβλεψη $\mathbf{y}_k$ για την k-οστή τούπλα ορίζεται ως η posterior πιθανότητα για όλες τις διαθέσιμες κλάσεις $\mathcal{C}$ δεδομένων των παραμέτρων του μοντέλου $\boldsymbol{\theta}$ και των δεδομένων του αντιπροσωπευτικού διανύσματος $\mathbf{x}_k$. Αν ορίζουμε ως $\mathbf{h}(\mathbf{x}_k, \boldsymbol{\theta})$ τις ενεργοποιήσεις του μοντέλου μας που χρησιμοποιούνται για την εξαγωγή της ετικέτας ενός δεδομένου δείγματος $\mathbf{x}_k$ μπορούμε να υποθέσουμε ότι το προβλεπόμενο διάνυσμα κλάσης $\hat{\mathbf{y}}_k$ είναι καθορισμένο στοιχείο προς στοιχείο από τις αντίστοιχες πιθανότητες posteriors:

$$\hat{\mathbf{y}}_k = [p(\hat{\mathbf{y}}_{k,1}|\boldsymbol{\theta}, \mathbf{x}_k), ..., p(\hat{\mathbf{y}}_{k,\mathcal{C}}|\boldsymbol{\theta}, \mathbf{x}_k)] \tag{2.2}$$

Επομένως, η συνάρτηση απώλειας για το δείγμα k εμφανίζεται στη συνέχεια:

$$\mathcal{J}(\boldsymbol{\theta})_k = -\mathbf{y}_k \cdot log(\hat{\mathbf{y}}_k) \tag{2.3}$$

Το διάνυσμα πρόβλεψης του μοντέλου μας αντιστοιχεί στην πιθανότητα των αντίστοιχων posteriors για κάθε κατηγορία. Εάν θέλουμε να βελτιστοποιήσουμε τις παραμέτρους του μοντέλου μας μεγιστοποιώντας την πιθανότητα του μοντέλου μας, μπορούμε εύκολα να το κάνουμε αντίστροφα, ελαχιστοποιώντας τον μέσο όρο της πιθανότητας αρνητικού λογαρίθμου από όλα τα διαθέσιμα δείγματα κατάρτισης N. Η συνάρηση στόχος εμφανίζεται στην ακόλουθη εξίσωση:

$$\mathcal{J}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{k=1}^{N} \mathbf{y}_k \cdot log(\hat{\mathbf{y}}_k) \tag{2.4}$$

Όπου N είναι ο αριθμός των δειγμάτων κατάρτισης. Ουσιαστικά, η προαναφερθείσα εξίσωση είναι ιδιαίτερα χρήσιμη όταν εκπαιδεύουμε ένα NN χρησιμοποιώντας την τιμή της συνάρτησης απώλειας $\mathcal{J}(\boldsymbol{\theta})$ για τον υπολογισμό του σφάλματος του NN όσον αφορά τις αποφάσεις κατάταξης που έκανε για τα N δείγματα. Στις περισσότερες περιπτώσεις δεν χρησιμοποιούμε ολόκληρο το σύνολο δεδομένων σε κάθε επανάληψη της διαδικασίας κατάρτισης, αλλά αντ 'αυτού υπολογίζουμε το σφάλμα πάνω από τα υποσύνολα του σετ εκπαίδευσης (αυτά τα σύνολα ονομάζονται συχνά παρτίδες (batches) όταν το μοντέλο μας είναι νευρωνικό δίκτυο). Για να μάθουμε τις βέλτιστες παραμέτρους για τα μοντέλα μας μπορούμε να χρησιμοποιήσουμε οποιοδήποτε εργαλείο βελτιστοποίησης παραμέτρων που βασίζεται στον υπολογισμό της κλίσης $\nabla_{\boldsymbol{\theta}}\mathcal{J}(\boldsymbol{\theta})$ για την εξεύρεση τοπικών ελάχιστων ή προσπαθεί να βρει τα τοπικά ελάχιστα με απευθείας μετακίνηση σημείων (βλ. επόμενη ενότητα 2.8). Αν το μοντέλο μας είναι NN ο πιο ευρέως χρησιμοποιούμενος αλγόριθμος για να βελτιστοποιήσουμε τα βάρη του είναι η πίσω-προώθηση (backpropagation) [90]. Εν συντομία, το σφάλμα που υπολογίζεται από τις μερικές παραγώγους κάθε στρώματος ρυθμίζει τη μεταβολή των βαρών της επόμενης στρώσης ακολουθώντας την ακριβή παρόμοια διαδικασία υπολογιστικών αλυσίδων παράγωγων μερών σε σχέση με τις παραμέτρους που πρόκειται να ρυθμιστούν.

Συνολικά, η συνάρτηση απώλειας $\mathcal{J}(\boldsymbol{\theta})$ είναι μόνο μια ευρετική μέθοδος για να είναι σε θέση να προσεγγίσει ένα τοπικό ελάχιστο για το πρόβλημα μη γραμμικής ελαχιστοποίησης το οποίο γενικά δεν μπορεί να λυθεί από κανένα αλγόριθμο σε πολυωνυμικό χρόνο (αυτοί οι τύποι προβλημάτων ονομάζονται προβλήματα NP). Επιπλέον, ο χώρος χαρακτηριστικών των δειγμάτων δεδομένων $\{\mathbf{x}_i\}_{i=1}^N$ είναι γενικά περίπλοκος και μη γραμμικός. Τα διανύσματα χαρακτηριστικών που εξάγονται ανήκουν σε υπο-χώρους όπου οι φορείς που βρίσκονται σε υποπεριοχές ανήκουν σε διαφορετικές κατηγορίες. Επομένως, το πρόβλημα της εύρεσης των βέλτιστων διακριτικών περιοχών του μεγάλου διαστάσεων χώρου εισόδου καθίσταται μη-εύκολο και απαιτεί προσεκτική εξέταση τόσο της εξαγωγής χαρακτηριστικών όσο και της βελτιστοποίησης του μοντέλου [59]. Για παράδειγμα, στην Εικόνα 1.7 είναι προφανές ότι τα διανύσματα χαρακτηριστικών που αντιστοιχούν στα ψηφία των "9" και "4" δεν είναι γραμμικά διαχωριζόμενα και επομένως η περιοχή διακρίσεων είναι πολύ πιο δύσκολο να βρεθεί άλλες περιοχές για άλλα ζεύγη ψηφίων όπως "0" και "7 " που φαίνεται να είναι εντελώς άσχετα με την αναπαράστασή τους στο δισδιάστατο επίπεδο.

Στηριζόμενοι σε αυτό, έχουν χρησιμοποιηθεί διάφορα μοντέλα ταξινόμησης προκειμένου να εντοπιστούν αυτές οι διαχωριστικές περιοχές σε οποιοδήποτε χώρο χαρακτηριστικών. Θα παράσχουμε τη μαθηματική διατύπωση των μοντέλων ταξινόμησης που χρησιμοποιούμε σε αυτό το έργο, καθώς και μια σύντομη επεξήγηση του τρόπου με τον οποίο λειτουργούν.

2.2.2 Υπολογιστικές μηχανές υποστήριξης (SVM)

Όπως αναφέρθηκε προηγουμένως, τα διανύσματα χαρακτηριστικών κάθε κατηγορίας αλληλοκαλύπτονται με διανύσματα χαρακτηριστικών από άλλες κατηγορίες και επομένως δεν είναι γραμμικά διαχωριζόμενα. Με άλλα λόγια, δεν μπορεί κανείς εύκολα να βρει ένα υποχώρο του χώρου των χαρακτηριστικών εισόδου που χρησιμεύει ως όριο ταξινόμησης για δεδομένα που ανήκουν σε κάθε κατηγορία του εκπαιδευτικού συνόλου. Τα SVM προσπαθούν να βρουν υπερεπίπεδα μέγιστου περιθωρίου διχωρισμού από τα διανύσματα που ανήκουν σε διαφορετικές κλάσεις [91]. Επομένως, κάθε δυαδικό πρόβλημα ταξινόμησης μπορεί να μειωθεί για να βρεθεί το διαχωριστικό υπερεπίπεδο το οποίο έχει το μέγιστο περιθώριο μεταξύ της απόστασης των πλησιέστερων σημείων κάθε κλάσης (Η απόσταση ενός σημείου από ένα υπερεπίπεδο είναι ακριβώς το μήκος της προβολής αυτού του σημείου κάθετα προς το υπερεπίπεδο).

Με πιο αυστηρό τρόπο μπορούμε να ορίσουμε ένα πρόβλημα δυαδικής ταξινόμησης για μια ακολουθία N δειγμάτων εκπαίδευσης $\{(\mathbf{x}_k, \mathcal{Y}_k)\}_{k=1}^N$ όπου $\mathcal{Y}_k \in \{-1, 1\}$. Μπορούμε να υποθέσουμε ότι οι παράμετροι $\boldsymbol{\theta}$ ενός SVM είναι οι συντελεστές του υπερεπίπεδου απόφασης που πρόκειται να βρεθούν. Τα SVM επιλύουν το ακόλουθο πρόβλημα:

$$\begin{aligned} &\underset{\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{\zeta}}{min} \tfrac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} + C\sum_{i=1}^N \zeta_i \\ &constrains: \qquad\qquad \mathcal{Y}_i(\boldsymbol{\theta}^T\phi(\mathbf{x}_i) + b) \geq 1 - \zeta_i \end{aligned} \tag{2.5}$$

όπου $C > 0$ είναι ένας όρος τακτοποίησης για τη διαμόρφωση του όρου ποινής κακώς ταξινομημένων στιγμιότυπων (πρακτικά χαρακτήρες vectors $\mathbf{x}$ που δεν ακολουθούν τους περιορισμούς της εξίσωσης 2.5 και $\phi\cdot$) είναι ένας μη γραμμικός χάρτης που χρησιμοποιείται για τη μετατροπή των δεδομένων εισόδου. Για το σκοπό αυτό, προκειμένου να αποκτήσουμε μη γραμμικές σχέσεις μεταξύ των σημείων δεδομένων, μπορούμε να εκτελέσουμε το τρίκ πυρήνα [92] χρησιμοποιώντας τους προαναφερθέντες χάρτες $\phi(\cdot)$. Αυτό επιτρέπει στον αλγόριθμο εκπαίδευσης SVM να ταιριάζει με το υπερεπίπεδο μεγίστου περιθωρίου σε ένα μετασχηματισμένο χώρο χαρακτηριστικών, το οποίο τα όρια μεταξύ των δύο κλάσεων είναι πολύ ευκολότερο να βρεθούν από ό, τι στον αρχικό χώρο εισόδου. τα όρια ταξινόμησης είναι υπερεπίπεδα τότε οι φορείς στήριξης θα είναι παράλληλοι με την μεταβολή της σταθεράς ($\boldsymbol{\theta}^T\phi(\mathbf{x}_i) + b = 1$ και $\boldsymbol{\theta}^T\phi(\mathbf{x}_i) + b = -1$), τότε η απόσταση μεταξύ αυτών των φορέων υποστήριξης θα είναι $\frac{2}{||\boldsymbol{\theta}||}$. Αυτή είναι μια επιπλέον επαλήθευση γιατί η εξίσωση 2.5 χρησιμεύει ως συνάρτηση στόχου ενός SVM.

Εφόσον εφαρμόζουμε το τέχνασμα του πυρήνα $\mathcal{K}(\mathbf{x_i}, \mathbf{x_j}) = \phi(\mathbf{x_i})^T\phi(\mathbf{x_j}))$ μπορούμε να μετατρέψουμε το πρόβλημα που ορίζεται παραπάνω στο δυϊκό του και επίσης να χρησιμοποιήσει τους νέους

μετασχηματισμούς του χώρου χαρακτηριστικών.

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \tfrac{1}{2}\boldsymbol{\alpha}^T \mathbf{Q}\boldsymbol{\theta} - \mathbf{1}_N^T \boldsymbol{\alpha} \\ &constrains: \qquad \sum_{i=1}^{N} \mathcal{Y}_i \alpha_i = 0 \\ &\qquad\qquad\quad 0 \leq \alpha_i \leq C, \;\; i \in \{1, .., N\} \end{aligned} \tag{2.6}$$

όπου $\mathbf{Q}_{i,j} = \mathcal{Y}_i \mathcal{Y}_j \mathcal{K}(\mathbf{x_i}, \mathbf{x_j})$ και $\mathbf{1}_N = [1, 1, ..., 1]$ είναι ένας διάνυσμα άσων που βρίσκονται στο χώρο $\mathbb{R}^N$.

Ως εκ τούτου, το υπερεπίπεδο απόφασης θα ορίζεται από την ακόλουθη εξίσωση:

$$sgn(\sum_{i=1}^{N} \mathcal{Y}_i \alpha_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}) + b)) \tag{2.7}$$

όπου b είναι η τομή του καθορισμένου μοντέλου και $sgn(\cdot)$ είναι η συνάρτηση προσήμου όπου -1 αντιστοιχεί σε αρνητικές τιμές και 1 σε θετικές και 0 για μηδενικές τιμές.

Μια λίστα με συναρτήσεις πυρήνα $\mathcal{K}(\mathbf{x_i}, \mathbf{x_j})$ που χρησιμοποιούνται ευρέως είναι η κατόθι:

- Γραμμικός πυρήνας: $\langle \mathbf{x_i}, \mathbf{x_j} \rangle$
- Ο πυρήνας πολυώνυμο με βαθμού d και μια παράμετρος προκατάλειψης-bias r: $(\gamma \langle \mathbf{x_i}, \mathbf{x_j} \rangle + r)^d$
- Η παράμετρος ρύθμισης του πυρήνα ακτινικής βάσης (RBF) γ: $exp(-\gamma ||\mathbf{x_i} - \mathbf{x_j}||^2)$
- Σιγμοειδή πυρήνα με παράμετρο ρύθμισης γ και παράμετρο bias r: $tanh(\gamma \langle \mathbf{x_i}, \mathbf{x_j} \rangle + r)$

Μπορούμε εύκολα να επεκτείνουμε την προηγούμενη διατύπωση των SVM δυαδικών αποφάσεων σε προβλήματα πολλαπλών τάξεων με την απλή εκπαίδευση ξεχωριστών δυαδικών ταξινομητών για όλες τις διαθέσιμες κατηγορίες στα δεδομένα εκπαίδευσης. Μετά από αυτό, ο δυαδικός ταξινομητής με τη μέγιστη βαθμολογία εμπιστοσύνης επιλέγεται ως η τελική απόφαση του πολυεθνικού SVM. Αυτή η μέθοδος ονομάζεται επίσης "ένα εναντίων όλων" ή "one-versus-rest".

2.2.3 Λογιστική Παλινδρόμηση (LR)

Ένας απλούστερος αλγόριθμος ταξινόμησης είναι ο ταξινομητής Λογιστικής Παλινδρόμησης Logistic Regression (LR) στον οποίο οι παράμετροι του μοντέλου $\boldsymbol{\theta}$ αντιστοιχούν και πάλι στους συντελεστές ενός υπερεπιπέδου ομοίως στα SVMs (βλέπε Ενότητα 2.2.2). Μπορούμε να χρησιμοποιήσουμε τη διατύπωση που ορίζεται στις προηγούμενες υποενότητες της ενότητας 2.2 για να παρουσιάσουμε τα μαθηματικά θεμέλια του ταξινομητή LR.

Θα περιγράψουμε και πάλι το απλό πρόβλημα για έναν δυαδικό ταξινομητή που έχει παρόμοια ρύθμιση και διαμόρφωση όπως περιγράφεται στην ενότητα 2.2.2. Τα δείγματα κατάρτισης N $\{(\mathbf{x}_k, \mathcal{Y}_k)\}_{k=1}^{N}$ αποτελούνται από τα διανύσματα και τις ετικέτες χαρακτηριστικών εισόδου όπου $\mathcal{Y} \in \{0, 1\}$. Η ενεργοποίηση του ταξινομητή LR προσδιορίζεται εφαρμόζοντας μια sigmoid λειτουργία (βλέπε Εξίσωση 2.8 παρακάτω) πάνω από την προσαρμοσμένη γραμμή για να πάρουμε την τελική απόφαση ταξινόμησης.

$$sigmoid(z) = \frac{1}{1 + e^{-z}} \tag{2.8}$$

Η ενεργοποίηση του LR για ένα δεδομένο διάνυσμα $\mathbf{x}$ θα ορίζεται ως εξής:

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = sigmoid(\boldsymbol{\theta}^T \mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}} \tag{2.9}$$

Η συνάρτηση απώλειας που πρέπει να ελαχιστοποιηθεί κατά την μάθηση των παραμέτρων του LR στα δεδομένα εκπαίδευσης είναι η εξής:

$$\frac{1}{2}||\boldsymbol{\theta}||_q + C \sum_{i=1}^{N} log(exp(-\mathcal{Y}_i(\boldsymbol{\theta}^T \mathbf{x}_i + b)) + 1) \tag{2.10}$$

όπου τα $C > 0$ και b αντιπροσωπεύουν τους συντελεστές της τιμωρίας των λανθασμένων ταξινομήσεων και της τομής του υπερεπιπέδου (ίδιο με το τμήμα 2.2.2). Το $||\cdot||_q$ είναι η νόρμα που χρησιμοποιείται για τον ορισμό της απόστασης μεταξύ των διανυσμάτων χαρακτηριστικών που πρέπει να τιμωρείται κατά τη διάρκεια της διαδικασίας κατάρτισης (για $q = 1$ και $q = 2$ υπολογίζουμε την Manhattan ℓ_1 και την Ευκλείδια ℓ_2, αντίστοιχα). Το δεύτερο μέρος της παραπάνω αντικειμενικής συνάρτησης είναι ακριβώς το ίδιο με αυτό που περιγράφεται στην ενότητα 2.2.1 ενώ το πρώτο ρυθμίζει το εύρος των παραμέτρων, προκειμένου να αποφευχθούν σφάλματα αριθμητικής σταθερότητας λόγω των μεγάλων αριθμητικών τιμών που παράγονται κατά τη διάρκεια του βέλτιστου συμπεράσματος συντελεστλων της γραμμής. Για να βρούμε τις τιμές των παραμέτρων για τον ταξινομητή LR τότε πάλι πρέπει να ελαχιστοποιήσουμε τη συνάρτηση απώλειας (βλ. Εξίσωση 2.10) υπολογίζοντας την κλίση αυτής της συνάρτησης απώλειας και προσπαθούμε να εφαρμόσουμε LR με μέγιστη πιθανοφάνειας με στοχαστική πτώση κλίσης [93].

2.2.4 Κ-Πλησιέστεροι Γείτονες (KNNs)

Το KNN είναι ένας μη παραμετρικός ταξινομητής. Με αυτό εννοούμε ότι δεν υπάρχουν διαθέσιμες παραμέτροι αυτού του μοντέλου για να συντονιστούν ή να βελτιστοποιηθούν καθώς και ότι το KNN δεν κάνει υποθέσεις σχετικά με την κατανομή πιθανότητας των δεδομένων εισόδου. Στην πραγματικότητα δεν υπάρχει καμία διαδικασία εκπαίδευσης. Για ένα δεδομένο διάνυσμα $\mathbf{x}'$ θα θέλαμε να υπολογίσουμε την αντίστοιχη ετικέτα $\mathcal{Y}'$. Για να γίνει αυτό, πρέπει να υπολογίσουμε τους πλησιέστερους γείτονες K στο διάνυσμα δοκιμής $\mathbf{x}'$. Μπορούμε να καθορίσουμε το σύνολο των πλησιέστερων γειτόνων του K, όπως φαίνεται παρακάτω:

$$\mathcal{U} = \underset{i}{argmin}||\mathbf{x}_i - \mathbf{x}'||_q \tag{2.11}$$

όπου $\mathbf{x}_i$ είναι το i th δείγμα των δεδομένων εκπαίδευσης. $||\cdot||_q$ είναι η νόρμα που χρησιμοποιείται για τον ορισμό της μέτρησης απόστασης μεταξύ του διανύσματος δοκιμής και κάθε δείγματος εκπαίδευσης (για $q = 1$, $q = 2$ ή $q = \infty$ υπολογίζουμε την Μανχάταν, Ευκλείδια ή Νόρμα Απείρου - Supremum, αντίστοιχα).

Τώρα για το προαναφερθέν σύνολο εγγύτερων γειτόνων βρίσκουμε την ετικέτα που αντιπροσωπεύεται από την πλειοψηφία των διανυσμάτων δεδομένων $\{\mathbf{x}_i \mid i \in \mathcal{U}\}$. Επομένως, η ετικέτα του διανύσματος υπό εξέταση υπολογίζονται χρησιμοποιώντας τις ετικέτες των πλησιέστερων γειτόνων, από την άποψη της μέτρησης απόστασης που επιλέξαμε, μέσα στο χώρο των φορέων χαρακτηριστικών εκπαίδευσης.

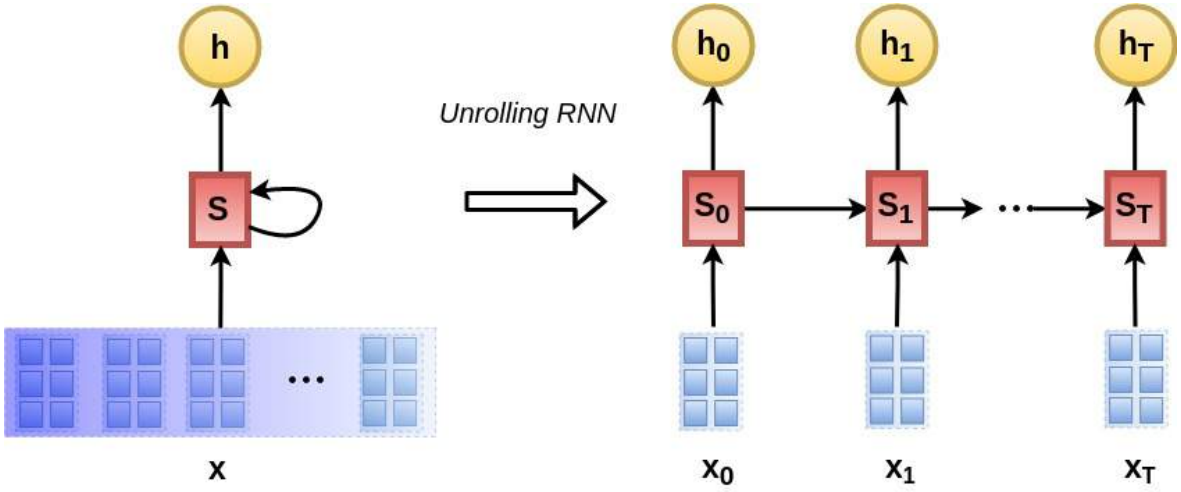
2.3 Ανατροφοδοτούμενα Νευρωνικά Δίκτυα (RNN)

Πηγαίνοντας ένα βήμα πέρα από τα απλά μοντέλα που απαιτούν σταθερές παραστάσεις των δεδομένων εισόδου, τα RRN δείχνουν την υπεροχή τους όταν χρησιμοποιούνται με ακολουθίες εισόδου. Συγκεκριμένα, αλλάζουμε τον ορισμό του επιτηρούμενου προβλήματος μετατρέποντας την είσοδο για το δείγμα i από ένα διάνυσμα $\mathbf{x}_i$ σε μια ακολουθία διανυσμάτων $\{\mathbf{x}_{ij}\}_{j=1}^{T}$ όπου T είναι ο αριθμός των χρονικών βημάτων ή το μήκος της ακολουθίας εισόδου των διανυσμάτων χαρακτηριστικών. Αυτό είναι ιδιαίτερα χρήσιμο όταν διαμορφώνουμε δεδομένα εισόδου, όπως τον ήχο και το κείμενο, όπου οι υποκείμενες εξαρτήσεις χρόνου αλλοιώνουν εντελώς τη διαδικασία εξαγωγής πληροφοριών σε ρυθμίσεις ταξινόμησης και περισσότερο [59]. Η βασική λειτουργικότητα ενός RNN είναι να συμπεράνει την υποκείμενη δυναμική της χρονικής συμπεριφοράς της ακολουθίας εισόδου διατηρώντας μια εσωτερική κατάσταση και ενημερώνοντάς την για διαφορετικές χρονικές στιγμές.

Στην ουσία, τα RNN δημιουργούν αντίγραφα μιας γεννήτριας κυττάρων με διαφορετικές παραμέτρους σε διαφορετικές χρονικές στιγμές. Με αυτό τον τρόπο, οι συνδέσεις ανατροφοδότησης θα μπορούσαν να θεωρηθούν ως συνδέσεις μεταξύ κρυφών κόμβων του ίδιου επιπέδου ενώ η είσοδος και η έξοδος αντιπροσωπεύουν συνδέσεις με τα επόμενα στρώματα ή προηγούμενα στρώματα. Αυτή

είναι ακριβώς η τοπολογία οποιουδήποτε τυπικού DNN, εκτός από το ότι το RNN μπορεί να βελτιστοποιηθεί χρησιμοποιώντας συγκεκριμένα χρονικά σημεία και όχι σε όλο το μήκος χρόνου του ξετυλιγμένου μοντέλου. Τα RNNs πρέπει να ρυθμιστούν με το μέγιστο μήκος που θα μπορούσε να χρησιμοποιηθεί είτε ως δοκιμή είτε ως εκπαίδευση πριν από τη χρήση του ίδιου του μοντέλου.

Μια απλή απεικόνιση ενός RNN εμφανίζεται στην εικόνα 2.1. Το διάνυσμα εισόδου $\mathbf{x}$ θα μπορούσε επίσης να θεωρηθεί ως μια ακολουθία διανυσμάτων $\{\mathbf{x}_j\}_{j=0}^{T}$ που είναι μέρη του αρχικού φορέα χαρακτηριστικών. Είναι φανερό ότι οι ανατροφοδοτούμενες συνδέσεις μεταφέρουν πληροφορίες για τις καταστάσεις $\mathbf{S}$ ως νέα δεδομένα που υπολογίζονται. Με αυτό τον τρόπο, κάθε ενεργοποίηση που αντιστοιχεί στην περίπτωση του κυττάρου RNN στο χρονικό βήμα t υπολογίζεται χρησιμοποιώντας πληροφορίες από την προηγούμενη κατάσταση $\mathbf{S}_{t-1}$ καθώς και το τρέχον διάνυσμα εισόδου $\mathbf{x}_t$. Επομένως, η αποδίπλωση ενός RNN είναι απλώς ένας τρόπος καλύτερης απεικόνισης της συμπεριφοράς και των υποκείμενων μηχανισμών αυτού του NN αλλά η βασική λειτουργικότητα είναι η ίδια στις δύο παραστάσεις. Η συνολική αρχιτεκτονική ενός RNN σχηματίζεται ως κατευθυνόμενο γράφημα κατά μήκος της ακολουθίας των διανυσμάτων εισόδου.

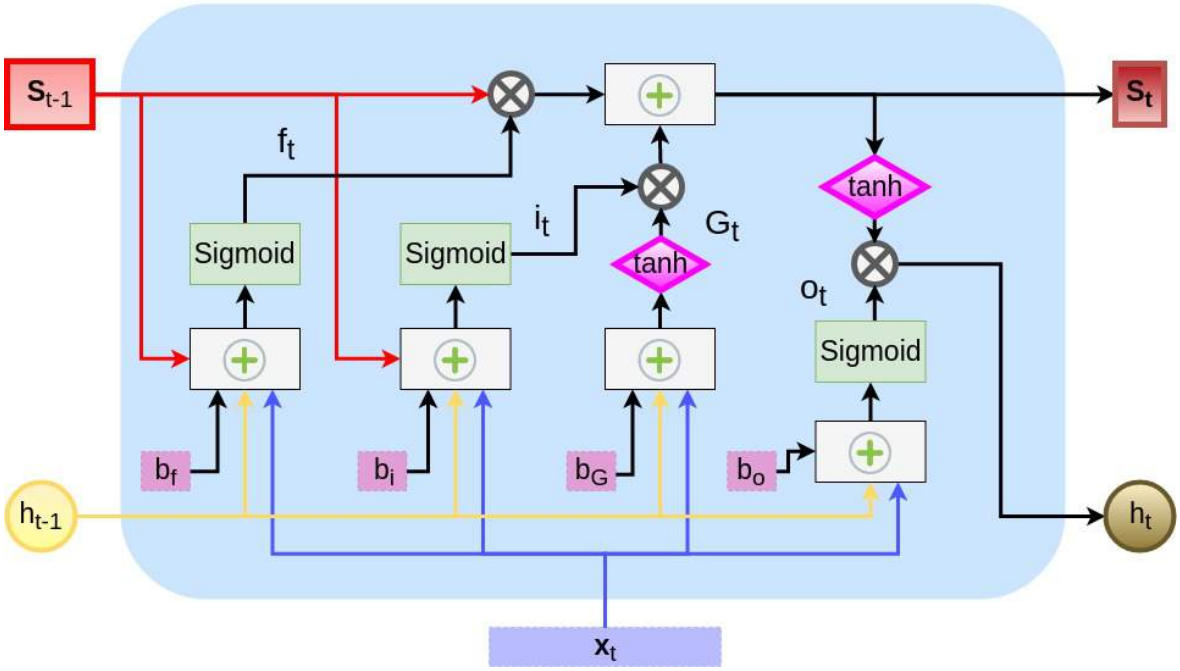


Σχήμα 2.1: Αποδίπλωση ενός ανατροφοδοτουμένου νευρικού δικτύου με ακολουθία εισόδου $t+1$ χρονικών βημάτων

Είναι εύκολο να συμπεράνουμε ότι θα μπορούσαμε να στοιβάζουμε σωρούς από στρώματα RNN το ένα πάνω στο άλλο, απλά συνδέοντας την ενεργοποίηση κάθε κελιού στο χρονικό βήμα t, το οποίο είναι $\mathbf{h}_t$, ως είσοδος στο επόμενο RNN για το ίδιο χρονικό διάστημα. Με αυτόν τον τρόπο δημιουργούμε βαθύτερα δίκτυα τα οποία είναι σε θέση να κωδικοποιήσουν καλύτερα την ροή πληροφοριών μεταξύ των διανυσμάτων της ακολουθίας εισόδου και να συμπεράνουν χαρακτηριστικά υψηλότερου επιπέδου που ίσως να μην μπορούν να συναχθούν επαρκώς χρησιμοποιώντας ρηχά δίκτυα [59].

2.3.1 Μονάδα μακράς βραχυπρόθεσμης μνήμης (LSTM)

Με βάση την τοπολογία RNN, είναι σαφές ότι κάθε πτυχή αυτού του NN διευκολύνει τη χρονική μοντελοποίηση της ακολουθίας εισόδου. Ωστόσο, οι μακροπρόθεσμες εξαρτήσεις της ακολουθίας εισαγωγής $\{\mathbf{x}_j\}_{j=0}^{T}$ μπορεί να είναι προβληματικές καθώς ο υπολογισμός που εκτελείται δημιουργεί εξαιρετικά πολύπλοκη σύνθεση συναρτήσεων και μη γραμμική συμπεριφορά. Η επαναλαμβανόμενη τοπολογία του δικτύου συνεπάγεται τον υπολογισμό της παραγώγου και της ροής της σε πολλαπλές χρονικές στιγμές που αποδίδει το διάσημο πρόβλημα της εξαφάνισης ή έκρηξης κλίσης όταν το σφάλμα προσπαθεί να μεταδοθεί προς τα πίσω για τη βελτιστοποίηση των παραμέτρων δικτύου [72]. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος και η δυνατότητα ενός αποτελεσματικού τρόπου κατασκευής και εκπαίδευσης ενός RNN, είναι η μονάδα LSTM [64], η οποία διαθέτει μια πύλη ξεχάσματος για να είναι σε θέση να κόβει πολύ μικρές και πολύ υψηλές τιμές κλίσης. Στο σχήμα 2.2 εμφανίζεται το μπλοκ διάγραμμα ενός κυττάρου LSTM το οποίο θα αναλυθεί παρακάτω.



Σχήμα 2.2: Μπλοκ διάγραμμα μίας μονάδας μακράς βραχυπρόθεσμης μνήμης (LSTM)

Οι πολλαπλασιαστές που αντιπροσωπεύονται με το γκρι "x" που είναι εγγεγραμμένοι σε κύκλους αποτελούνται από τα βασικά στοιχεία της αρχιτεκτονικής LSTM είναι τα εξής:

1. Η πύλη ξεχάσματος ελέγχει την ροή πληροφοριών για το ιστορικό από προηγούμενες χρονικές στιγμές $\mathbf{S}_{t-1}$
2. Η πύλη εισόδου ελέγχει την ροή πληροφοριών για το διάνυσμα εισόδου $\mathbf{x}_t$ στο τρέχον χρονικό διάστημα
3. Η πύλη εξόδου ελέγχει τη ροή πληροφοριών από την τρέχουσα κατάσταση $\mathbf{S}_t$ (η οποία έχει ήδη υπολογιστεί) στην τελική ενεργοποίηση του στοιχείου $\mathbf{h}_t$ στο τρέχον χρονικό βήμα

Πηγαίνοντας ένα βήμα πιο βαθιά στη λειτουργικότητα του LSTM κυττάρου που θα χρησιμοποιηθεί εκτενώς για τα πειράματα αυτής της εργασίας θα αναλύσουμε την αρχιτεκτονική που εμφανίζεται στο Σχήμα 2.2.

Πρώτα απ 'όλα, οι πολλαπλασιασμοί εκτελούνται στοιχείο προς στοιχείο για τους φορείς εισόδου. Επιπλέον, το ορθογώνιο με πράσινο συν αντιπροσωπεύει οποιαδήποτε συνάρτηση συσσωμάτωσης για δύο ή περισσότερα διανύσματα εισόδου όπως: συγκόλληση, μέσος όρος στοιχείου, πολλαπλασιασμός, προσθήκη, μέγιστο ή ελάχιστο. Η εφαρμογή των μη γραμμικών συναρτήσεων $sigmoid(\cdot)$ και $tanh(\cdot)$ αντιπροσωπεύεται από ένα πράσινο ορθογώνιο και ένα πορφυρό ρόμβο, αντίστοιχα. Οι ενεργοποιήσεις για το τρέχον χρονικό βήμα $\mathbf{h}_t$ και το προηγούμενο χρονικό βήμα $\mathbf{h}_{t-1}$ χρησιμεύουν ως έξοδος και μία από τις εισόδους αυτού του στοιχείου, αντίστοιχα. Επιπλέον, απεικονίζονται με χρήση κίτρινων κύκλων. Τέλος, τα διανύσματα κατάστασης $\mathbf{S}_t$ και $\mathbf{S}_{t-1}$ αντιπροσωπεύουν τα βάρη του νευρικού δικτύου που μπορούν να αποθηκεύσουν εσωτερικά την κατάσταση για κάθε χρονική στιγμή και να ακολουθήσουν μια παρόμοια είσοδο / εξήγηση εξόδου όπως προηγουμένως. Επιπλέον, τα τρέχοντα χρονικά βήματα είναι χρωματισμένα με πιο σκούρα χρώματα σε σύγκριση με τα αντίστοιχα διανύσματα που χρησιμεύουν ως είσοδος για το κελί στο χρονικό βήμα t για να δείξουν την εξάρτηση χρόνου μεταξύ αυτών των ζευγών. Τα βάρη προκατάλειψης **b** είναι μόνο διανύσματα με αριθμούς που προστίθενται στην τελική αναπαράσταση φορέα όταν αυτά μεταβιβάζονται από μια συνάρτηση συσσωμάτωσης. Η ίδια σχηματική παράσταση ακολουθείται σε όλα τα σχήματα αυτής της ενότητας. Θα επεξεργαστούμε επίσης κάθε στοιχείο του μπλοκ διαγράμματος χρησιμοποιώντας αυστηρή μαθηματική διατύπωση παρακάτω.

Η πύλη ξεχάσματος διαμορφώνεται ως εξής:

$$\mathbf{f}_t = sigmoid(\mathbf{W}_f \cdot (\mathbf{h}_{t-1}||\mathbf{x}_t) + \mathbf{b}_f) \tag{2.12}$$

όπου $\mathbf{W}_f$ αντιπροσωπεύει τα διανύσματα βάρους που αντιστοιχούν στην πύλη ξεχάσματος και $\mathbf{b}_f$ είναι η αντίστοιχη προκατάληψη. Με τον τρόπο αυτό, η έξοδος θα είναι ένας αριθμός μεταξύ του μηδέν και του πρώτου που αντιστοιχεί στο πρώτο για να ξεχάσει την είσοδο της προηγούμενης ενεργοποίησης. Αντίθετα, μια τιμή 1 σε αυτήν την πύλη αντιπροσωπεύει ότι οι πληροφορίες της προηγούμενης ενεργοποίησης $\mathbf{h}_{t-1}$ μαζί με τις πληροφορίες από το τρέχον διάνυσμα εισόδου $\mathbf{x}_t$ θα να ληφθούν πλήρως υπόψη για τον υπολογισμό της κατάστασης αυτής της LSTM. Το ίδιο ισχύει για όλες τις άλλες πύλες που εμφανίζονται στο Σχήμα 2.2.

Η πύλη εισόδου ελέγχει τις πληροφορίες που θα προκύψουν από την ενεργοποίηση του προηγούμενου χρονομέτρου $\mathbf{h}_{t-1}$ παράλληλα με τις πληροφορίες από το τρέχον διάνυσμα εισόδου $\mathbf{x}_t$ όταν προσπαθούμε για να ενημερώσετε τα τρέχοντα βάρη κατάστασης $\mathbf{S}_t$. Αυτά διαμορφώνονται χρησιμοποιώντας τις ακόλουθες εξισώσεις:

$$\mathbf{i}_t = sigmoid(\mathbf{W}_i \cdot (\mathbf{h}_{t-1}||\mathbf{x}_t) + \mathbf{b}_i) \tag{2.13}$$

$$\mathbf{G}_t = tanh(\mathbf{W}_G \cdot (\mathbf{h}_{t-1}||\mathbf{x}_t) + \mathbf{b}_G) \tag{2.14}$$

όπου $\mathbf{W}_G$ και $\mathbf{W}_i$ αντιπροσωπεύουν τα βάρη του δικτύου για την πύλη του ελέγχου πληροφοριών εισόδου καθώς και τις αντίστοιχες προκαταλήψεις $\mathbf{b}_i$ και $\mathbf{b}_G$.

Για να μπορέσουμε να υπολογίσουμε τις ενημερώσεις για την τρέχουσα κατάσταση $\mathbf{S}_t$, προσθέτουμε τις πληροφορίες που ελέγχθηκαν από τις προαναφερθείσες πύλες (εισόδους και ξεχασμένες πύλες). Συγκεκριμένα, η τρέχουσα κατάσταση υπολογίζεται όπως φαίνεται παρακάτω:

$$\mathbf{S}_t = \mathbf{i}_t \odot \mathbf{G}_t + \mathbf{f}_t \odot \mathbf{S}_{t-1} \tag{2.15}$$

Στη συνέχεια, για να υπολογίσουμε την έξοδο στο τρέχον χρονικό βήμα t, πρέπει να συνδυάσουμε το προυπολογισμένο διάνυσμα κατάστασης $\mathbf{S}_t$ αφού εφαρμοστεί μια μη γραμμική συνάρτηση $tanh(\cdot)$ καθώς και πληροφορίες από το εισόδου και την ενεργοποίηση από το προηγούμενο χρονικό διάστημα στον πολλαπλασιαστή πύλης εξόδου. Συγκεκριμένα, οι ακόλουθες εξισώσεις εκφράζουν το διάνυσμα ενεργοποίησης στο τρέχον χρονικό διάστημα:

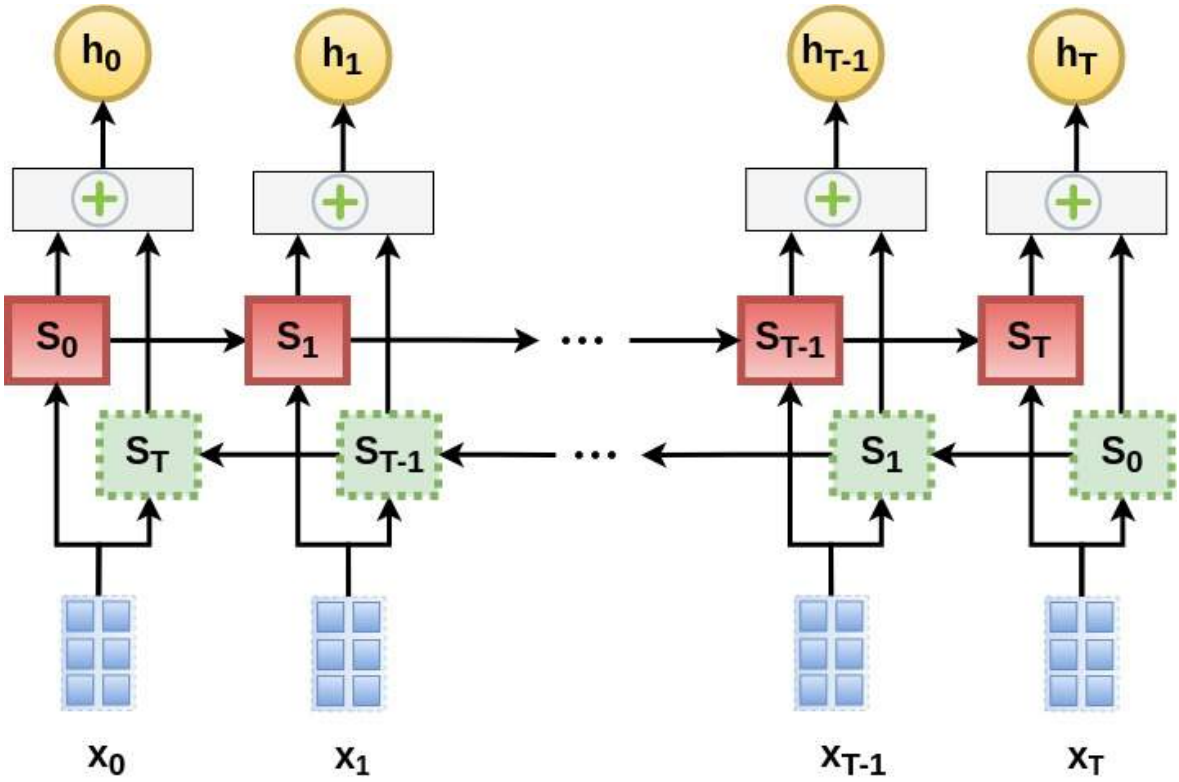
$$\mathbf{o}_t = tanh(\mathbf{W}_o \cdot (\mathbf{h}_{t-1}||\mathbf{x}_t) + \mathbf{b}_o) \tag{2.16}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot tanh(\mathbf{C}_t) \tag{2.17}$$

2.3.2 Αμφίδρομο LSTM

Αν και τα LSTMs είναι σε θέση να κωδικοποιούν τις χρονικές εξαρτήσεις από τις ακολουθίες εισόδου των φορέων, είναι αρκετά κοινό το γεγονός ότι το μήκος της ακολουθίας εισόδου είναι αρκετά μακρύς ώστε να μην εκπαιδεύεται επαρκώς με τη χρήση οπίσθιας προώθησης. Το κύριο πρόβλημα είναι ότι η χρονική εξάρτηση κατά την εκπαίδευση πολύ μακρών ακολουθιών μειώνεται καθώς οι διαδοχικοί υπολογισμοί σε πολλαπλές χρονικές στιγμές μεταβάλλουν τη ροή πληροφορίας μέσω της κλίσης [65] και κατά συνέπεια την ανανέωση των βαρών του δικτύου. Για να αντιμετωπιστεί αυτό το πρόβλημα μπορούμε να κωδικοποιήσουμε την ακολουθία εισόδου από την αρχή μέχρι το τέλος (προς τα εμπρός RNN) και αντίστροφα (πίσω RNN) και τελικά να συνδυάσουμε τις ενεργοποιήσεις των δύο RNN στρωμάτων για να βρούμε την ενεργοποίηση εξόδου για κάθε χρονική στιγμή. Θα επικεντρωθούμε μόνο στα αμφίδρομα LSTMs (BLSTMs) επειδή χρησιμοποιούμε μόνο τη μονάδα LSTM για τα μοντέλα μας RNN.

Στο σχήμα 2.3, απεικονίζεται μια αρχιτεκτονική BLSTM με συνδυασμένες στρώσεις LSTM προς τα εμπρός και προς τα πίσω προκειμένου να παραχθεί η αντίστοιχη ενεργοποίηση. Τα κόκκινα ορθογώνια με στερεή περίμετρο αντιστοιχούν στην αρχική αρχιτεκτονική, η οποία είναι επίσης εμφανής από την κατεύθυνση των επαναλαμβανόμενων ενεργοποιήσεων και συνεπώς από την πληροφορική ροή. Από την άλλη πλευρά, τα προς τα πίσω LSTM κύτταρα για κάθε χρονική στιγμή αντιπροσωπεύονται χρησιμοποιώντας πράσινα ορθογώνια με διακεκομμένες γραμμές ως περίμετρο.



Σχήμα 2.3: Αμφίδρομο LSTM

Συγκεκριμένα, συνδυάζουμε τις δύο αρχιτεκτονικές με ξεχωριστό υπολογισμό της προώθησης $\overrightarrow{\mathbf{h}}_t$ στο timestep t και τη σύζευξη τους για τον υπολογισμό της τελικής ενεργοποίησης σε κάθε χρονική στιγμή. Για το σκοπό αυτό, η ενεργοποίηση στο χρονικό βήμα t είναι απλά η συνένωση των προαναφερθέντων διανυσμάτων, δηλαδή: $\mathbf{h}_t = \overrightarrow{\mathbf{h}}_t || \overleftarrow{\mathbf{h}}_{T-t}$. Το ίδιο ισχύει για όλα τα χρονικά βήματα $T+1$ της ακολουθίας εισόδου. Μπορούμε να επεκτείνουμε την τοπολογία της αρχιτεκτονικής BLSTM στοιβάζοντας επιπλέον BLSTM στρώματα στην κορυφή, συνδέοντας απλώς τις ενεργοποιήσεις από την αρχική αρχιτεκτονική $\overrightarrow{\mathbf{h}}_t$ στην είσοδο του επόμενου στρώματος LSTM προς τα εμπρός. Κάνουμε το ίδιο για την αρχική αρχιτεκτονική συνδέοντας $\overleftarrow{\mathbf{h}}_{T-t}$ στην είσοδο ενός στρώματος LSTM που είναι ένα επίπεδο υψηλότερο. Τέλος, αναβάλλουμε τη συγκόλληση και των δύο ενεργοποιήσεων στο τρέχον χρονικό διάστημα μέχρι το τελευταίο επίπεδο της αρχιτεκτονικής BLSTM.

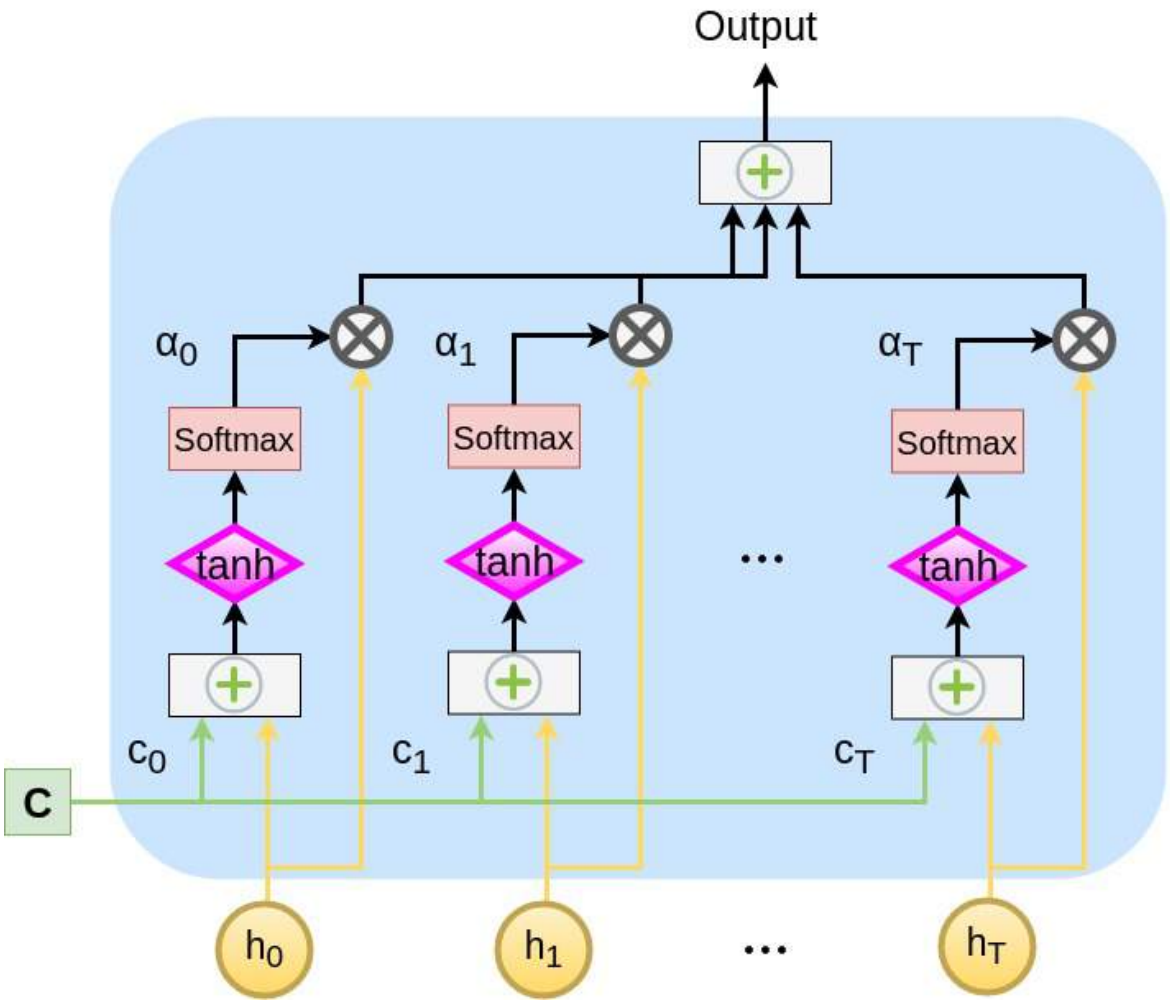
2.3.3 Μηχανισμός Προσοχής

Ένας μηχανισμός προσοχής εφαρμόζεται πάνω στην ακολουθία του φορέα κρυφούς κατάστασης ο οποίος αντλείται από το ανώτερο στρώμα ενός RNN προκειμένου να παρέχεται ένας μηχανισμός για εστίαση σε συγκεκριμένα τμήματα της ακολουθίας εισόδου που μπορεί να είναι πιο ενδεικτική για την τελική ταξινόμηση ή παλινδρόμηση. Συγκεκριμένα, ένα διάνυσμα πλαισίου $\mathbf{C}$ χρησιμοποιείται για να επιτρέψει στον μηχανισμό προσοχής να μάθει τι πρέπει να παρακολουθήσει με βάση την τρέχουσα ακολουθία εισόδου, καθώς και τις κωδικοποιημένες πληροφορίες μέχρι τώρα [67]. Ως εκ τούτου, η έξοδος αυτού του στρώματος θα είναι σε θέση να παραμελήσει άχρηστες ενεργοποιήσεις

από χρονικά βήματα οι οποίες είναι εντελώς μη ενημερωτικές για τον σκοπό ταξινόμησης ή παλινδρόμησης που χρησιμοποιούμε την αρχιτεκτονική RNN. Παραδείγματος χάριν, στο SER αναμένουμε ότι οι ακολουθίες εισόδου των διανυσμάτων χαρακτηριστικών μικρού τμήματος θα περιέχουν ανεξάρτητες πληροφορίες στην αρχή και στα τελειωτικά χρονικά σημεία που θα μπορούσαν να παραμεληθούν από τον μηχανισμό προσοχής με την εκμάθηση βαρών που είναι κοντά στο μηδέν. Αυτά τα τμήματα μπορεί να αντιστοιχούν σε διάρκειες σιωπής που είναι τυπικές για τους χρόνους έναρξης και λήξης των προτάσεων.

Στο σχήμα 2.4 απεικονίζεται ένας μηχανισμός προσοχής χρησιμοποιώντας ένα διάνυσμα περιβάλλοντος **C** και την ακολουθία εισόδου των ενεργοποιήσεων RNN $\{\mathbf{h}_t\}_{t=0}^{T}$. Αντιπροσωπεύουμε τη συνάρτηση softmax με ένα ελαφρώς κόκκινο ορθογώνιο χωρίς περίμετρο. Η τιμή softmax ενός διάνυσματος εισόδου **x** με μήκος $T+1$ θα δίδεται ως στοιχείο-στοιχείο από την ακόλουθη εξίσωση:

$$softmax(\mathbf{x})_k = \frac{e^{x_k}}{\sum_{t=0}^{T} e^{x_t}}, \quad 0 \leq k \leq T \tag{2.18}$$



Σχήμα 2.4: Μηχανισμός Προσοχής για δεδομένες ενεργοποιήσεις από ένα RNN

Για να κατανοήσουμε πραγματικά πώς λειτουργεί ο μηχανισμός προσοχής, είναι σημαντικό να καθορίσουμε τον τρόπο με τον οποίο υπολογίζεται το διάνυσμα πλαισίου. Το διάνυσμα πλαισίου **C** υπολογίζεται ως ο μέσος όρος όλων των ενεργοποιήσεων από όλες τις χρονικές στιγμές, τυπικά θα

γράψαμε:

$$\mathbf{C} = \frac{1}{T+1}\sum_{t=0}^{T}\mathbf{h}_t \tag{2.19}$$

Για κάθε χρονικό βήμα συνδυάζουμε το διάνυσμα πλαισίου και το αντίστοιχο διάνυσμα κρυφής κατάστασης εισόδου ως εξής για να πάρουμε την εσωτερική αναπαράσταση r:

$$\mathbf{r} = tanh(\mathbf{W}_c \cdot (\mathbf{h}_t||\mathbf{c}_t) + \mathbf{b}_c) \tag{2.20}$$

όπου $\mathbf{W}_c$ και $\mathbf{b}_c$ είναι οι παράμετροι εκπαίδευσης του στρώματος προσοχής.

Το στρώμα προσοχής μαθαίνει τα αντίστοιχα προσεκτικά βάρη α για όλες τις χρονικές στιγμές χρησιμοποιώντας το προηγουμένως καθορισμένο εσωτερικό διάνυσμα $\mathbf{r}$, ως εξής:

$$\mathbf{a} = softmax(\mathbf{r}) \tag{2.21}$$

Επομένως το διάνυσμα α θα περιέχει τα αντίστοιχα βάρη για να πολλαπλασιαστεί για να πάρει την τελική αναπαράσταση εξόδου του στρώματος προσοχής. Δηλαδή, η έξοδος θα ορίζεται ως: $\sum_{t=0}^{T} a_t \odot \mathbf{h}_t$. Είναι εύκολο να συμπεράνουμε ότι αν αθροίσουμε όλες τις τιμές από το διάνυσμα προσοχής α θα έχουμε ένα ως αποτέλεσμα. Αυτό είναι ενδεικτικό της διαδικασίας που ακολουθούμε όπου όλοι οι φορείς της αλληλουχίας εισόδου κανονικοποιούνται σε ένα σαν έναν φορέα πιθανότητας.

2.3.4 Αμφίδρομα LSTM με μηχανισμό προσοχής

Αν συνδυάσουμε την αρχιτεκτονική BLSTM που παρουσιάστηκε στην ενότητα 2.3.2 με μηχανισμό προσοχής (βλ. Προηγούμενη ενότητα 2.3.4), μπορούμε να δημιουργήσουμε ένα BLSTM). Συγκεκριμένα, πρέπει να συνδέσουμε τις έξοδους του τελικού στρώματος ενός BLSTM στην είσοδο του στρώματος προσοχής προκειμένου να φτάσουμε τον τελικό φορέα αναπαράστασης εξόδου για την ακολουθία εισόδου. Τώρα μπορούμε να προωθήσουμε το διάνυσμα εξόδου της συνολικής αρχιτεκτονικής σε επόμενα στρώματα προκειμένου να εκτελέσουμε ταξινόμηση ή παλινδρόμηση χρησιμοποιώντας διακριτές ή συνεχείς ετικέτες, αντίστοιχα. Μετά από τη σημείωση από τα προηγούμενα τμήματα, συνδυάζουμε την κρυφή κατάσταση του τελικού χρονικού βήματος $\mathbf{h}_T = \overrightarrow{\mathbf{h}}_T||\overleftarrow{\mathbf{h}}_0$ (το οποίο είναι σε θέση να κωδικοποιήσει τις πληροφορίες από την ακολουθία εισόδου) με το διάνυσμα εξόδου του μηχανισμού προσοχής για να πάρει την τελική αναπαράσταση της ακολουθίας εισόδου.

2.4 Ανακατασκευή Χώρου Φάσης

2.4.1 Ορισμός

Έχοντας ένα πλαίσιο ομιλίας με N δείγματα $\{s(i)\}_{i=1}^{N}$ ανασυνθέτουμε την αντίστοιχη τροχιά στον Χώρο Φάσης (PS) υπολογίζοντας d_e χρονικά καθυστερημένες εκδοχές του αρχικού πλαισίου ομιλίας με πολλαπλάσια χρονικής καθυστέρησης τ και δημιουργία των διανυσμάτων που βρίσκονται στον χώρο εμβύθινσης $\mathbb{R}^{d_e}$ όπως φαίνεται παρακάτω:

$$\mathbf{x}(i) = [s(i), s(i+\tau), ..., s(i+(d_e-1)\tau)] \tag{2.22}$$

όπου d_e είναι η διάσταση εμβύθινσης του ανακατασκευασμένου PS και τ είναι η χρονική υστέρηση. Αν το θεώρημα εμβύθινσης ισχύει και οι προαναφερθείσες παράμετροι ρυθμιστούν κατάλληλα, τότε η τροχιά που ορίζεται από τα σημεία $\{\mathbf{x}(i)\}_{i=1}^{N}$ θα διατηρούν τις αμετάβλητες ποσότητες από την πραγματική υποκείμενη δυναμική που θεωρείται άγνωστη [82]. Στη μελέτη των δυναμικών συστημάτων, το θεώρημα εμβύθινσης καθυστέρησης καθορίζει τις συνθήκες κάτω από τις οποίες η δυναμική του αρχικού συστήματος $\mathbf{s}^{\star}$ μπορεί να ανακατασεκυαστεί από μια ακολουθία παρατηρήσεων $\{\mathbf{x}(i)\}_{i=1}^{N}$ της κατάστασης του δυναμικού συστήματος. Μια επιτυχημένη ανακατασκευή διατηρεί τις ιδιότητες

του πραγματικού δυναμικού συστήματος $\mathbf{s}^\star$ που δεν αλλάζουν κάτω από ομαλές αλλαγές στις συντεταγμένες του όπως οι διαφορομορφισμοί. Ωστόσο, δεν είναι βέβαιο ότι οι γεωμετρικές ιδιότητες του ελκυστή θα ίδιες με εκείνες της προσεγγιστικής πολλαπλότητας $\mathcal{M}$.

Πολυάριθμες έρευνες έχουν προσπαθήσει να εντοπίσουν τις βέλτιστες παραμέτρους για την ανασυγκρότηση του PS. Σύμφωνα με την παράμετρο [78], οι παράμετροι τ και d_e για κάθε πλαίσιο ομιλίας μπορούν να εκτιμηθούν μεμονωμένα με τη χρήση της μέσης αμοιβαίας πληροφορίας (AMI) [94] και του αλγορίθμου ψευδών πλησιέστερων γειτόνων (FNN) [95], αντίστοιχα. Τόσο οι προσεγγίσεις όσο και η εκτεταμένη ανάλυση για την επιλογή των προαναφερθέντων παραμέτρων θα παρουσιαστούν στα επόμενα τμήματα.

2.4.2 Μέση Αμοιβαία Πληροφορία (AMI)

Πρώτα από όλα, η Μέση Αμοιβαία Πληροφορία (AMI) είναι ένα μέτρο της μη γραμμικής συσχέτισης μεταξύ του δοθέντος σήματος και μιας χρονικά καθυστερημένης έκδοσης αυτού του σήματος από $\{s(i)\}_{i=1}^{N}$ και μια χρονικά καθυστερημένη έκδοση αυτού του σήματος κατά τ δείγματα π.χ. $\{s(i+\tau)\}_{i=1}^{N-\tau}$. Αν θέλουμε να είμαστε ακριβείς, πρέπει να χρησιμοποιήσουμε τον ίδιο αριθμό δειγμάτων για κάθε σήμα και συνεπώς θεωρούμε τα σήματα $\{s(i)\}_{i=\tau+1}^{N}$ και $\{s(i+\tau)\}_{i=1}^{N-\tau}$.

Επίσημα, το AMI για ένα σήμα $\{s(i)\}_{i=1}^{N}$ και ένα συγκεκριμένο χρονικό διάστημα τ όπως φαίνεται παρακάτω:

$$\mathcal{I}(\{s(i)\}_{i=1}^{N},\tau)=\sum_{i=1}^{N-\tau} p_b(s(i),s(i+\tau))\cdot log_2[\frac{p_b(s(i),s(i+\tau))}{p_b(s(i))\cdot p_b(s(i+\tau))}] \tag{2.23}$$

όπου το $p_b(s(i),s(i+\tau))$ ορίζει τη συνάρτηση της απο κοινού πιθανότητας που προέρχεται από το ιστόγραμμα τιμών του σήματος N καθώς και τις περιθωριακές πιθανότητες $p_b(s(i))$ και $p_b(s(i+\tau))$. Συγκεκριμένα, οι τιμές αυτών των πιθανοτήτων υπολογίζονται σε ορισμένο αριθμό κουβάδων. Σε αυτήν την εργασία, χρησιμοποιούμε $N_{bins}=32$ κουβάδες. Οποιοδήποτε ψηφιακό σήμα έχει διακεκριμένες ακέραιες τιμές, έτσι σημειώνουμε τις μέγιστες και ελάχιστες ακέραιες τιμές που θα μπορούσε να πάρει ένα διακριτοποιημένο σήμα, αντίστοιχα. Χωρίς απώλεια της γενικότητας θα μπορούσαμε να υποθέσουμε ότι οι αξίες που θα μπορούσε να πάρει ένα ψηφιακό σήμα είναι εξίσου διανεμημένες για αρνητικούς και θετικούς ακέραιους αριθμούς. Προκειμένου να εξάγουμε την κατανομή πιθανότητας χρειαζόμαστε το ιστόγραμμα του αρχικού σήματος $\mathcal{H}(\{s(i)\}_{i=1}^{N},s_{max})$ που ορίζεται για κάθε κουβά ως τον αριθμό δεικτών του σήματος εισόδου για τον οποίο βρίσκονται οι τιμές του σήματος μέσα σε αυτόν τον κουβά. Πριν από αυτό ορίζουμε τα ακόλουθα σύνολα που διαιρούν το σήμα στο προαναφερθέν N_{bins}:

$$\mathcal{A}_k=\{i\ \ \forall i\in\{1,...,N\}\ \mid\ [s(i)+s_{max}]\ \ div\ \ 2s_{max}=k-1\} \tag{2.24}$$

όπου $a\ \ div\ \ b$ ορίζει την ακέραια διαίρεση μεταξύ των αριθμών a και b και $k\in\{1,...,N_{bins}\}$ είναι το ευρετήριο του αντίστοιχου κουβά. Επιπλέον, για να διατηρήσουμε μια ευκολότερη ανάγνωση, ορίζουμε τον αντίστοιχο δείκτη του κουβά για κάθε δείγμα του εισερχόμενου σήματος $s(i)$ χρησιμοποιώντας την ακόλουθη συμβολική τιμή:

$$ind_{\mathcal{A}}(s(i))=[s(i)+s_{max}]\ \ div\ \ 2s_{max}+1 \tag{2.25}$$

Συνεπώς, περιγράφουμε ισοδύναμα τα σύνολα $\{\mathcal{A}_k\}_{k=1}^{N_{bins}}$,

$$\mathcal{A}_k=\{i\ \ \forall i\in\{1,...,N\}\ \mid\ ind_{\mathcal{A}}(s(i))=k\} \tag{2.26}$$

Επιπλέον, θα μπορούσαμε να καθορίσουμε την τιμή του κουβά k ως εξής:

$$\mathcal{H}(\{s(i)\}_{i=1}^{N},s_{max},N_{bins},k)=card(\mathcal{A}_k) \tag{2.27}$$

όπου το $card(\mathcal{A}_k)$ ορίζει τον αριθμό των στοιχείων του συνόλου $\mathcal{A}_k$ και πάλι $k \in \{1, ..., N_{bins}\}$ είναι ο δείκτης του αντίστοιχου κουβά. Ωστόσο, στην περίπτωση μας πρέπει επίσης να δημιουργήσουμε το αντίστοιχο σύνολο για όλα τα διαθέσιμα ζεύγη δεικτών i και $i + \tau$ τα οποία αντανακλούν τη συσχέτιση των τιμών του σήματος για την καθορισμένη χρονική υστέρηση. Ομοίως, καθορίζουμε ένα σύνολο για όλα τα διαθέσιμα ζεύγη του σήματος εισόδου όπως ορίζεται παρακάτω:

$$\mathcal{B}_{jk} = \{i \ \ \forall i \in \{1, ..., N\} \ \mid \ ind_{\mathcal{A}}(s(i)) = j, \ \ ind_{\mathcal{A}}(s(i + \tau)) = k\} \tag{2.28}$$

όπου $k \in \{1, ..., N_{bins}\}$ είναι πάλι ο δείκτης του αντίστοιχου κουβά.

Σε αυτό το πλαίσιο, μπορούμε εύκολα να ορίσουμε τις περιθωριακές πιθανότητες $p_b(s(i))$ για όλους τους δείκτες του σήματος $\{s(i)\}_{i=1}^{N}$ εισόδου, χρησιμοποιώντας τα προαναφερθέντα σύνολα $\{\mathcal{A}_k\}_{k=1}^{N_{bins}}$, όπως φαίνονται παρακάτω:

$$p_b(s(i)) = \frac{card(\mathcal{A}_{ind_{\mathcal{A}}(s(i))})}{\sum_{k=1}^{N_{bins}} card(\mathcal{A}_k)} \tag{2.29}$$

Ομοίως στην προηγούμενη εξίσωση μπορούμε να προσδιορίσουμε επίσης τις κοινές πιθανότητες $p_b(s(i), s(i + \tau))$ για όλους τους δείκτες του σήματος εισόδου και τα αντίστοιχα ζευγάρια δεικτών για καθυστέρηση κατά τ όπως φαίνεται παρακάτω:

$$p_b(s(i), s(i + \tau)) = \frac{card(\mathcal{B}_{ind_{\mathcal{A}}(s(i)), ind_{\mathcal{A}}(s(i+\tau))})}{\sum_{j=1}^{N_{bins}} \sum_{k=1}^{N_{bins}} card(\mathcal{B}_{jk})} \tag{2.30}$$

2.4.3 Θεώρημα Takens

Προκειμένου να ανακατασκευάσουμε επιτυχώς τη δυναμική του συστήματος, πρέπει να καθορίσουμε τη διάσταση εμβύθισης d_e για τα σημεία του ανασυγκροτημένου PS. Για να γίνει αυτό, θα πρέπει να προσέξουμε ότι για την επιλεγμένη διάσταση εμβύθισης το προαναφερθέν "θεώρημα εμβύθισης" ισχύει. Αυτό το θεώρημα ονομάζεται επίσης "θεώρημα Takens", το οποίο υποδηλώνει ότι αν αυξήσουμε τη διάσταση εμβύθισης d_e ή ισοδύναμα τον αριθμό των χρονικά καθυστερημένων εκδόσεων του δεδομένου σήματος τότε η δυναμική της προσεγγιστικής πολλαπλότητας $\mathcal{M}$ που αποτελείται από τα σημεία $\{\mathbf{x}(i)\}_{i=1}^{N}$ γίνεται πολύ πιο ντετερμινιστική. Συγκεκριμένα, θα πρέπει να υπάρχει ένας διαφορομορφισμός $\phi : \mathcal{M} \to \mathbf{R}^{d_e}$ που χαρτογραφεί την προσεγγιστική πολλαπλότητα των σημείων $\{\mathbf{x}(i)\}_{i=1}^{N}$ σε $\mathbf{R}^{d_e}$ διατηρώντας ταυτόχρονα ότι η κλίση του χαρτογράφου ϕ έχει πλήρη τάξη [82]. Οι προαναφερθείσες δηλώσεις διατηρούνται εάν:

$$d_e \geq 2d_{BCD}(\mathcal{M}) + 1 \tag{2.31}$$

όπου d_{BCD} αντιστοιχεί στη διάσταση μέτρησης κουτιού της προσεγγιστικής ομαλής πολλαπλότητας $\mathcal{M}$:

$$d_{BCD}(\mathcal{M}) = \lim_{\epsilon \to 0} \frac{log(N(\epsilon))}{log(\frac{1}{\epsilon})} \tag{2.32}$$

όπου $N(\epsilon)$ είναι ο αριθμός κουτιών με μήκος πλευράς ίσο με ϵ τα οποία απαιτούνται για την κάλυψη της πολλαπλότητας $\mathcal{M}$.

Με άλλα λόγια, η διάσταση εμβύθισης d_e πρέπει να είναι αρκετά μεγάλη ώστε να ξετυλίγει με επιτυχία τη δυναμική του πραγματικού χώρου φάσης του αρχικού συστήματος. Διαφορετικά, θα ήταν δυνατό οι τροχιές του ανακατασκευασμένου χώρου φάσης να καταρρεύσουν σε κοντινές περιοχές και συνεπώς τα σημεία $\{\mathbf{x}(i)\}_{i=1}^{N}$ που βρίσκονται στην προκύπτουσα πολλαπλότητα $\mathcal{M}$ θα μπορούσαν να εμφανιστούν λανθάνοντες γείτονες, ενώ αυτό δεν ισχύει για το αρχικό σύστημα.

2.4.4 Αλγόριθμος Λανθασμένων Πλησιέστερών Γειτόνων (FNN)

Για να καθορίσουμε τη διάσταση εμβύθινσης d_e θα χρησιμοποιήσουμε το κριτήριο των ψευδών πλησιέστερων γειτόνων [95]. Χρησιμοποιώντας αυτό το κριτήριο, αυξάνουμε σταδιακά τη διάσταση εμβύθινσης και υπολογίζουμε έναν λόγο που είναι ενδεικτικός του αριθμού των γειτόνων για κάθε σημείο της τροχιάς. Στην ουσία, υπολογίζουμε τις αποστάσεις $\mathbf{D}_{\hat{d_e}}(\mathbf{x}(i), \mathbf{x}(j))$. Αυτές οι αποστάσεις αντιστοιχούν στις τιμές των μετρήσεων απόστασης μεταξύ των σημείων $\mathbf{x}(i)$ και $\mathbf{x}(j)$ της ανακατασκευασμένης τροχιάς του χώρου φάσης για διαφορετικές διαστάσεις εμβύθινσης. Αν υποθέσουμε ότι η προσεγγισμένη πολλαπλότητα $\mathcal{M}$ θα ενσωματωθεί στον ευκλείδειο χώρο $\mathbb{R}^{\hat{d_e}}$ τότε θα μπορούσαμε να ορίσουμε την απόσταση μετρήσεων μεταξύ των σημείων $\mathbf{x}(i)$ και $\mathbf{x}(j)$ ως εξής:

$$\mathbf{D}_{\hat{d_e}}(\mathbf{x}(i), \mathbf{x}(j)) = ||\mathbf{x}(i) - \mathbf{x}(j)||_2 = \sqrt{\sum_{k=0}^{\hat{d_e}-1} [s(i + k \cdot \tau) - s(i + k \cdot \tau)]^2} \tag{2.33}$$

Τώρα για όλα τα σημεία υπολογίζουμε το σχετικό ποσοστό αλλαγής για την απόσταση μεταξύ οποιουδήποτε ζεύγους σημείων στην τροχιά χώρου φάσης όταν χρησιμοποιούμε μια επιπλέον διάσταση για τη διάσταση ενσωμάτωσης. Δηλαδή υπολογίζουμε:

$$R^{\hat{d_e}}_{FNN}(\mathbf{x}(i), \mathbf{x}(j)) = \frac{\mathbf{D}_{\hat{d_e}+1}(\mathbf{x}(i), \mathbf{x}(j)) - \mathbf{D}_{\hat{d_e}}(\mathbf{x}(i), \mathbf{x}(j))}{\mathbf{D}_{\hat{d_e}}(\mathbf{x}(i), \mathbf{x}(j))} \tag{2.34}$$

Όταν η μέση τιμή του παραπάνω πίνακα $\mathbf{R}^{\hat{d_e}}_{FNN}$ σε όλες τις τιμές υπερβαίνει μια τιμή $(0.10 - 0.15)$ τότε αυτό σημαίνει ότι υπάρχει μια μεγάλη διαφορά στην απόσταση μεταξύ των σημείων της τροχιάς PS $\mathbf{x}(i)$ και $\mathbf{x}(j)$ όταν προσθέτουμε μια επιπλέον διάσταση. Ως αποτέλεσμα, αυτό σημαίνει ότι το σημείο $\mathbf{x}(i)$ έχει έναν ψεύτικο γείτονα όταν χρησιμοποιούμε μια διάσταση ενσωμάτωσης του $\hat{d_e}$ αντί ενός μεγαλύτερου $\hat{d_e} + 1$. Στην ιδανική περίπτωση, θα θέλαμε να υπολογίσουμε μια διάσταση ενσωμάτωσης d_e η οποία επιτυγχάνει έναν αριθμό ψεύτικων πλησιέστερων γειτόνων που είναι όσο το δυνατόν πιο κοντά στο μηδέν. Παρόλο που υπάρχει αντιστάθμιση επειδή, καθώς αυξάνουμε τον αριθμό των διαστάσεων που χρησιμοποιούνται για την ενσωμάτωση, αυξάνουμε επίσης τις διαστάσεις των δεδομένων εισόδου που πρέπει να αναλυθούν. Αυτό είναι επίσης ένα παράδειγμα του γενικού προβλήματος της "κατάρας της διαστασιμότητας" (βλ. Ενότητα 1.4.3). Σε γενικές γραμμές, μια αναλογία αξίας των ψευδών πλησιέστερων γειτόνων για μια επιλεγμένη διάσταση ενσωμάτωσης γύρω στα 10% φαίνεται επαρκής για την ανακατασκευή της δυναμικής [78].

2.5 Γραφήματα Επαναληψιμότητας (RPs)

Με δεδομένη την τροχιά PS $\{\mathbf{x}(i)\}_{i=1}^{N}$ αναλύουμε τις ιδιότητες επαναληψιμότητας αυτών των καταστάσεων υπολογίζοντας τις αποστάσεις ανά ζεύγη και οριοθετώντας αυτές τις τιμές, αντίστοιχο RP [96].

Μερικές φορές δεν προχωρούμε με το κατώτατο όριο προκειμένου να παράγουμε οικόπεδα με συνεχείς τιμές που είναι βασικά η απεικόνιση τοπικά κανονικοποιημένων πινάκων απόστασης. Αυτά τα γραφήματα μερικές φορές ονομάζονται επίσης μη-κατωφλιωμένα γραφήματα επαναληψιμότητας ή διαγράμματα συνεχούς επαναληψιμότητας [84]. Από τυπική άποψη, θα μπορούσαμε να γράψουμε την απόσταση μεταξύ των δύο σημείων ως εξής:

$$\mathbf{D}_q(\mathbf{x}(i), \mathbf{x}(j)) = ||\mathbf{x}(i) - \mathbf{x}(j)||_q \tag{2.35}$$

όπου $||\cdot||_q$ είναι ο κανόνας που χρησιμοποιείται για τον ορισμό της απόστασης μεταξύ οποιωνδήποτε δύο σημείων τροχιάς $\mathbf{x}(i)$ και $\mathbf{x}(j)$. Συγκεκριμένα, για $q = 1$, $q = 2$ ή $q = \infty$ υπολογίζουμε τον κανόνα του Μανχάταν, Ευκλείδειας ή Supremum, αντίστοιχα. Σε γενικές γραμμές, ο πίνακας απόστασης μπορεί να αντιπροσωπεύει οποιοδήποτε έγκυρη νόρμα.

Τα RPs είναι δυαδικές τετραγωνικές μήτρες και ορίζονται στοιχείο προς στοιχείο ως εξής:

$$\mathbf{R}_{i,j}(\epsilon, q) = \Theta(\epsilon - ||\mathbf{x}(i) - \mathbf{x}(j)||_q) \tag{2.36}$$

όπου $\Theta(\cdot)$ είναι η συνάρτηση Heaviside και ϵ είναι η τιμή κατωφλίου. Έτσι, το matrix $\mathbf{R}$ αποτελείται από αυτά που βρίσκονται σε περιοχές όπου οι καταστάσεις της τροχιάς είναι κοντά και μηδενικές αλλού. Το μέτρο της εγγύτητας ορίζεται από το όριο ϵ για το οποίο έχουν μελετηθεί πολλαπλά κριτήρια επιλογής [97]. Θεωρούμε τρία κριτήρια ανάλογα με: 1) μια σταθερή ad-hoc τιμή κατωφλίου, 2) ένα σταθερό ποσοστό επαναληψιμότητας (RR) όπως ορίζεται στην επόμενη εξίσωση 2.40 ϵ σύμφωνα με μια σταθερή πιθανότητα των ζευγών αποστάσεων των σημείων PS $P(||\mathbf{x}(i) - \mathbf{x}(j)||_q < \epsilon) = 0.15,\ \ 1 \leq i, j, \leq N$, μια σταθερή αναλογία της τυπικής απόκλισης σ των σημείων $\{\mathbf{x}(i)\}_{i=1}^{N}$, π.χ., $\epsilon = 5\sigma$ [98]. Για τις σταθερές τιμές των ϵ και q υποδηλώνουμε ως $\mathbf{R}_{i,j}$ την αντίστοιχη καταχώρηση του πίνακα RP για απλότητα.

Οι κύριες δομές που αναδύονται σε RPs και είμαστε σε θέση να συλλάβουμε αποτελούνται από γραμμές και σημεία της δυαδικής μήτρας. Είναι σημαντικό να κατανοήσουμε τις κύριες κατηγορίες γραμμών που είναι σε θέση να ποσοτικοποιήσουν πληροφορίες σχετικά με ένα RP:

Μια διαγώνιος γραμμή μήκους L (από άσους) ορίζεται από:

$$(1 - \mathbf{R}_{i-1,j-1})(1 - \mathbf{R}_{i+L+1,j+L+1}) \prod_{k=1}^{k=L} \mathbf{R}_{i+k,j+k} = 1 \tag{2.37}$$

Μια κατακόρυφη γραμμή ύψους L περιγράφεται από:

$$(1 - \mathbf{R}_{i,j-1})(1 - \mathbf{R}_{i,j+L+1}) \prod_{k=1}^{k=L} \mathbf{R}_{i,j+k} = 1 \tag{2.38}$$

Μία λευκή κατακόρυφη γραμμή μήκους L (με μηδενικά) ορίζεται ως:

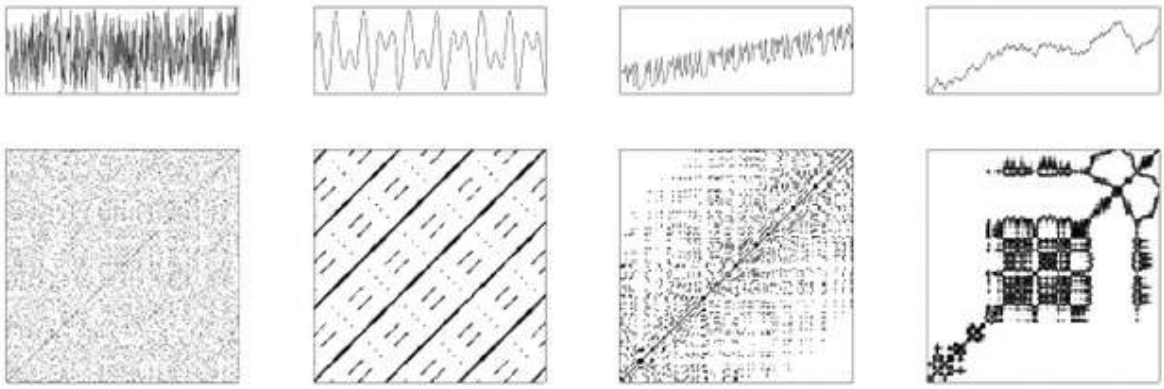
$$\mathbf{R}_{i,j-1}\mathbf{R}_{i,j+L+1} \prod_{k=1}^{k=L} (1 - \mathbf{R}_{i,j+k}) = 1 \tag{2.39}$$

Φυσικά ένα μόνο σημείο θα αντιστοιχούσε σε μια είσοδο του δυαδικού πίνακα $\mathbf{R}$ αλλά χωρίς να ανήκει σε μία από τις προαναφερθείσες κατηγορίες γραμμών. Επομένως, οποιαδήποτε δομή RPs που βασίζεται σε μεμονωμένα σημεία θα αντιστοιχούσε σε μια θορυβώδη εικόνα γεμάτη με απομονωμένα σημεία παρόμοια με τις εικόνες που παράγονται όταν επιβάλεται θόρυβος πιπέρι.

Επίσης, υποδηλώνουμε με τις κατανομές ιστογράμμων των διαγώνων, κάθετων και λευκών κατακόρυφων γραμμών $P_d(l)$, $P_v(l)$ και $P_w(l)$ αντίστοιχα. Συνεπώς, ο συνολικός αριθμός αυτών των γραμμών είναι αντίστοιχα $P_v(l)$ and $N_d = \sum_{l \geq d_m} P_d(l)$, $N_v = \sum_{l \geq v_m} P_v(l)$ και $N_w = \sum_{l \geq w_m} P_w(l)$, όπου $d_m = 2$, $v_m = 2$ και $w_m = 1$ καθορίζουν τα ελάχιστα μήκη για κάθε τύπο γραμμής [84].

Όλες αυτές οι δομές αντικατοπτρίζουν τη δυναμική επαναληψιμότητας του υπό ανάλυση συστήματος από τα διαφορετικά πρότυπα για κάθε τύπο συστήματος. Στην Εικόνα 2.5 [1] μπορούμε να δούμε τέσσερις διαφορετικούς τύπους συστημάτων που συλλαμβάνονται από κάθε χρονοσειρλα και τον τρόπο εμφάνισης των αντίστοιχων RP. Είναι προφανές ότι τα RP αντανακλούν τη δυναμική κάθε τύπου συστήματος από διαφορετικές δομές σε γραμμές και σημεία. Τα θορυβώδη δεδομένα θα οδηγούσαν επίσης σε RPs με μη συσχετισμένη δομή (θορυβώδεις εικόνες). Η περιοδικότητα είναι επίσης εμφανής στη δεύτερη εικόνα και για τις δύο συχνότητες παρακολουθώντας το παράλληλο προς τις διαγώνιες γραμμές που αντανακλούν την συν-εξέλιξη των τροχιών των τροχιών του PS για τα επόμενα χρονικά πλαίσια. Τέλος, η αυτορυθμιζόμενη διαδικασία και ο ντετερμινισμός της χρονοσειράς είναι εμφανείς από την ύπαρξη ακραίων γεγονότων γύρω από το δυαδικό σχέδιο από ορισμένες περιοχές που ενεργοποιούνται και άλλες που δεν είναι. Σε γενικές γραμμές, η υφή αυτών των εικόνων είναι αντιπροσωπευτική του υποκείμενου συστήματος.

[1] Το Σχήμα βρέθηκε στο: https://en.wikipedia.org/wiki/Recurrence_plot



Σχήμα 2.5: Ιστορικά Αναδρομή για διαφορετικούς τύπους συστημάτων. Από αριστερά προς τα δεξιά: τυχαίος θόρυβος, περιοδικές ταλαντώσεις με δύο συχνότητες, ντετερμινιστικό χαοτικό σύστημα και αυτορυθμιζόμενη διαδικασία.

2.6 Ποσοτική Ανάλυση Επαναληψιμότητας (RQA)

Όπως αναφέρθηκε και στην προηγούμενη ενότητα, τα RPs αντανακλούν κάποιες ιδιότητες των υπό ανάλυση συστημάτων που σχετίζονται στενά με τη δυναμική επαναληψιμότητας που παρουσιάζει το σύστημα εξελισσόμενο προς το χρόνο. Ωστόσο, αυτές οι εικόνες θα μπορούσαν να χρησιμοποιηθούν για την ποιοτική ανάλυση της πραγματικής ταυτότητας του συστήματος, αλλά θα μπορούσαν επίσης να χρησιμοποιηθούν για την εξαγωγή ποσοτικών μετρήσεων ώστε να διακρίνουν τα υποκείμενα συστήματα που θα αναλυθούν.

2.6.1 Μέτρα RQA

Τα μέτρα που εξάγουμε από τα RP βασίζονται σε πρότυπα από διαγώνιες, κάθετες γραμμές παρόμοιες με τους ορισμούς της προηγούμενης ενότητας 2.5. Σύμφωνα με αυτό, ορίζουμε μερικά μέτρα RQA που θα εξαχθούν από κάθε $N \times N$ RP και θα χρησιμοποιηθούν σε αυτό το έργο.

Λόγος Επαναληψιμότητας (RR): το οποίο είναι ένα μέτρο της πυκνότητας των σημείων στο RP. Η RR ορίζει την πιθανότητα ανάκτησης παρόμοιας κατάστασης στη γειτονιά της στο PS.

$$RR = \frac{1}{N^2} \sum_{i,j=1}^{N} \mathbf{R}_{i,j} \tag{2.40}$$

Ντετερμινισμός (DET): ο λόγος των διαγώνιων γραμμών τουλάχιστον του μήκους l_{min} με τον συνολικό αριθμό των σημείων του RP. Οι διαδικασίες με λιγότερο συσχετισμένες χρονολογικές σειρές ή χρονικές σειρές με χαοτική συμπεριφορά θα παρουσίαζαν πολύ σύντομες διαγώνιες στα RP τους. Από την άλλη πλευρά, οι καθαρά ντετερμινιστικές διαδικασίες θα οδηγούσαν σε μεγαλύτερες διαγώνιες και λιγότερο απομονωμένες μονάδες στα RPs τους.

$$DET = \frac{\sum_{l=d_m}^{N} lP_d(l)}{\sum_{l=1}^{N} lP_d(l)} \tag{2.41}$$

Μέγιστο μήκος Διαγώνιας Γραμμής ($\mathbf{L_{max}}$): το μέγιστο μήκος μιας διαγωνίου γραμμής που βρέθηκε. Το L_{max} αντικατοπτρίζει το αντίστροφο της εκθετικής απόκλισης των τροχιών του PS. Για μακρύτερες διαγώνιες, εάν οι τροχιές του PS αποκλίνουν γρήγορα τότε αναμένουμε να δούμε κοντύτερες διαγώνιες δομές και κατά συνέπεια χαμηλότερες τιμές για L_{max}.

$$L_{max} = max(\{l_i\}_{i=1}^{N_d}) \tag{2.42}$$

Μέσο μήκος Διαγώνιας Γραμμής (L): το μέσο μήκος όλων των διαγώνων γραμμών που βρίσκονται στο RP. Διαισθητικά, μια διαγώνια γραμμή μήκους l σημαίνει ότι οι τροχιές αναπτύσσονται

από κοινού κατά τη διάρκεια δειγμάτων l αλλά αντιστοιχούν σε διαφορετικούς χρόνους της εξέλιξης του συστήματος. Αυτές οι γραμμές υποδεικνύουν τον τρόπο με τον οποίο διαφορετικές τροχιές αποκλίνουν κατά την εξέλιξη του συστήματος όσο περνά ο χρόνος. Αυτό το μέσο μήκος είναι πραγματικά ο μέσος χρόνος που μπορούμε να προβλέψουμε την επόμενη επανάληψη μιάς κατάστασης από την κατάσταση που παρατηρούμε τώρα.

$$L = \frac{\sum_{l=d_m}^{N} lP_d(l)}{\sum_{l=d_m}^{N} P_d(l)} \tag{2.43}$$

Διαγώνια Εντροπία (DENTR): αναφέρεται στην εντροπία Shannon της πιθανότητας να βρεθεί μια διαγώνια γραμμή μήκους ακριβώς ίση με l. Αυτή η εντροπία των διαγώνιων γραμμών αντικατοπτρίζει την πολυπλοκότητα του RP. Ως εκ τούτου, θα περίμενε κανείς ότι για τα τμήματα ομιλίας χωρίς φωνή ή τις περιοχές λευκού θορύβου η τιμή αυτή θα είναι σχεδόν μηδενική, καθώς και η πολυπλοκότητα του υπό ανάλυση συστήματος.

$$DENTR = \sum_{l=d_m}^{N} \frac{P_d(l)}{N_d} ln(\frac{N_d}{P_d(l)}) \tag{2.44}$$

Λαμιναρότητα (LAM): η αναλογία των κάθετων γραμμών τουλάχιστον του μήκους u_{min} στο συνολικό αριθμό των σημείων του RP. Αυτές οι κατακόρυφες δομές αντιπροσωπεύουν την ύπαρξη στρωματοποιημένων φάσεων που παρουσιάζει το υποκείμενο σύστημα χωρίς να υποδεικνύει τις χρονικές περιόδους που αυτές εμφανίζονται.

$$LAM = \frac{\sum_{l=v_m}^{N} lP_v(l)}{\sum_{l=1}^{N} lP_v(l)} \tag{2.45}$$

Μέγιστο μήκος καθέτου γραμμής ($\mathbf{V_{max}}$): το μέγιστο μήκος όλων των κάθετων γραμμών που βρίσκονται στο RP. Υποδεικνύει τη μέγιστη χρονική περίοδο που θα παραμείνει το υποκείμενο σύστημα σε ελασματοποιημένη κατάσταση.

$$V_{max} = max(\{l_i\}_{i=1}^{N_v}) \tag{2.46}$$

Χρόνος παγίδευσης (TT): το μέσο μήκος όλων των κάθετων γραμμών που βρίσκονται στο RP. Αντανακλά τον μέσο χρόνο των στρωματοποιημένων φάσεων του συστήματος που επαναλαμβάνονται κάτω από το καθορισμένο πλαίσιο ανάλυσης.

$$TT = \frac{\sum_{l=v_m}^{N} lP_v(l)}{\sum_{l=v_m}^{N} P_v(l)} \tag{2.47}$$

Εντροπία κατακόρυφων γραμμών (VENTR): αναφέρεται στην εντροπία Shannon της πιθανότητας να βρεθεί μια κάθετη γραμμή μήκους ακριβώς ίση με l. Αυτή η εντροπία κάθετων γραμμών αντικατοπτρίζει τη διανομή των χρονικών περιόδων για τις οποίες το σύστημα ακολουθεί τις στρωματοειδείς φάσεις.

$$VENTR = \sum_{l=v_m}^{N} \frac{P_v(l)}{N_v} ln(\frac{N_v}{P_v(l)}) \tag{2.48}$$

Μέγιστο μήκος λευκής καθέτου γραμμής ($\mathbf{W_{max}}$): το μέγιστο μήκος όλων των λευκών γραμμών που βρίσκονται στο RP. Οι λευκές περιοχές ή ζώνη είναι ενδεικτικές της απότομης αλλαγής της δυναμικής. Έτσι, οι μακρύτερες λευκές κατακόρυφες γραμμές υποδεικνύουν την ύπαρξη σπάνιων συμβάντων στο RP που ίσως να μην προβλεφθούν χρησιμοποιώντας πληροφορίες από προηγούμενες τροχιές του PS.

$$W_{max} = max(\{l_i\}_{i=1}^{N_w}) \tag{2.49}$$

Μέσο Μήκος Λευκών Κάθετων Γραμμών (AWVL): το μέσο μήκος όλων των λευκών γραμμών που βρέθηκαν στο RP. Οι μέσες χρονικές περίοδοι μεταξύ των απότομων αλλαγών του συστήματος, από την μία κατάσταση στην άλλη.

$$AWVL = \frac{\sum_{l=w_m}^{N} lP_w(l)}{\sum_{l=w_m}^{N} P_w(l)} \tag{2.50}$$

Λευκή κατακόρυφη εντροπία (WENTR): αναφέρεται στην εντροπία Shannon της πιθανότητας να βρεθεί μια λευκή κάθετη γραμμή μήκους ακριβώς ίση με l. Αυτή η εντροπία της λευκής κάθετης αντανακλά τη διανομή χρονικών περιόδων που εμφανίζονται απότομες μεταβολές της δυναμικής του υποκείμενου συστήματος.

$$WENTR = \sum_{l=w_m}^{N} \frac{P_w(l)}{N_w} ln(\frac{N_w}{P_w(l)}) \tag{2.51}$$

2.7 Πολυδιάστατη κλιμάκωση (MDS)

2.7.1 Κλασική MDS

Τα κλασσικό MDS παρουσιάστηκαν για πρώτη φορά στο [99] και μπορεί να τυποποιηθεί ως εξής. Λαμβάνοντας υπόψη τον πίνακα που αποτελείται από αποστάσεις ανά ζεύγη ή ανομοιότητες $\{\delta_{ij}\}_{1\leq i,j\leq N}$ μεταξύ N σημείων, η λύση για την κλασική MDS δίνεται από ένα σύνολο σημείων $\{\mathbf{x}_i\}_{i=1}^{N}$ που βρίσκονται στην πολλαπλότητα $\mathcal{M} \in \mathbb{R}^L$ και οι αποστάσεις ανά ζεύγη είναι σε θέση να διατηρήσουν όσο το δυνατόν πιστότερα τις δεδομένες ανομοιότητες $\{\delta_{ij}\}_{1\leq i,j\leq N}$. Κάθε σημείο $\mathbf{x}_i \in \mathbb{R}^L,\ 1 \leq i \leq N$ αντιστοιχεί σε μια στήλη του πίνακα $\mathbf{X}^T \in \mathbb{R}^{L\times N}$. Η διάσταση εμβλυθινσης L επιλέγεται όσο το δυνατόν μικρότερη ώστε να επιτυγχάνεται η μέγιστη μείωση των διαστάσεων αλλά και να μπορεί να προσεγγίσει τις δεδομένες ανομοιότητες δ_{ij} με τις ευκλείδειες αποστάσεις $d_{ij}(\mathbf{X}) = ||\mathbf{x}_i - \mathbf{x}_j||_2 = \sqrt{\sum_{k=1}^{L}(x_{ik} - x_{jk})^2}$ στον εμβυθισμένο χώρο $\mathbb{R}^L$.

Ο προτεινόμενος αλγόριθμος χρησιμοποιεί ένα πλέγμα κεντραρίσματος $\mathbf{H} = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N^T\mathbf{1}_N$ για να αφαιρέσει των στηλών και των σειρών για κάθε στοιχείο. Όπου $\mathbf{1}_N = [1, 1, ..., 1]$ ένα διάνυσμα από αυτά στο $\mathbb{R}^N$ space. Εφαρμόζοντας το διπλό κεντράρισμα στο προϊόν Hadamard των δεδομένων ανισοτήτων, ο πίνακας Gram $\mathbf{B}$ κατασκευάζεται ως εξής:

$$\mathbf{B} = -\frac{1}{2}\mathbf{H}^T(\mathbf{\Delta} \odot \mathbf{\Delta})\mathbf{H} \tag{2.52}$$

Μπορούμε να δείξουμε (κ. 12 [100]) ότι το κλασικό MDS ελαχιστοποιεί το αλγεβρικό κριτήριο Στέλεχος στο Equation 2.53 παρακάτω:

$$||\mathbf{X}\mathbf{X}^T - \mathbf{B}||_F^2 \tag{2.53}$$

Η ιδιοδιάσπαση της συμμετρικής μήτρας $\mathbf{B}$ μας δίνει $\mathbf{B} = \mathbf{V}\mathbf{V}^T$ και έτσι το νέο σύνολο σημείων η ενσωμάτωση στο $\mathbb{R}^L$ δίνεται από τις πρώτες L θετικές ιδιοτιμές του , δηλαδή $\mathbf{X} = \mathbf{V}_L$. Αυτή η λύση παρέχει το ίδιο αποτέλεσμα με την ανάλυση σε κυρίαρχες συνιστώσες (PCA) που εφαρμόζεται στον φορέα στον χώρο υψηλής διαστάσεων [101].Το κλασικό MDS προτάθηκε αρχικά για μήτρες ανομοιότητας οι οποίες μπορούν να ενσωματωθούν με καλή ακρίβεια προσέγγισης σε έναν ευκλείδειο χώρο χαμηλού διαστάσεων. Ωστόσο, πίνακες που αντιστοιχούν σε ενσωματώσεις σε ευκλείδειους υποτομείς [102], έχουν επίσης μελετηθεί οι δίσκοι Poincare [103] και οι σταθεροί καμπυλότητα Riemannian spaces [104].

2.7.2 Μετρικό MDS

Το μετρικό MDS περιγράφει μια υπερσύνολο προβλημάτων βελτιστοποίησης που περιέχουν κλασσικό MDS. Ο Shepard έχει εισαγάγει ευρετικές μεθόδους για να επιτρέψει μετασχηματισμούς των δοσμένων ανισοτήτων, αλλά δεν έδωσε καμία λειτουργία απώλειας προκειμένου να τις μοντάρει [107].

Το Kruskal στο [108] και το [109] επισημοποίησε το μετρικό MDS ως πρόβλημα βελτιστοποίησης ελαχίστων τετραγώνων ελαχιστοποιώντας την μη κυρτή συνάρτηση Stress-1 που ορίζεται στην εξίσωση 2.54

$$\sigma_1(\mathbf{X}, \hat{\mathbf{D}}) = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(\hat{d_{ij}} - d_{ij}(\mathbf{X}))}{\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2(\mathbf{X})}} \tag{2.54}$$

όπου ο πίνακας $\hat{\mathbf{D}}$ με στοιχεία $\hat{d_{ij}}$ αντιπροσωπεύει όλα τα ζεύγη μετασχηματισμένων ανισοτήτων δ_{ij} που χρησιμοποιούνται για να χωρέσουν τα ενσωματωμένα ζεύγη αποστάσεων $d_{ij}(\mathbf{X})$.

Στην ουσία, $\mathcal{F}$ είναι συνήθως ένας αφινικός μετασχηματισμός $\hat{d_{ij}} = \alpha + \beta\delta_{ij}$ για άγνωστο α και β. Επιπρόσθετα, χρησιμοποιούνται μετασχηματισμοί μονοτονικής και πολυωνυμικής παλινδρόμησης για μη-μετρικούς-MDS, καθώς και μια ευρύτερη οικογένεια μετασχηματισμών [110]. Ο Kruskal πρότεινε έναν επαναληπτικό αλγόριθμο με βάση την κλίση για την ελαχιστοποίηση του σ_1 αφού η λύση δεν μπορεί να εκφραστεί σε κλειστή μορφή.

Υποθέτοντας ότι $d_{ij} \hat{=} \delta_{ij}$ ο αλγόριθμος προσπαθεί με απαναλήψεις να βρει τις συντεταγμένες των σημείων $\mathbf{X}$ που βρίσκονται στο χαμηλό χώρο ενσωμάτωσης $\mathbb{R}^L$. Οι τετριμένες λύσεις ($\mathbf{X} = \mathbf{0}$ και $\hat{\mathbf{D}} = \mathbf{0}$) αποφεύγονται από τον όρο που βρίσκεται στον παρονομαστή της Εξίσωσης 2.54.

Μια σταθμισμένη συνάρτηση MDS raw Stress ορίζεται ως:

$$\sigma_{raw}^2(\mathbf{X}, \hat{\mathbf{D}}) = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(\hat{d_{ij}} - d_{ij}(\mathbf{X}))^2 \tag{2.55}$$

όπου τα βάρη w_{ij} περιορίζονται να είναι μη αρνητικά, για τα ελλείποντα δεδομένα, τα βάρη ορίζονται ως μηδέν. Με τη ρύθμιση $w_{ij} = 1, \ \forall 1 \leq i, j \leq N$ μπορούμε να διαμορφώσουμε ίση συμβολή στην λύση Μετρικού-MDS για όλα τα στοιχεία.

2.7.3 SMACOF

Το SMACOF, το οποίο αντιπροσωπεύει την κλιμάκωση, με την επικέντρωση σε μια σύνθετη συνάρτηση είναι ένας αλγόριθμος για την επίλυση του μετρικού MDS που συνήθως χρησιμοποιείται ως ο καλύτερος αλγόριθμος για να λύσει το MDS και παρουσιάστηκε στο [111]. Με τη ρύθμιση $\hat{d_{ij}} = \delta_{ij}$ σε συνάρτηση Stress που ορίζεται στην Εξίσωση 2.55, ο αλγόριθμος SMACOF ελαχιστοποιεί την προκύπτουσα συνάρτηση Stress $\sigma_{raw}^2(X)$.

$$\sigma^2(\mathbf{X}) = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(\delta_{ij}^2 - 2\delta_{ij}d_{ij}(\mathbf{X}) + d_{ij}^2(\mathbf{X})) \tag{2.56}$$

Ο αλγόριθμος προχωράει επαναληπτικά και μειώνει την Stress συνάρτηση μονοτονικά μέχρι ένα σταθερό σημείο βελτιστοποιώντας μια κυρτή συνάρτηση που χρησιμεύει ως ανώτερο όριο για τη μη κυρτή συνάρτηση τάσης (εξίσωση 2.56). Μια εκτεταμένη περιγραφή του SMACOF βρίσκεται στο [100] ενώ η σύγκλιση του για ένα ευκλείδειο ενσωματωμένο χώρο $\mathbb{R}^L$ έχει αποδειχθεί στο [112].

Έστω ότι οι πίνακες $\mathbf{U}$ και $\mathbf{R}(\mathbf{X})$ ορίζονται στοιχείο προς στοιχείο ως εξής:

$$u_{ij} = \begin{cases} -w_{ij} & i \neq j \\ \sum_{k \neq i} w_{ik} & i = j \end{cases} \tag{2.57}$$

$$r_{ij} = \begin{cases} -w_{ij}\delta_{ij}d_{ij}^{-1}(\mathbf{X}) & i \neq j, d_{ij}(\mathbf{X}) \neq 0 \\ 0 & i \neq j, d_{ij}(\mathbf{X}) = 0 \\ \sum_{k \neq i} r_{ik} & i = j \end{cases} \tag{2.58}$$

Η συνάρτηση Stress στην εξίσωση 2.56 μετατρέπεται στην ακόλουθη τετραγωνική μορφή:

$$\sigma^2(\mathbf{X}) = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\delta_{ij}^2 - 2tr(\mathbf{X}^T\mathbf{R}(\mathbf{X})\mathbf{X}) + tr(\mathbf{X}^T\mathbf{U}\mathbf{X}) \tag{2.59}$$

Η προκύπτουσα τετραγωνική μορφή μπορεί να ελαχιστοποιηθεί επαναληπτικά ως εξής:

$$T(\mathbf{X}, \hat{\mathbf{X}}^{(k)}) = \sum_{j=1}^{N} w_{ij}\delta_{ij}^2 - 2tr(\mathbf{X}^T\mathbf{R}(\hat{\mathbf{X}}^{(k)})\hat{\mathbf{X}}^{(k)}) + tr(\mathbf{X}^T\mathbf{U}\mathbf{X}) \tag{2.60}$$

$$\hat{\mathbf{X}}^{(k+1)} = \underset{\mathbf{X}}{\operatorname{argmin}}\ T(\mathbf{X}, \hat{\mathbf{X}}^{(k)}) = \mathbf{U}^{\dagger}\mathbf{R}(\hat{\mathbf{X}}^{(k)})\hat{\mathbf{X}}^{(k)} \tag{2.61}$$

όπου $\hat{\mathbf{X}}^{(k)}$ είναι η εκτίμηση των τιμών του πίνακα $\mathbf{X}$ στην k-οστή επανάληψη και $\mathbf{U}^{\dagger}$ είναι η ψευδο-στροφή μήτρα Moore-Penrose του πίνακα $\mathbf{U}$. Σημειωτέον ότι ο όρος $\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\delta_{ij}^2$ αντιστοιχεί σε μια σταθερά που ορίζεται μόνο από την αρχική το γράφημα βάρους $\mathbf{W}$ και το δεδομένο disimilaritires δ_{ij}. Στην επανάληψη k η κυρτή κυρτή κυρτή συνάρτηση αγγίζει την επιφάνεια του σ στο σημείο $\hat{\mathbf{X}}^{(k)}$. Με την ελαχιστοποίηση αυτής της απλής τετραγωνικής συνάρτησης στην Εξίσωση 2.60 βρίσκουμε την επόμενη ενημέρωση που χρησιμεύει ως αφετηρία για την επόμενη επανάληψη $k+1$. Η λύση στο πρόβλημα ελαχιστοποίησης παρουσιάζεται στην Εξίσωση 2.61. Ο αλγόριθμος σταματά όταν η νέα ενημερωμένη έκδοση αποφέρει μια μείωση $\sigma^2(\hat{\mathbf{X}}^{(k+1)}) - \sigma^2(\hat{\mathbf{X}}^{(k)})$ που είναι μικρότερη από μια τιμή κατωφλίου.

2.8 Μέθοδοι Γενικής Αναζήτησης Προτύπων (GPS)

2.8.1 Σύνταξη GPS

Το μη περιορισμένο πρόβλημα της ελαχιστοποίησης μιας συνεχώς διαφοροποιήσιμης συνάρτησης $f : \mathbb{R}^n \to \mathbb{R}$ τυπικά περιγράφεται ως:

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}}\ f(\mathbf{x}) \tag{2.62}$$

Στη συνέχεια παρουσιάζουμε μια σύντομη περιγραφή της επαναληπτικής ελαχιστοποίησης που πραγ-ματοποιούν οι μέθοδοι GPS της Εξίσωσης 2.62 που βασίζεται στα [113, 114]. Πρώτα πρέπει να ορίσουμε τα ακόλουθα στοιχεία:

- Μία μήτρα βάσης που θα μπορούσε να είναι οποιοσδήποτε κανονικός (nonsingular) πίνακας $\mathbf{B} \in \mathbb{R}^{n\times n}$.
- Ένας πίνακας $\mathbf{C}^{(k)}$ για τη δημιουργία όλων των δυνατών κινήσεων για την kοστή επανάληψη του αλγόριθμου ελαχιστοποίησης

  $$\mathbf{C}^{(k)} = [\mathbf{M}^{(k)} \quad -\mathbf{M}^{(k)} \quad \mathbf{L}^{(k)}] = [\mathbf{\Psi}^{(k)} \quad \mathbf{L}^{(k)}] \tag{2.63}$$

  όπου οι στήλες του πίνακα $\mathbf{M}^{(k)} \in \mathbb{Z}^{n\times n}$ σχηαμτίζουν ένα θετικό ανάπτυγμα (span) του χώρου $\mathbb{R}^n$ και ο πίνακας $\mathbf{L}^{(k)}$ περιέχει τουλάχιστον το μηδενικό διάνυσμα του χώρου αναζήτησης $\mathbb{R}^n$.
- Ένας πίνακας προτύπων $\mathbf{P}^{(k)}$ ορίζεται ως

  $$\mathbf{P}^{(k)} = \mathbf{B}\mathbf{C}^{(k)} = [\mathbf{B}\mathbf{M}^{(k)} \quad -\mathbf{B}\mathbf{M}^{(k)} \quad \mathbf{B}\mathbf{L}^{(k)}] \tag{2.64}$$

  όπου η υπομήτρα $\mathbf{B}\mathbf{M}^{(k)}$ σχηματίζει μια βάση του χώρου $\mathbb{R}^n$.

Σε κάθε επανάληψη k, ορίζουμε ένα σύνολο βημάτων $\{\mathbf{s}_i^{(k)}\}_{i=1}^{m}$ που δημιουργούνται από το πρό-τυπο matrix $\mathbf{P}^{(k)}$ όπως φαίνεται παρακάτω:

$$\mathbf{s}_i^{(k)} = \Delta^{(k)}\mathbf{p}_i^{(k)}, \quad \mathbf{P}^{(k)} = [\mathbf{p}_1^{(k)}, ..., \mathbf{p}_m^{(k)}] \in \mathbb{R}^{n\times m} \tag{2.65}$$

όπου $\mathbf{p}_i^{(k)}$ είναι η i-οστή στήλη του πίνακα προτύπων $\mathbf{P}^{(k)}$ και καθορίζει την κατεύθυνση του νέου βήματος, ενώ το $\Delta^{(k)}$ ρυθμίζει το μήκος προς αυτήν την κατεύθυνση. Εάν ο πίνακας προτύπων $\mathbf{P}^{(k)}$

περιέχει m στήλες, τότε πρέπει $m \geq n+1$ για να επεκταθεί θετικά ο χώρος αναζήτησης $\mathbb{R}^n$. Έτσι, ένα νέο σημείο δοκιμής του αλγορίθμου GPS προς το βήμα αυτό θα είναι $\mathbf{x}_i^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{s}^{(k)}$ όπου αξιολογούμε την τιμή της συνάρτησης f για να ελαχιστοποιήσουμε. Η επιτυχία ενός νέου δοκιμαστικού σημείου αποφασίζεται με βάση την προϋπόθεση ότι κάνει ένα βήμα προς την περαιτέρω ελαχιστοποίηση της συνάρτησης f, δηλαδή $f(\mathbf{x}^{(k)} + \mathbf{s}_i^{(k)}) > f(\mathbf{x}_i^{(k+1)})$. Τα βήματα μιας μεθόδου GPS παρουσιάζονται στον αλγόριθμο 1.

Algorithm 1 Αναζήτηση γενικών προτύπων (GPS)

1: **procedure** GPS_SOLVER($\mathbf{x}^{(0)}$, $\Delta^{(0)}$, $\mathbf{C}^{(0)}$, $\mathbf{B}$)
2: $k = -1$
3: **do**
4: $k = k + 1$
5: $\mathbf{s}^{(k)} =$ EXPLORE_MOVES($\mathbf{BC}^{(k)}$, $\mathbf{x}^{(k)}$, $\Delta^{(k)}$)
6: $\rho^{(k)} = f(\mathbf{x}^{(k)} + \mathbf{s}^{(k)}) - f(\mathbf{x}^{(k)})$
7: **if** $\rho^{(k)} < 0$ **then**
8: $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{s}^{(k)}$ ▷ Επιτυχής επανάληψη
9: **else**
10: $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)}$ ▷ Ανεπιτυχής επανάληψη
11: $\Delta^{(k+1)}, \mathbf{C}^{(k+1)} =$ UPDATE($\mathbf{C}^{(k)}$, $\Delta^{(k)}$, $\rho^{(k)}$)
12: **while** Κριτήριο σύγκλισης δεν ικανοποιείται

Για να αρχικοποιήσουμε τον αλγόριθμο επιλέγουμε ένα σημείο $\mathbf{x}^{(0)} \in \mathbb{R}^n$ και μια θετική παράμετρο μήκους βήματος $\Delta^{(0)} > 0$. Σε κάθε επανάληψη k, διερευνάμε ένα σύνολο κινήσεων που ορίζονται από την `EXPLORE_MOVES()` υπορουτίνα στη γραμμή 5 του αλγορίθμου που περιγράφονται με τη χρήση ενός φορμαλισμού μεθόδων GPS. Όλες αυτές οι μέθοδοι διαφέρουν μεταξύ τους κυρίως στους τρόπους που χρησιμοποιούνται για την επιλογή διερευνητικών κινήσεων. Αν ένα νέο ερευνητικό σημείο μειώνει την αξία της συνάρτησης f, τότε η επανάληψη k είναι επιτυχής και το σημείο εκκίνησης της επόμενης επαναληψιμότητας ενημερώνεται κατάλληλα $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{s}^{(k)}$ όπως φαίνεται στη γραμμή 8, αλλιώς δεν υπάρχει ενημέρωση. Η παράμετρος μήκους βήματος $\Delta^{(k)}$ τροποποιείται από την `UPDATE()` υπορουτίνα στη γραμμή 11. Για επιτυχείς επαναλήψεις, δηλ., $\rho^{(k)} < 0$, το μήκος βημάτων αναγκάζεται να αυξηθεί κατά οριστικό τρόπο ως εξής:

$$\begin{aligned} &\Delta^{(k+1)} = \lambda^{(k)}\Delta^{(k)}, \;\; \lambda^{(k)} \in \Lambda = \{\tau^{w_1}, ..., \tau^{w_{|\Lambda|}}\} \\ &\tau > 1, \;\; \{w_1, ..., w_{|\Lambda|}\} \subset \mathbb{N}, \;\; |\Lambda| < +\infty \end{aligned} \tag{2.66}$$

όπου τ και w_i είναι προκαθορισμένες σταθερές που χρησιμοποιούνται για την επιτυχημένη επανάληψη i. Για μη επιτυχείς επαναλήψεις, η παράμετρος μήκους βήματος μειώνεται, δηλ. $\Delta^{(k+1)} \leq \Delta^{(k)}$ ως εξής:

$$\Delta^{(k+1)} = \theta\Delta^{(k)}, \;\; \theta = \tau^{w_0}, \;\; \tau > 1, \;\; w_0 < 0, \tag{2.67}$$

όπου τ και ο αρνητικός ακέραιος w_0 καθορίζουν τον σταθερό λόγο της μείωσης βημάτων. Σημειώνουμε επίσης ότι η γεννήτορας μήτρα $\mathbf{C}^{(k+1)}$ θα μπορούσε επίσης να ενημερωθεί για ανεπιτυχείς / επιτυχείς επαναλήψεις για να περιέχει περισσότερες / λιγότερες κατευθύνσεις αναζήτησης, αντίστοιχα.

2.8.2 Σύγκλιση GPS

Οι μέθοδοι GPS στο πλαίσιο του προαναφερθέντος καθορισμένου πλαισίου έχουν ορισμένες σημαντικές ιδιότητες σύγκλισης που παρουσιάζονται στο παρόν κεφάλαιο και έχουν αποδειχθεί στα [113, 114, 115, 116, 117]. Για κάθε μέθοδο GPS που ικανοποιεί τις προδιαγραφές της Υπόθεσης 1 στις εξερευνητικές κινήσεις, μπορεί κανείς να δείξει σύγκλιση για τον αλγόριθμο 1.

Υπόθεση 1 (Ασθενής υπόθεση στις εξερευνητικές κινήσεις): *Η υπορουτίνα* `EXPLORE_MOVES()` *όπως ορίζεται στον αλγόριθμο 1, γραμμή 5 εγκυάται τα κατόθι:*

- *Η διερευνητική βηματική κατεύθυνση για την επανάληψη k επιλέγεται από τις στήλες του μήτρας προτύπου $\mathbf{P}^{(k)}$ όπως ορίζεται στο βήμα δοκιμής εξισώσεων 2.65 και το μήκος εξερευνητικού βήματος είναι $\Delta^{(k)}$ όπως ορίζεται στις εξισώσεις 2.66, 2.67.*

- *Εάν ανάμεσα στις ερευνητικές κατευθύνσεις $\mathbf{a}^{(k)}$ στην επανάληψη k επιλεγμένα από τις στήλες του πίνακα $\Delta^{(k)}\mathbf{B}[\mathbf{M}^{(k)} - \mathbf{M}^{(k)}]$ υπάρχει τουλάχιστον μία κατεύθυνση που οδηγεί σε επιτυχημένη επανάληψη, δηλαδή, $f(\mathbf{x}^{(k)} + \mathbf{a}) < f(\mathbf{x}^{(k)})$, τότε η υπορουτίνα* `EXPLORE_MOVES()` *πρέπει να επιστρέψει μία κίνηση $\mathbf{s}^{(k)}$ για την οποία ισχύει ότι:*
  $f(\mathbf{x}^{(k)} + \mathbf{s}^{(k)}) < f(\mathbf{x}^{(k)})$.

Η υπόθεση 1 επιβάλλει κάποιους ήπιους περιορισμούς στη διαμόρφωση των εξερευνητικών κινήσεων που παράγονται από το αλγόριθμο 1 γραμμή 5. Ουσιαστικά, το προτεινόμενο βήμα $\mathbf{s}^{(k)}$ προέρχεται από το πίνακα προτύπων $\mathbf{P}^{(k)}$, ενώ ο αλγόριθμος πρέπει να παρέχει μια απλή μείωση για την συνάρτηση στόχο f. Συγκεκριμένα, ο μόνος τρόπος αποδοχής μιας μη επιτυχημένης επαναληψιμότητας θα ήταν εάν κανένα από τα βήματα από τις στήλες του πίνακα $\Delta^{(k)}\mathbf{B} = [\mathbf{M}^{(k)} - \mathbf{M}^{(k)}]$ οδηγεί σε μείωση της συνάρτησης στόχου f. Βάσει αυτής της υπόθεσης μπορεί κανείς να διατυπώσει το Θεώρημα. 1 ως εξής:

Θεώρημα 1: *Έστω $L(\mathbf{x}^{(0)}) = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^{(0)})\}$ να είναι κλειστό και φραγμένο και η συνάρτηση f να είναι συνεχώς παραγωγίσιμη σε μία γειτονιά $L(\mathbf{x}^{(0)})$, συγκεκριμένα στην ένωση των ανοιχτών σφαιρών $\bigcup_{\mathbf{a} \in L(\mathbf{x}^{(0)})} B(\mathbf{a}, \eta)$ όπου $\eta > 0$. Αν μια μέθοδος GPS έχει διατυπωθεί όπως περιγράφεται στην ενότητα 2.8.1 και η Υπόθεση 1 ισχύει, τότε για την ακολουθία των επαναλήψεων $\{\mathbf{x}^{(k)}\}$ που παράγονται από τον αλγόριθμο 1 θα ισχυεί το εξής:*

$$\lim_{k \to +\infty} \inf \, ||\nabla f(\mathbf{x}^{(k)})|| = 0$$

Για την απόδειξη αυτού του θεωρήματος παραπέμπουμε τον αναγνώστη στο [113].

Όπως φαίνεται στο [118], μπορεί κανείς να κατασκευάσει μια συνεχώς διαφοροποιήσιμη συνάρτηση στόχου και μια μέθοδο GPS με άπειρα πολλά οριακά σημεία με μη μηδενικές κλίσεις και επομένως ακόμη και αν ισχύει το Θεώρημα 1, η σύγκλιση του $||\nabla f(\mathbf{x}^{(k)})||$ δεν είναι εξασφαλισμένη. Ωστόσο, οι ιδιότητες σύγκλισης των μεθόδων GPS μπορούν να ενισχυθούν περαιτέρω εάν πληρούνται πρόσθετα κριτήρια. Συγκεκριμένα, μια ισχυρότερη υπόθεση σχετικά με τις εξερευνητικές κινήσεις είναι η παρακάτω υπόθεση 2 η οποία ουσιαστικά ρυθμίζει το μέτρο μείωσης της αντικειμενικής συνάρτησης για κάθε βήμα που παράγεται από τη μέθοδο GPS, ως εξής:

Υπόθεση 2 (Ισχυρή υπόθεση στις εξερευνητικές κινήσεις): *Η υπορουτίνα* `EXPLORE_MOVES()` *όπως ορίζεται στον αλγόριθμο 1, γραμμή 5 εγκυάται τα κατόθι:*

- *Η διερευνητική βηματική κατεύθυνση για την επανάληψη k επιλέγεται από τις στήλες του μήτρας προτύπου $\mathbf{P}^{(k)}$ όπως ορίζεται στο βήμα δοκιμής εξισώσεων 2.65 και το μήκος εξερευνητικού βήματος είναι $\Delta^{(k)}$ όπως ορίζεται στις εξισώσεις 2.66, 2.67.*

- *Εάν ανάμεσα στις ερευνητικές κατευθύνσεις $\mathbf{a}^{(k)}$ στην επανάληψη k επιλεγμένα από τις στήλες του πίνακα $\Delta^{(k)}\mathbf{B}[\mathbf{M}^{(k)} - \mathbf{M}^{(k)}]$ υπάρχει τουλάχιστον μία κατεύθυνση που οδηγεί σε επιτυχημένη επανάληψη, δηλαδή, $f(\mathbf{x}^{(k)} + \mathbf{a}) < f(\mathbf{x}^{(k)})$, τότε η υπορουτίνα* `EXPLORE_MOVES()` *πρέπει να επιστρέψει μία κίνηση $\mathbf{s}^{(k)}$ για την οποία ισχύει ότι:*
  $f(\mathbf{x}^{(k)} + \mathbf{s}^{(k)}) \leq \min_{\mathbf{a}^{(k)}} f(\mathbf{x}^{(k)} + \mathbf{a}^{(k)})$.

Η υπόθεση 2 επιβάλλει τον πρόσθετο ισχυρό περιορισμό στη διαμόρφωση των εξερευνητικών κινήσεων, δηλαδή ότι η υπορουτίνα EXPLORE_MOVES() δεν θα επιδεινώσει την τιμή της συναρτησης στόχου περισσότερο από την καλύτερη διερευνητική κίνηση από το τις στήλες της μήτρας $\Delta^{(k)}\mathbf{B}[\mathbf{M}^{(k)} - \mathbf{M}^{(k)}]$.

Βάσει αυτής της υπόθεσης και προσθέτοντας απαιτήσεις που περιορίζουν την κατεύθυνση και το μήκος του βήματος εξερεύνησης για τη μέθοδο GPS, μπορεί κανείς να διατυπώσει Θεώρημα 2 που παρουσιάζεται και εδώ χωρίς απόδειξη.

Θεώρημα 2: *Έστω $L(\mathbf{x}^{(0)}) = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^{(0)})\}$ να είανι κλειστό και φραγμένο και η συνάρτηση f να είναι συνεχώς παραγωγίσιμη σε μία γειτονιά $L(\mathbf{x}^{(0)})$, συγκεκριμένα στην ένωση των ανοιχτών σφαιρών $\bigcup_{\mathbf{a}\in L(\mathbf{x}^{(0)})} B(\mathbf{a}, \eta)$ όπου $\eta > 0$. Αν μια μέθοδος GPS έχει διατυπωθεί όπως περιγράφεται στην ενότητα 2.8.1, $\lim_{k\to+\infty} \Delta^{(k)} = 0$, οι στήλες της γεννήτορος μήτρας $\mathbf{C}^{(k)}$ φράζονται από μία νόρμα και η υπόθεση 2 ισχύει, τότε για την ακολουθία των επαναλήψεων $\{\mathbf{x}^{(k)}\}$ που παράγονται από τον αλγόριθμο 1 θα ισχυεί το εξής:*

$$\lim_{k\to+\infty} ||\nabla f(\mathbf{x}^{(k)})|| = 0$$

Οι πρόσθετες απαιτήσεις ορίζουν ότι: 1) η γεννήτορα μήτρα $\mathbf{C}^{(k)}$ θα πρέπει να οριοθετείται από νόρμα για να παράγει δοκιμαστικά βήματα από την εξίσωση 2.65 τα οπία οριοθετούνται από την παράμετρος μήκους βήματος $\Delta^{(k)}$ και 2) $\lim_{k\to+\infty} \Delta^{(k)} = 0$ που μπορεί εύκολα να επιτευχθεί επιλέγοντας $\Lambda = \{1\}$ στην Εξίσωση 2.67, αυτό εγγυάται επίσης και μια μη αυξανόμενη ακολουθία $\Delta^{(k)}$ βημάτων [113]. Παρόλο που τα κριτήρια αυτά παρέχουν πολύ ισχυρότερες ιδιότητες σύγκλισης, αντιμετωπίζουμε ένα trade-off μεταξύ της θεωρητικής απόδειξης σύγκλισης και της πραγματικής αποτελεσματικότητας των ευρεστικών μεθόδων στον χώρο αναζήτησης για την εξεύρεση ενός τοπικού βέλτιστου.

Και τα δύο θεωρήματα 1 και 2 παρέχουν μια εγκύηση για την σύγκλιση των GPS μεθόδων εάν τηρούνται οι προδιαγραφές τους. Παρόλο που το τελευταίο τεστ θεώρησης έχει πολύ ισχυρότερη σύγκλιση, η παράμετρος ελέγχου του μήκους βήματος $\Delta^{(k)}$, παρέχει ένα αξιόπιστο ασυμπτωτικό μέτρο της στασιμότητας πρώτης τάξης όταν μειώνεται μετά από ανεπιτυχείς επαναλήψεις [114].

Κεφάλαιο 3

Χρονικές Κλίμακες Απόφασης για Αναγνώριση Συναισθημάτων από Όμιλία

Το κεφάλαιο αυτό είναι μια εκτεταμένη έκδοση του δημοσιευμένου άρθρου [2] στο Διεθνές Συνέδριο Affective Computing and Intelligent Interaction (ACII) που πραγματοποιήθηκε στο Σαν Αντόνιο, Τέξας, ΗΠΑ στις 23-26 Οκτωβρίου 2017. Εάν ο αναγνώστης πρέπει να αναφέρει τμήματα αυτού του κεφαλαίου τότε θα ήταν προτιμότερο να χρησιμοποιήσει την ακόλουθη αναφορά:

- *Efthymios Tzinis and Alexandros Potamianos. "Segment-based speech emotion recognition using recurrent neural networks." In Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on, pp. 190-195. IEEE, 2017.*

3.1 Κίνητρο

Όπως έχουμε συζητήσει στο Τμήμα 1.4, ένα σήμα ομιλίας περικλείει πληροφορίες σε ποικίλες χρονικές κλίμακες. Εστιάζοντας στο έργο στην αναγνώριση συναισθήματος από ομιλία, μπορούμε εύκολα να συμπεράνουμε ότι διαφορετικά σύνολα ακουστικών χαρακτηριστικών θα μπορούσαν να ενεργοποιηθούν σε διαφορετικες χρονικές κλίμακες. Ως εκ τούτου, η απόδοση του χρησιμοποιούμενου συστήματος SER είναι ευαίσθητη σε αυτή τη διαμόρφωση. Παρόλο που οι χρονικές κλίμακες κάτω από τις οποίες εξάγονται τα LLDs ή τα στατιστικά χαρακτηριστικά που βασίζονται στην εφαρμογή στατιστικών συναρτήσεων πάνω απο LLDs έχει σημαντικές επιπτώσεις στην απόδοση SER, έχει σημειωθεί μικρή πρόοδος προς την κατεύθυνση αυτή [53], [54]. Λαμβάνοντας υπόψη την εκφραστικότητα των RNNs όταν μοντελοποιούν σειρές διανυσμάτων χαρακτηριστικών και τη στενή τους σχέση με την χρονική κλίμακα που αντιστοιχεί στα διανύσματα χαρακτηριστικών εισόδου, είναι απαραίτητο να αναλυθεί η αποτελεσματικότητα των διαφόρων χρονικών κλιμάκων στα συστήματα SER που βασίζονται σε RNN. Συγκεκριμένα, η κατάλληλη χρονική κλίμακα απόφασης από την οποία τα RNNs μπορούν να διερευνήσουν μια πιο αφηρημένη αναπαράσταση είτε από τοπικά είτε από στατιστικά χαρακτηριστικά.

3.2 Ακουστικά χαρακτηριστικά

Σε αυτή την ενότητα θα παρουσιάσουμε μερικές μεθόδους που χρησιμοποιήσαμε για την εξαγωγή των ακουστικών τοπικών και παγκόσμιων συνόλων χαρακτηριστικών για την προσέγγισή μας, όπως περιγράφεται στα τμήματα 3.2.3 και 3.2.4. Τα LLDs μοντελοποιούν κάθε έκφραση χρησιμοποιώντας διανύσματα χαρακτηριστικών άμεσα από μικρά χρονικά παράθυρα. Από την άλλη πλευρά, για να εξάγουμε παγκόσμια χαρακτηριστικά πρέπει πρώτα να εξαγάγουμε LLDs και μετά να εφαρμόσουμε στατιστικές συναρτήσεις πάνω σε τμήματα ομιλίας. Τώρα, σε αυτό το πλαίσιο, κάθε συναισθηματική φράση θα μοντελοποιείται χρησιμοποιώντας μια ακολουθία από μία ή περισσότερες στατιστικές αναπαραστάσεις των συμπεριλαμβανόμενων τμημάτων. Στην Ενότητα 3.2.1 επισημαίνουμε μερικές στρατηγικές προεπεξεργασίας που ακολουθούμε και δοκιμάζουμε, πριν από τη διαδικασία εξαγωγής χαρακτηριστικών που βασίζεται σε LLD από πλαίσια φωνής. Επιπρόσθετα, στο τμήμα 3.2.2 αναλύουμε τον μηχανισμό της ανίχνευσης φωνητικής δραστηριότητας (VAD) για το OpenSMILE και

άλλα προγραμματιστικά περιβάλλοντα. Για να εξαγάγουμε τα παραπάνω χαρακτηριστικά χρησιμοποιήσαμε το προγραμματιστικό περιβάλλον OpenSMILE [119] που χρησιμοποιείται επίσης σε πολλά άλλα έργα στη βιβλιογραφία.

3.2.1 Προεπεξεργασία

Οι συμβατικές τεχνικές εξαγωγής ακουστικών χαρακτηριστικών βασίζονται στο σπάσιμο του σήματος το οποίο πρόκειται να αναλυθεί σε επικαλυπτόμενα πλαίσια (10ms-100ms) προκειμένου να εξαχθούν τα απαιτούμενα LLDs [120]. Για κάθε πλαίσιο ομιλίας εφαρμόζουμε παράθυρο Hamming του ιδίου μήκους όπως φαίνεται παρακάτω.

$$s_w(i) = \begin{cases} s(i) \cdot w(i) & c - \frac{W}{2} \leq i \leq c + \frac{W}{2} \\ 0 & otherwise \end{cases} \tag{3.1}$$

όπου s_w αντιστοιχεί στο παραθυροποιημένο σήμα, c είναι ο κεντρικός δείκτης γύρω από τον οποίο θα εφαρμοστεί η συνάρτηση του παραθύρου. Το αντίστοιχο παράθυρο έχει μήκος W δείγματα. Είναι εύκολο να συμπεράνουμε ότι το αντίστοιχο πλαίσιο θα έχει επίσης ένα μήκος ίσο με W καθώς όλες οι άλλες περιοχές θα είναι μηδέν. Παρόλο που έχουν προταθεί ποικίλα παράθυρα για ψηφιακή επεξεργασία σήματος, σε αυτή την εργασία θα χρησιμοποιήσουμε ένα παράθυρο hamming που ορίζεται ως εξής:

$$w_H(i) = \begin{cases} 0.54 + 0.46 \cdot cos(\frac{\pi \cdot i}{W}) & c - \frac{W}{2} \leq i \leq c + \frac{W}{2} \\ 0 & otherwise \end{cases} \tag{3.2}$$

Όπως και ένα Γκαουσιανό παράθυρο που θα μπορούσε να προσεγγιστεί όπως φαίνεται παρακάτω:

$$w_G(i) = \begin{cases} G(i) - \frac{G(-0.5)[G(i+\frac{W}{2}) + G(i-\frac{W}{2})]}{G(-0.5+\frac{W}{2}) + G(-0.5-\frac{W}{2})} & c - \frac{W}{2} \leq i \leq c + \frac{W}{2} \\ 0 & otherwise \end{cases} \tag{3.3}$$

όπου $G(\cdot)$ ορίζεται ως η Γκαουσιανή στην επόμενη Εξίσωση:

$$G(i) = e^{(\frac{i - \frac{W}{4} + 0.5}{0.2})^2} \tag{3.4}$$

Και πάλι, το W ορίζει το μήκος του καθορισμένου παραθύρου.

Μετά την προαναφερθείσα διαδικασία μπορούμε να συσσωρεύσουμε τα παράθυρα πλαισίων ομιλίας και να τα επεξεργαστούμε ξεχωριστά. Σε αυτό το πλαίσιο, υποθέτουμε ότι αυτά τα παράθυρα περιέχουν πληροφορία σε συχνότητας κάτω από 4kHz, η οποία είναι μια αποδεκτή παραδοχή για τα σήματα ομιλίας [6]. Χρησιμοποιούμε ένα φίλτρο προ-έμφασης προκειμένου να ενισχύσουμε τις πληροφορίες που μεταφέρονται σε υψηλότερες συχνότητες και να καταργήσουμε τις αντίστοιχες στα χαμηλότερα, οι οποίες θα μπορούσαν να είναι περιοχές σιωπής ή θορύβου. Το φιλτράρισμα πριν από την έμφαση ορίζεται ως εξής:

$$s_{pe}(i) = s(i) - \alpha \cdot s(i-1) \tag{3.5}$$

όπου το α θεωρείται σταθερό και ίσο με 0.975 [23].

Μετά από αυτή τη διαδικασία είμαστε σε θέση να εκτελέσουμε διάφορες μεθόδους εξαγωγής χαρακτηριστικών οι οποίες βασίζονται στην εξαγωγή πληροφορίας από το φάσμα του σήματος ή χαρακτηριστικών ποιότητας της ομιλίας για να λάβουμε μια στατική αναπαράσταση χαρακτηριστικών LLDs για κάθε πλαίσιο. Η διαδικασία που περιγράψαμε παραπάνω είναι παρόμοια για όλους τους τοπικούς περιγραφητές LLDs που εξάγονται σε όλες τις επόμενες ενότητες αυτής της διπλωματικής, αλλά οι διαμορφώσεις του μήκους του πλαισίου ή η επικάλυψη μεταξύ των πλαισίων θα μπορούσαν να αλλάξουν. Έτσι, θα αναφερθούμε σε αυτή τη διαμόρφωση αλλάζοντας τις απαιτούμενες παραμέτρους.

3.2.2 Ανίχνευση Φωνητικής Δραστηριότητας σε ένα Ακουστικό Σήμα

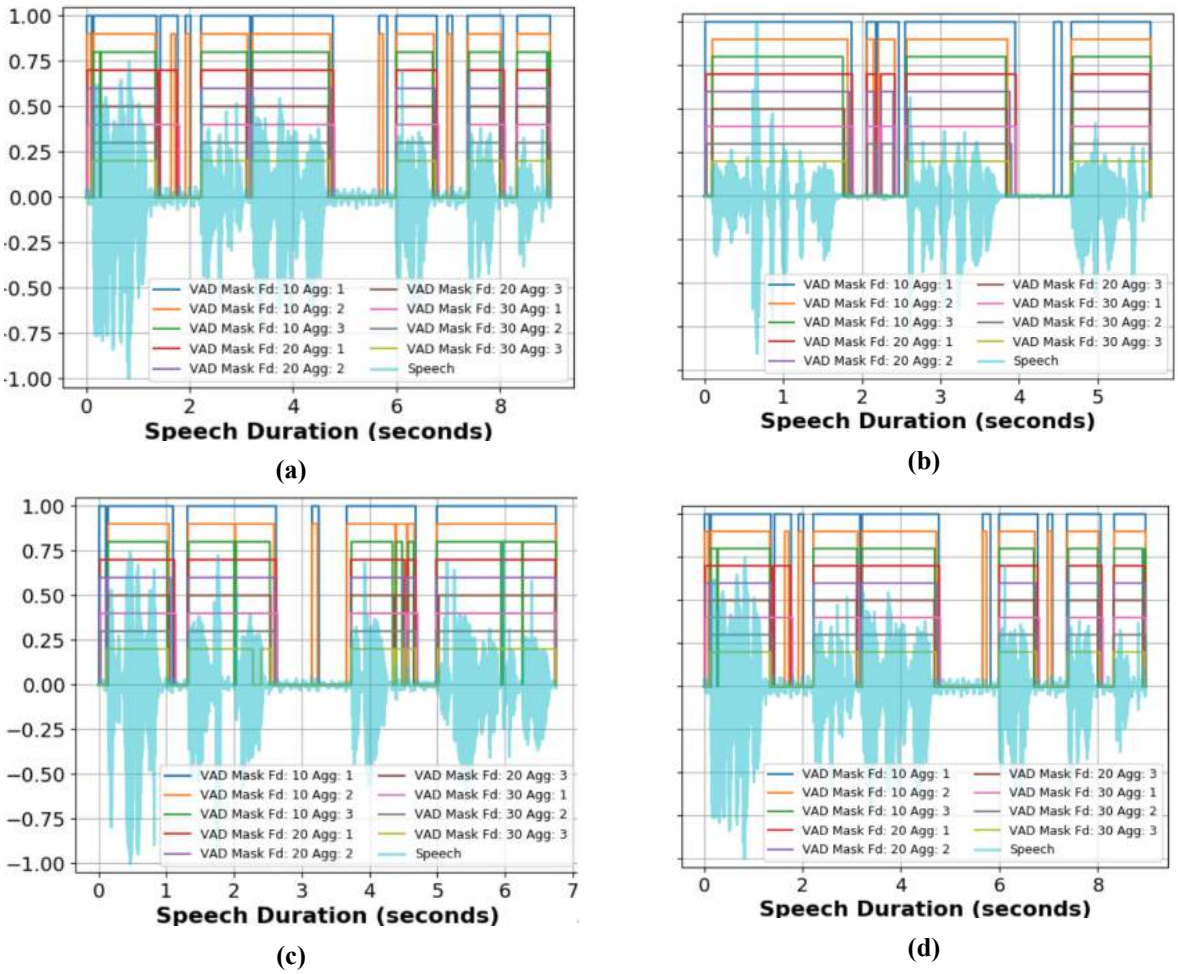
Τα πλαίσια σιωπής που υπάρχουν επίσης μέσα σε ένα σήμα ομιλίας το οποίο μπορεί να παράγουν παραπλανητικές τιμές στις διανυσματικές αναπαραστάσεις των ακουστικών εξαγωμένων χαρακτηριστικών και έτσι μπορεί να οδηγήσουν το σύστημα SER να ταξινομεί λάθος ολόκληρη τη φράση. Προκειμένου να ακυρωθεί η επίδραση των πλαισίων σιωπής στην ακολουθία των διανυσμάτων χαρακτηριστικών, χρειαζόμαστε έναν μηχανισμό ανίχνευσης φωνητικής δραστηριότητας (VAD). Υπάρχουν πολλές εφαρμογές του VAD όπως το WebRTC Voice Activity Detector [1] ή αυτός που συμπεριλαμβάνεται στο OpenSMILE toolkit [119].

Μπορούμε να δούμε την αποτελεσματικότητα του WebRTC της Google όταν το αξιολογούμε σε ομιλίες από τη βάση δεδομένων συναισθηματικής ομιλίας BERLIN (EmoDB) [121] στο σχήμα 3.1. Στο προαναφερθέν σχήμα, εμφανίζονται διάφορες μάσκες VAD που παράγονται με διαφορετικές επιλογές των παραμέτρων «Fd» και «Agg». Αυτές οι παράμετροι αντιστοιχούν στον χρόνο σε ms των πλαισίων ομιλίας και στην επιθετικότητα επί των οποίων ένα πλαίσιο θα θεωρείται ένα πλαίσιο ομιλίας ή ένα σιωπηλό πλαίσιο. Μπορούμε να δούμε ότι διαφορετικές παραμετροποιήσεις παράγουν διαφορετικές μάσκες VAD για κάθε μία από τις 4 ομιλίες που εμφανίζονται. Ένα μακρύτερο παράθυρο λαμβάνει υπόψη μια ευρύτερη και κατά συνέπεια μια πιο ομαλή περιοχή του σήματος ομιλίας, προκειμένου να εξαχθεί το συμπέρασμα εάν συμπεριλαμβάνεται η ομιλία και συνεπώς είναι πιθανότερο να παράγει θετικό αποτέλεσμα. Αντίθετα, τα πλαίσια που έχουν μικρότερη χρονική διάρκεια ενδέχεται να μην περιέχουν ομιλία καθόλου και έτσι ορισμένες περιοχές μη-ομιλίας θα μπορούσαν να αντικατοπτρίζονται στη μάσκα VAD παρότι μπορεί να εκπέμπεται ένα σύμφωνο. Μια επιθετική τεχνική θα παράσχει πολύ περισσότερες περιοχές φωνής, επειδή τα τρισδιάστατα και χαμηλής ενέργειας συγγενικά στοιχεία θα μπορούσαν να προκαλέσουν σύγχυση στο σύστημα VAD. Από την άλλη πλευρά, μια λιγότερο επιθετική τεχνική μπορεί να παράγει περιοχές με μη ομιλία ως φωνητικές περιοχές. Το τελευταίο αποτέλεσμα θα μπορούσε να μετριαστεί με την εισαγωγή φιλτραρίσματος εξομάλυνσης πριν από το τελικό αποτέλεσμα της μάσκας VAD.

Όπως αναφέρθηκε προηγουμένως, στο έργο μας χρησιμοποιούμε το OpenSMILE για να εξαγάγουμε τα σύνολα ακουστικών χαρακτηριστικών μας. Για το σκοπό αυτό, χρησιμοποιούμε δύο από τους ενσωματωμένους μηχανισμούς VAD. Ο πρώτος βασίζεται σε μια μονάδα LSTM για την εξαγωγή συμπερασμάτων χρησιμοποιώντας LLD στο βήμα και την ενέργεια σε επίπεδο πλαισίου και έχει εκπαιδευτεί σε ένα σύνολο δεδομένων ταινιών. Από την άλλη πλευρά, μια πιο αφελής προσέγγιση είναι να εξαγάγουμε την πιθανότητα φωνής για κάθε πλαίσιο και στη συνέχεια να το χρησιμοποιήσουμε για να αποκτήσετε τις περιοχές ομιλίας μέσα στο σήμα. Και οι δύο αυτές αρχιτεκτονικές VAD απεικονίζονται στο σχήμα 3.2. Είναι προφανές ότι η μέθοδος που βασίζεται σε LSTM υποφέρει από τα ελλείποντα δεδομένα αυτού του συνόλου δεδομένων που αξιολογείται. Συγκεκριμένα, το LSTM VAD έχει εκπαιδευτεί σε διαφορετικά δείγματα με πιθανώς διαφορετικές συνθήκες θορύβου και εγγραφής αντί για εκείνα που υπάρχουν στο EmoDB. Αυτός είναι ο λόγος για τον οποίο επιλέγουμε τον πιο αφηρημένο μηχανισμό VAD, ο οποίος βασίζεται στην πιθανότητα έκφρασης και ένα κατώφλι, προκειμένου να συναχθεί η τελική μάσκα για φωνητικές / αθόρυβες περιοχές κάθε συναισθηματικής ομιλίας. Επιπλέον, παρατηρούμε ότι ο περιγραφικός συντελεστής βήματος σε επίπεδο πλαισίου είναι υψηλότερος στις ίδιες περιοχές με την πιθανότητα φωνής. Αυτό υποδεικνύει ότι ακόμη και ad-hoc κατώτατα όρια που εφαρμόζονται στην τελευταία τιμή μπορούν να περιγράψουν καταλλήλως περιοχές ομιλίας και να διακρίνουν μεταξύ των περιοχών της σιωπής. Το επιλεγμένο όριο αποκοπής πιθανότητας φωνής ορίζεται ειδικά τόσο ψηλά (0.85) στο Σχήμα 3.2, προκειμένου να παρασχεθεί η χρονοσειρά του τελευταίου χαρακτηριστικού κάτω από μια μη βέλτιστη τιμή. Είναι σαφές ότι με μια ποιοτική ανάλυση ενός κατωφλίου μπορούμε να προσαρμόσουμε κατάλληλα την τιμή του κατωφλίου για τους σκοπούς μας.

Παρουσιάζουμε επίσης μια σύντομη ανάλυση για κάποιες τιμές κατωφλίων για τον τοπικό περιγραφητή της πιθανότητας να υπάρχει ομιλία μεσα στην ίδια φράση. Στο Σχήμα 3.3 μπορούμε να δούμε το σήμα ομιλίας σχεδιασμένο παράλληλα με την αντίστοιχη σειρά χαρακτηριστικών πιθανό-

[1] Ο πηγαίος κώδικας έχει αναπτυχθεί από την Google και είναι ελεύθερα διαθέσιμος για λήψη και μη εμπορική χρήση εδώ: https://github.com/wiseman/py-webrtcvad



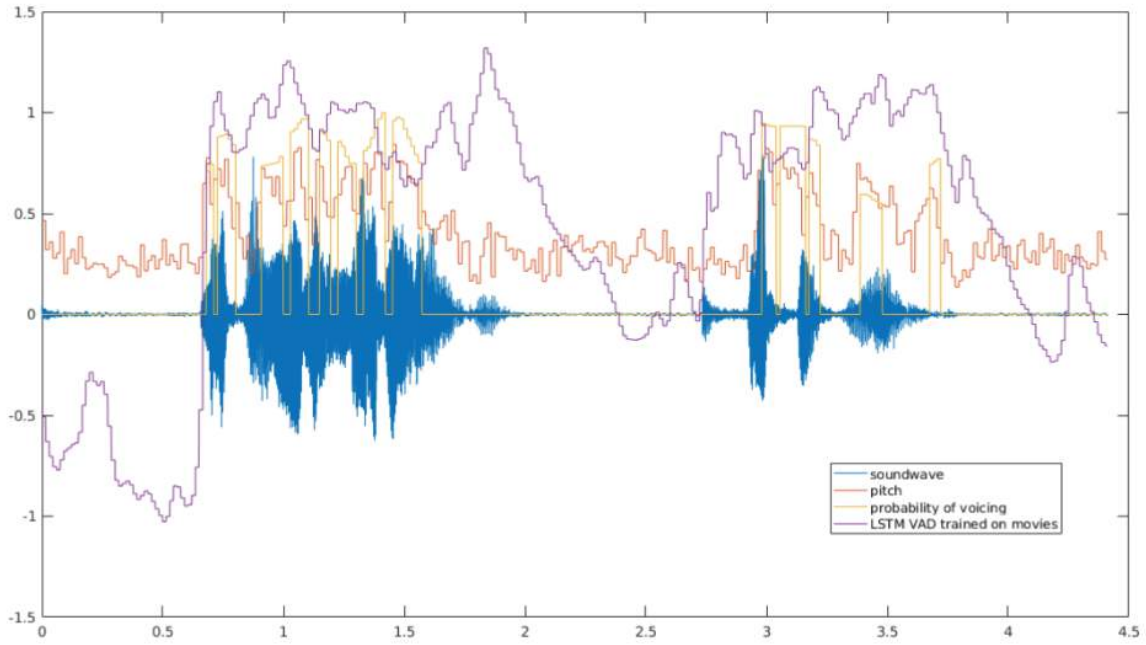
Σχήμα 3.1: Αξιολόγηση του WebRTC VAD σε ομιλίες της βάσης δεδομένων EmoDB.

τητας για όλα τα πλαίσια του (30ms). Καθώς αυξάνουμε το κατώτατο όριο μπορούμε να δούμε ότι οι περιοχές ομιλίας παραμελούνται και η παραγόμενη μάσκα VAD, με βάση την αξία της πιθανότητας ομιλίας δεν περιλαμβάνει αυτές τις περιοχές. Στο πλαίσιο αυτό πιστεύουμε ότι μια τιμή κατωφλίου κοντά στο 0.35 θα ικανοποιούσε επαρκώς τον σκοπό μας. Θα πρέπει να σημειωθεί ότι αυτή είναι μια άλλη ευρεστική μέθοδος για την εύρεση ενός κατωφλίου όπου θα μπορούσε να υποφέρει σημαντικά για άλλες βάσεις δεδομένων ομιλίας. Ωστόσο, ο σκοπός αυτής της μελέτης δεν σχετίζεται με την αναζήτηση κάθε παραμέτρου στο χώρο των παραμέτρων που προκύπτει από τη διαδικασία εξαγωγής χαρακτηριστικών και έτσι δεν θα αναλύσουμε περαιτέρω αυτό το θέμα.

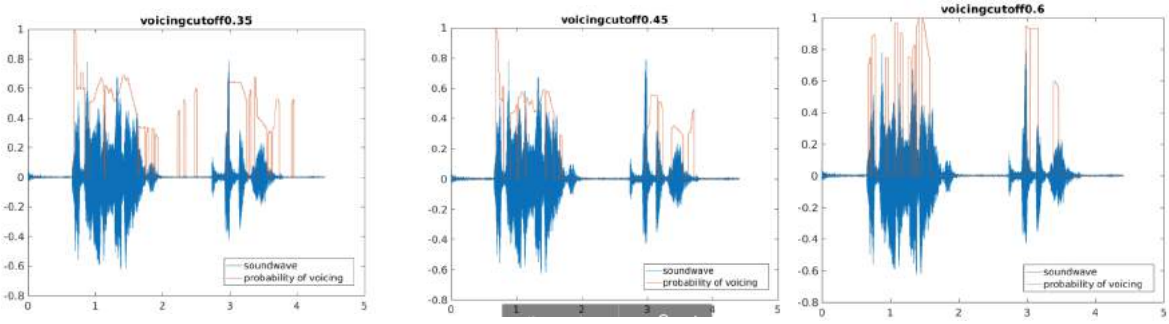
3.2.3 Τοπικά Χαρακτηριστικά - Περιγραφητές LLDs

Σε αυτό το κεφάλαιο αναφερόμαστε στα τοπικά χαρακτηριστικά ως τα ακατέργαστα LLD που εξάγονται σε επίπεδο πλαισίου για να μοντελοποιήσουν κάθε συναισθηματική φράση. Σε αυτή την ενότητα θα παρουσιάσουμε μια σύντομη ανάλυση του τρόπου με τον οποίο εξάγονται διαφορετικοί LLDs και ποιοι από αυτούς θα χρησιμοποιηθούν για τα πειράματά μας. Είναι σημαντικό να σημειωθεί ότι σε αυτό το έργο εξάγονται πολλαπλά LLD. Σε αυτή την ενότητα εξετάζονται μόνο οι LLD που χρησιμοποιούνται απευθείας ως είσοδος στα μοντέλα SER.

Διαχωρίζουμε κάθε συναισθηματική φράση σε πλαίσια 30ms με μέγεθος βήματος 15ms και παράθυρο τους (βλ. Εξίσωση 3.1) με παράθυρο Hamming όπως ορίζεται στην Εξίσωση 3.2. Μετά το άνοιγμα του παραθύρου εφαρμόζουμε επίσης ένα φιλτράρισμα πριν από την έμφαση όπως ορίζεται στην Εξίσωση 3.5.



Σχήμα 3.2: Μηχανισμοί OpenSMILE VAD με βάση την πιθανότητα ομιλίας και προ-εκπαιδευμένου LSTM



(a) Κατώφλι πιθανότητας ομιλίας 0,35

(b) Κατώφλι πιθανότητας ομιλίας 0,45

(c) Κατώφλι πιθανότητας ομιλίας 0,6

Σχήμα 3.3: OpenSMILE VAD βασισμένο σε διαφορετικά κατώφλια για πιθανότητα φωνής

Εξάγουμε 25 LLDs για κάθε πλαίσιο και εξάγουμε επίσης τα παράγωγα για όλα αυτά (εκτός jitter και shimmer). Τα τελευταία χαρακτηριστικά περιγράφουν ήδη μέτρα διαταραχής της αστάθειας συχνότητας και πλάτους, αντίστοιχα. Έτσι, συνδέονται στενά με την έννοια της διαφοράς που είναι το διακριτό ισοδύναμο του παραγώγου (πρακτικά μια τιμή δέλτα). Η λήψη ενός δέλτα αξίας δέλτα μπορεί να παράγει πολύ θορυβώδεις και υπερβολικές τιμές που μπορεί να είναι καταστροφικές για τη διαδικασία κατάρτισης. Ωστόσο, για το jitter εξάγουμε ένα παρόμοιο μέτρο που υπολογίζεται ως διαφορά διαφοράς για την τιμή του jitter σε διαφορετικές περιόδους.

Τα επιλεγμένα LLDs προέρχονται από την θεμελιώδη συχνότητα, το σπεκτρόγραμμα, την τοπική ενέργεια του πλαισίου και την ποιότητα φωνής, τα οποία είναι τα πιο διαδεδομένα σύνολα ακουστικών χαρακτηριστικών. Συγκεκριμένα, εξάγουμε: την Ροή μέσου τετραγωνικού στρώματος (RMS) ενέργειας του πλαισίου παραθύρου, την ποιότητα φωνής, το Ποσοστό μηδενικής διασταύρωσης (ZCR), το ποσοστό Αρμονικών προς θόρυβο (HNR) με τη χρήση της λειτουργίας αυτόματης συσχέτισης (ACF), την διαφορά διαφοράς περιόδων (DDP). Επιπλέον, εξάγουμε: τοπική λάμψη, θεμελιώδη συχνότητα (F0) με υπο-αρμονικό άθροισμα (SHS), ένταση, πιθανότητα έκφρασης και ζώνη συχνότητας

Πίνακας 3.1: Τοπικοί περιγραφητές LLDs

Selected LLDs	Including 1st Delta
RMS Energy	✓
Quality of Voice	✓
ZCR	✓
Jitter Local	✗
Jitter DDP	✗
Shimmer Local	✗
F0 by SHS	✓
Loudness	✓
Probability of Voicing	✓
HNR by ACF	✓
MFCCs[0-14]	✓
Σύνολο: 25	Σύνολο: 22

Mel (MFB) και MFCCs. Οι προαναφερόμενες επιλεγμένες λειτουργίες παρουσιάζονται επίσης στον Πίνακα 3.1 σε πιο συμπαγή μορφή. Η δεύτερη στήλη υποδεικνύει εάν προσθέτουμε επίσης την τιμή της διακριτής παραγώγου (delta) του αντίστοιχου χαρακτηριστικού μεταξύ αυτής της τιμής και της αντίστοιχης που περιέχεται στο προηγούμενο πλαίσιο. Ένα σύμβολο ελέγχου ✓ υποδεικνύει ότι η τιμή δέλτα είναι επίσης προσαρτημένη στο τελικό διάνυσμα χαρακτηριστικών ενώ ένα x-mark ✗ αντικατοπτρίζει το αντίθετο.

Τώρα, κάθε διανυσματικό χαρακτηριστικό πλαισίου είναι εφοδιασμένο με 47 τοπικά χαρακτηριστικά. Λόγω του μεγάλου αριθμού τοπικών χαρακτηριστικών, επιλέγονται οι πιο σημαντικές LLDs [61], [65], [27]. παραπέμπουμε τον αναγνώστη σε εκτεταμένες περιγραφές για το πώς εξάγονται τα προαναφερθέντα LLDs [19], [20], [120]. Κατά συνέπεια, κάθε συναισθηματική φράση θα εκπροσωπείται ως μια ακολουθία LLD-φορέων που εξάγονται από όλα τα πλαίσια που περιλαμβάνονται σε αυτήν. Φυσικά, η διάρκεια αυτής της ακολουθίας φορέων χαρακτηριστικών δεν θα ήταν η ίδια σε διαφορετικές δηλώσεις.

3.2.4 Παγκόσμια-Γενικά στατιστικά χαρακτηριστικά

Σε αυτό το κεφάλαιο αναφέρουμε τα Παγκόσμια Χαρακτηριστικά ως τις τιμές των στατιστικών πάνω σε τμήματα ομιλίας οποιασδήποτε διάρκειας που εφαρμόζονται σε LLD που εξάγονται από πλαίσια. Προφανώς, η διαδικασία εξαγωγής χαρακτηριστικών σε επίπεδο πλαισίου είναι παρόμοια με αυτή που περιγράφηκε στην προηγούμενη ενότητα (βλ. Ενότητα 3.2.3). Μετά την εξαγωγή χαρακτηριστικών σε επίπεδο πλαισίου εφαρμόζουμε σύνολα στατιστικών προκειμένου να έχουμε την τελική ακολουθία των παραστάσεων του διανύσματος στατιστικών χαρακτηριστικών. Το μήκος των τμημάτων και το βήμα μεταξύ των διαδοχικών πλαισίων θα ορίζουν τον αριθμό των διανυσμάτων χαρακτηριστικών σε κάθε ακολουθία. Στην ακραία περίπτωση όπου οι διάρκειες του τμήματος ομιλίας θα ήταν ίσες με τη συνολική διάρκεια κάθε ομιλίας, τότε η ακολουθία θα γίνει μόνο ένα διάνυσμα χαρακτηριστικών. Σε γενικές γραμμές, ένα τμήμα ομιλίας θα μπορούσε να θεωρηθεί οποιοδήποτε χρονικό διάστημα μέσα σε μια έκφραση η οποία είναι μικρότερη από την ίδια τη πρόταση και μεγαλύτερη από μια διάρκεια που αντιστοιχεί σε ένα τυπικό πλαίσιο 10ms-100ms ομιλίας.

Ακολουθούμε μια παρόμοια διαδικασία με το [23] που εξήγαγε ένα ισχυρό σύνολο συναισθηματικών χαρακτηριστικών. Τα χαρακτηριστικά για τα αντίστοιχα τμήματα ομιλίας βασίζονται σε LLD που εξάγονται από πλαίσια, εμπλουτισμένα με άλλα LLDs επίσης. Αυτή η επέκταση των περιεχόμενων LLDs είναι εφαρμόσιμη σε αυτή την περίπτωση λόγω της χρονικής συσσωμάτωσης αυτών των τιμών χρησιμοποιώντας στατιστικά. Δεν θα μπορούσαμε να συμπεριλάβουμε όλα αυτά τα LLDs στο προηγούμενο σύνολο τοπικών χαρακτηριστικών λόγω της αυξημένης διαστάσης της ακολουθίας εισόδου

Πίνακας 3.2: Σύνολα στατιστικών για την εξαγωγή παγκόσμιων-γενικών στατιστικών χαρακτηριστικών

Στατιστική τιμή	Σύνολο
Θέση μεγαλύτερου στοιχείου Θέση μικρότερου στοιχείου Αριθμητικός μέσος Τυπική Απόκλιση Μέτρο Λοξότητας Κύρτωση Γραμμικό συντελεστής παλινδρόμησης 1/2 Σφάλμα τετραγωνικής παλινδρόμησης Απόλυτο σφάλμα γραμμικής παλινδρόμησης τεταρτημόριο 1/2/3 εύρη τεταρτημόριων 2-1/3-2/3-1 ποσοστό 99 χρόνος ανύψωσης 75/90	A
ποσοστό 1 εύρος μεταξύ ποσοστών 1-99	B
Αριθμός ενεργοποιήσεων Διάρκεια	C

που θα δημιουργηθεί.

Διαφορετικά LLDs απαιτούν διαφορετικά μεγέθη παραθύρων για την καλύτερη δυνατή εξαγωγή τους. Εν ολίγοις, εξάγουμε την θεμελιώδη συχνότητα χρησιμοποιώντας ένα Γκαουσιανό παράθυρο μεγέθους 60ms και χρονικού βήματος 10ms, ενώ για όλους τους άλλους LLDs χρησιμοποιούμε παράθυρο Hamming μεγέθους 25ms και χρονικού βήματος 10ms. Αυτή η διαδικασία ακολουθείται επειδή η εξαγωγή της θεμελιώδους συχνότητας απαιτεί ένα πολύ μεγαλύτερο μήκος παραθύρου για μια καλή εκτίμηση της από το σήμα ομιλίας. Δεν μπορούσαμε να ακολουθήσουμε μια παρόμοια προσέγγιση στην προηγούμενη ενότητα (Τμήμα 3.2.3), κυρίως επειδή η διαμορφωμένη ακολουθία που αποκτήθηκε από τα τοπικά χαρακτηριστικά έπρεπε να εξαχθεί κάτω από ένα προκαθορισμένο μήκος παραθύρου. Παρόλο που σε αυτή την προσέγγιση μπορούμε εύκολα να εξαγάγουμε LLDs χρησιμοποιώντας προσεγγίσεις πολλαπλών χρονικών κλιμάκων για να προσαρμόσουμε κάθε μέγεθος παραθύρου σε ένα συγκεκριμένο LLD το οποίο θα κάνει βέλτιστη την εξαγωγή της τιμής του τελευταίου.

Για το σκοπό αυτό, εφαρμόζουμε στατιστικά πάνω σε όλα τα LLDs σε επίπεδο πλαισίου. Για να αποκτήσουμε τα παγκόσμια χαρακτηριστικά ανά τμήμα ομιλίας, χρησιμοποιούμε μια ποικιλία στατιστικών τα οποία εφαρμόζονται σε κάθε τιμή LLD καθ όλη τη διάρκεια του επιλεγμένου τμήματος. Τα στατιστικά που χρησιμοποιούνται αντιστοιχούν σε εύρη, θέση ακραίων τιμών, εκατοστημορίων και άλλες στατιστικές τιμές, οι οποίες είναι σε θέση να καταγράψουν πώς κάθε τοπικό χαρακτηριστικό ή LLD σε επίπεδο πλαισίου μεταβάλλει την αξία του μέσα στο επιλεγμένο τμήμα ομιλίας. Μια επισκόπηση των επιλεγμένων στατιστικών που εφαρμόζονται, παρουσιάζεται στον πίνακα 3.2 μαζί με την αντίστοιχη ομαδοποίησή τους. Θα χρησιμοποιήσουμε το σύμβολο κάθε ομαδοποίησης για να καθορίσουμε αργότερα το σύνολο των στατιστικών λειτουργιών που εφαρμόζονται σε κάθε LLD σε επίπεδο πλαισίου.

Τώρα που έχουμε τα σύνολα των στατιστικών λειτουργιών που θα εφαρμοστούν για την εξαγωγή των στατιστικών αναπαραστάσεων των τμημάτων ομιλίας, μπορούμε να τα εφαρμόσουμε πάνω σε LLDs σε επίπεδο πλαισίου καθώς και στις αντίστοιχες τιμές τους στις διακριτές τους παραγώγους. Συγκεκριμένα, οι τιμές δέλτα υπολογίζονται πρώτα μεταξύ των διαδοχικών πλαισίων εντός του τμήματος ομιλίας το οποίο πρόκειται να αναλυθεί και μετά από αυτό μία στατιστική τιμή αντιστοιχεί σε μια τιμή από την ακολουθία των LLD σε επίπεδο πλαισίου σε αυτό το τμήμα ομιλίας.

Χρησιμοποιούμε παρόμοια LLDs σε επίπεδο πλαισίου όπως στην προηγούμενη ενότητα 3.2.3

Πίνακας 3.3: Παγκόσμια-Γενικά στατιστικά χαρακτηριστικά

Επιλεγμένα LLDs	1η Διακριτή Παράγωγος	Σύνολα Στατιστικών*
Jitter Τοπικό	✗	A
Jitter DDP	✗	A
Shimmer Τοπικό	✗	A
F0 by SHS	✓	A,C
Ένταση	✓	A,B
Πιθανότητα Ομιλίας	✓	A,B
HNR με ACF	✓	A,B
MFCCs[0-14]	✓	A,B
LSP Συχνοτήτων [0-7]	✓	A,B
log MFB [0-7]	✓	A,B
F0 Σκελετός	✓	A,B

**Σύνολα Στατιστικών που εφαρμόζονται (A,B,C) έχουν οριστεί στον Πίνακα 3.2.*

και στη συνέχεια εφαρμόζουμε τα στατιστικά. Ωστόσο, για να ακολουθήσουμε τη μέθοδο εξαγωγής χαρακτηριστικών που περιγράφεται στο [23], χρησιμοποιούμε τα ίδια LLDs και αποκλείσαμε την RMS βραχυχρόνια ενέργεια, την ποιότητα της φωνής και το ZCR. Εξηγάγαμε επίσης και άλλα LLD, όπως τα πρώτα 8 Γραμμικά Ζεύγη (LSP), τους πρώτους 8 συντελεστές από τις Μελ τράπεζες φίλτρων Mel (MFB) και τον σκελετό της χρονοσειράς της εκτιμημένης θεμελιώδους συχνότητας F0. Για ορισμένα από τα προαναφερθέντα LLD εξάγουμε επίσης τα δελτάρια τους από τα διαδοχικά τους πλαίσια τα οποία είναι επίσης εμφανή στον Πίνακα 3.3. Η διαδικασία της εξαγωγής χαρακτηριστικών είναι παρόμοια με αυτή που περιγράφηκε προηγουμένως ενώ μόνο το τελευταίο μέρος αλλάζει εδώ όπου εφαρμόζονται οι στατιστικές λειτουργίες για το τελευταίο βήμα της διαδικασίας εξαγωγής χαρακτηριστικών.

Ο κατάλογος των επιλεγμένων LLDs μαζί με τα στατιστικά τους λειτουργικά σύνολα εμφανίζονται στον Πίνακα 3.3. Στην πρώτη στήλη, εμφανίζονται τα επιλεγμένα LLD που θα εξαχθούν από κάθε πλαίσιο. Η δεύτερη στήλη υποδεικνύει εάν προσθέτουμε επίσης την τιμή δέλτα του αντίστοιχου χαρακτηριστικού μεταξύ αυτής της τιμής και της αντίστοιχης που περιέχεται στο προηγούμενο πλαίσιο. Ένα σύμβολο ελέγχου ✓ υποδεικνύει ότι η τιμή δέλτα είναι επίσης προσαρτημένη στο τελικό διάνυσμα χαρακτηριστικών ενώ ένα x-mark ✗ αντικατοπτρίζει το αντίθετο. Κάτω από τη στήλη "Σύνολα Στατιστικών" μπορούμε να δούμε το αντίστοιχο σύμβολο που ορίζεται στον προηγούμενο Πίνακα 3.2 που υποδεικνύει το αντίστοιχο σύνολο στατιστικών που θα εφαρμοστούν σε κάθε ακολουθία LLD σε επίπεδο πλαισίου. Συνεπώς, κάθε τμήμα ομιλίας αντιπροσωπεύεται από ένα διάνυσμα σταθερού μήκους 1582 χαρακτηριστικών που είναι το ίδιο με αυτό που προτείνεται στο [23]. Για μια μέθοδο ταξινόμησης που βασίζεται αποκλειστικά σε ολόκληρη την ομιλία (βλέπε ενότητα 1.3.3) η ακολουθία των διανυσμάτων θα γίνει μόνο ένα διάνυσμα 1582-d.

3.3 Προσεγγίσεις σε διαφορετικές χρονικές κλίμακες

Ο πυρήνας της προσέγγισής μας έγκειται στη διερεύνηση των χρονικών κλιμάκων για την αποφάση συναισθημάτων για διαφορετικούς συνδυασμούς συνόλων χαρακτηριστικών. Τα διαδοχικά πλαίσια με τοπικά χαρακτηριστικά μπορούν να συνενωθούν και, κατά συνέπεια, να αλλάξουν την χρονική κλίμακα αποφάσεων των συναισθημάτων. Για παράδειγμα, αν θεωρήσουμε την συνένωση 5 διαδοχικών πλαισίων, αυτή μπορεί να αντιπροσωπεύει μια συναισθηματική κατάσταση *λύπη* αλλά κάθε πλαίσο ανεξάρτητα μπορεί να μην αντιπροσωπεύει το ίδιο συναίσθημα. Συνεπώς, η χρονική κλίμακα λήψης απόφασης για συναισθήματα είναι ευαίσθητη στη διάρκεια του τμήματος ομιλίας στην οποία εξάγονται τα παγκόσμια χαρακτηριστικά. Ένα τμήμα ομιλίας μπορεί να θεωρηθεί ως χρονικό διάστημα μεγαλύτερο από ένα πλαίσιο (π.χ. ολόκληρη η ομιλία θα μπορούσε να χρησιμεύσει ως τμήμα).

Εστιάζουμε στην επίδραση που έχει ο αριθμός των συνενωμένων πλαισίων και των τμημάτων σε συστήματα SER που βασίζονται σε αρχιτεκτονικές RNN και ιδιαίτερα LSTM. Επομένως, βρίσκουμε την κατάλληλη χρονική κλίμακα απόφασης για τοπικά και παγκόσμια χαρακτηριστικά, αντίστοιχα.

Στη συνέχεια, διαιρούμε τις προσεγγίσεις μας σε τρεις διαφορετικές παραμετροποιήσεις που αντιστοιχούν ουσιαστικά στις τρεις κύριες υποκατηγορίες των προσεγγίσεων ταξινόμησης SER όπως περιγράφονται στο τμήμα 1.3.3.

1. Στο πρώτο εξάγουμε τα τοπικά χαρακτηριστικά σε επίπεδο πλαισίου (όπως περιγράφεται στην Ενότητα 3.2.3) και προσπαθούμε να συμπεράνουμε τη συναισθηματική εκφράση απευθείας χρησιμοποιώντας την ακολουθία εισόδου για να τροφοδοτήσουμε το LSTM μας. Αυτή η προσέγγιση περιγράφεται στο τμήμα 3.3.1 και αντιστοιχεί στην προσέγγιση "Βασισμένη σε πλαίσια" σύμφωνα με την ορολογία του τμήματος 1.3.3.

2. Στη δεύτερη, εξάγουμε τα παγκόσμια-γενικά στατιστικά χαρακτηριστικά (όπως περιγράφονται στην Ενότητα 3.2.4) σε διάφορες διάρκειες μιας έκφρασης και προσπαθήσαμε να συμπεράνουμε τη συναισθηματική φράση χρησιμοποιώντας την ακολουθία εισόδου των στατιστικών διανυσμάατων για να τροφοδοτήσουμε το LSTM μας. Αυτή η προσέγγιση περιγράφεται στο τμήμα 3.3.2 και αντιστοιχεί στην προσέγγιση "Βασισμένη σε τμήματα ομιλίας" ακολουθώντας την ορολογία του τμήματος 1.3.3.

3. Στην τελευταία, εξάγουμε τα Παγκόσμια χαρακτηριστικά (όπως περιγράφεται στην Ενότητα 3.2.4) σε ολόκληρη τη συζήτηση. Τώρα έχουμε μόνο ένα διάνυσμα χαρακτηριστικών που χρησιμοποιούμε ως είσοδο για έναν ταξινομητή SVM. Αυτή η προσέγγιση περιγράφεται στο τμήμα 3.3.3 και αντιστοιχεί στην προσέγγιση "Βασισμένη σε ολόκληρη την ομιλία" ακολουθώντας την ορολογία του τμήματος 1.3.3.

3.3.1 Βασισμένη σε πλαίσια

Χρησιμοποιούμε ένα LSTM RNN όπως περιγράφεται στο [64] και το εκπαιδεύουμε με πολλαπλές χρονικές στιγμές που αντιστοιχούν σε συνεστραμμένα LLDs. Στην ουσία, η ακολουθία εισόδου για το LSTM αποτελείται από διαδοχικούς LLDs σε επίπεδο πλαισίου οι οποίοι συνενώνονται σε μία επιλαχούσα χρονική κλίμακα. Για παράδειγμα, αν θέλουμε να συγκεντρώσουμε τα LLDs σε μία χρονική κλίμακα που αντιστοιχεί σε ένα φωνήμα, τότε μετά την εξαγωγή των τοπικών χαρακτηριστικών (βλ. Ενότητα 3.2.3) για κάθε πλαίσιο, ενώνουμε όλα τα LLD σε επίπεδο πλαισίου χρησιμοποιώντας τα παράθυρα των 100ms. Συγκεκριμένα, επειδή τα τοπικά χαρακτηριστικά εξάγονται χρησιμοποιώντας πλαίσια 30ms με επικάλυψη 15ms, κάθε παράθυρο θα είναι ένα συνενωμένο διάνυσμα 6 πλαισίων LLD.

Κάθε παράθυρο συνεκτικοποιημένων LLDs θα αντιστοιχούσε σε ένα χρονικό βήμα της ακολουθίας εισόδου. Αυτοί οι φορείς σταθερού μήκους χαρακτηριστικών αντιστοιχούν σε διαφορετικές χρονικές στιγμές και ποικίλλουν μεταξύ των εκφράσεων. Για κάθε ακολουθία χρονικών βημάτων, που αντιπροσωπεύει μια συναισθηματική φράση, η αναμενόμενη έξοδος είναι μια ετικέτα συναισθήματος (αυτό συχνά ονομάζεται εκπαίδευση πολλών προς μία - many to one) [66]). Κατά τη διάρκεια της εκπαιδευτικής διαδικασίας εκπαιδεύουμε το LSTM χρησιμοποιώντας όλα τα διαθέσιμα συναισθήματα. Ομοίως, σε οποιοδήποτε πρόβλημα supervised Machine Learning (ML), στη διαδικασία εκπαίδευσης μας δίνεται η ακολουθία των συνεκτικοποιημένων LLDs $\{\mathbf{x}_0, ..., \mathbf{x}_{T-1}\}$ ως είσοδος και ένα συναίσθημα y που είναι η αντίστοιχη ετικέτα. Θεωρούμε ότι το T είναι ίσο με τον αριθμό των ομαδοποιημένων LLD-διανυσμάτων που περιλαμβάνονται σε μια συγκεκριμένη έκφραση. Χρησιμοποιώντας πολλαπλές ακολουθίες εισόδου και υποστηρίζοντας εκ νέου το σφάλμα κάθε πρόβλεψης κατά τη διάρκεια της εκπαιδευτικής διαδικασίας, είμαστε σε θέση να εκπαιδεύσουμε το RNN μας και να το δοκιμάσουμε σε δείγματα που δεν χρησιμοποιήθηκαν για εκπαίδευση.

Τα RNNs μπορούν να κωδικοποιήσουν επαρκώς τις πληροφορίες που περικλείονται σε μια ακολουθία χρονικών βημάτων και να παράγουν την αναμενόμενη έξοδο στο τελευταίο χρονικό βήμα. Στην περίπτωσή μας, οι πληροφορίες για το συναισθηματικό περιεχόμενο θα κωδικοποιούνται πάνω

σε όλους τους συνδεμένους φορείς LLD και στο τελικό χρονικό βήμα, το μοντέλο κάνει μια πρόβλεψη για το πιο εμφανές συναίσθημα μιας δεδομένης έκφρασης. Όταν τα στρώματα LSTM στοιβάζονται, η έξοδος κάθε στρώματος τροφοδοτείται ως είσοδος στο επόμενο στρώμα. Το LSTM έχει 2 κρυμμένα στρώματα και σχεδόν κάθε στρώμα ενεργοποιεί την επόμενη σε αυτή την ακολουθία. Στην κορυφή της τελικής παράστασης του χρονικού βήματος υπάρχει ένα πυκνό στρώμα εξόδου το οποίο οδηγεί σε ένα στρώμα softmax προκειμένου να συναχθεί η συναισθηματική κατηγορία της ακολουθίας εισόδου.

Μια επισκόπηση της όλης διαδικασίας παρουσιάζεται στο Σχήμα 3.4. Κάθε μικρό μπλε τετράγωνο αντιστοιχεί σε ένα διάνυσμα τοπικών χαρακτηριστικών, που εξάγεται από ένα πλαίσιο. Το κυανό ορθογώνιο που περιέχει πολλαπλά διανύσματα LLDs (6 στο σχήμα) αντιστοιχεί σε ένα χρονικό βήμα της ακολουθίας εισόδου $\mathbf{x}_i$. Ένα κόκκινο τετράγωνο με τις συνδέσεις με την προηγούμενη και τις επόμενες στρώσεις καθώς και τις ανατροφοδοτούμενες συνδέσεις υποδεικνύει μια ξετυλιγμένη μονάδα LSTM πάνω από τα χρονικά βήματα της ακολουθίας εισόδου. Οι ενεργοποιήσεις του τελευταίου χρονικου βήματος δίδονται ως είσοδος σε ένα πυκνό στρώμα με αριθμό κρυφών κόμβων N_h ίσο με τον αριθμό συναισθηματικών τάξεων. Η πρόβλεψη βασίζεται στην έξοδο του τελευταίου στρώματος (που είναι πρακτικά N_h posterior πιθανοτήτων που αντιστοιχούν σε κάθε συναισθηματική κλάση) μετά την εφαρμογή μιας συνάρτησης softmax προκειμένου να αποκτηθεί ένα one-hot διάνυσμα πρόβλεψης. Η προβλεπόμενη κλάση θα ενεργοποιηθεί (1 στην αντίστοιχη καταχώρηση του διανύσματος) ενώ όλες οι άλλες κλάσεις θα έχουν μηδενικά στους υπόλοιπους δείκτες.

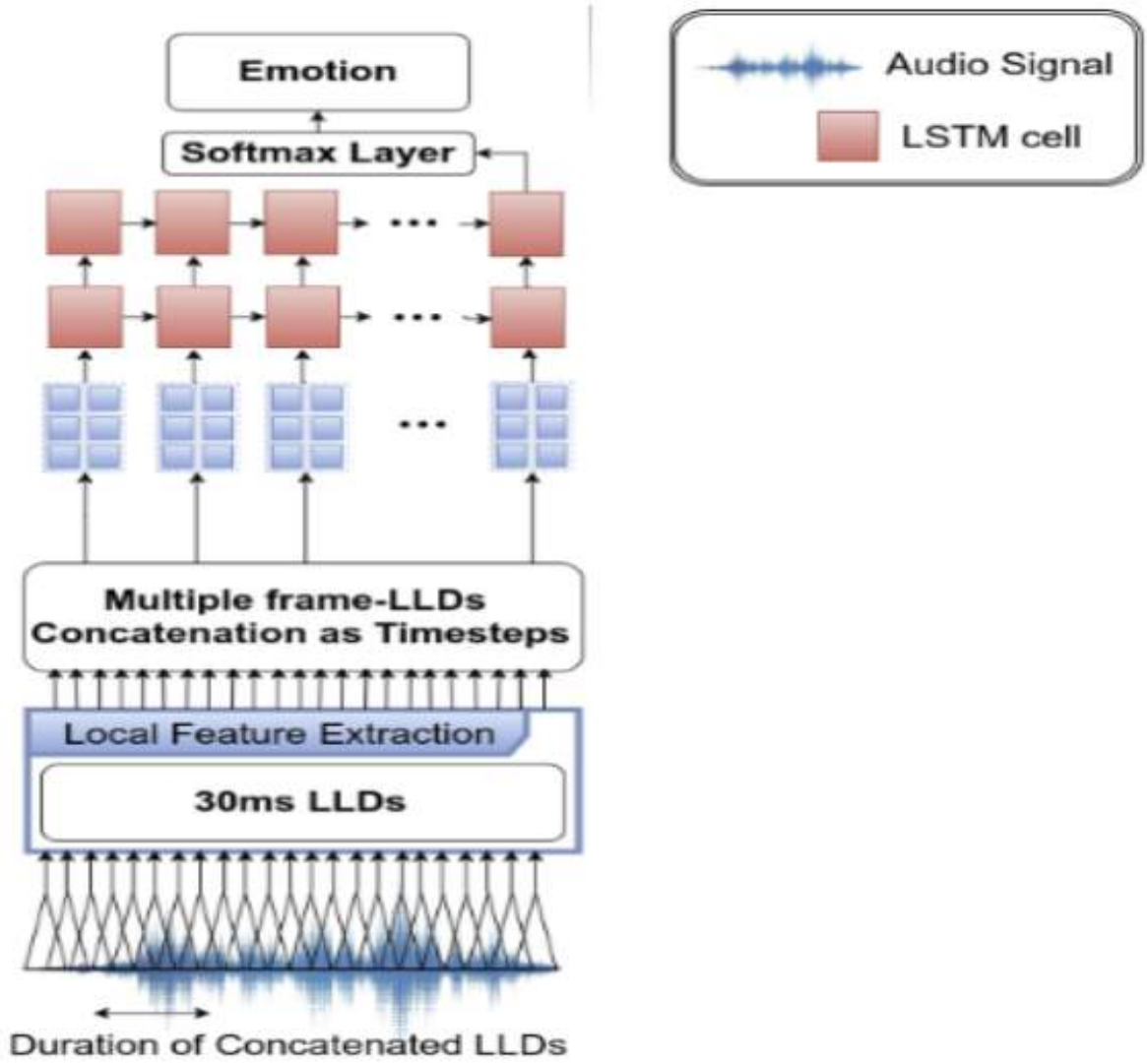
3.3.2 Βασισμένο σε τμήματα ομιλίας

Ένα LSTM εκπαιδεύεται σε αυτή την προσέγγιση που έχει την ίδια αρχιτεκτονική που περιγράφηκε στην προηγούμενη ενότητα 3.3.1. Αν και οι διαφορές μεταξύ των δύο προσεγγίσεων περιστρέφονται γύρω από τη διαδικασία εξαγωγής χαρακτηριστικών και τον σχηματισμό της ακολουθίας εισόδου. Στην προσέγγιση βασισμένη σε τμήματα ομιλίας, δημιουργούμε αρχικά επικαλυπτόμενα τμήματα από την έκφραση μακρύτερης διάρκειας από αυτά που χρησιμοποιούνται για την εξαγωγή LLD. Στη συνέχεια, ευθυγραμμίζουμε όλα τα πλαίσια που ανήκουν σε κάθε τμήμα και εξάγουμε LLD παρόμοια με εκείνα που περιγράφονται στην προηγούμενη ενότητα. Συγκεκριμένα, κάθε τμήμα αντιπροσωπεύεται με ένα διάνυσμα χαρακτηριστικών από στατιστικές τιμές μετά την εφαρμογή τους πάνω στα προυπολογισμένα LLDs (βλ. Ενότητα 3.2.4).

Το συνολικό σχήμα ταξινόμησης με το σπάσιμο σε τμήματα, η εξαγωγή γενικών στατιστικών χαρακτηριστικών και η τελική ταξινόμηση με χρήση LSTM, απεικονίζεται στο σχήμα 3.5. Είναι προφανές ότι η αρχιτεκτονική LSTM είναι παρόμοια με αυτήν που παρουσιάστηκε για το LSTM που εκπαιδεύτηκε σε τοπικά χαρακτηριστικά από συντετμημένα LLDs (βλέπε Εικόνα 3.4) αλλά ο σχηματισμός της ακολουθίας εισόδου έχει τροποποιηθεί. Συγκεκριμένα, κάθε χρονικό βήμα εισόδου αντιστοιχεί σε ένα διάνυσμα διαστάσεων 1582 όπως περιγράφεται στο 3.2.4 το οποίο αντιπροσωπεύεται με ένα πράσινο ορθογώνιο. Παρατηρούμε ότι η χρονική κλίμακα κατά την οποία λαμβάνεται η απόφαση του συναισθηματικού περιεχομένου, μετά την κωδικοποίηση της ακολουθίας εισόδου, είναι ακριβώς η ίδια με εκείνη κάτω από την οποία εξάγονται τα παγκόσμια χαρακτηριστικά.

3.3.3 Βασιζόμενη σε ολόκληρη την ομιλία

Σε αυτή την προσέγγιση εξάγουμε παγκόσμια χαρακτηριστικά όπως αναφέρθηκε προηγουμένως (βλ. Τμήματα 3.3.2, 3.2.4), αλλά δεν διασπάμε κάθε έκφραση σε επικαλυπτόμενα τμήματα. Εφαρμόζουμε στατιστικές λειτουργίες πάνω στην ακολουθία των LLD-vectors της όλης έκφρασης. Τώρα κάθε συναισθηματική φράση αντιπροσωπεύεται με ένα μόνο διάνυσμα χαρακτηριστικών που αποτελείται από τα χαρακτηριστικά 1582. Ένας πυρήνας SVM με πυρήνα ακτινικής βάσης (RBF) κατασκευάζεται και εκπαιδεύεται σε αυτά τα διανύσματα χαρακτηριστικών. Το συνολικό σχήμα ταξινόμησης παρουσιάζεται στο Σχήμα 3.6.



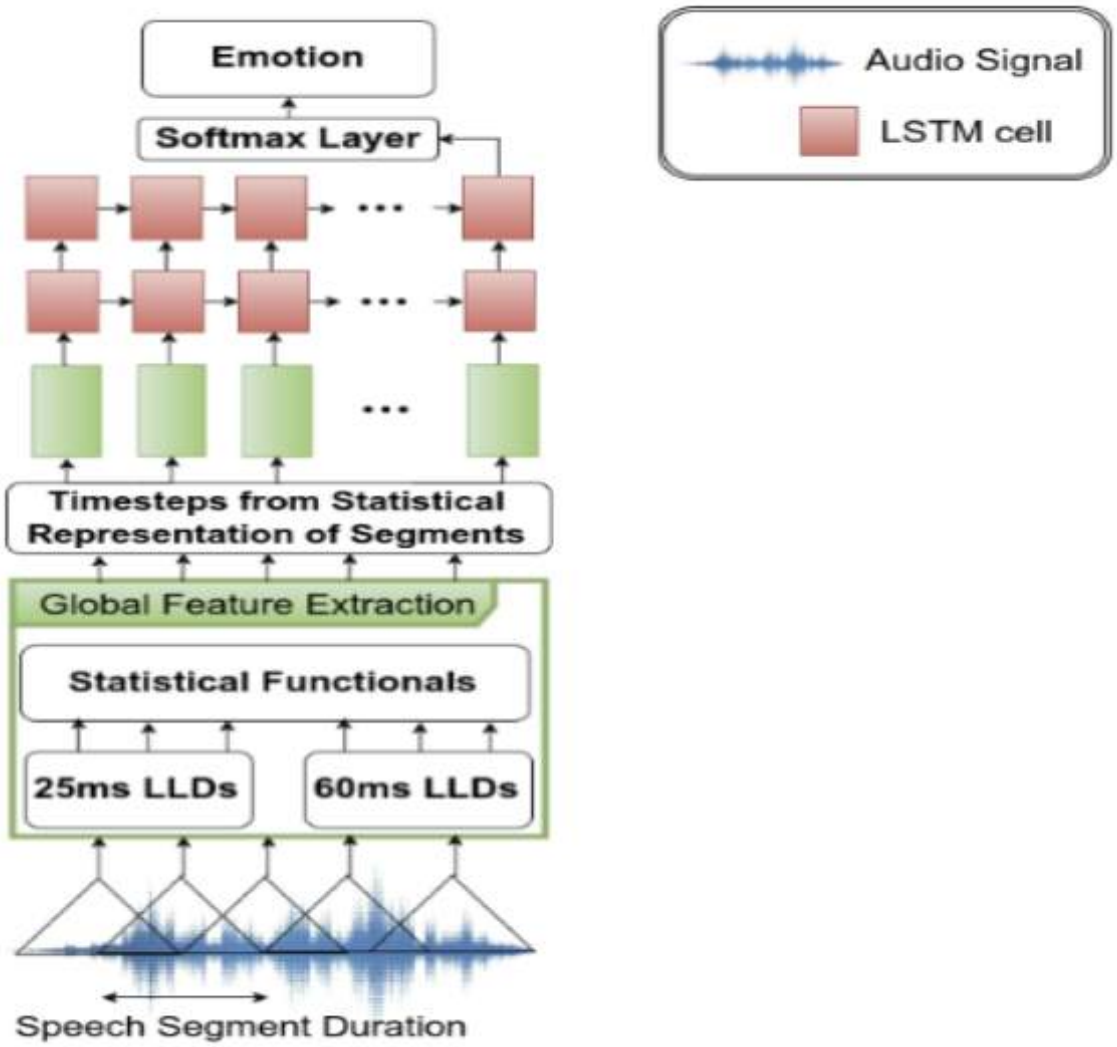
Σχήμα 3.4: Προσέγγιση βασισμένη σε πλαίσιο χρησιμοποιώντας LSTM εκπαιδευμένο σε ακολουθίες συνεπτυγμένων LLDs

3.4 Πειραματική ρύθμιση

Σε αυτή την ενότητα παρέχουμε την πειραματική ρύθμιση που ακολουθούμε στα πειράματά μας για να αξιολογήσουμε τις μεθόδους μας σε διαφορετικά χρονικά πλαίσια. Περιγράφουμε το σύνολο δεδομένων που χρησιμοποιούμε και τις μετρικές αξιολόγησης που χρησιμοποιούμε για να μετρήσουμε την απόδοση των συστημάτων μας στα τμήματα 3.4.1 και 3.4.2, αντίστοιχα. Επιπλέον, αναπτύσουμε το πειραματικό πλαίσιο για κάθε μία από τις τρεις προσεγγίσεις των χρονικών κλιμακίων, που περιγράφονται στην προηγούμενη ενότητα 3.3, στα τμήματα 3.4.3, 3.4.4 και 3.4.5.

3.4.1 Σύνολο δεδομένων

Για τα πειράματα χρησιμοποιούμε τη βάση δεδομένων IEMOCAP. Αυτή η βάση δεδομένων περιέχει δεδομένα ήχου και εικόνας από 5 συνεδρίες. Σε κάθε συνεδρία, 2 άτομα (ένας Άνδρας και ένας Γυναίκα) εκτελούν ενεργό ή αυτοσχέδιο διάλογο. Για κάθε έκφραση, 3 ανθρώπινοι σχολιαστές το χαρακτήρισαν με μια κατηγορηματική ετικέτα συναισθήματος. Επιλέγουμε μόνο τις ομιλίες για τις οποίες τουλάχιστον 2 από τους 3 σχολιαστές είχαν την ίδια άποψη. Συγκεκριμένα, το σύνολο δεδομένων περιλαμβάνει ηχητικά σήματα από 4 συναισθηματικές κατηγορίες: *Θυμός* (1103 ομιλίες), *Λύπη* (1084 ομιλίες), *Χαρά* (595 ομιλίες) και *Ουδετερότητα* (1708 ομιλίες). Συνολικά έχουμε συναι-



Σχήμα 3.5: Προσέγγιση βασιζόμενη σε τμήματα ομιλίας χρησιμοποιώντας ακολουθίες LSTM εκπαιδευμένες για τα γενικά στατιστικά χαρακτηριστικά

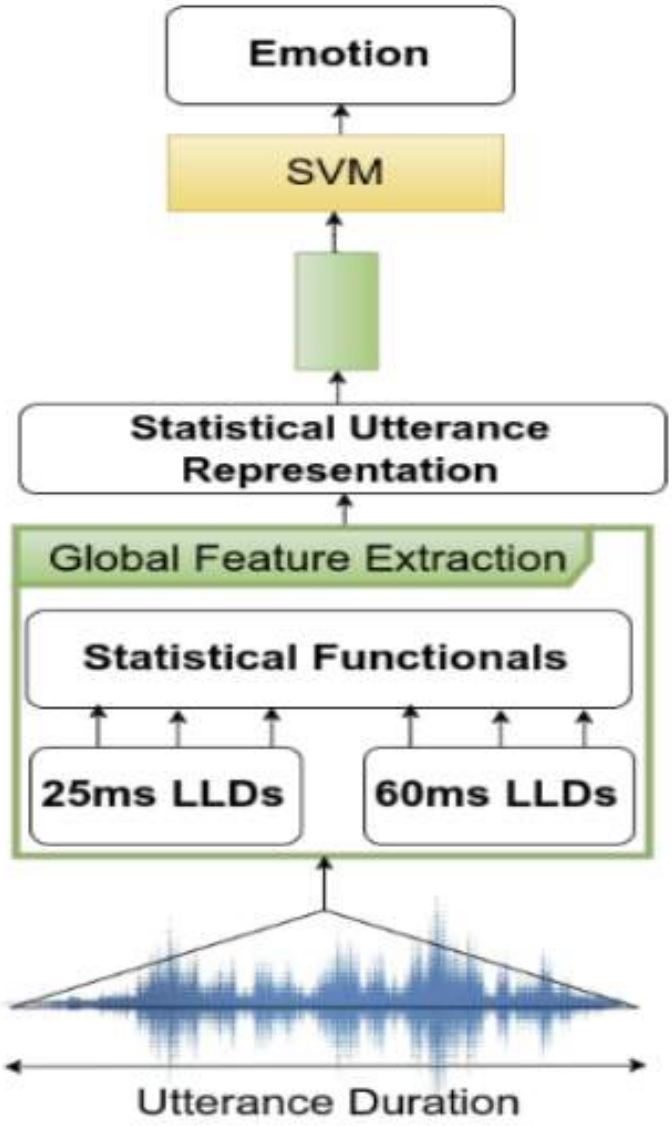
σθηματικές δηλώσεις 4490. Τα συναισθήματα που διατηρήσαμε είναι τα ίδια με τα άλλα πειραματικά συστήματα στη βιβλιογραφία [66], [36], [61].

3.4.2 Μέτρηση επίδοσης

Χρησιμοποιούμε το σχήμα "Άσε έξω μία συνεδρία" (LOSO) για τη δοκιμή των μοντέλων μας, όπως τα [66], [61], [36]. Σε κάθε συνεδρία (5 συνολικά), μία συνεδρία χρησιμοποιείται για δοκιμή και τα υπόλοιπα 4 για εκπαίδευση. Για κάθε συνεδρία που τεστάρουμε, ο ένας ομιλητής χρησιμοποιείται για την επικύρωση και ο άλλος για έλεγχο. Επαναλαμβάνουμε το πείραμα αναστρέφοντας τα σύνολα επικύρωσης και δοκιμών. Η μέση ακρίβεια από τους δύο ομιλητές περιλαμβάνεται στην τελική αξιολόγηση. Όλα τα χαρακτηριστικά ομαλοποιούνται από τον γενικό μέσο όρο και την τυπική απόκλιση από τα δείγματα του συνόλου δεδομένων. Η αξιολόγηση πραγματοποιείται χρησιμοποιώντας:

1. **Weighted Accuracy (WA)** το οποίο είναι το ποσοστό των σωστών αποφάσεων ταξινόμησης πάνω από το σετ δοκιμών

2. **Unweighted Accuracy (UA)** το οποίο είναι το μέσο ποσοστό ακρίβειας που προκύπτει από κάθε συναισθηματική κλάση ξεχωριστά



Σχήμα 3.6: Προσέγγιση βασιζόμενη στη συμπεριφορά χρησιμοποιώντας SVM εκπαιδευμένο σε στατιστικά χαρακτηριστικά από όλη την ομιλία

Η τελευταία μέτρηση απόδοσης χρησιμοποιείται για να παράσχει μια πολύ πιο αξιόπιστη μέτρηση ακρίβειας για ένα μη ισορροπημένο σύνολο δεδομένων όπως αυτό που χρησιμοποιούμε εδώ. Για παράδειγμα, στην ακραία περίπτωση που έχουμε έναν ταξινομητή που ψευδώς ψηφίζει «ουδέτερο» για όλες τις συναισθηματικές δηλώσεις, αυτό ονομάζεται συχνά ταξινομητής πλειοψηφίας και χρησιμοποιείται ευρέως για δυαδικά προβλήματα ταξινόμησης [4]. Σε αυτή την περίπτωση, ένας ταξινομητής πλειοψηφίας για το προαναφερθέν υποσύνολο της βάσης δεδομένων IEMOCAP (βλ. Προηγούμενη ενότητα 3.4.1) θα αποδώσει 38.04% WA αλλά μόνο 25% UA.

3.4.3 Εκπαίδευση LSTM σε LLD πλαισίων

Όπως έχουμε ήδη αναφέρει στην Ενότητα 3.3.1, εκπαιδεύουμε ένα LSTM με δύο κρυφά στρώματα. Στην Ενότητα 2.3.1 προσφέραμε μια εκτενή εξήγηση για το κύριο μπλοκ LSTM και πώς μπορεί να στοιβάζεται προκειμένου να κατασκευαστούν βαθύτερες αρχιτεκτονικές όπως αυτές που χρησιμοποιούμε εδώ. Τα δύο κρυμμένα στρώματα έχουν κυτταρικά μεγέθη 512 και 256 νευρώνων το καθένα. Ένα στρώμα softmax για την ταξινόμηση των συναισθηματικών κατηγοριών 4 τοποθετείται στην κορυφή του τελικού χρονομέτρου όπως φαίνεται στο σχήμα 3.4. Παρατηρήσαμε ότι η απόδοση δεν αυξά-

νεται όταν προσθέτουμε επιπλέον κρυμμένα στρώματα. Αν αυξήσουμε τον αριθμό των νευρώνων σε κάθε κύτταρο, αυξάνουμε επίσης τετραγωνικά την πολυπλοκότητα του μοντέλου μας λόγω των συνδέσεων μεταξύ των επόμενων στρωμάτων καθώς και της πιθανότητας υπερεκπαίδευσης (overfitting) του μοντέλου μας [122]. Η υπερεκπαίδευση είναι σχεδόν βέλτιστη ρύθμιση των παραμέτρων του LSTM για το σετ εκπαίδευσης και έτσι οι περιοχές που εισάγουν διακρίσεις των μαθημένων παραμέτρων σφίγγονται στενά για τη βελτιστοποίηση των περιοχών του εκπαιδευτικού σετ. Με αυτόν τον τρόπο, τα μαθησιακά βάρη του LSTM μας θα χάσουν την δυνατότητα γενίκευσης των περιοχών που μαθαίνουν για άλλα σετ δεδομένων πλην αυτού που χρησιμοποιήσαμε για την εκπαίδευση. Πρακτικά, οι περιπτώσεις του συνόλου δοκιμών θα ταξινομούνται λανθασμένα επειδή οι διακριτικές περιοχές των κατηγοριών είναι κατάλληλα εκπαιδευμένες για το σύνολο εκπαίδευσης, αλλά δεν μπορούν να επεκταθούν ώστε να συμπεριληφθούν οι περιπτώσεις του συνόλου δοκιμών.

Προκειμένου να αποφευχθεί η υπερεκπαίδευση του LSTM, επιλέξαμε να κανονικοποιήσουμε το μοντέλο μας, εφαρμόζοντας πρόωρη εγκατάλειψη (early stopping) στις συνδέσεις του Νευρικού μας Δικτύου (NN). Η εφαρμογή της ρίψης συνδέσεων (dropout) στις συνδέσεις του NN μας μπορεί να θεωρηθεί ως μια διαδικασία όπου ένα υποσύνολο των βαρών NN επιλέγεται τυχαία και όλες οι μονάδες σε αυτό το σετ πέφτουν καθώς και οι συνδέσεις τους κατά τη διάρκεια της εκπαίδευσης. Έχει αποδειχθεί ότι η ρίψη συνδέσεων είναι μια αποτελεσματική τεχνική για την αποτροπή ενός NN από την υπερεκπαίδευση [123]. Παρόλα αυτά, η εφαρμογή του λόγου ρίψης συνδέσεων σε RNNs δεν είναι τόσο εύκολη όσο όλες οι άλλες αρχιτεκτονικές, όπως DNN και CNN, όπου υπάρχουν μόνο συνδέσεις εμπρόσθιας τροφοδότησης. Σύμφωνα με το [124], κανονικοποιήσουμε το LSTM εφαρμόζοντας ρυθμό ρίψης συνδέσεων *Dr=0.5*, μόνο στις συνδέσεις μη ανατροφοδότησης του LSTM. Επιπλέον, η απόδοση του LSTM δεν αυξάνεται περαιτέρω όταν δοκιμάζουμε διαφορετικά ποσοστά ρίψης συνδέσεων.

Το μοντέλο βελτιστοποιείται με την ελαχιστοποίηση της κατηγορικής διασταυρούμενης εντροπίας, χρησιμοποιώντας το εργαλείο Nadam οπτιμιστή [125], το οποίο βασικά είναι ο βελτιστοποιητής του Adam με την ορμή Nesterov. Η βασική τιμή του ρυθμού εκμάθησης ορίζεται στο $\alpha_{base} = 0.002$. Ο αριθμός των εποχών εκπαίδευσης ορίζεται σε 100 αλλά η πρόωρη διακοπή εφαρμόζεται όταν η συνάρτηση απώλειας του συνόλου επικύρωσης δεν βελτιώνεται για 10 διαδοχικές εποχές. Η ρύθμιση που δίνει το καλύτερο άθροισμα των WA και UA στο σύνολο επικύρωσης επιλέγεται. Επιπλέον, τα επίπεδα ομαλοποίησης της παρτίδας (batch-normalization) απορρίπτονται λόγω των ανεπιθύμητων επιπτώσεών τους στην ταχύτερη σύγκλιση όταν εφαρμόζονται σε συνδέσεις επαναληψιμότητας ενός LSTM [126]. Η κανονικοποίηση παρτίδων είναι ουσιαστικά η ιδέα της χρήσης στατιστικών τιμών των ενεργοποιήσεων που παράγονται στο εμπρός βήμα μεταξύ παρτίδων δεδομένων κατάρτισης ή δοκιμών. Σε κάθε στρώση χρησιμοποιούνται αυτές οι στατιστικές προκειμένου να ομαλοποιηθούν οι ενεργοποιήσεις του τρέχοντος στρώματος και κατά συνέπεια να οδηγηθούν όλες οι τιμές σε ένα εύρος όπου οι υπολογισμένες κλίσεις δεν εκρήγνυνται ούτε συγκλίνουν στο μηδέν. Τα δείγματα που χρησιμοποιούνται για την εκπαίδευση δεν είναι τεράστια για να χρησιμοποιήσουμε όλες αυτές τις τεχνικές για την ταχύτερη σύγκλιση του δικτύου μας. Αντίθετα, όπως περιγράφουμε στη συνέχεια, το βασικό μας πρόβλημα είναι πώς να αποφύγουμε την υπερεκπαίδευση, επειδή το LSTM μας συγκλίνει υπερβολικά γρήγορα. Επιπλέον, το μέγεθος παρτίδας έχει οριστεί ίσο με 400, προκειμένου να χωρέσει στη μνήμη της Μονάδας Γραφικής Επεξεργασίας (GPU). Η προαναφερθείσα αρχιτεκτονική LSTM εφαρμόστηκε χρησιμοποιώντας Keras [127] πάνω από το πίσω-μέρος Theano [128].

Μετά την εξαγωγή των LLDs όπως περιγράφεται στην Ενότητα 3.2.3, τα συνενώνουμε σε κομμάτια που αντιστοιχούν σε μεγαλύτερες διάρκειες (βλ. Εικόνα 3.4). Τα διαδοχικά πλαίσια συναρμολογούνται σε κομμάτια διαφόρων μηκών και τροφοδοτούνται στο LSTM. Το μήκος κάθε τεμαχίου είναι ο αριθμός των διαδοχικών διανυσμάτων φορέων πλαισίων που περιέχει. Η αξιολόγηση πραγματοποιείται για μήκη κομματιών που αντιστοιχούν σε χρονικές κλίμακες πλαισίου (30ms), φωνήματος (90ms-300ms) και μεγαλύτερες χρονικές κλίμακες (400ms-8sec) για την έκφραση της συναισθηματικής ετικέτας. Ο ρυθμός εκμάθησης για τοπικά εξαγόμενα LLDs δίνεται από την ακόλουθη εξίσωση:

$$\alpha_{local} = \frac{\alpha_{base}}{N_{Frames}} \tag{3.6}$$

όπου ο αρχικός ή ο βασικός ρυθμός εκμάθησης τίθεται σε $\alpha_{base} = 0.002$ και N_{Frames} αντιστοιχεί στον αριθμό πλαισίων που περιλαμβάνονται σε κάθε κομμάτι LLD-διανυσμάτων.

Αυτός ο ρυθμός εκμάθησης χρησιμοποιείται επειδή όταν αυξάνουμε τον αριθμό των LLDs σε κάθε κομμάτι τότε μια αντίστροφη αναλογική μείωση θα αντανακλάται στο μήκος των χρονικών βημάτων που δίνεται ως είσοδος στο LSTM μας. Σε αυτό το πλαίσιο, ο αριθμός των συνδέσεων στο NN μας σταθεροποιείται και μειώνουμε τον αριθμό των χρονικών βημάτων. Κάνοντας αυτό, παρατηρούμε ότι το LSTM μας συγκλίνει υπέρμετρα γρήγορα. Ως εκ τούτου, χρησιμοποιήσαμε την προαναφερθείσα ευρετική τεχνική για να αποφύγουμε αυτό το πρόβλημα σε όλα τα αντίστοιχα χρονοδιαγράμματα στα οποία συγκολλούμε τα LLD σε χρονική κλίμακα πλαισίου.

3.4.4 LSTM εκπαιδευμένο στα γενικά-στατιστικά χαρακτηριστικά

Για αυτό το πείραμα, τα γενικά στατιστικά χαρακτηριστικά εξάγονται όπως περιγράφεται στην Ενότητα 3.2.4. Χρησιμοποιούμε τμήματα ομιλίας μήκους (0.5s-8s) με αναλογία επικάλυψης *OL=0.5* μεταξύ τους. Η διαδικασία κατάρτισης που ακολουθείται εδώ είναι παρόμοια με εκείνη που περιγράφηκε στο προηγούμενο τμήμα 3.4.3. Η αρχιτεκτονική του LSTM που χρησιμοποιείται είναι στην πραγματικότητα η ίδια με αυτή που χρησιμοποιούμε για την κατάρτιση με συντετμημένα LLDs σε επίπεδο πλαισίου, η οποία είναι επίσης εμφανής συγκρίνοντας τις Εικόνες 3.4 και 3.5. Ωστόσο, επιλέγουμε μια διαφορετική μέθοδο για να προσαρμόσουμε τον ρυθμό εκμάθησης για χαρακτηριστικά παγκόσμιας στατιστικής. Χρησιμοποιώντας το ίδιο $\alpha_{base} = 0.002$, ο αρχικός ρυθμός εκμάθησης δίνεται από την ακόλουθη εξίσωση:

$$\alpha_{global} = \frac{\alpha_{base} \cdot OL}{100 \cdot T_{segment}} \tag{3.7}$$

όπου $T_{segment}$ είναι η διάρκεια σε δευτερόλεπτα κάθε τμήματος ομιλίας που χρησιμοποιείται για την εξαγωγή των παγκόσμιων χαρακτηριστικών από.

3.4.5 SVM εκπαιδευμένο σε στατιστικά χαρακτηριστικά από όλη την ομιλία

Όπως περιγράφεται στην Ενότητα 3.2.4, τα στατιστικά χαρακτηριστικά εξάγονται από ολόκληρη την ομιλία η οποία τώρα χρησιμεύει ως το μόνο τμήμα. Χρησιμοποιούμε έναν πυρήνα RBF και ρυθμίζουμε το συντελεστή γάμμα γ με τον αριθμό των χαρακτηριστικών για κάθε φράση. Και συγκεκριμένα:

$$\gamma = \frac{1}{1582} \tag{3.8}$$

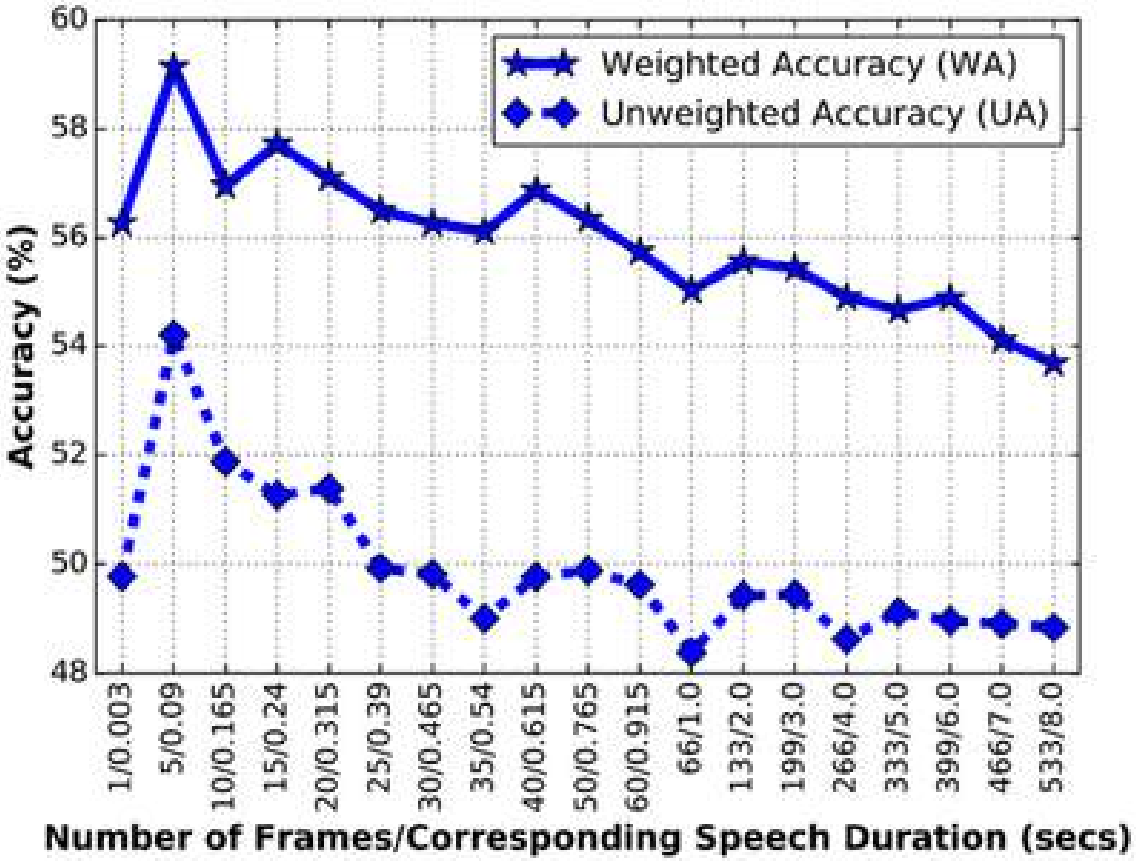
Για μια εκτενή εξήγηση για τον τρόπο με τον οποίο όλες αυτές οι παράμετροι ρυθμίζουν τον ταξινομητή πολλαπλών κλάσεων SVM, παραπέμπουμε τον αναγνώστη στην προηγούμενη ενότητα 2.2.2. Προκειμένου να βελτιστοποιήσουμε το μοντέλο μας, επιλέγουμε το χώρο αναζήτησης των παραμέτρων του SVM μας να είναι ίσο με το σύνολο τιμών για τον συντελεστή κόστους C που βρίσκονται στο διάστημα $[0.001, 60]$. Για κάθε ομιλητή, επιλέγεται ο συντελεστής κόστους C που αποδίδει το καλύτερο άθροισμα των WA και UA για το αντίστοιχο ομιλητή επικύρωσης (validation). Όλα οι αλλές υπερπαράμετροι έχουν οριστεί στις προεπιλεγμένες τιμές τους (αναφέρουμε τον αναγνώστη στην αρχική δημοσίευση του LibSVM [129]).

3.5 Πειραματικά αποτελέσματα και συζήτηση

3.5.1 Βέλτιστες χρονικές κλίμακες για LSTM εκπαιδευμένα απευθείας σε LLDs

Τα αποτελέσματα για το πείραμα εκπαίδευσης με τοπικούς περιγραφητές εμφανίζονται στο Σχήμα 3.7. Είναι προφανές ότι μια συνένωση των LLDs των 5 πλαισίων αποδίδει την καλύτερη απόδοση και στις μετρήσεις WA και UA. Αυτό αντιστοιχεί σε απόφαση χρονικού διαστήματος φωνήματος (100ms)

για κάθε συναισθηματικό πλαίσιο της χρονικής ακολουθίας. Καθώς αυξάνουμε τον αριθμό των πλαισίων σε κάθε χρονικό βήμα, αλλάζουμε την κλίμακα απόφασης σε μεγαλύτερες κλίμακες (1-8s) και παρατηρούμε μια σταδιακή μείωση τόσο της μετρικής WA όσο και της UA. Υποθέτουμε ότι αυτή η πτώση στις μετρήσεις απόδοσης μπορεί να συμβαίνει επειδή οι LLDs παραποιούν τα σιωπηρά πλαίσια ομιλίας και όταν συνενώνονται για μακρές διάρκειες ομιλίας, οδηγούν το RNN σε παραπλανητικές αφαιρέσεις για τη συναισθηματική εκδήλωση των τμημάτων αυτών. Επιπλέον, όταν η απόφαση λαμβάνεται σε επίπεδο πλαισίου, τα σιωπηλά πλαίσια λανθασμένα χαρακτηρίζονται από την ίδια ετικέτα συναισθήματος με τα υψηλότερα πλαίσια ενέργειας. Είναι επίσης προφανές ότι λόγω των ανισορροπιών κατανομών των συναισθηματικών δηλώσεων επιτυγχάνουμε πάντα ελαφρώς καλύτερη WA σε σύγκριση με την UA.



Σχήμα 3.7: Σταθμισμένη ακρίβεια (WA) και μη-σταθμισμένη ακρίβεια (UA) LSTM εκπαιδευμένα απευθείας σε LLDs με σύντηξη σε διαφορετικές χρονικές κλίμακες

Θα μπορούσαμε να υποθέσουμε ότι η κορυφή και για τις δύο μετρήσεις απόδοσης των WA 59.14% και UA 54.2%, κατά την εκπαίδευση του LSTM με κομμάτια LLDs που αντιστοιχούν σε χρονικές διάρκειες 100ms συνιστούν μια βέλτιστη χρονική κλίμακα απόφασης για αυτό το είδος χαρακτηριστικών. Πιθανότατα, όταν χρησιμοποιούμε κομμάτια μόνο ενός πλαισίου σε κάθε χρονικό βήμα, τότε διευρύνουμε το μήκος της ακολουθίας εισόδου $\{\mathbf{x}_i\}_{i=0}^{T}$ για μια ακολουθία με $T+1$ βήματα. Αυτό θα μπορούσε να προκαλέσει το πρόβλημα που συζητήσαμε προηγουμένως στα τμήματα 2.3.3 και 2.3.2 όπου η ροή πληροφοριών για τη συναισθηματική δυναμική εξαφανίζεται μέσω του μεγάλου πλήθους των χρονικών βημάτων που οδηγούν πιθανότατα σε ένα πολύ λιγότερο ανθεκτικό σύστημα SER εξαιτίας της απώλειας όταν το RNN προσπαθεί να συντονίσει τα βάρη χρησιμοποιώντας οπισθοδιάδωση σε όλα αυτά τα χρονικά περιθώρια. Αν και η βασική ιδέα της προσέγγισής μας δεν είναι να εφαρμόσουμε την πιο ανθεκτική αρχιτεκτονική RNN με επιπλέον μηχανισμούς (μηχανισμούς προσοχής ή πολλαπλά RNNs) για να αντιμετωπίσουμε αυτό το πρόβλημα. Αυτά τα πειραματικά αποτελέσματα προσφέρουν μια πολύτιμη ποιοτική εικόνα για το πώς μπορούμε να επιτύχουμε ανώτερη απόδοση

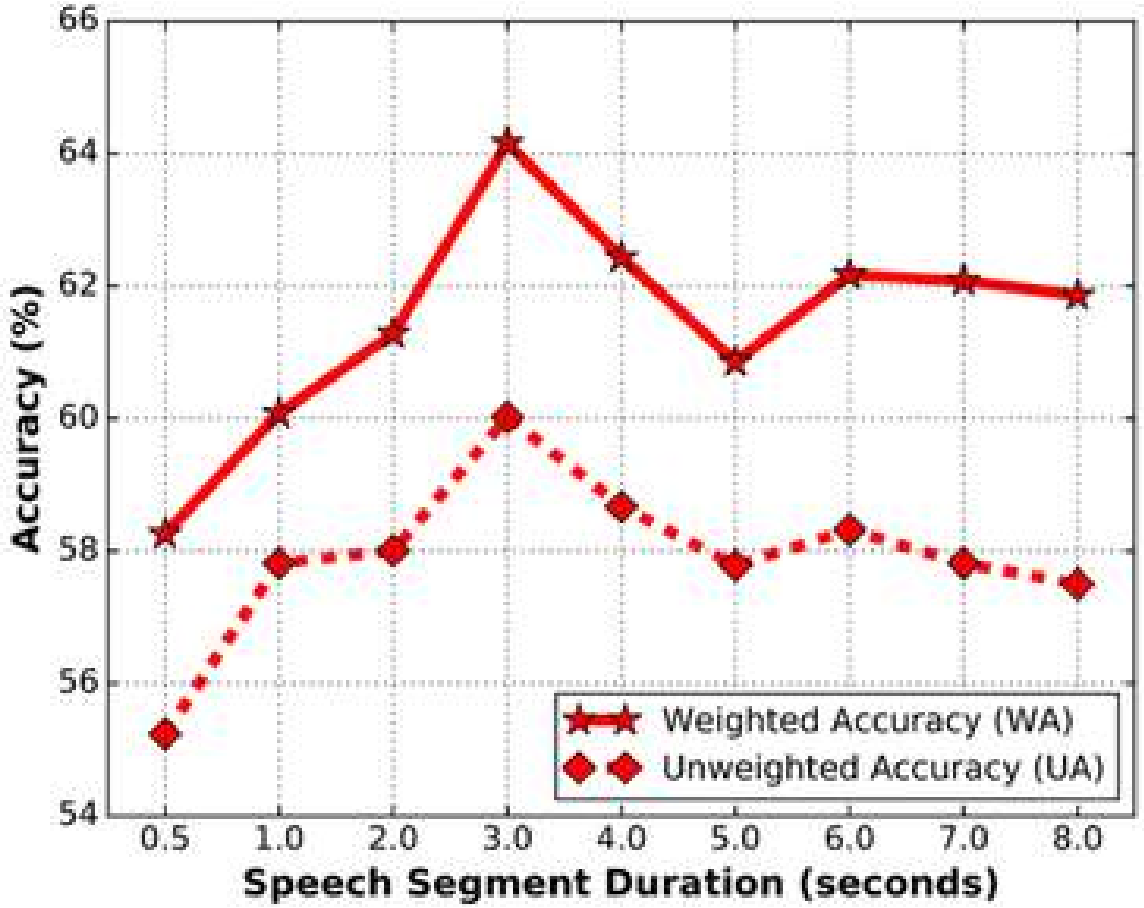
όταν χρησιμοποιούμε απλά μοντέλα RNN. Για το σκοπό αυτό, ο συνδυασμός των προαναφερθέντων μηχανισμών και αρχιτεκτονικής με τη βέλτιστη χρονική κλίμακα των χαρακτηριστικών εισόδου θα μπορούσε ενδεχομένως να οδηγήσει σε μια αξιόπιστη εκτίμηση για την επιλογή του μοντέλου SER όταν χρησιμοποιούνται LLD σε επίπεδο πλαισίου.

Δεν πρέπει επίσης να αγνοούμε το γεγονός ότι επειδή ο υπολογισμός της κλίσης και η ενημέρωση των βαρών πραγματοποιείται διαδοχικά μέσα από τα χρονικά σημεία της ακολουθίας εισόδου, θα ήταν πολύ πιο βολικό για τη διαδικασία εκπαίδευσης να έχουμε όσο λιγότερα χρονικά διαστήματα μπορούσαμε. Ωστόσο, όταν το διάνυσμα κάθε χρονικού βήματος είναι πολύ υψηλό σε διαστάσεις αναμένουμε να δούμε ότι το δίκτυο δεν μπορεί να κωδικοποιήσει εύστοχα τις εξαρτήσεις μεταξύ όλων αυτών των χαρακτηριστικών σε διαφορετικά χρονικά βήματα με τον ίδιο αριθμό παράνομων παραμέτρων. Αυτό είναι επίσης εμφανές από τα αποτελέσματα που παρουσιάζονται στο σχήμα 3.7 όπου η αρχιτεκτονική του RNN παραμένει σταθερή ενώ συγκολλούμε πολλαπλά διανύσματα από LLDs σε επίπεδο πλαισίου για κάθε χρονικό βήμα. Στην άκρη της αλληλουχίας 533 συναθροίσεων βλέπουμε ότι έχουμε πολύ χαμηλές επιδόσεις και στις δύο μετρήσεις WA και UA κυρίως λόγω του λόγου που αναφέραμε προηγουμένως. Αντίθετα, οι κλίμακες φωνημάτων φαίνεται να διαφυλάσσουν επαρκώς τις συναισθηματικές πληροφορίες και να παρέχουν μια καλή λύση για το αντιστάθμισμα μεταξύ του μήκους της ακολουθίας μοντελοποίησης και της ακρίβειας του μοντέλου.

3.5.2 Βέλτιστα χρονοδιαγράμματα για LSTM εκπαιδευμένα σε παγκόσμια-γενικά στατιστικά χαρακτηριστικά

Τα αποτελέσματα του πειράματος όπου το LSTM εκπαιδεύεται σε παγκόσμια-γενικά στατιστικά χαρακτηριστικά φαίνονται στο Σχήμα 3.8. Τα αποτελέσματα στο Σχήμα 3.8 καταδεικνύουν την επίδραση της εκπαίδευσης ενός LSTM με παγκόσμια χαρακτηριστικά σε διαφορετικές χρονικές κλίμακες. Τα στατιστικά χαρακτηριστικά που αντιστοιχούν σε χρονικές κλίμακες συλλαβών (0,5s) και εκφράσεων (6-8)s δεν συμπεριφέρονται τόσο καλά σε σύγκριση με χρονικές κλίμακες λέξεων (3-4 δευτερολέπτων). Τα τμήματα ομιλίας που προσεγγίζουν τις διάρκειες των φωνημάτων δεν περιλαμβάνουν επαρκείς συναισθηματικές πληροφορίες όταν τις εκπροσωπούμε με παγκόσμια στατιστικά χαρακτηριστικά. Από την άλλη πλευρά, όταν εξάγουμε παγκόσμια-γενικά στατιστικά χαρακτηριστικά από σχεδόν ολόκληρη τη φράση, αποκτώνται κακά αποτελέσματα. Είναι σημαντικό να παρατηρήσουμε ότι αν μια φράση υπερβαίνει τη χρονική διάρκεια του επιλεγμένου τμήματος, μηδενίζουμε την πρώτη για να δημιουργήσουμε φορείς σήματος με μήκος τουλάχιστον ίσο με εκείνο του επιλεγμένου τμήματος. Η ψευδής παρουσίαση των παγκόσμιων-γενικών στατιστικών χαρακτηριστικών για μεγαλύτερες διάρκειες πιθανώς οφείλεται στη στατιστική εσφαλμένη παρουσίαση ολόκληρης της συναισθηματικής ομιλίας όταν χρησιμοποιούμε τόσο λίγα χρονικά διαστήματα για τις εισερχόμενες ακολουθίες από τις οποίες θέλουμε να συναγάγουμε το συναισθηματικό περιεχόμενο. Αν ο αριθμός των χρονικών βημάτων εισόδου είναι σχετικά χαμηλός τότε θα ήταν αδύνατο να μάθουμε τις μακροχρόνιες και βραχυπρόθεσμες εξαρτήσεις μεταξύ αυτών των διανυσμάτων χαρακτηριστικών για κάθε χρονική στιγμή. Πρόκειται ουσιαστικά για το ίδιο αντιστάθμισμα, το οποίο συζητήσαμε προηγουμένως, όπου θα θέλαμε να αποκτήσουμε τον ελάχιστο αριθμό χρονικών βημάτων που μπορούν να διευκολύνουν την ταχύτερη σύγκλιση αλλά χωρίς να οδηγούν σε ψευδείς δηλώσεις του συναισθηματικού περιεχομένου για την εκάστοτε ομιλία.

Η εξαγωγή στατιστικών συναρτήσεων υψηλότερου επιπέδου από πολλαπλά LLDs πάνω σε τμήματα ομιλίας, οδηγεί σε μια πιο σημαντική αναπαράσταση της υποκείμενης συναισθηματικής δυναμικής. Επιπλέον, συνδυάζοντας την αποτελεσματικότητα ενός LSTM για την ερμηνεία των μακροπρόθεσμων εξαρτήσεων και της προαναφερθείσας στατιστικής αναπαράστασης πάνω στα τμήματα ομιλίας σε χρονοδιαγράμματα λέξεων, επιτυγχάνουμε την καλύτερη απόδοση σε μετρήσεις WA και UA. Αυτή είναι μια άλλη ένδειξη ότι η έννοια της επιλογής των κατάλληλων χρονικών κλιμάκων είναι το κλειδί για να φτιάξουμε κατάλληλα συστήματα SER, όταν ακολουθούν μια προσέγγιση βασισμένη σε τμήματα με παγκόσμια-γενικά στατιστικά χαρακτηριστικά.



Σχήμα 3.8: Σταθμισμένη ακρίβεια (WA) και μη σταθμισμένη ακρίβεια (UA) για LSTMs εκπαιδευμένα σε παγκόσμια-γενικά στατιστικά χαρακτηριστικά σε διάφορες χρονικές κλίμακες

3.5.3 Σύγκριση μεταξύ διαφορετικών χρονικών κλιμάκων

Όπως έχουμε συζητήσει εκτενώς σε προηγούμενες ενότητες, ακολουθούμε τις τρεις προσεγγίσεις ενός πιο άμεσου LSTM που έχει εκπαιδευτεί σε LLDs σε επίπεδο πλαισίου, μια προσέγγιση που βασίζεται σε τμήματα σε παγκόσμια-γενικά στατιστικά χαρακτηριστικά που εξάγονται από τμήματα ομιλίας, καθώς και μια προσέγγιση βασισμένη σε ολόκληρη την ομιλία χρησιμοποιώντας SVMs. Τα αποτελέσματα των καλύτερων μοντέλων για κάθε κατηγορία παρουσιάζονται στον Πίνακα 3.4.

Πίνακας 3.4: Ακρίβεια των προτεινόμενων μοντέλων σε διαφορετικές χρονικές κλίμακες

Μοντέλο	Τύπος Χαρακτηριστικών Εισόδου	WA (%)	UA(%)
SVM	Γενικά στατιστικά χαρακτηριστικά από όλη την ομιλία	53.54	49.23
LSTM	LLDs συντηγμένα ανά 90ms	59.14	54.2
LSTM	Γενικά στατιστικά χαρακτηριστικά για τμήματα ομιλίας 3s	**64.16**	**60.02**

Είναι προφανές ότι η προσέγγιση με βάση τμήματα ομιλίας παρέχει ένα πολύ πιο ανθεκτικό μοντέλο SER. Συνολικά, η σύμπτηξη πολλαπλών LLDs για SER καθιστά αναγκαία την ύπαρξη ενός μηχανισμού που περιορίζει την επίδραση των πλαισίων που δεν περιέχουν συναισθηματική πληροφορία [60], [66]. Η ανάγκη ενός τέτοιου μηχανισμού μπορεί να αποφευχθεί με τη χρήση στατιστικής αναπαράστασης σε κατάλληλη χρονική κλίμακα. Διαισθητικά, αυτό μοιάζει με το ανθρώπινο σύστημα συναισθηματικής αφαίρεσης λαμβάνοντας υπόψη πληροφορίες από μερικές λέξεις.

Το προτεινόμενο μοντέλο LSTM, εκπαιδευμένο σε στατιστικά χαρακτηριστικά για ένα τμήμα ομιλίας 3 δευτερολέπτων, επιτυγχάνει σχετική βελτίωση 5.02 % σε WA (59.14 % $\rightarrow$ 64.16 %) και

Πίνακας 3.5: Ακρίβεια Μοντέλων στη Βιβλιογραφία και στα προτεινόμενα μοντέλα

Μοντέλο	Τύπος Χαρακτηριστικών Εισόδου	WA (%)	UA(%)
Καλύτερο LSTM [35]	Σπεκτόγραμμα	61.71	58.05
BLSTM-SUA [66]	LLDs	59.33	49.96
BLSTM-WPA [65]	LLDs	63.5	58.8
BLSTM-ELM [61]	LLDs συντμημένα ανά 250ms	62.85	**63.89**
LSTM	Στατιστικά για τμήματα ομιλίας 3s	**64.16**	60.02

Weighted Pooling Attention (WPA), Sub-Utterance Attention (SUA)

5.82 (54.2% → 60.02%) στο UA από το LSTM με τις καλύτερες επιδόσεις που έχει εκπαιδευτεί απευθείας σε LLD και 10.62 % στην WA (53.54% → 64.16%) και 10.79% (49.23% → 60.02%) σε UA από μια απόφαση σε επίπεδο ολόκληρης της ομιλίας με SVM. (βλ. Πίνακα 3.4).

3.5.4 Σύγκριση με τη Βιβλιογραφία

Συγκριτικά αποτελέσματα με αντίστοιχα μοντέλα της βιβλιογραφίας παρουσιάζονται στον Πίνακα 3.5. Το μοντέλο μας ξεπερνά όλες τις αρχιτεκτονικές με RNN με ενσωματομένο μηχανισμό προσοχής: 1) με σταθμισμένη προσοχή BLSTM [65] τόσο σε WA όσο και UA κατά 0,66 % και 1,22 %, 2) με βάση την προσοχή σε υποπροτασιακό επίπεδο BLSTM [66] σε WA και UA κατά 4,83 % και 10,06 % αντίστοιχα. Οι συντάκτες στο [61] δοκιμάζουν το μοντέλο τους μόνο στους αυτοσχέδιους συναισθηματικούς διαλόγους τα οποία αποτελούν υποσύνολο της βάσης δεδομένων IEMOCAP και όπως φαίνεται στο [36] τείνουν να δίνουν υψηλότερες βαθμολογίες σε WA και UA. Επιπλέον, δοκιμάζονται μόνο σε έναν από τους ομιλητές για κάθε συνεδρία IEMOCAP χωρίς να προσδιοριστεί ποιος ομιλητής χρησιμοποιείται για την τελική αξιολόγηση. Παρ 'όλα αυτά, η απλή μας αρχιτεκτονική LSTM είναι 1.31 % υψηλότερη σε WA από το προτεινόμενο μοντέλο BLSTM-ELM. Τέλος, η τρέχουσα προσέγγιση της τεχνολογίας στο [35] αποκτήθηκε χρησιμοποιώντας αποκλειστικά την τελική συνεδρία ως σύνολο αξιολόγησης και ελέγχοντας μόνο τις άλλες 4. Θα πρέπει επίσης να παρατηρήσουμε ότι σε αυτή τη δημοσίευση, οι συγγραφείς έχουν συγχωνεύσει την κλάση ενθουσιασμού με την κλάσση της χαράς που εξισσοροπεί την α-πριόρι κατανομή στις κλάσεις και κατα συνέπεια είναι ευεργετικό όσον αφορά την απόδοση στην μετρική UA. Παρ 'όλα αυτά, η προτεινόμενη αρχιτεκτονική του CNN δίνει πολύ παρόμοια αποτελέσματα (64.78 % WA και 60.89 % UA) με τα δικά μας (64.16 % WA και 60.02 % UA). Ανεξάρτητα από τον αποκλεισμό των συνεδριών, το μοντέλο μας ξεπερνά σαφώς όλες τις απλές αρχιτεκτονικές RNN (LSTM, BLSTM) που αναφέρονται στη βιβλιογραφία. Συγκρίνοντας με την καλύτερη αρχιτεκτονική RNN-LSTM που αναφέρθηκε στο [35], επιτυγχάνουμε σχετική βελτίωση 2.45 % σε WA και 1.97 % σε UA.

Κεφάλαιο 4

Μοντελοποίηση μη γραμμικών υποτροπιάζουσων δυναμικών για αναγνώριση συναισθημάτων από φωνή

Το κεφάλαιο αυτό είναι μια εκτεταμένη έκδοση του εγγράφου [1] το οποίο έχει επιλεχθεί για να παρουσιαστεί στο παγκόσμιο συνέδριο Interspeech 2018 που θα πραγματοποιηθεί στο Hyderabad της Ινδίας στις 2-6 Σεπτεμβρίου 2018. Εάν ο αναγνώστης χρειάζεται να αναφέρει τμήματα αυτού του κεφαλαίου τότε θα ήταν προτιμότερο να χρησιμοποιήσει την παρακάτω αναφορά (ή την πιο ενημερωμένη με τον ίδιο τίτλο και τους ίδους συγγραφείς):

- *Efthymios Tzinis †, Giorgos Paraskevopoulos †, Christos Baziotis, and Alexandros Potamianos, "Integrating recurrence dynamics for speech emotion recognition," in Proceedings of INTERSPEECH, (in press), 2018.*
  †Οι δύο συγγραφείς συμμετείχαν εξίσου σε άυτή την δουλειά

4.1 Κίνητρο

Όπως έχει ήδη αναφερθεί προηγουμένως στην Ενότητα 1.4.2, τα μη γραμμικά φαινόμενα στους μηχανισμούς παραγωγής ομιλίας είναι βασικές πτυχές της κατανόησης των παραλλαγών στη διαδικασία παραγωγής ομιλίας και την αξιοποίηση τους για τους σκοπούς μας στην αναγνώριση συναισθήματος. Συγκεκριμένα, οι μη γραμμικές ταλαντώσεις των φωνητικών χορδών συνιστούν την ανάλυση της μη γραμμικής δυναμικής της φωνής ως βασική πτυχή για την κατανόηση του λόγου και των συναισθημάτων που μπορεί να φέρει η ομιλία. Μερικά από αυτά τα φαινόμενα, όπως οι διακυμάνσεις της ροής του αέρα ομιλίας και των διαφόρων στροβιλισμών, άλματα στις οκτάβες, συγκεντρώσεις θορύβου καθώς και διφωνία που είναι στην πραγματικότητα η ύπαρξη υπο-αρμονικών στον τομέα των συχνοτήτων περιλαμβάνουν μερικά από τα πιο ενδεικτικά γεγονότα που επαληθεύουν το ότι η παραδοχή γραμμικότητας της πηγής -φίλτου ενδέχεται να μην ισχύει. Συνεπώς, η στασιμότητα των πλαισίων ομιλίας για την εκτέλεση της ανάλυσης Fourier είναι ζωτικής σημασίας για την τελευταία προσέγγιση, αλλά γενικά αμφισβητήσιμη. Μπορεί να είναι ιδιαίτερα ωφέλιμο για τα συστήματα SER να εκμεταλλεύονται αυτό το είδος μη γραμμικών πληροφοριών για να επιτύχουν καλύτερες επιδόσεις.

Βασιζόμενοι στη σύλληψη της μη γραμμικής δυναμικής της ομιλίας σε βραχυπρόθεσμα παράθυρα, είναι συμαντικό να ακολουθήσουμε τη διαδικασία της μη γραμμικής ανάλυσης από τα σήματα ομιλίας. Η μη γραμμική ανάλυση ενός σήματος ομιλίας πραγματοποιείται μέσω της ανακατασκευής του αντίστοιχου χώρου φάσης (PS) έγκειται στην εμβύθινση του σήματος σε ένα υψηλότερο χώρο διαστάσεων όπου η δυναμική του ξεδιπλώνεται [78]. Ο χώρος κατάστασης κατασκευάζεται χρησιμοποιώντας χρονικά καθυστερημένες εκδοχές του σήματος, ευθυγραμμίζοντας τες, συντημήτοντάς τες και στη συνέχεια χρησιμοποιώντας το συντμημένο διάνυσμα ως την ανακατασκευασμένη αναπαράσταση του PS. Το ανακατασκευασμένο PS του σήματος είναι απλώς μια άλλη προσέγγιση της πραγματικής δυναμικής του σήματος για την οποία πάντα προσπαθούμε να εκτιμήσουμε καλύτερα τις παραμέτρους της χρονικής καθυστέρησης και της διάστασης εμβύθινσης προκειμένου να κάνουμε μια επιτυχημένη αποδίπλωση της υποκείμενης δυναμικής. Τα επαναλαμβανόμενα μοτίβα αυτών των τροχιών είναι ενδεικτικά χαρακτηριστικά της συμπεριφοράς του συστήματος (βλ. Συζήτηση στην ενότητα 1.4.2 και ειδικότερα το σχήμα 1.5 για την ανακατασκευή της υποτροπιάζουσας δυναμικής ή

δυναμικής επαναληψιμότητας ενός πλαισίου ομιλίας) και μπορούν να αναλυθούν με τη χρήση γραφημάτων επαναληψιμότητας (RPs) [96]. Τα RPs είναι απλώς τοπικοί πίνακες απόστασης οι οποίοι μπορούν να διατηρήσουν τις πληροφορίες επαναληψιμότητας του υπό ανάλυση συστήματος. Για το σκοπό αυτό, η μέθοδος της εξαγωγής ποσοτικών μετρήσεων για την ανάλυση της επαναλαψιμότητας (RQA) του σήματος παρέχει μέτρα πολυπλοκότητας για ένα RP που είναι ικανά να αναγνωρίσουν τις μεταβάσεις του συστήματος μεταξύ περιοχών χαοτικής δυναμικής και αρμονίας [84].

4.2 Σχετική δουλειά

Παρά τις μεγάλες εξελίξεις στην SER, οι περισσότερες σύγχρονες προσεγγίσεις παραμελούν πλήρως τη μη γραμμική πτυχή της φωνητικής δυναμικής για να καταστήσουν τη διαδικασία εξαγωγής χαρακτηριστικών αρκετά μικρή σε χρονικό κόστος και ακολουθώντας πιο άμεσες προσεγγίσεις με ελάχιστη επεξεργασία ομιλίας. Ωστόσο, υπάρχουν διάφορα έργα όπου η μη γραμμική ανάλυση χρησιμοποιώντας PS και άλλες μεθόδους έχει χρησιμοποιηθεί για SER οι οποίες περιγράφονται εν συντομία παρακάτω.

Εκτός από την ανάλυση PS, άλλες προσεγγίσεις ακολουθούν διαφορετικές μεθόδους προκειμένου να εξάγουν μη γραμμικά χαρακτηριστικά από συναισθηματικές εκφράσεις. Στο [130] έχουν εξαχθεί φασματικές απεικονείσεις των σημάτων προκειμένου να συλλάβει τόσο την ακουστική συχνότητα όσο και τα συστατικά της συχνότητας χρονικής διαμόρφωσης πάνω από την στιγμιαία φάση και πλάτος για τις εργασίες SER. Ο συνδυασμός με συμβατικά σύνολα χαρακτηριστικών ήταν αρκετά ευεργετικός για την απόδοση των μοντέλων. Για το σκοπό αυτό, τα χαρακτηριστικά μικροδιαμορφώσεως [43] από το στιγμιαίο εύρος και φάση έχουν διερευνηθεί και αξιολογηθεί στη βάση δεδομένων EmoDB [121] καθώς και ο συνδυασμός τους με MFCCs. Στο ίδιο πλαίσιο των διαμορφώσεων των σημάτων ομιλίας ο τελεστής ενέργειας Teager (TEO) έχει χρησιμοποιηθεί για SER [40]. Ο TEO είναι ένας μη γραμμικός τελεστής που εφαρμόζεται σε σήματα χρονοσειράς και ειδικά ομιλίας. Έχει πολλές ενδιαφέρουσες εφαρμογές σε διάφορους τομείς της ψηφιακής επεξεργασίας σημάτων καθώς και απλό και αποτελεσματικό ορισμό και υλοποίηση. Από τις πιο επιτυχημένες εφαρμογές του TEO υπήρξε η αποδιαμόρφωση των σημάτων AM-FM, η οποία υπήρξε ένας καρποφόρος τομέας έρευνας κατά τα προηγούμενα έτη [41].

Λαμβάνοντας υπόψη την κλασική ανάλυση των χρονοσειρών που χρησιμοποιούν ανακατασκευασμένους χώρους φάσης, μπορούμε να δούμε ένα τεράστιο όγκο εργασιών που επικεντρώθηκε στον τρόπο εφαρμογής αυτής της ανάλυσης για το SER. Στο [131] εκπαιδεύτηκε ένα νευρικό δίκτυο χρησιμοποιώντας ένα σύνολο χειροποίητων χαρακτηριστικών τα οποία αντλούνται από την αναπαράσταση του σήματος στον ανακατασκευασμένο PS. Τα χαρακτηριστικά που χρησιμοποιούνται για την εκπαίδευση του NN είναι στην πραγματικότητα στατιστικά που εφαρμόζονται σε μη γραμμικούς περιγραφείς σε επίπεδο πλαισίου. Οι περιγραφείς που χρησιμοποιήθηκαν ήταν μερικοί από αυτούς που έχουν χρησιμοποιηθεί ευρέως στην ASR [78]. Η εντροπία συσχέτισης, η πολυπλοκότητα Lempel-Ziv, το πρώτο ελάχιστο της αμοιβαίας πληροφορίας, η εντροπία Shannon, η διάσταση συσχετισμού και ο εκθέτης Hurst χρησιμοποιήθηκαν για τη διάκριση μεταξύ θυμού, φόβου και ουδέτερων συναισθηματικών δηλώσεων από τη βάση δεδομένων EmoDB [121]. Ένα ποιοτικό εύρημα αυτής της μελέτης ήταν ότι οι συναισθηματικές φράσεις που σχετίζονται με το φόβο και την οργή δείχνουν πιο θορυβώδεις και σύνθετες δομές από εκείνες που σχετίζονται με την ουδέτερη κατάσταση. Επιπλέον, στο [42] μια ποικιλία γεωμετρικών μετρήσεων από τις ανασχηματισμένες τροχιές PS ανέφεραν σημαντική βελτίωση σε SER όταν συνδυάστηκαν με συμβατικά σύνολα χαρακτηριστικών. Συγκεκριμένα, από κάθε PS σε επίπεδο πλαισίου έχουν υπολογιστεί μερικά μέτρα όπως: απόσταση από το κέντρο του ελκυστή, μήκος τμημάτων της τροχιάς, τρέχουσα γωνία μεταξύ των διαδοχικών τμημάτων τροχιάς και απόσταση των σημείων από τη ταυτοική γραμμή και μετά από αυτά κάποιες στατιστικές συναρτήσεις εφαρμόστηνκαν για την τελική ταξινόμηση χρησιμοποιώντας SVM.

Παρόλο που οι τελευταίες προσεγγίσεις είναι αρκετά ενδιαφέρουσες και παρέχουν ενδιαφέροντα αποτελέσματα σχετικά με τη μη γραμμική συμπεριφορά των σημάτων ομιλίας συναισθηματικών ρημάτων, οι ιδιότητες επαναληψιμότητας των τελευταίων δεν έχουν χρησιμοποιηθεί ακόμα για SER.

Τα RPs από τους ανακατασκευασμένους χώρους φάσης είναι σε θέση να συλλάβουν μερικές από τις πιο σημαντικές πτυχές του υποκείμενου δυναμικού συστήματος και ειδικά της υποτροπιάζουσας δυναμικής των εμβυθισμένων τροχιών [84] και έτσι θα μπορούσαν να χρησιμοποιηθούν και για την ανάλυση συναισθηματικών προτάσεων ομιλίας. Ωστόσο, οι πληροφορίες από τα RP δεν έχουν ακόμη χρησιμοποιηθεί για εργασίες SER. Ως εξαίρεση από την προηγούμενη δήλωση στο [132], τα μέτρα RQA που εξάγονται από RPs των ανακατασκευασμένων χώρων φάσης των πλαισίων ομιλίας έχουν αποδειχθεί στατιστικά σημαντικά για τη διάκριση των συναισθημάτων. Αν και η στατιστική ανάλυση περιελάμβανε διαφορετικές παραμέτρους των ανακατασκευασμένων μέτρων PS και RQA που χρησιμοποιούνται ευρέως για την ανάλυση RPs [79], λείπει μια πραγματική πειραματική διάταξη για να αξιολογηθεί το πόσο καλοί είναι αυτοί οι τοπικοί περιγραφητές για SER. Αυτό είναι βασικό επειδή η στατιστική σημασία που παρουσιάζεται σε μια μελέτη μπορεί να είναι αρκετά παραπλανητική για μια πραγματική περίπτωση όπου κάποιος χρειάζεται ένα αυτόνομο σύστημα που εκτελεί ταξινόμηση συναισθημάτων. Συγκεκριμένα, η στατιστική ανάλυση των ζευγαριών δεν μπορεί να υποκαταστήσει την πραγματική πειραματική ρύθμιση χρησιμοποιώντας όλα τα χαρακτηριστικά και συγκεκριμένα μοντέλα που μπορούν να χρησιμοποιηθούν σε πραγματικά προβλήματα.

Αντίθετα, οι πληροφορίες επανάληψιμότητας από RPs και ιδιαίτερα το χαρακτηριστικό που λαμβάνεται από την εκτέλεση της μεθόδου RQA έχουν χρησιμοποιηθεί εκτενώς σε πολλούς άλλους τομείς και εφαρμογές που σχετίζονται επίσης με την αναγνώριση συναισθημάτων αλλά όχι από την ομιλία. Για παράδειγμα, στο [133] παρουσιάστηκε μια ανάλυση σειρών χρονοσειρών και σύνθετων δικτύων που χρησιμοποιούν γραφλήματα επαναληψιμότητας. Συγκεκριμένα μερικά μέτρα έδειξαν ότι αντιπροσωπεύουν δομικές πτυχές δυναμικών συστημάτων που είναι συμπληρωματικά προς εκείνα που χαρακτηρίζονται από άλλες μεθόδους ανάλυσης χρονοσειρών. Επιπλέον, στο [134] χρησιμοποιήθηκε RQA για τη μέτρηση της δομικής σύζευξης και του συγχρονισμού σε φυσικές και κλινικές λεκτικές αλληλεπιδράσεις. Τα στοιχεία κίνησης του σώματος [135] αναλύθηκαν για την εύρεση εξαπάτησης χρησιμοποιώντας δυναμικές υπογραφές που εξάγονται από RPs. Τα αποτελέσματά αυτής της έρευνας έδειξαν ότι οι συνεχείς διακυμάνσεις των παραπλανητικών κινήσεων τόσο στην άνω πλευρά όσο και στους βραχίονες χαρακτηρίζονται από δυναμικές ιδιότητες με λιγότερη σταθερότητα και μεγαλύτερη πολυπλοκότητα που αντανακλάται στα αντίστοιχα μέτρα της μεθόδου RQA. Επιπλέον, στο [136] το RQA έχει χρησιμοποιηθεί για την αναγνώριση των δυναμικών μοτίβων εγκεφαλικών ερεθισμάτων και την ταξινόμηση των συναισθημάτων των αντίστοιχων σημάτων ηλεκτροεγκεφαλογράμματος (EEG). Είναι ενδιαφέρον ότι οι τροχιές του PS εμφανίζουν υψηλότερη περιοδικότητα κατά τη διάρκεια της συναρπαστικής αρνητικής, ενώ στα θετικά συναισθήματα το υποκείμενο σύστημα είναι εξαιρετικά χαοτικό. Τα ευρήματα της προαναφερθείσας μελέτης υποδεικνύουν ότι το RQA έχει τη δυνατότητα να ανακαλύψει διαφορές χαρακτηριστικών σημάτων εγκεφάλου αντίστοιχες με ένα συναισθηματικό ερέθισμα το οποίο είναι αρκετά ενδιαφέρον εάν σκεφτούμε αν αυτό θα μπορούσε να εφαρμοστεί και στα συναισθηματικά σήματα ομιλίας.

Στις σύγχρονες προσεγγίσεις, έχουν μελετηθεί διάφορες άλλες προσεγγίσεις για την αποτελεσματική εξαγωγή των ιδιοτήτων επαναληψιμότητας των σημάτων. Για παράδειγμα, στο [137] εξάγονται νέα RPs για διαφορετικές κλίμακες συχνότητας προκειμένου να αναλυθούν οι ιδιότητες επαναληψιμότητας του σήματος εισόδου χρησιμοποιώντας μια προσέγγιση πολλαπλών χρονικών κλιμάκων. Με άλλα έργα, οι πληροφορίες μη γραμμικής επαναληψιμότητας και υποτρπιάζουσας δυναμικής έχουν επίσης εξαχθεί απευθείας από RPs χωρίς τη χρήση μέτρων RQA. Συγκεκριμένα, στο [138] ένα CNN εκπαιδεύτηκε χρησιμοποιώντας RPs για γενική εργασία ταξινόμησης χρονοσειρών και έδειξε ότι αυτή η μέθοδος παρέχει μια άριστη μέθοδο για την ταξινόμηση τόσο των στατικών όσο και των μη στατικών χρονικών σειρών.

4.3 Ανακατασκευή χώρου φάσης από σήματα ομιλίας

4.3.1 Αναδίπλωση της Αληθινής Δυναμικής της Ομιλίας

Γενικά, υποτίθεται ότι τα σήματα ομιλίας, δεδομένου ότι παράγονται από ένα τόσο περίπλοκο και άκρως μη γραμμικό σύστημα όπως ο μηχανισμός παραγωγής ομιλίας του ανθρώπου (βλέπε ενότητα 1.4.2), φέρουν κρυφά χαρακτηριστικά της υποκείμενης δυναμικής του συστήματος γεννήτριας που είναι ακόμα άγνωστο [78] . Με άλλα λόγια, υποτίθεται ότι το δειγματοληπτημένο σήμα ομιλίας $\{s(i)\}_{i=1}^{N}$ το οποίο είναι μια συνάρτηση του χρόνου σε 1 διάσταση θα μπορούσε να αντιπροσωπεύει μια προβολή του πραγματικού συστήματος $\{\mathbf{s}^{\star}(i)\}_{i=1}^{N}$ όπου $\mathbf{s}^{\star}(i) \in \mathbb{R}^{d^{\star}} \ \ \forall 1 \leq i \leq N$. Υποθέτουμε ότι το $d_{\star}$ αντιστοιχεί στις διαστάσεις του διανύσματος που αντιπροσωπεύει το πραγματικό δυναμικό σύστημα υπό ανάλυση. Ουσιαστικά, θα θέλαμε να αναδημιουργήσουμε τη δυναμική και τις αμετάβλητες ιδιότητες του άγνωστου συστήματος $\{\mathbf{s}^{\star}(i)\}_{i=1}^{N}$ χρησιμοποιώντας μόνο την μονοδιάστατη προβολή του $\{s(i)\}_{i=1}^{N}$.

Παρομοίως, θα μπορούσαμε να αποκτήσουμε μια πιο διαισθητική εικόνα για τις έννοιες της ανακατασκευής της πραγματικής δυναμικής, παρουσιάζοντας ένα παράδειγμα στην πραγματική ζωή. Φανταστείτε μια σφαίρα με ακτίνα ρ που βρίσκεται σε ένα τρισδιάστατο χώρο και στη συνέχεια προβάλλοντάς την σε ένα επίπεδο 2 διαστάσεων. Αυτό θα είχε ως αποτέλεσμα έναν κύκλο με ακτίνα ρ. Αν κάποιος μπορούσε να δει μόνο την προβολή στο επίπεδο, δεν θα ήταν δυνατόν να καταλάβουμε ότι αυτή η προβολή αντιστοιχεί σε μια σφαίρα που βρίσκεται σε ένα τρισδιάστατο χώρο. Επιπλέον, αν προβάλλουμε κάθετα ένα φλιτζάνι τσάι με έναν κυκλικό πυθμένα ακτίνας ρ που βρίσκεται σε ένα χώρο 3 διαστάσεων στο ίδιο επίπεδο και ελέγχοντας τον προκύπτοντα κύκλο θα καταλήξουμε σε ένα ίδιο σχήμα όπως αυτό που παράγεται από την προβολή του σφαίρα. Θα ήταν επίσης αδύνατο να διακρίνουμε το αρχικό αντικείμενο διαστάσεων 3 από το οποίο προέκυψε αυτός ο κύκλος ως προβολή του. Τώρα φανταστείτε ότι το αρχικό αντικείμενο των τριών διαστάσεων αντιπροσωπεύει την πραγματική δυναμική του κάθε τύπου συστήματος και η προβολή στο δισδιάστατο επίπεδο είναι αυτό που μπορούμε να δούμε. Στην περίπτωση αυτή, έχουμε δύο τύπους συστημάτων: το σύστημα κυπέλλου και το σύστημα σφαίρας αλλά η παρατηρούμενη αναπαράσταση είναι κοινή (κύκλος στο επίπεδο) και δεν μπορούμε να είμαστε σίγουροι για το ποιο σύστημα δημιούργησε αυτή την προβολή. Παρόλο που θα ήταν πολύ χρήσιμο να αποκτήσουμε πληροφορίες σχετικά με το αντικείμενο που χρησιμοποιήθηκε για την προβολή με την παρατήρηση μόνο των προβολών που μπορούν να χρησιμοποιηθούν για τον προσδιορισμό του αντίστοιχου πρωτότυπου συστήματος.

Επιστρέφοντας στο σκοπό μας, επιδιώκουμε να αποκαλύψουμε τις πραγματικές δυναμικές πτυχές της παραγωγής ομιλίας όταν υπάρχει έκφραση συναισθημάτων κατά την διάρκεια της ομιλίας. Προφανώς, αν μπορούσαμε να καταγράψουμε δυναμικές πτυχές όπως τα πρότυπα υποτροπιαζουσών δυναμικών από τους ελκυστές των πλαισίων ομιλίας και να τα συγκεντρώσουμε μπορούμε να δημιουργήσουμε καλύτερα συστήματα SER. Σε αυτό το πλαίσιο, πρέπει να βρούμε τρόπους για να μπορέσουμε να ανακατασκευάσουμε, τουλάχιστον εν μέρει, την πραγματική δυναμική του συστήματος, ξεδιπλώνοντάς τις σε ανώτερες διαστάσεις. Είναι προφανές ότι αυτή η διαδικασία δεν είναι τετριμμένη και πρέπει να γίνει προσεκτική εξέταση όλων των μεταβλητών για την ανασυγκρότηση αυτή. Όπως φαίνεται παρακάτω, θα καθορίσουμε πώς μπορούμε να ανακατασκευάσουμε τοπικά το χώρο φάσης κάθε πλαισίου ομιλίας προκειμένου να χρησιμοποιήσουμε αυτές τις πληροφορίες για την εξαγωγή συναισθηματικού περιεχομένου από την άποψη των δυναμικών υπογραφών της εκτίμησης του πραγματικού συστήματος.

4.3.2 Ορισμός

Λαμβάνοντας υπόψη ένα πλαίσιο ομιλίας με N δείγματα $\{s(i)\}_{i=1}^{N}$ ανασυνθέτουμε την αντίστοιχη τροχιά του PS με υπολογισμό d_e χρονικά καθυστερημένων εκδόσεων του αρχικού πλαισίου ομιλίας με τ πολλαπλάσια χρονικής καθυστέρησης και δημιουργώντας διανύσματα που βρίσκονται στον πραγματικό Ευκλείδιο χώρο $\mathbb{R}^{d_e}$ όπως ορίζεται στην εξίσωση 2.22.

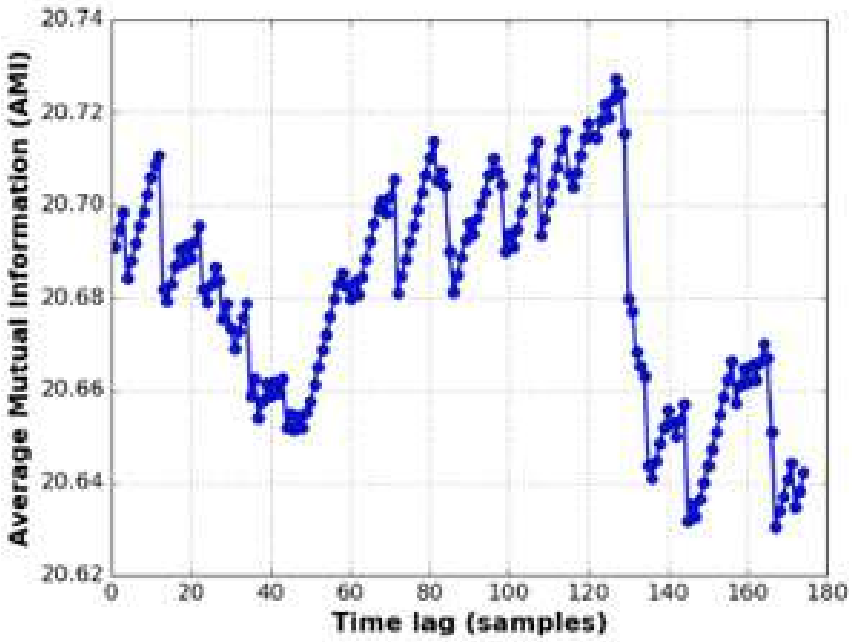
Για την ανασυγκρότηση του PS κάθε πλαισίου ομιλίας πρέπει να καθορίσουμε τις δύο παραμέ-

τρους της χρονικής καθυστέρησης τ και της διάστασης εμβύθινσης d_e. Στα επόμενα τμήματα 4.3.3 και 4.3.5 θα αναλύσουμε πλήρως τον τρόπο με τον οποίο επιλέγονται αυτές οι παράμετροι καθώς και ορισμένα ποιοτικά αποτελέσματα αλλάζοντας αυτές τις παραμέτρους για διαφορετικές διάρκειες των ομιλιών. Στην Ενότητα 4.3.6 παρέχουμε κάποια απεικόνιση PS για διάφορα τμήματα.

4.3.3 Εκτίμηση της παραμέτρου χρονικής καθυστέρησης

Προκειμένου να εκτιμηθεί η παράμετρος χρονικής καθυστέρησης τ για τον αριθμό των δειγμάτων καθυστέρησης που θα εφαρμοστούν για το σήμα εισόδου $\{s(i)\}_{i=1}^{N}$ χρησιμοποιούμε το κριτήριο AMI $\mathcal{I}(\tau)$ (βλ. Τον ορισμό στο τμήμα 2.4.2) για να επιλέξουμε ένα επαρκές χρονικό διάστημα τ δειγμάτων για το σήμα εισόδου. Η συνάρτηση $\mathcal{I}(\tau) : \mathbb{N} \to \mathbb{R}$ θα χαρτογραφήσει όλες τις επιλεγμένες ακέραιες τιμές για τ σε πραγματικούς αριθμούς, υποδεικνύοντας την αντίστοιχη αμοιβαία πληροφορία για κάθε χρονική υστέρηση. Μ M $\mathcal{M}$

Στο σχήμα 4.1 παρέχουμε ένα παράδειγμα της συνάρτησης AMI για ένα τμήμα ομιλίας 200ms. Είναι προφανές ότι για μικρές χρονικές υστερήσεις υπάρχει γενικά υψηλότερη αμοιβαία πληροφορία, ενώ για πολύ μεγάλες τιμές μπορούμε να δούμε ότι το αντίθετο ισχύει. Επιπλέον, η περιοδικότητα στο γράφημα AMI αντιστοιχεί σε ιδιότητες επαναληψιμότητας του τμήματος ομιλίας για συγκεκριμένες χρονικές υστερήσεις που δεικνύουν όχι μόνο γραμμικούς αλλά και μη γραμμικούς συσχετισμούς.

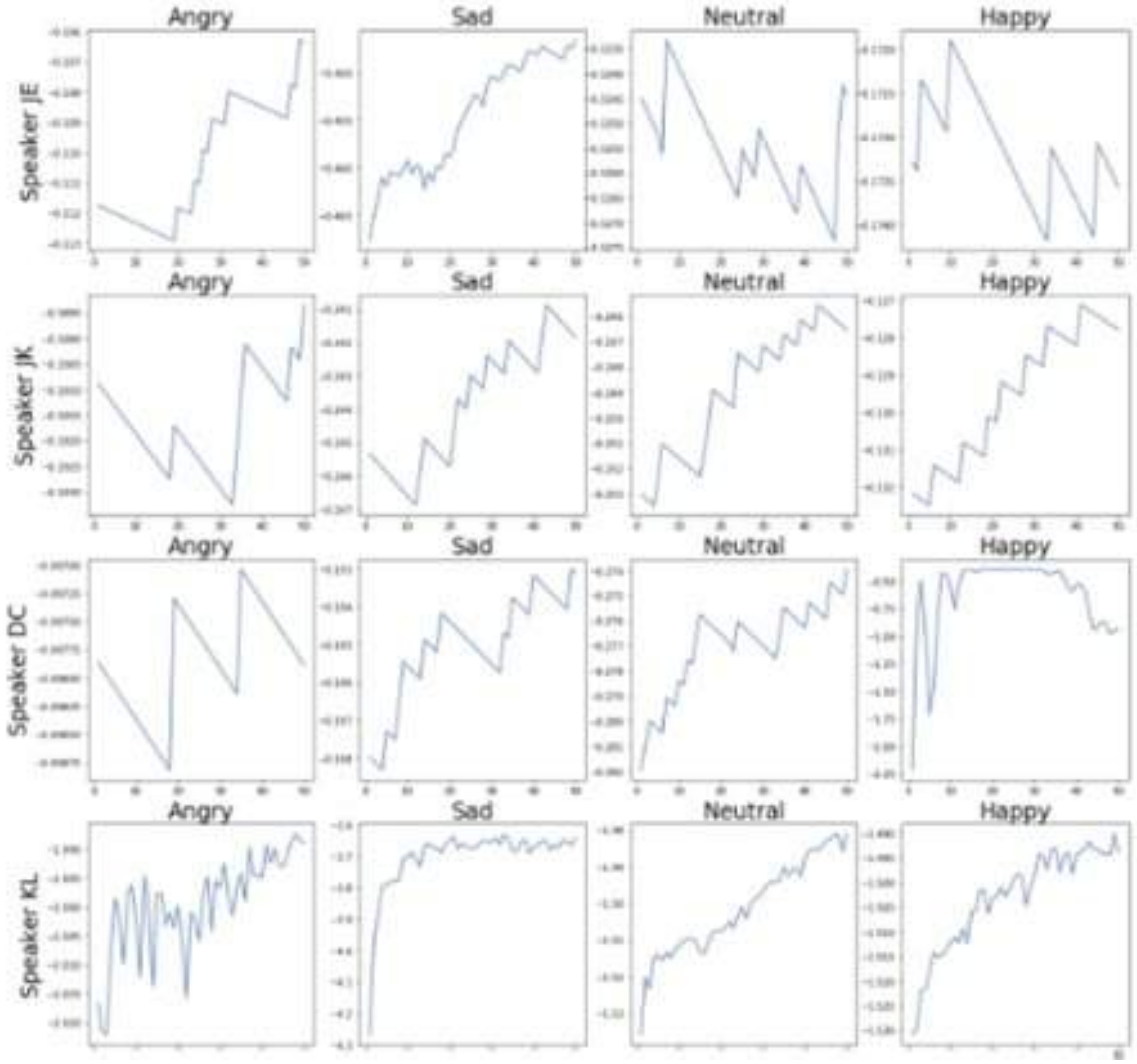


Σχήμα 4.1: Μέση αμοιβαία πληροφορία για ένα τμήμα ομιλίας από 0.2 δευτερόλεπτα

4.3.4 Αναλύοντας την AMI για διαφορετικούς ομιλητές και συναισθήματα

Στο σχήμα 4.2 εμφανίζεται μια ποικιλία από αντίστροφα παράθυρα AMI για διαφορετικά ηχεία και συναισθήματα για το φώνημα /αε/. Οι δηλώσεις που χρησιμοποιούνται για αυτό το Σχήμα αντλούνται από την βάση δεδομένων Surrey Audio-Visual Express Emotion (SAVEE) [139] ακολουθώντας τα ίδια ψευδώνυμα για τα ηχεία και τα διαθέσιμα συναισθήματα. Μπορούμε να δούμε ότι η AMI μειώνεται (στο γράφημα εμφανίζεται το αντίστροφο των AMIs) εξαιτίας των φθίνουσων συσχετισμών των σημάτων εισόδου. Οι ομοιότητες μεταξύ των ομιλητών ή των συναισθημάτων της λειτουργίας AMI είναι εμφανείς. Για παράδειγμα, για τον ομιλητή *JK* οι *Λύπη* και *Ουδετερότητα* εκφράσεις του /αε/ είναι πολύ πιο όμοια από τα άλλα δύο συναισθήματα που εμφανίζονται. Επιπλέον, οι δηλώσεις *Θυμού* και για τους ομιλητές *DC* και *JK* φαίνονται αρκετά παρόμοια. Ωστόσο, αυτές οι λειτουργίες

εξάγονται ως μέτρα για την εκτίμηση των παραμέτρων του PS. Ως εκ τούτου, αυτοί οι ευρεστικοί αλγόριθμοι για την εκτίμηση παραμέτρων δεν μπορούν να χρησιμοποιηθούν άμεσα για την ταξινόμηση των συναισθημάτων.



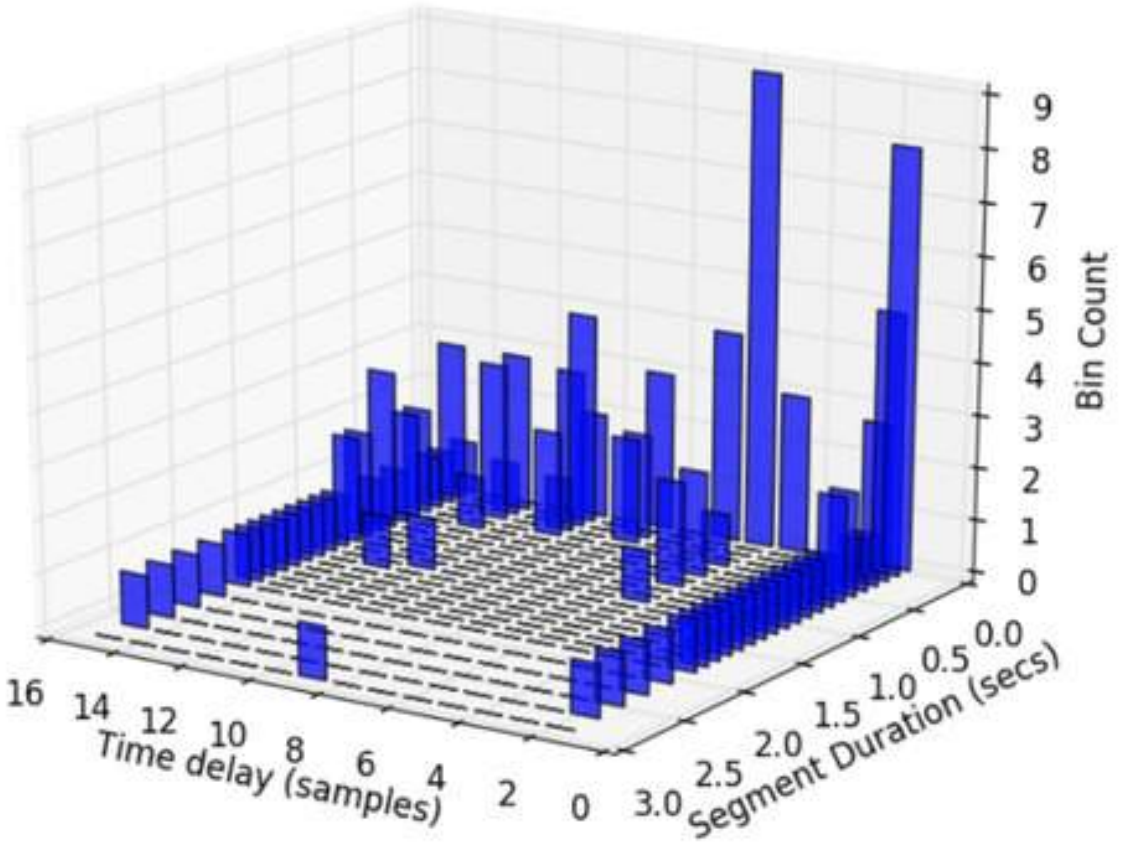
Σχήμα 4.2: Αντίστροφες συναρτήσεις AMI για διαφορετικούς ομιλητές και συναισθήματα για το φώνημα /αε/

Προκειμένου να αποφευχθούν τα ακραία συμβάντα των τελικών συσχετισμών των τιμών $\{s(i)\}_{i=1}^{N}$ και $\{s(i+\tau)\}_{i=1}^{N}$ ή περιπτώσεις όπου τα δύο σήματα είναι εντελώς ασυσχέτιστα. Αυτοί οι δύο τύποι περιπτώσεων αντιστοιχούν συνήθως σε πολύ μικρές ή πολύ μεγάλες τιμές αντίστοιχα. Προκειμένου να αποφευχθούν τέτοιες ακραίες περιπτώσεις, επιλέγουμε την χρονική καθυστέρηση τ με βάση το πρώτο τοπικό ελάχιστο της συνάρτησης AMI $\mathcal{I}(\tau)$. Πιο τυπικά, επιλέγουμε την ακόλουθη χρονική υστέρηση:

$$\tau = min\{\hat{\tau} \mid \mathcal{I}(\hat{\tau}) < min(\mathcal{I}(\hat{\tau}+1), \mathcal{I}(\hat{\tau}) < \mathcal{I}(\hat{\tau}-1))\} \tag{4.1}$$

Αυτή η διαδικασία επιλογής του πρώτου τ που τοπικά ελαχιστοποιεί τη συνάρτηση AMI $\mathcal{I}(\tau)$ θα παράγει διαφορετικά αποτελέσματα για κάθε μήκος ενός τμήματος ομιλίας. Έτσι, εάν προσπαθήσουμε να σπάσουμε κάθε έκφραση σε πλαίσια ή τμήματα ομιλίας, θα εργαστούμε ανεξάρτητα σε κάθε μία από αυτές τις χρονικές περιόδους για την ανακατασκευή του αντίστοιχου PS. Επομένως θα ήταν σημαντικό να αναλυθεί περαιτέρω ο τρόπος με τον οποίο εκτιμάται η παράμετρος χρονικής υστέρησης χρησιμοποιώντας διάφορες διάρκειες για τη διάσπαση σε τμήματα ομιλίας / πλαίσια.

Στην Εικόνα 4.3 έχουμε επιλέξει μια έκφραση *Θυμού* από 3 δευτερόλεπτα και τρέχουμε τον προηγούμενο αλγόριθμο για να υπολογίσουμε το βέλτιστο χρονικό διάστημα για την ανακατασκευή των χώρων φάσης για διάφορες χρονικές διάρκειες των ομιλιών και πλαισίων. Το αρχικό σήμα υποβλήθηκε σε δειγματοληψία στο 44100 kHz επειδή είναι επίσης ένα δείγμα του συνόλου δεδομένων SAVEE. Περιορίσαμε την αναζήτηση για το βέλτιστο τ θέτοντας το μέγιστο $\tau_{max} = 16$. Καθώς αυξάνεται η χρονική διάρκεια των τμημάτων, μπορούμε να δούμε ότι είτε το $\tau = 1$ είτε το $\tau = 16$ επειδή η συσχέτιση μεταξύ του ίδιου σήματος ορίζεται από μια μονότονη συνάρτηση η οποία δεν δημιουργεί κανένα τοπικό ελάχιστο σε έναν τόσο μικρό αριθμό διαθέσιμων χρονικών υστερήσεων για αναζήτηση. Ωστόσο, τα πιο εμφανή αποτελέσματα αυτού του σχήματος είναι ότι για τις χρονικές κλίμακες που αντιστοιχούν σε χρονικά πλαίσια περίπου 20ms (που είναι αυτά που χρησιμοποιούνται επίσης στην παραδοσιακή επεξεργασία ομιλίας [6]) η κατανομή των επιλεγμένων χρονικών υστερήσεων δεν είναι διασκορπισμένη στις δύο ακραίες τιμές για χρονικές υστερήσεις. Αυτό είναι ενδεικτικό των διαφορετικών χρονικών υστερήσεων που θα επιλεχτούν για διαφορετικά πλαίσια που θα μπορούσαν να περιέχουν περιόδους σιωπής, σύμφωνα ή τρεμάμενη ηχός.



Σχήμα 4.3: Επιλεγμένες χρονικές υστερήσεις για μια έκφραση *Θυμού* των 3 δευτερολέπτων για διάφορες χρονικές κλίμακες

4.3.5 Εκτίμηση της παραμέτρου διάστασης εμβύθινσης

Για αυτή την ενότητα θα υποθέσουμε ότι η χρονική καθυστέρηση τ εκτιμάται επίσης όπως έχει αποδειχθεί στην προηγούμενη ενότητα 4.3.3 ή εκτιμάται διαφορετικά. Τώρα που έχουμε ρυθμίσει την χρονική καθυστέρηση τ θα πρέπει επίσης να καθορίσουμε τον αριθμό του πλήθους των χρονικά υστερημένων αντιγράφων του σήματος που θα χρησιμοποιήσουμε για την ανασυγκρότηση του χώρου φάσης του δεδομένου σήματος $\{s(i)\}_{i=1}^{N}$. Αυτό είναι κρίσιμο, διότι όπως αναφέρθηκε προηγουμένως

(βλέπε Ενότητες 2.4.4 και 4.3), η δυναμική του ανακατασκευασμενου χώρου του σήματος εισόδου θα πρέπει να διατηρεί καταλλήλως τις αμετάβλητες ποσότητες του αρχικού γεννήτορα συστήματος $\mathbf{s}^\star$. Αυτό ισχύει μόνο εάν οι επιλεγμένες διαστάσεις d_e του ανακατασκευασμένου χώρου φάσης είναι σε θέση να ξεδιπλώσουν τη δυναμική χωρίς να συρρικνωθούν οι τροχιές της προσεγγιστικής πολλαπλότητας $\mathcal{M}$ του ελκυστή.

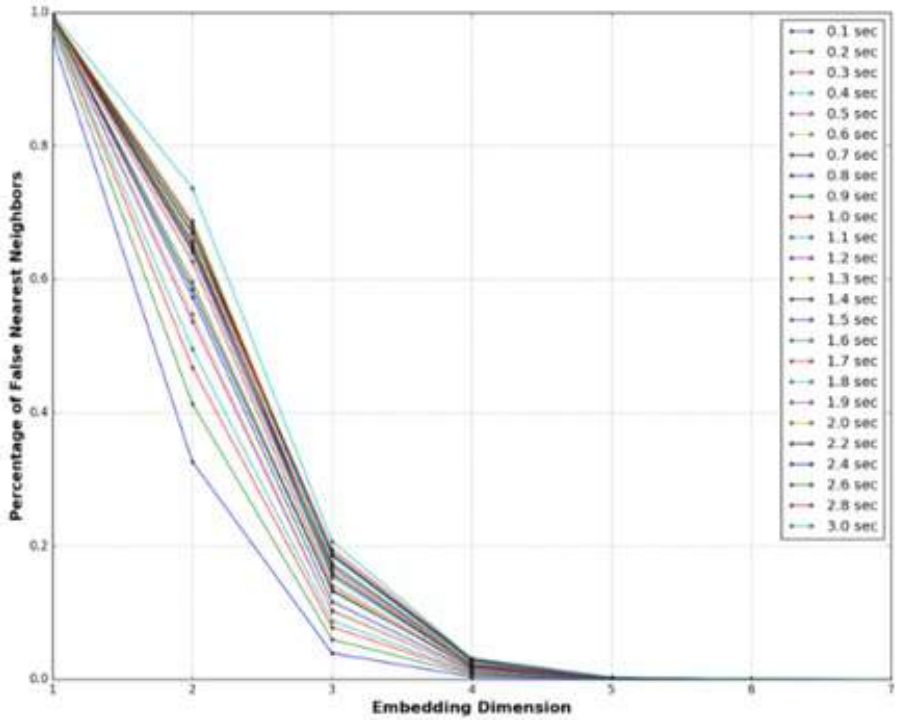
Για την εκτίμηση των διαστάσεων ενσωμάτωσης θα χρησιμοποιήσουμε την ευρετική μέθοδο Λανθάνοντες κοντινότεροι γείτονες (FNN) όπως περιγράφηκε στην προηγούμενη ενότητα 2.4.4. Προκειμένου να αποκτήσουμε μια εικόνα για το πώς αλλάζει το ποσοστό των ψευδών πλησιέστερων γειτόνων καθώς προσθέτουμε περισσότερες διαστάσεις για την ανασυγκρότηση του χώρου φάσης, σχεδιάζουμε αυτούς τους αριθμούς για διαφορετικά μήκη ομιλιών. Στο σχήμα 4.4 παρουσιάζουμε τα ποσοστά του συνολικού αριθμού των πλησιέστερων πλησιέστερων γειτόνων καθώς προσθέτουμε μια επιπλέον διάσταση στην ανασυγκρότηση του PS από τον προηγούμενο αριθμό διαστάσεων. Αυτή η γραφική παράσταση αναφέρεται στην κατάτμηση της αρχικής συναισθηματικής έκφρασης των 3 δευτερολέπτων που χρησιμοποιήθηκε επίσης στον προηγούμενο πειραματισμό μας για την εκτίμηση της επιλεγμένης χρονικής καθυστέρησης τ για κάθε τμήμα χωριστά (βλ. Εικόνα 4.3). Συγκεκριμένα, στις Εικόνες 4.4a και 4.4b παρουσιάζουμε τα ποσοστά του FNN για μια εκτίμηση της χρονικής καθυστέρησης από το πρώτο τοπικό ελάχιστο της AMI και μια ad hoc επιλογή χρονικής καθυστέρησης $\tau = 1$, αντίστοιχα. Είναι ξεκάθαρο ότι στην πρώτη μέθοδο 4.4a τα ποσοστά των πλησιέστερων πλησιέστερων γειτόνων είναι αρκετά μικρότερα από αυτά που παρουσιάζονται στο 4.4b λόγω της εκτίμησης της παράμετρος χρονικής καθυστέρησης τ με βάση το τοπικό ελάχιστο της AMI αντί της χρήσης ad-hoc τιμής. Είναι αξιοσημείωτο ότι αυτή η δήλωση ισχύει για τις περισσότερες από τις διάρκειες των παρουσιαζόμενων τμημάτων ομιλίας. Και στα δύο διαγράμματα μπορούμε να δούμε ότι καθώς αυξάνουμε τον αριθμό των διαστάσεων, ο αριθμός των FNN συγκλίνει στο μηδέν, ο οποίος είναι σύμφωνος με τη διαίσθησή μας καθώς και με τη μαθηματική περιγραφή του αλγορίθμου (βλέπε Ενότητα 2.4.4). Επιπλέον, μπορούμε να διαβεβαιώσουμε ποιοτικά την πεποίθησή μας ότι για το $d_e = 3$ έχουμε επαρκή αριθμό FNN που είναι μικρότερο από 20%.

Σε γενικές γραμμές, δεν μας ενδιαφέρουν πλέον τμήματα ομιλίας εκτός από εκείνα που αντιστοιχούν στα χρονικά πλαίσια του εκάστοτε τμήματος υπό ανάλυση, επειδή αυτό είναι το χρονικό διάστημα από το οποίο θα θέλαμε να εξαγάγουμε τις τοπικές τροχιές PS των πλαισίων ομιλίας. Αυτό είναι αρκετά καθησυχαστικό καθώς η μέθοδος μας δείχνει ότι θα μπορούσαμε πραγματικά να ανακατασκευάσουμε την τοπική δυναμική επαναληψιμότητας από κάθε πλαίσιο ομιλίας χρησιμοποιώντας μόνο 3 διαστάσεις για την ανακατασκευή του τοπικού PS. Με άλλα λόγια, καθώς ο αριθμός των FNN δεν μειώνεται σημαντικά με την προσθήκη μιας επιπλέον 4ης διάστασης μπορούμε να υποθέσουμε ότι η δυναμική του πραγματικού συστήματος $\mathbf{s}^\star$ θα ήταν κατάλληλη να περιγράψει την δυναμική του πραγματικού ελκυστή χρησιμοποιώντας μια εμβυθισμένη πολλαπλότητα $\mathcal{M} \in \mathbf{R}^3$. Το τελευταίο αποτέλεσμα είναι επίσης σημαντικό, καθώς μπορούμε να απεικονίσουμε αυτή την ξεδιπλωμένη δυναμική.

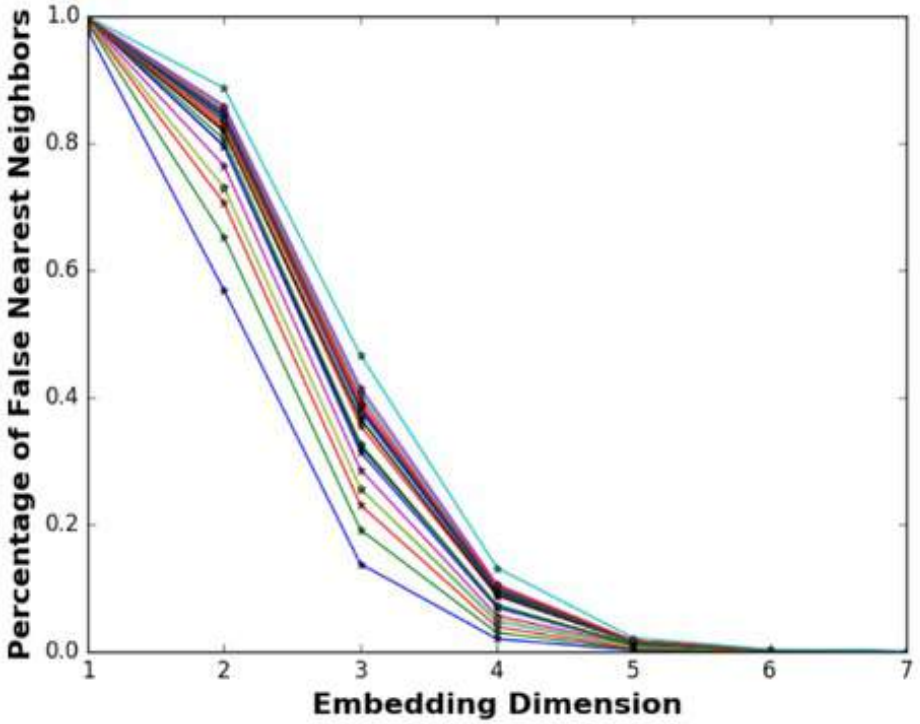
4.3.6 Ανακατασκευή χώρων φάσης για τμήματα ομιλίας διαφορετικής χρονικής κλίμακας

Αφού θέσουμε τις παραμέτρους κάθε PS που θέλουμε να προσεγγίσουμε, θα πρέπει να προχωρήσουμε στην πραγματική ανακατασκευή αυτού του ελκυστή. Χρησιμοποιούμε την εξίσωση 2.22 και χρησιμοποιούμε τις προκαθορισμένες παραμέτρους της χρονικής καθυστέρησης τ και τις διαστάσεις ενσωμάτωσης d_e προκειμένου να αναπαραστήσουμε κάθε δείγμα με ένα διάνυσμα d_e διαστάσεων που βρίσκεται στην προκύπτουσα πολλαπλότητα $\mathcal{M} \in \mathbf{R}^3$.

Στο σχήμα 4.5, εμφανίζεται μια ποικιλία διαφορετικών ανακατασκευασμένων τροχιών διαστήματος φάσης για διάφορα φωνήματα ομιλίας. Από αυτό το διάγραμμα μπορούμε να δούμε ξεκάθαρα τους ανακατασκευασμένους ελκυστές καθώς υπάρχουν περιοχές στις οποίες οι τροχιές αναπτύσσονται ασυμπτωτικά γύρω αυτών των ελκυστών. Κάθε ελκυστής υποδεικνύει επίσης την υποτρπιάζουσα υποκείμενη δυναμική κάθε φωνήματος που είναι στην πραγματικότητα πολύ παρόμοια από την τοπολογική άποψη αλλά εντελώς διαφορετική από την προοπτική των δυναμικών συστημάτων. Με άλλα



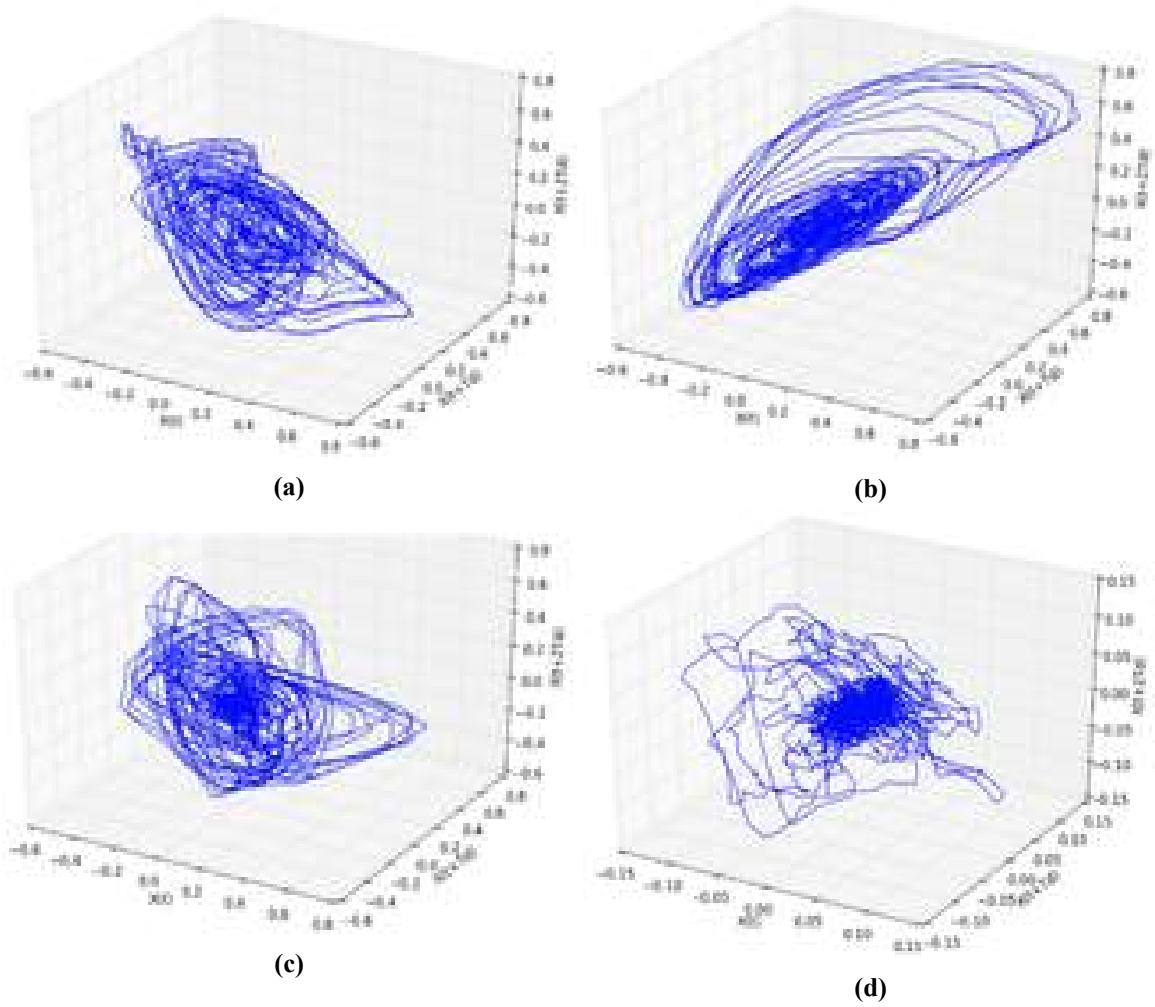
(a) Επιλεγμένη χρονική καθυστέρηση τ με βάση το πρώτο τοπικό ελάχιστο AMI



(b) Αυθαίρετη επιλογή χρονικής καθυστέρησης $\tau = 1$

Σχήμα 4.4: Ποσοστά λανθανόντων πλησιέστερων γειτόνων για διάφορες διαστάσεις εμβύθινσης και τμήματα ομιλίας ποικίλων χρονικών κλιμάκων

λόγια, μερικές τοπολογικές υπογραφές, όπως τα κενά που σχηματίζονται σε κάθε ελκυστήρα, μπορεί να παραμείνουν οι ίδιες μεταξύ όλων των ανακατασκευασμένων χώρων φάσης 4, αλλά αυτές οι δομές είναι αρκετά ανόμοιες από την άποψη του πώς εξελίσσονται οι τροχιές στο χρόνο. Για παράδειγμα, στο Σχήμα 4.5b οι καταστάσεις του PS εξελίσσονται παράλληλα μεταξύ τους δημιουργώντας



Σχήμα 4.5: Ανακατασκευασμένοι χώροι φάσης πλαισίων ομιλίας από φωνήεντα

μία δείνη στο κέντρο του ελκυστήρα που θα μπορούσε επίσης να συλληφθεί με μια προβολή σε ένα δισδιάστατο επίπεδο. Αντίθετα, η ανασυγκροτημένη δυναμική των Εικόνων 4.5a και 4.5c έχει καταστάσεις που εξελίσσονται παράλληλα, αλλά και κατακόρυφα παρουσιάζουν μια διαφορετική πλευρά της δυναμικής από το προηγούμενο παράδειγμα. Τέλος, μπορούμε να δούμε ότι το ανασυσταθέν PS του Σχήματος 4.5 έχει τροχιές που δεν εμφανίζουν αρμονικές ταλαντώσεις ή εξέλιξη. Πρόκειται για ενδεικτική μέτρηση ενός θορυβώδους δυναμικού συστήματος ή, στην χειρότερη περίπτωση, μιας ανεπιτυχούς αποδίπλωσης της δυναμικής του υποκείμενου συστήματος.

Και στις δύο περιπτώσεις, είναι προφανές ότι οι ιδιότητες επαναληψιμότητας όλων αυτών των σημείων που βρίσκονται στην προκύπτουσα πολλαπλότητα $\mathcal{M} \in \mathbf{R}^3$ είναι πολύ ενδεικτικές της φύσης κάθε δυναμικού συστήματος και μπορούν να αξιοποιηθούν για την εξόρυξη μη γραμμικών εξαρτήσεων για σήματα ομιλίας εν γένει, συνεπώς και για SER.

4.4 Γραφήματα επαναληψιμότητας (RPs) από τα πλαίσια ομιλίας

Αφού έχουμε αποκτήσει μια εικόνα για το πώς οι ανακατασκευασμένες τροχιές PS θα μοιάζουν στον ενσωματωμένο χώρο, θα πρέπει να εξετάσουμε τρόπους για την εξαγωγή χρήσιμων πληροφοριών σχετικά με αυτές, αξιοποιώντας τις αναδυόμενες δομές επαναληψιμότητας. Ένα από τα καταλληλότερα εργαλεία για την απεικόνιση των ιδιοτήτων επαναληψιμότητας που εκδηλώνει κάθε PS είναι τα Γραφήματα επαναληψιμότητας (RPs) [84]. Σε αυτή τη σύντομη ανάλυση του τρόπου εξαγωγής των RP από τα τοπικά ανακατασκευασμένα PS, θα αναλύσουμε τόσο τα συνεχή RPs και τα

δυαδικά-κατωφλιωμένα RPs (βλ. Ορισμό στο τμήμα 2.5).

Στην ουσία, τα συνεχόμενα διαγράμματα είναι ίσα με το μήτρα απόστασης κάτω από οποιονδήποτε καθορισμένο κανόνα του PS (π.χ. Ευκλείδια, Supremum, Μανχάταν). Η συνέχεια προκύπτει από το γεγονός ότι τα τελευταία διαγράμματα αποτελούνται από τιμές οι οποίες είναι συνεχείς με βάση την επιλεγμένη μετρική. Συγκεκριμένα, θα θέλαμε να ομαλοποιήσουμε την έξοδο αυτών των συνεχών RPs στην περιοχή $[0, 1]$. Για να το κάνουμε αυτό, μπορούμε να ομαλοποιήσουμε κάθε συνεχές RP με τη μέγιστη απόσταση από όλες τις αποστάσεις μεταξύ οποιωνδήποτε δύο σημείων τροχιάς $\mathbf{x}(i)$ και $\mathbf{x}(j)$. Τυπικά, σημειώνουμε το συνεχές RP με **CR** και το ορίζουμε να είναι ίσο με τον παρακάτω συμμετρικό πίνακα ως εξής:

$$\mathbf{CR}_{i,j} = \frac{1 - \mathbf{D}_q(\mathbf{x}(i), \mathbf{x}(j))}{\underset{1 \leq \hat{i}, \hat{j} \leq N-(d_e-1)\tau}{max} \{\mathbf{D}_q(\mathbf{x}(\hat{i}), \mathbf{x}(\hat{j}))\}} = \frac{1 - ||\mathbf{x}(i) - \mathbf{x}(j)||_q}{\underset{1 \leq \hat{i}, \hat{i} \leq N-(d_e-1)\tau}{max} \{||\mathbf{x}(\hat{i}) - \mathbf{x}(\hat{i})||_q\}} \tag{4.2}$$

όπου $|| \cdot ||_q$ είναι η νόρμα που χρησιμοποιείται για τον ορισμό της απόστασης μεταξύ οποιωνδήποτε δύο σημείων τροχιάς $\mathbf{x}(i)$ και $\mathbf{x}(j)$. Συγκεκριμένα, για τα $q = 1$, $q = 2$ ή $q = \infty$ υπολογίζουμε την νόρμα Μανχάταν, Ευκλείδεια ή Supremum, αντίστοιχα. Εάν θέλουμε να εξαγάγουμε τα δυαδικά RPs από τα προηγούμενα συνεχή RPs, απλά κατωφλιώνουμε αυτές τις τιμές ανάλογα με ένα από τα προαναφερθέντα κριτήρια για την επιλογή της τιμής κατωφλίου ϵ (δείτε Τμήμα 2.6 για μια εκτενέστερη συζήτηση σχετικά με διάφορες μεθόδους για τον καθορισμό του τρόπου καθώς και την τιμή του κατωφλίου).

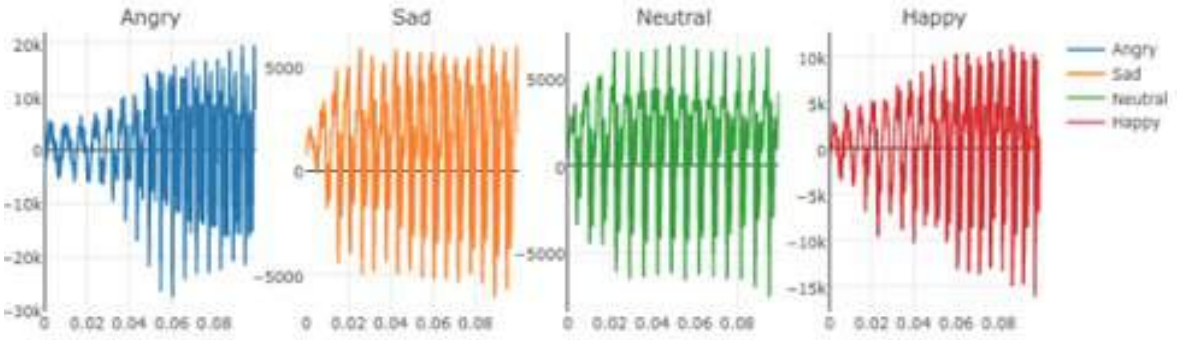
4.4.1 Από τα φωνήματα στα γραφήματα επαναληψιμότητας

Εστιάζουμε στην απεικόνιση των συνεχόμενων RPs για χρονικές κλίμακες πλαισίων ομιλίας που αντιστοιχούν σε φωνήματα και πιο συγκεκριμένα σε φωνήεντα που παρουσιάζουν μεγάλη δυναμική επαναληψιμότητας. Προκειμένου να αποκτήσουμε μια πιο διαισθητική εικόνα για το πώς αυτά τα σήματα ομιλίας αναπαρίστανται στον τομέα χρόνου και στις συνεχείς εικόνες RPs τους, πρέπει να τα συγκρίνουμε δίπλα-δίπλα.

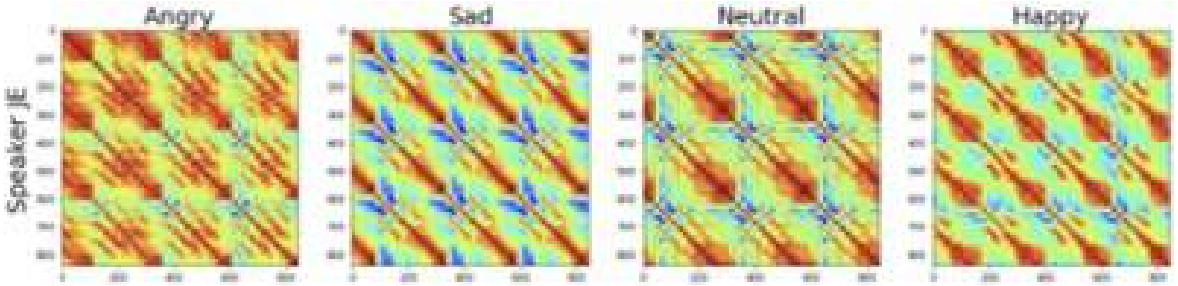
Στο Σχήμα 4.6, συγκρίνουμε τις αναπαραστάσεις της συλλαβής /αε/ τόσο για τις παραστάσεις του χρονικού τομέα όσο και για τις συνεχείς RPs αναπαπραστάσεις για δύο ομιλητές του συνόλου δεδομένων SAVEE [139] για τα διαφορετικά συναισθήματα 4, δηλαδή: *Θυμός*, *Θλίψη*, *Ουδετερότητα* και *Χαρά* που αντιστοιχούν στις αντίστοιχες στήλες το πλέγμα των εικόνων. Είναι προφανές ότι και για τους δύο ομιλητές η αναπαράσταση στον τομέα του χρόνου είναι λίγο πολύ ίδια. Αυτό ισχύει και για τις ομοιότητες των παραστάσεων του χρονικού πεδίου μεταξύ διαφορετικών συναισθημάτων.

Από την άλλη πλευρά, οι συνεχείς αναπαραστάσεις RP των σημάτων ομιλίας είναι αρκετά διαφορετικές και για τους δύο ομιλητές *JK* και *JE*. Παρατηρούμε μερικά ενδιαφέροντα μοτίβα στην υφή των εικόνων που προκύπτουν τα οποία είναι περιοδικά με ένα μοτίβο που επαναλαμβάνεται σε όλη τη διαγώνια γραμμή. Είναι σαφές ότι το μικρό υπόγραμμα αυτών των εικόνων που επαναλαμβάνεται σε όλη την κύρια διαγώνιο παρουσιάζει μια χωρική περιοδικότητα ίση με τη θεμελιώδη συχνότητα της φωνής (pitch). Αυτό είναι υποδειγματικό του κύριου προβλήματος των αναπαραστάσεων RP που παραμένουν αυστηρά συσχετισμένες με τις θεμελιώδεις συχνότητες που υπάρχουν στο σήμα ομιλίας. Μπορούμε επίσης να δούμε κάποιες άλλες δομές που εμφανίζονται σε αυτές τις περιοδικές εικόνες υπογράμματος και να επιδείξουν την ύπαρξη υποθεμελιωδών ταλαντώσεων φωνής που συχνά ονομάζεται διφωνία και είναι ένα μη γραμμικό φαινόμενο στην φωνητική οδό (βλέπε Ενότητα 1.4.2). Ωστόσο, εάν εστιάσουμε στις αναπαραστάσεις συνεχών RP για τους ομιλητές *JE* και *JK* ξεχωριστά, είναι ξεκάθαρο ότι οι εμφανιζόμενες δομές είναι διαφορετικές μεταξύ των διαφορετικών συναισθημάτων.

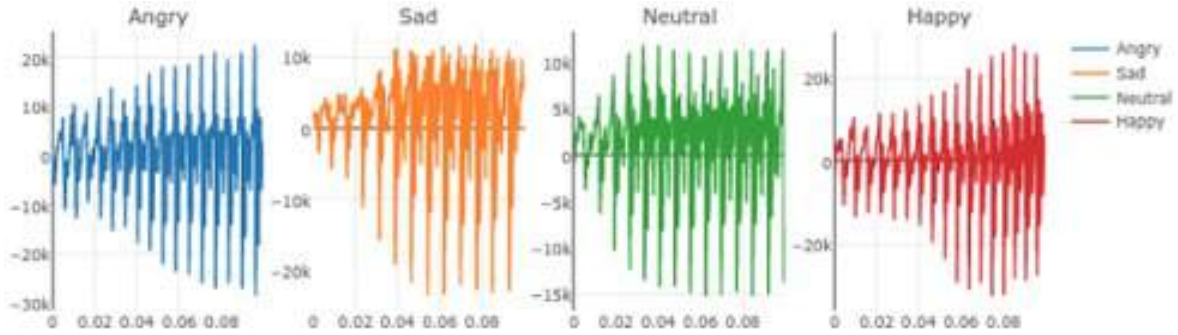
Η εμφανιζόμενη υποεικόνα που επαναλαμβάνεται περιοδικά σε όλη την εικόνα έχει δομή που συνδέεται στενά με την ταυτότητα του ομιλητή. Δηλαδή, για τον ομιλητή *JE* βλέπουμε ότι αυτή η υποδομή μοιάζει με μια μικρή περόνη, ενώ για τον άλλο ομιλητή *JK* είναι πολύ πιο παρόμοια με έναν κάκτο. Αν μπορούσαμε να βγάλουμε κάποια γενικά συμπεράσματα από αυτές τις απεικονίσεις, θα μπορούσαμε να δούμε ότι γενικά οι RP αναπαραστάσεις *Χαράς* και *Θυμού* και για τους δύο ομιλητές



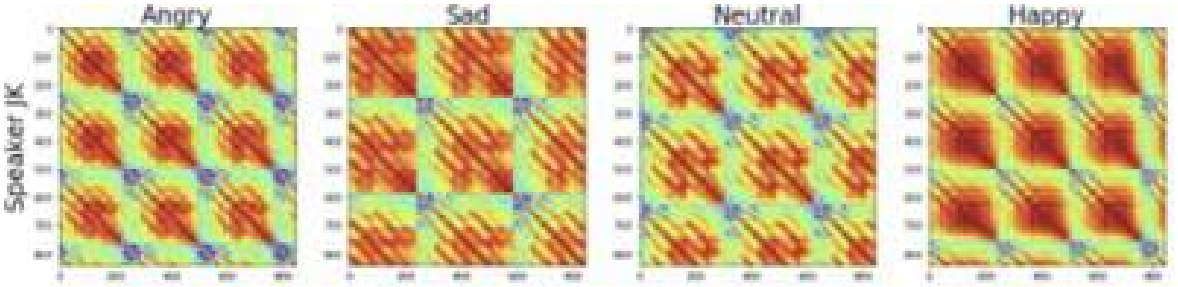
(a) /αε/ Φώνημα του ομιλητή *JE* στο πεδίο χρόνου για κάθε τύπο συναισθημάτων



(b) Συνεχή γραφήματα επαναληψιμότητας για τον ομιλητή *JE* για κάθε τύπο συναισθημάτων



(c) /αε/ Φώνημα του ομιλητή *JK* στο πεδίο χρόνου για κάθε τύπο συναισθημάτων



(d) Συνεχή γραφήματα επαναληψιμότητας για τον ομιλητή *JK* για κάθε τύπο συναισθημάτων

Σχήμα 4.6: Φώνημα /αε/ στο πεδίο χρόνου και το αντίστοιχο γράφημα επαναληψιμότητας για διαφορετικούς ομιλητές και τις συναισθηματικές τους εκφράσεις

φαίνεται να είναι αρκετά πιο πυκνές από τα άλλα συναισθήματα για τα οποία εκπέμπεται το φώνημα /αε/. Αυτό θα μπορούσε να είναι ενδεικτικό των ιδιοτήτων επαναληψιμότητας του τρόπου με τον οποίο το φώνημα αναπαρίσταται σε διαφορετικές συναισθηματικές συνθήκες και ως εκ τούτου ένα κλειδί για την κατασκευή ενός καλύτερου συστήματος SER χρησιμοποιώντας αυτές τις πληροφορίες.

4.4.2 Η επίδραση των παραμέτρων PS στην εξαγωγή των RPs

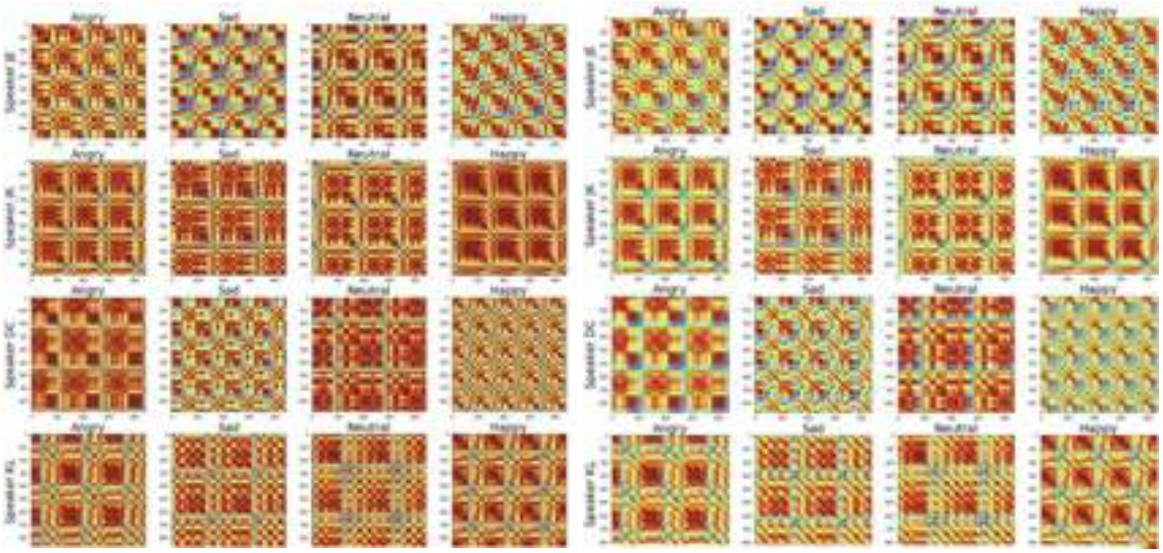
Θα πρέπει επίσης να αναλύσουμε την επίδραση των παραμέτρων του ανακατασκευασμένου PS που θα μπορούσαν επίσης να αντικατοπτριστούν στην εξαγωγή των αντίστοιχων RPs. Επειδή η διαδικασία εκτελείται διαδοχικά για ένα δεδομένο σήμα, πρέπει να είμαστε προσεκτικοί για το πώς οι παράμετροι κάθε μπλοκ επεξεργασίας δημιουργούν παραπλανητικές αναπαραστάσεις των επόμενων μπλοκς επεξεργασίας. Σε μια προηγούμενη μελέτη [42], δείχθηκε ότι η διάσταση εμβύθινσης d_e δεν είναι μια κρίσιμη παράμετρος για τις αναπαραστάσεις RP και ειδικά για τις δυαδικές RP αναπαραστάσεις. Πιθανότατα, αυτό συμβαίνει επειδή οι αναπαραστάσεις RP αποτυπώνουν τις ιδιότητες επαναληψιμότητας των υπό εξέταση χρονικών σειρών χωρίς να απαιτούν ακόμη την ανακατασκευή του PS. Ωστόσο, η εκτόνωση της δυναμικής της χρονοσειράς με τις υψηλότερες διαστάσεις αποπροσανατολίζει τα πρότυπα επαναληψιμότητας που εμφανίζονται σε RPs κυρίως επειδή τα περιοδικά τμήματα των τροχιών θα εμφανίζονταν πιο συσχετισμένα σε σύγκριση με τα μη συσχετισμένα μέρη των τροχιών εάν η εικόνα κανονικοποιηθεί. Για παράδειγμα, αν έχουμε αριθμούς d_e που συμβαίνουν περίπου με την ίδια περιοδικότητα με την χρονική καθυστέρηση τ, την οποία έχουμε ήδη επιλέξει για την ανασυγκρότηση του PS, τότε οι αντίστοιχες καταχωρήσεις του συνεχούς RP θα ήταν πολύ πιο κοντά στο ένα απότι πριν. Αυτή η ενίσχυση των άκρων της εικόνας γύρω από τις εγγραφές με υψηλότερη περιοδικότητα ενισχύεται περαιτέρω επειδή άλλα σημεία που γενικά δεν συσχετίζονται μεταξύ τους, η εκδήλωση της δυναμικής από την ανασυγκρότηση του PS ωθεί αυτές τις διαφορές να είναι πολύ πιο εμφανείς. Ωστόσο, η παράμετρος χρονικής καθυστέρησης των δειγμάτων τ μεταξύ των χρονικά υστερημένων αντιγράφων του σήματος που χρησιμοποιούνται για την ενσωμάτωση της δυναμικής παίζουν σημαντικό ρόλο στην τελική συνεχή αναπαράσταση RP.

Στο Σχήμα 4.7, χρησιμοποιούμε όλους τους ομιλητές από τη βάση δεδομένων SAVEE [139] για την ίδια φράση και κατατάσσουμε τη φράση για να αποκτήσουμε τα ευθυγραμμισμένα πλαίσια για τη διέγερση του φωνήματος /αε/ κάτω από διαφορετικές συναισθηματικές συνθήκες. Και πάλι, οι επιλεγμένες συναισθηματικές ετικέτες είναι οι ίδιες με αυτές που χρησιμοποιήσαμε στην προηγούμενη ανάλυση. Όλες οι συνεχείς αναπαραστάσεις RP εξάγονται χρησιμοποιώντας τις τροχιές ενός διακριτού PS 3 του πλαισίου εισόδου που αντιστοιχεί στο φωνήμα /αε/. Εάν δεν χρησιμοποιήσουμε χρονική καθυστέρηση (βλ. Εικόνα 4.7a) για να ενσωματώσουμε τη δυναμική, τότε οι αναπαραστάσεις RP φαίνεται να συσχετίζονται πολύ περισσότερο μεταξύ τους για όλα τα σημεία, παρόλο που πραγματοποιούμε κανονικοποίηση με τη μεγαλύτερη απόσταση αυτών των τοπικών ζευγών απόστασης. Επιπλέον, βλέπουμε ότι καθώς αυξάνουμε την επιλεγμένη παράμετρο χρονικής καθυστέρησης tau βλέπουμε ότι όλα τα σημεία γίνονται πολύ πιο ασυσέτιστα έως ότου μόνο τα πιο συσχετισμένα μοτίβα είναι κοντά σε ένα (καφέ περιοχές στο σχήμα 4.7d) με όλες τις άλλες να είναι πολύ πλησιέστερα στο μηδέν (μπλε περιοχές στο σχήμα 4.7d) σε σχέση με τα άλλα σχήματα 4.7a, 4.7b και 4.7c. Πιθανώς αυτή είναι η ένδειξη ότι η πραγματική δυναμική του σήματος εισόδου εξακολουθεί να καταρρέει χωρίς την κατάλληλη ενσωμάτωση. Αν αυξήσουμε περαιτέρω την καθυστέρηση, τότε περιμένουμε να δούμε μόνο στην κύρια διαγώνιο που είναι η αντίθετη περίπτωση. Ως εκ τούτου, η παράμετρος χρονικής καθυστέρησης έχει σημαντικό αντίκτυπο στον τρόπο εμφάνισης του προκύπτοντος RP.

4.4.3 Αναδυόμενες χαοτικές δομές στα RPs των φωνημάτων;

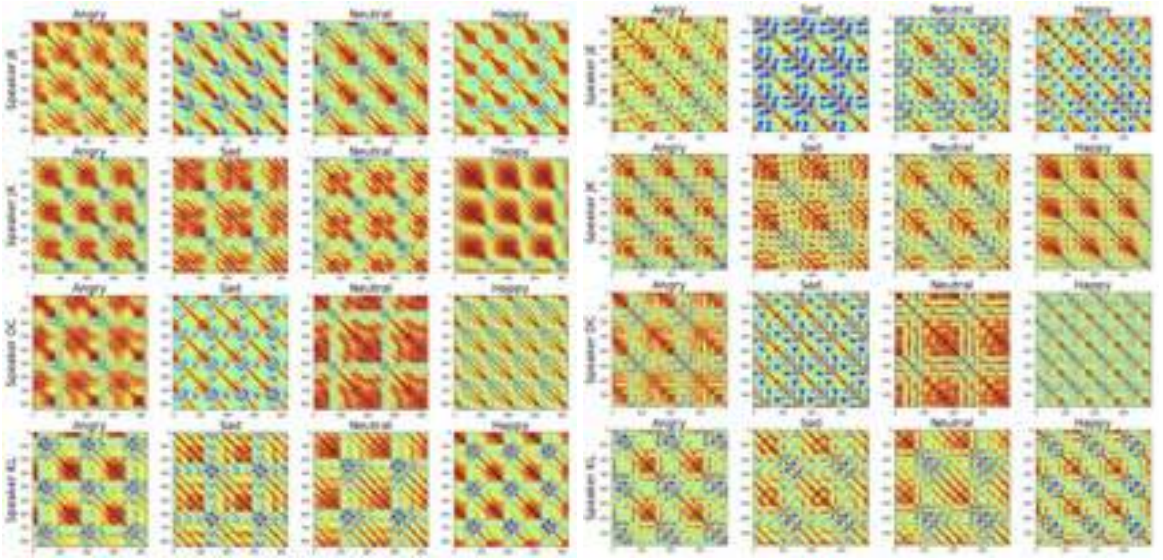
Όπως αναφέρθηκε προηγουμένως, οι υπογραφές επαναληψιμότητας των χρονοσειρών είναι σημαντικές για την κατανόηση της πραγματικής υποκείμενης δυναμικής και ταυτότητας αυτών των σημάτων. Όπως αναφέρθηκε προηγουμένως στο τμήμα 2.5, αυτές οι κατηγορίες συστημάτων θα μπορούσαν να χωριστούν σε αυτορυθμιζόμενες διαδικασίες, τυχαίους θορύβους, περιοδική και χαοτική δυναμική. Έτσι λοιπόν το ερώτημα είναι πώς λειτουργεί η συνολική διαδικασία της εξαγωγής RP και εάν μπορούμε πραγματικά να αποκτήσουμε πληροφορίες από αυτά τα πρότυπα επαναληψιμότητας σχετικά με τη δυναμική του πραγματικού συστήματος $\mathbf{s}^\star$;

Στο σχήμα 4.8, εμφανίζεται η συνολική διαδικασία εξαγωγής ενός δυαδικού RP καθώς και μια σύγκριση δύο RPS που εξάγονται από διαφορετικά συστήματα. Συγκεκριμένα, κατατάσσουμε μια συναισθηματική φράση *angry* και απομονώσαμε ένα πλαίσιο ομιλίας 30ms όπως φαίνεται στο Σχήμα



(a) Γραφήματα Επαναληψιμότητας χρησιμοποιώντας $\tau = 0$

(b) Γραφήματα Επαναληψιμότητας χρησιμοποιώντας $\tau = 7$



(c) Γραφήματα Επαναληψιμότητας χρησιμοποιώντας $\tau = 25$

(d) Γραφήματα Επαναληψιμότητας χρησιμοποιώντας $\tau = 50$

Σχήμα 4.7: Εξαγωγή συνεχόμενων RP για το φώνημα /αε. για όλους τους ομιλητές της βάσης δεδομένων SAVEE και τις συναισθηματικές τους εκδηλώσεις

4.8a που περιλαμβάνεται μέσα στο χρονικό διάστημα της διέγερσης του φωνήματος /ε/. Μετά από αυτό, ανασυνθέσαμε το PS του δεδομένου πλαισίου όπως εξηγήθηκε στις προηγούμενες ενότητες 2.5 και 4.3 με εκτίμηση των παραμέτρων της χρονικής καθυστέρησης τ και της διάστασης ενσωμάτωσης d_e. Επιπλέον, εφαρμόζουμε την εξίσωση των δυαδικών RPs (βλ. Εξίσωση 2.36) επιλέγοντας αυθέραιτα τιμή κατωφλίου $\epsilon = 0.15$ για να παραχθεί η τελική δυαδική εικόνα. Αυτή η διαδικασία μέχρι εδώ παρουσιάζει μια πλήρη απεικόνιση ολόκληρης της διαδοχικής διαδικασίας της εξαγωγής μη γραμμικών χαρακτηριστικών που εισάγεται σε αυτή την Ενότητα.

Έχουμε επίσης προσομοιώσει ένα γνωστό σύστημα για τη σύνθετη δυναμική που εμφανίζει κάτω από συγκεκριμένες παραμετροποιήσεις και δημιουργήσαμε τη δυαδική αναπαράσταση RP χρησιμοποιώντας την εξέλιξη του χρόνου των καταστάσεων των μεταβλητών του συστήματος. Χρησιμοποιήσαμε το σύστημα Lorenz96 [140] χρησιμοποιώντας $N_{Lor} = 50$ μεταβλητές για μοντελοποίησή του με αρχική θέση μηδέν και δύναμη μονάδων $F = 5$ που εφαρμόζεται σε μια καθορισμένη θέση όταν το

σύστημα βρίσκεται σε ισορροπία. Εάν αφήσουμε τη δυναμική του συστήματος να εξελιχθεί μέσα στο χρόνο, αναμένουμε να δούμε μια χαοτική συμπεριφορά μετά από μερικές μονάδες διακριτού χρόνου. Το δυναμικό μοντέλο του Lorenz96 διαμορφώνεται χρησιμοποιώντας τις εξισώσεις που προκύπτουν από την ακόλουθη εξίσωση:

$$\frac{dz(i)}{dt} = [z(i+1) - z(i-2)]z(i-1) - z(i) + F \tag{4.3}$$

Αν θέλουμε να προσομοιώσουμε σε έναν υπολογιστή την παραπάνω εξίσωση, πρέπει να μετατρέψουμε το χρονικό παράγωγο σε διακριτή διαφορά διαδοχικής εισαγωγής ενός δείκτη σχετικά με τον αριθμό των μονάδων χρόνου που κάθε μεταβλητή υπάρχει. Συγκεκριμένα, εάν βρισκόμαστε στον χρόνο t του συστήματος, το διάνυσμα των μεταβλητών θα αντιπροσωπευόταν ως $\mathbf{z}_t$. Επιπλέον, η διακριτοποιημένη εξίσωση θα ορίζεται παρομοίως από την ακόλουθη έκφραση:

$$z_t(i) = [z_{t-1}(i+1) - z_{t-1}(i-2)]z_{t-1}(i-1) + F \tag{4.4}$$

όπου διατηρούνται οι ακόλουθες αρχικές συνθήκες

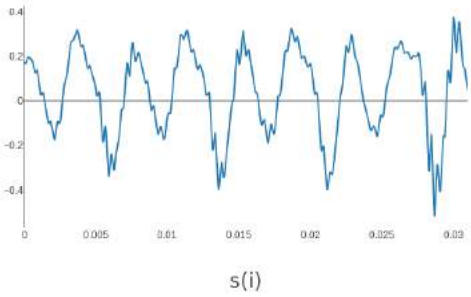
$$z(i) = z(i \mod N_{Lor}), \quad i \in \mathbb{N} \cap [0, N_{Lor}] \tag{4.5}$$

όπου $a \mod b$ ορίζει το υπόλοιπο της ακέραιας διαίρεσης μεταξύ των αριθμών a και b. Τέλος, το αντίστοιχο δυαδικό RP με ad-hoc κατώφλι $\epsilon = 0.1$ εμφανίζεται στο Σχήμα 4.8d. Όπως έχει δειχθεί, οι ιδιότητες επαναληψιμότητας αυτής της δυαδικής εικόνας εμφανίζουν μικρές διακεκομμένες διαγώνιες παράλληλες προς την κύρια διαγώνιο που είναι ενδεικτική της ύπαρξης χαοτικών δυναμικών που διέπεται το σύστημα Lorenz96 [140].

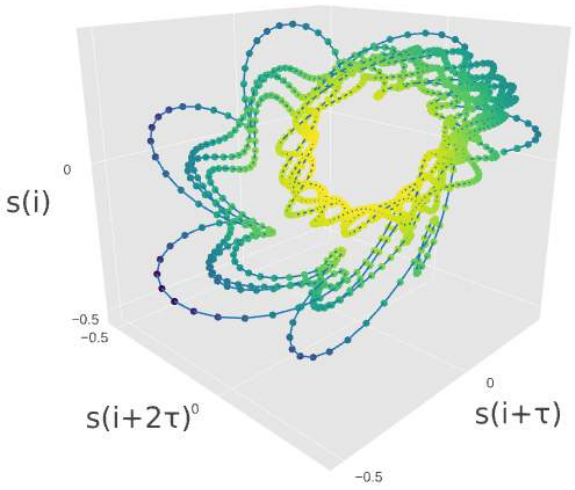
Συνολικά, οι αναδυόμενες δομές μικρής κλίμακας που βασίζονται σε γραμμές ή σε μηδενικά, αντανακλούν τη δυναμική συμπεριφορά του συστήματος. Για παράδειγμα, οι διαγώνιες γραμμές υποδεικνύουν τόσο παρόμοια εξέλιξη καταστάσεων για διάφορα τμήματα της τροχιάς του PS όσο και ντετερμινιστική χαοτική δυναμική του συστήματος [84]. Αυτό απεικονίζεται επίσης στα σχήματα 4.8c και 4.8d κατά τη σύγκρισή τους δίπλα-δίπλα. Είναι προφανές ότι και τα δύο συστήματα έχουν διαγώνιες παράλληλες προς την κύρια διαγώνιο με βασική συχνότητα που είναι εμφανής και για τα δύο συστήματα. Επιπλέον, αυτές οι διαλείπουσες διαγώνιες γραμμές και στα δύο συστήματα είναι ενδείξεις εξαιρετικά μη γραμμικών φαινομένων στις ταλαντώσεις και των δύο συστημάτων. Έτσι, οι αναπαραστάσεις RP αποκαλύπτουν κοινές πτυχές για τη δυναμική των διαφορετικών συστημάτων χρησιμοποιώντας μια παρόμοια απεικόνιση στη δυαδική εικόνα. Με άλλα λόγια, τα διαφορετικά δυναμικά συστήματα εμφανίζουν παρόμοια συμπεριφορά όσον αφορά τις ιδιότητες δυναμικής και επαναληψιμότητας που εμφανίζονται εύστοχα σε αναπαραστάσεις RP. Στην περίπτωση αυτή, η παραγωγή της ανθρώπινης ομιλίας υπό μια "θυμωμένη" συναισθηματική διέγερση φαίνεται να παρμοιάζει την δυναμική του θεωρητικού συστήματος Lorenz96 που εμφανίζει χαοτική φύση στη δυναμική του.

4.5 Εξαγωγή Ακουστικών Χαρακτηριστικών

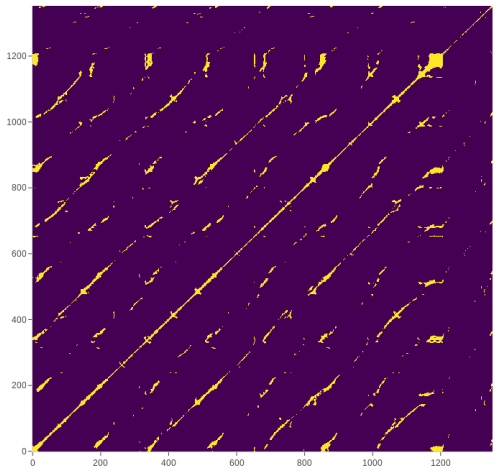
Ομοίως με το προηγούμενο Κεφάλαιο 3 χρησιμοποιούμε ακουστικά σύνολα χαρακτηριστικών προκειμένου να εκτελέσουμε SER. Εξάγουμε τα ακουστικά χαρακτηριστικά για δύο μεθόδους διαφορετικών χρονικών κλιμάκων, συγκεκριμένα: σε επίπεδο τμήματος ομιλίας και ολόκληρης της ομιλίας. Και οι δύο μέθοδοι ταξινόμησης είναι πανομοιότυπες με αυτές που περιγράψαμε στις προηγούμενες ενότητες 3.3.2 και 3.3.3 όσον αφορά την εξαγωγή χαρακτηριστικών, αλλά σε αυτή τη ρύθμιση χρησιμοποιούμε διαφορετικά μοντέλα ταξινόμησης. Εισάγουμε το νέο σύνολο μη γραμμικών χαρακτηριστικών (σύνολο χαρακτηριστικών RQA) με βάση τα μέτρα RQA που περιγράφονται στην Ενότητα 4.5.2. Ως σύνολο βασικών χαρακτηριστικών εξάγουμε τα ίδια χαρακτηριστικά όπως περιγράφονται στην Ενότητα 3.2.4 και το ονομάζουμε "IS10" σύνολο [23]. Τέλος, το συνεπτυγμένο σύνολο χαρακτηριστικών περιγράφεται στην Ενότητα 4.5.3 και περιλαμβάνει τις λειτουργίες και από τα δύο προαναφερθέντα σύνολα χαρακτηριστικών.



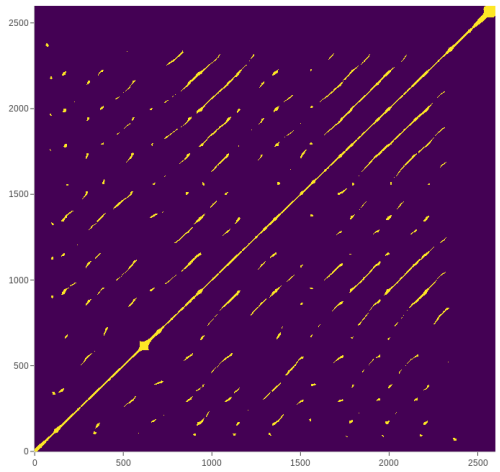
(a) Αναπαράσταση στον χρόνο του πλαισίου ομιλίας



(b) Ανακατασευασμένος χώρος φάσης ($m = 3, \tau = 7$)



(c) Δυαδικό Γράφημα Επαναληψιμότητας χρησιμοποιώντας ένα κατώφλι $\epsilon = 0.15$ και νόρμα Μανχάταν για τη μέτρηση της απόστασης μεταξύ των σημείων στο PS



(d) Δυαδικό Γράφημα Επαναληψιμότητας του συστήματος Lorenz96 που εμφανίζει χαοτική συμπεριφορά

Σχήμα 4.8: RP ενός πλαισίου 30ms που περιλαμβάνεται στη εκφώνηση του φωνήματος /ε/ και κατανόηση της υποκείμενης δυναμικής με σύγκριση

4.5.1 Βασικό σύνολο χαρακτηριστικών αναφοράς (IS10 Σύνολο)

Χρησιμοποιούμε το σύνολο χαρακτηριστικών IS10 [23], στο οποίο εξάγονται 1582 στατιστικά χαρακτηριστικά που αντιστοιχούν που εφαρμόζονται σε διάφορους τοπικούς περιγραφητές LLD ακουστικών πλαισίων. Η εξαγωγή γίνεται για προσεγγίσεις που βασίζονται σε τμήματα και ολόκληρες ομιλίες χρησιμοποιώντας το openSMILE προγραμματιστικό περιβάλλον [119]. Μια περαιτέρω ανάλυση για αυτό το σετ χαρακτηριστικών παρέχεται σε μια προηγούμενη ενότητα 3.2.4.

4.5.2 Προτεινόμενο σύνολο μη γραμμικών χαρακτηριστικών (σύνολο RQA)

Το προτεινόμενο σύνολο χαρακτηριστικών RQA που έχει οριστεί για ένα δεδομένο τμήμα ομιλίας ή ολόκληρη την ομιλία και εξάγεται όπως περιγράφεται στη συνέχεια.

Αρχικά, σπάμε το δεδομένο σήμα ομιλίας σε αλληλεπικαλυπτόμενα πλαίσια με αναλογία επικα-

λύψεως $OL = 0.5$. Για κάθε πλαίσιο ανασυνθέτουμε την τροχιά του PS όπως ορίζεται στην Εξίσωση 2.22. Επιλέγουμε για την ανακατασκευή του κάθε PS εκτιμούμε την παράμετρο χρονικής καθυστέρησης τ και τη διάσταση ενσωμάτωσης d_e όπως περιγράφεται στα τμήματα 2.4.2 και 2.4.4, αντίστοιχα.

Στη συνέχεια, για κάθε τροχιά PS, το αντίστοιχο RP υπολογίζεται ως δυαδική εικόνα από τις αποστάσεις μεταξύ των σημείων της τροχιάς όπως εξηγείται στο τμήμα 2.5 με ένα επιβαλώμενο κατώφλι. Προκειμένου να επιλέξουμε την τιμή κατωφλίου ϵ για να δημιουργήσουμε τη δυαδική εικόνα, χρησιμοποιούμε ένα από τα παρακάτω κριτήρια με βάση: 1) σταθερή αυθαίρετη τιμή κατωφλίου, 2) σταθερό ρυθμό επαναληψιμότητας (RR) ή 3) αναλογία της τυπικής απόκλισης σ των σημείων στην τροχιά PS (μια πιο μακροσκελή εξήγηση των κριτηρίων δόθηκε στο τμήμα 2.5 αλλά μπορεί να βρεθεί και στη βιβλιογραφία [84]).

Για να ποσοτικοποιηθούν οι σύνθετες δομές του RP, εξάγεται ένας κατάλογος των μέτρων RQA για να δημιουργηθεί μια αναπαράσταση RQA στατικού μήκους για κάθε πλαίσιο. Τα επιλεγμένα μέτρα RQA παρουσιάζονται στον πίνακα 4.1 ενώ η αντίστοιχη ποιοτική ανάλυση των μέτρων αυτών έχει ήδη δοθεί στο τμήμα 2.6.1. Τώρα, κάθε πλαίσιο ομιλίας αντιπροσωπεύεται από διανυσμα 12 χαρακτηριστικών που αντιστοιχεί στις τιμές των προαναφερθέντων μέτρων RQA.

Πίνακας 4.1: Μέτρα RQA που εξάγονται από κάθε RP

Name	Formulation
Λόγος Επαναληψιμότητας (RR)	$\frac{1}{N^2}\sum_{i,j=1}^{N}\mathbf{R}_{i,j}$
Ντετερμινισμός (DET)	$\frac{\sum_{l=d_m}^{N} lP_d(l)}{\sum_{l=d_m}^{N} lP_d(l)}$
Μέγιστο μήκος Διαγώνιας Γραμμής ($\mathbf{L_{max}}$)	$max(\{l_i\}_{i=1}^{N_d})$
Μέσο μήκος Διαγώνιας Γραμμής (L)	$\frac{\sum_{l=d_m}^{N} lP_d(l)}{\sum_{l=d_m}^{N} P_d(l)}$
Διαγώνια Εντροπία (DENTR)	$\sum_{l=d_m}^{N} \frac{P_d(l)}{N_d} ln(\frac{N_d}{P_d(l)})$
Λαμιναρότητα (LAM)	$\frac{\sum_{l=v_m}^{N} lP_v(l)}{\sum_{l=v_m}^{N} lP_v(l)}$
Μέγιστο μήκος καθέτου γραμμής ($\mathbf{V_{max}}$)	$max(\{l_i\}_{i=1}^{N_v})$
Χρόνος παγίδευσης (TT)	$\frac{\sum_{l=v_m}^{N} lP_v(l)}{\sum_{l=v_m}^{N} P_v(l)}$
Εντροπία κατακόρυφων γραμμών (VENTR)	$\sum_{l=v_m}^{N} \frac{P_v(l)}{N_v} ln(\frac{N_v}{P_v(l)})$
Μέγιστο μήκος λευκής καθέτου γραμμής ($\mathbf{W_{max}}$)	$max(\{l_i\}_{i=1}^{N_w})$
Μέσο Μήκος Λευκών Κάθετων Γραμμών (AWVL)	$\frac{\sum_{l=w_m}^{N} lP_w(l)}{\sum_{l=w_m}^{N} P_w(l)}$
Λευκή κατακόρυφη εντροπία (WENTR)	$\sum_{l=w_m}^{N} \frac{P_w(l)}{N_w} ln(\frac{N_w}{P_w(l)})$

Οι διανυσματικές αναπαραστάσεις για τμήματα ομιλίας και εκφράσεις που περιέχουν πολλαπλά πλαίσια γίνονται με την εφαρμογή ενός συνόλου 18 στατιστικών πάνω από όλα τα μέτρα RQA που εξάγονται σε επίπεδο πλαισίου. Πιό συγκεκριμένα αυτά τα στατιστικά υπολογίζονται για κάθε ένα από τα 12 χαρακτηριστικά των πλαισίων ομιλίας καθώς και οι διακριτές τους παράγωγοι από πλαίσιο σε πλαίσιο. Τα επιλεγμένα στατιστικά που χρησιμοποιούνται παρουσιάζονται στον Πίνακα 4.2. Έτσι, ένας φορέας χαρακτηριστικών $432 = 18 \cdot 24$ λαμβάνεται για κάθε τμήμα ομιλίας ή ολόκληρη την ομιλία.

4.5.3 Συντηγμένο σύνολο λειτουργιών (σύνολο χαρακτηριστικών RQA + IS10)

Επειδή τα δύο σύνολα χαρακτηριστικών (RQA, IS10) εξάγονται για οποιαδήποτε δεδομένη διάρκεια ενός τμήματος ομιλίας ή ολόκληρης της έκφρασης, μπορούν να συνδυαστούν χρησιμοποιώντας

Πίνακας 4.2: Σύνολα στατιστικών λειτουργιών για το σύνολο χαρακτηριστικών RQA

Στατιστικά
ελάχιστο
μέγιστο
αριθμητικός μέσος
μέσος
συνδιακύμανση
μέτρα λοξότητας
κύρτωση
εύρος
$1, 5, 25, 50, 75, 95, 99$ ποσοστά
$25-50$, $50-75$ και $25-75$ εύρη τεταρτημορίων

μια απλή συνένωση των δύο στατιστικών αναπαραστάσεων. Συγκεκριμένα, όταν αναφερόμαστε στο συντηγμένο σύνολο χαρακτηριστικών γνωρισμάτων (RQA + IS10 Σύνολο χαρακτηριστικών), θα αναφερόμαστε σε ένα διανυσμα με 2014 χαρακτηριστικά που έχει δημιουργηθεί από την συνένωση των χαρακτηριστικών IS10 (1582 χαρακτηριστικά) και RQA (432 χαρακτηριστικά).

4.6 Μέθοδοι ταξινόμησης

Προκειμένου να εκτιμηθούν οι επιδόσεις του προτεινομένου συνόλου χαρακτηριστικών RQA και IS10 καθώς και η σύντηξη τους, δοκιμάζουμε προσεγγίσεις σε ποικίλες χρονικές κλίμακες. Συγκεκριμένα, ερευνούμε τόσο SER που βασίζεται σε ολοκληρη την ομιλία όσο και σε τμήματα, όπως περιγράφεται παρακάτω.

4.6.1 Βασιζόμενη σε ολόκληρη την ομιλία

Για κάθε ομιλία λαμβάνουμε τη στατιστική αναπαράστασή του εξαγομένου αντίστοιχου σετ χαρακτηριστικών όπως περιγράφεται στην Ενότητα 4.5. Για την τελική ταξινόμηση συναισθημάτων χρησιμοποιούμε ένα SVM με RBF πυρήνα και έναν ταξινομητή LR. Μια εκτεταμένη ανάλυση καθώς και η μαθηματική διατύπωση και των δύο αυτών μοντέλων μπορούν να βρεθούν στο τμήμα 2.2.2 για το SVM και στην ενότητα 2.2.3 για τον ταξινομητή LR. Ο συντελεστής κόστους C βρίσκεται στο διάστημα $[0.001, 30]$ και για τα δύο μοντέλα SVM και LR που είναι η μόνη υπερπαράμετρος που πρέπει να ρυθμιστεί. Και τα δύο μοντέλα υλοποιούνται χρησιμοποιώντας την βιβλιοθήκη scikit [141].

4.6.2 Βασισμένο σε τμήματα ομιλίας

Διαχωρίζουμε κάθε έκφραση σε τμήματα διάρκειας 1.0 δευτερολέπτου και βήμα 0.5 δευτερολέπτων, σύμφωνα με την προσέγγιση βάσει τμήματος ομιλίας του προηγούμενου τμήματος 3.3.2. Για κάθε τμήμα ομιλίας εξάγουμε τα σύνολα χαρακτηριστικών που περιγράφονται στην ενότητα 4.5 και ως αποτέλεσμα κάθε έκφραση αντιπροσωπεύεται τώρα από μια ακολουθία διανυσματων με στατιστικά που αντιστοιχούν σε διαφορετικά χρονικά βήματα. Αυτή η ακολουθία τροφοδοτείται ως είσοδος σε ένα LSTM για την ταξινόμηση των συναισθημάτων. Μια εκτενής περιγραφή ενός μοντέλου LSTM μπορεί να βρεθεί στο 2.3.1. Το SER μπορεί να διατυπωθεί ως μια αλληλουχία μάθησης πολλαπλών προς μία, όπου η αναμενόμενη έξοδος κάθε ακολουθίας χαρακτηριστικών τμήματος είναι μια συναισθηματική ετικέτα προερχόμενη από τις ενεργοποιήσεις του τελευταίου κρυμμένου στρώματος [66].

Συγκεκριμένα, χρησιμοποιούμε μια αμφίδρομη αρχιτεκτονική LSTM (A-BLSTM) με μηχανισμό προσοχής στην κορυφή [65] όπου η απόφαση για τη συναισθηματική ετικέτα προέρχεται από μια

σταθμισμένη συνάθροιση των ενεργοποιήσεων του τελευταίου στρώματος όλων των χρονικών βημάτων. Μια εκτεταμένη ανάλυση αυτού του μοντέλου μπορεί να βρεθεί στα τμήματα 2.3.4. Επιπλέον, η τοπολογία δικτύου ενός αμφίδρομου LSTM περιγράφεται στην ενότητα 2.3.2 καθώς και ο τρόπος με τον οποίο λειτουργεί ο μηχανισμός προσοχής όταν εφαρμόζεται πάνω από μια αρχιτεκτονική RNN που αναλύθηκε στο τμήμα 2.3.3. Εφαρμόζουμε αυτήν την αρχιτεκτονική στην βιβλιοθήκη pytorch [142]. Επιπλέον, ο χώρος αναζήτησης των βελτίστων υπερπαραμέτρων αποτελείται από τις ακόλουθες παραμέτρους:

- αριθμός στρωμάτων $\{1, 2\}$
- αριθμός κρυφών κόμβων $\{128, 256\}$
- ποσοστό θορύβου εισόδου $[0.3, 0.8]$
- ποσοστό διαγραφής βαρών $[0.3, 0.8]$
- ρυθμός μάθησης $[0.0002, 0.002]$.

4.7 Πειράματα

Αξιολογούμε το προτεινόμενο σύνολο χαρακτηριστικών μας σε τρεις διαφορετικές πειραματικές διατάξεις για SER που περιγράφονται στη συνέχεια. Συγκρίνουμε επίσης τα αποτελέσματά μας με τις πιο σχετικές πειραματικές ρυθμίσεις που αναφέρονται στη βιβλιογραφία. Για όλες τις εργασίες, αναφέρουμε: σταθμισμένη ακρίβεια (WA) και μη σταθμισμένη ακρίβεια (UA) που περιγράφονται στην προηγούμενη ενότητα 3.4.2.

Δοκιμάζουμε ποικίλες διάρκειες πλαισίου $\{20, 30, 50\}$ ms από τις οποίες υπολογίζονται τα RPs. Για τον υπολογισμό του RP, εξετάζουμε τα σύνολα παραμέτρων που περιγράφονται λεπτομερώς στην Ενότητα 2.5. Συγκεκριμένα, δοκιμάζουμε τις νόρμες Μανχάταν, την Ευκλείδεια και την Supremum, καθώς και πολλαπλά κριτήρια επιλογής για την τιμή κατωφλίου ϵ, ανάλογα με τη ρύθμιση κατωφλίου αυθαίρετα, το σταθερό ποσοστό επαναληψιμότητας και το σταθερό σ. Η αντίστοιχη παράμετρος για τα τρία προαναφερθέντα κριτήρια βρίσκεται στο $[0.05, 0.5]$.

Μετά από μια εκτεταμένη μελέτη των παραμέτρων διαμόρφωσης των χαρακτηριστικών RQA, καταλήγουμε στο συμπέρασμα ότι τα καλύτερα αποτελέσματα για SER αποκτώνται χρησιμοποιώντας μια διάρκεια πλαισίου 20 ms για την εξαγωγή RPs. Επιπλέον, οι παράμετροι με τις καλύτερες επιδόσεις για τη διαμόρφωση RP φαίνονται να είναι η νόρμα Μανχάταν με μια ρύθμιση κατωφλίου που εξαρτάται από ένα σταθερό ποσοστό επαναληψιμότητας που βρίσκεται στο $[0.1, 0.2]$.

4.7.1 Δεδομένα

Οι παρακάτω βάσεις δεδομένων χρησιμοποιούνται στα πειράματά μας:

1. **SAVEE:** Η βάση δεδομένων Surrey Audio-Visual Express (SAVEE) αποτελείται από συναισθηματική ομιλία που εκφράζεται από άντρες ηθοποιούς 4. Το SAVEE περιλαμβάνει 480 εκφράσεις (120 εκφράσεις ανά ηθοποιό) συναισθημάτων 7 δηλ. 60 θυμός, 60 αηδία, 60 φόβος, 60 ευτυχία, 60 θλίψη, 60 έκπληξη και 120 ουδετερότητα.

2. **Emo-DB:** Βάση δεδομένων της συναισθηματικής ομιλίας του Βερολίνου (EmoDB) [121] περιέχει 535 συναισθηματικές φράσεις στα γερμανικά, που εκφράζονται από τους φορείς 10 (5 αρσενικά και 5 θηλυκά). Συγκεκριμένα συμπεριλαμβάνονται συναισθήματα 7, δηλ. Οργή 127, αποστροφή 45, φόβος 70, χαρά 71, θλίψη 60, πλήξη 81 και ουδετερότητα 70.

3. **IEMOCAP:** Η βάση δεδομένων IEMOCAP [68] περιέχει 12 ώρες δεδομένων βίντεο σεναριογραφικού και αυτοσχεδιασμένου διαλόγου που έχουν εγγραφεί από 10 ηθοποιούς. Οι συμπεριφορές οργανώνονται σε 5 συνεδρίες των διαδραστικών αλληλεπιδράσεων μεταξύ 2 ομιλητών.

Για τα πειράματά μας θεωρούμε 5531 αποκομένες ομιλίες που περιλαμβάνουν 4 συναισθήματα (1103 θυμός, 1636 ευτυχία, 1708 ουδετερότητα και 1084 θλίψη), όπου συγχωνεύουμε την κλάση ενθουσιασμού και της ευτυχίας στην τελευταίο [27], [28], [35], [36].

4.7.2 Πειράματα Εξαρτημένου-Ομιλητή (SD)

Αξιολογούμε τα χαρακτηριστικά RQA στα SAVEE και Emo-DB ακολουθώντας την προσέγγιση που βασίζεται σε ολόκληρη την ομιλία που περιγράφεται στην ενότητα 4.6.1. Σε αυτή τη ρύθμιση εφαρμόζουμε z κανονικοποίηση ανά ομιλητή (PS-N) και διαχωρίζουμε τυχαία τις φράσεις σε σύνολα εκπαίδευσης και αξιολόγησης. Τα ποσοστά επιτυχίας και στις δύο μετρικές που αναφέρουμε αφορούν τη διασταυρωμένη επικύρωση 5-διπλώματα και συνοψίζονται στον Πίνακα 4.3 για τις τιμές υπερπαραμέτρων των ταξινομητών με τις καλύτερες επιδόσεις.

Το συντηγμένο σύνολο επιτυγχάνει σημαντική βελτίωση απόδοσης σε σχέση με το βασικό σύνολο χαρακτηριστικών IS10 που έχει οριστεί και για τα δύο σύνολα δεδομένων. Στο SAVEE, η WA βελτιώνεται με 3.1% (77.1% $\rightarrow$ 80.2%) και UA με 3.4% (74.5% $\rightarrow$ 77.9%). Επιτύχαμε επίσης μια βελτίωση 4.9% (88.4% $\rightarrow$ 93.3%) και 5.7% (87.2% $\rightarrow$ 92.9%) για WA και UA, αντίστοιχα στην βάση δεδομένων Emo-DB. Το σετ χαρακτηριστικών που χρησιμοποιείται στο [49] εξάγεται πάνω από LLDs από κέπστρουμ, σπεκτρόγραμμα και προσοδία παρόμοια με αυτά που χρησιμοποιούνται στο IS10 [23]. Είναι αξιοσημείωτο ότι επιτυγχάνουν παρόμοιες επιδόσεις με τη δική μας όταν χρησιμοποιούμε μόνο IS10, αλλά το συντηγμένο σύνολο χαρακτηριστικών με LR ξεπερνά στην Emo-DB (κατα 5% στην UA και κατά 4.6% στην WA) και στην SAVEE (κατά 4.5% στην UA και κατά 3.9% στην WA). Ο προτεινόμενος συνδυασμός χαρακτηριστικών και LR ξεπερνάει επίσης μια προσέγγιση με συνελλικτικό SAE [37] σε WA από 5% στην Emo-DB και από 4.8% στο SAVEE. Πιθανότατα, τα μέτρα RQA περιέχουν πληροφορίες που σχετίζονται στενά με την συναισθηματική δυναμική των ομιλητών, η οποία δεν καταγράφεται από συμβατικά χαρακτηριστικά.

Πίνακας 4.3: Τα αποτελέσματα SD για τα SAVEE και Emo-DB

Χαρακτηριστικά	Μοντέλο	SAVEE		Emo-DB	
		WA	UA	WA	UA
IS10	SVM	77.1	74.5	88.4	87.2
	LR	74.4	71.8	87.4	86.3
RQA	SVM	66.0	63.0	81.8	80.4
	LR	64.4	61.1	81.9	79.9
RQA+IS10	SVM	77.3	75.5	90.1	88.9
	LR	**80.2**	**77.9**	**93.3**	**92.9**
[37] Σπεκτρόγραμμα	SAE	75.4	-	88.3	-
[49] LLDs Στατιστικά	ESR	76.3	73.4	88.7	87.9

4.7.3 Πειράματα Ανεξαρτήτου-ομιλητή (SI)

Και πάλι, ακολουθούμε την προσέγγιση που βασίζεται σε ολόκληρη την ομιλία και περιγράφεται στην Ενότητα 4.6.1 και στα δύο σύνολα δεδομένων SAVEE και Emo-DB, αλλά εδώ δεν κάνουμε υποθέσεις για την ταυτότητα του ομιλιτή κατά τη διάρκεια της εκπαίδευσης. Χρησιμοποιούμε άσε-έναν-ομιλητή-έξω αξιολόγηση ανάμεσα σε όλα τα πιθανά διπλώματα (folds), όπου ένας ομιλητής διατηρείται για την αξιολόγηση και οι υπόλοιποι για εκπαίδευση. Η μέση και η τυπική απόκλιση υπολογίζονται μόνο σε δεδομένα εκπαίδευσης και χρησιμοποιούνται για κανονικοποίηση z σε όλα τα δεδομένα εκπαίδευσης. Από εδώ και πέρα, αναφερόμαστε σε αυτή την κανονικοποίηση ως Κανονικοποίηση ανά Δίπλωση (PF-N). Ο πίνακας 4.4 παρουσιάζει τις μέσες τιμές ακριβείας σε όλες τις διπλώσεις για τις τιμές υπερπαραμέτρων των ταξινομητών με τις καλύτερες επιδόσεις.

Σε σύγκριση με το σετ χαρακτηριστικών IS10 της βασικής γραμμής, το συγκρινόμενο σύνολο χαρακτηριστικών επιτυγχάνει απόλυτη βελτίωση κατά 5.5% και 8.2% στο SAVEE, καθώς και κατά 2.4%

και 3.2% στο Emo-DB όσον αφορά τα WA και UA, αντίστοιχα. Επιπλέον, το συντηγμένο σύνολο μας επιτυγχάνει υψηλότερες επιδόσεις στο SAVEE (3.5% σε WA και 4.5% σε UA) και ελαφρώς χαμηλότερο σε Emo-DB σε σύγκριση με το [49]. Στα [143] Σταθμισμένα Φασματικά Χαρακτηριστικά που βασίζονται σε Hu Moments (WSFHM) συγχωνεύονται με το IS10 σε επίπεδο ολόκληρης της ομιλίας που είναι παρόμοιο με την προσέγγισή μας. Σε απευθείας σύγκριση χρησιμοποιώντας το ίδιο μοντέλο (SVM) ξεπερνάμε την αναφερόμενη απόδοση σε όρους WA από 2.5% και 0.4% σε SAVEE και Emo-DB, αντίστοιχα. Επιπλέον, τόσο το RQA όσο και το IS10 καθορίζουν αρκετά χαμηλές επιδόσεις στο SAVEE. Ωστόσο, ο συνδυασμός τους αποδίδει μια εντυπωσιακή βελτίωση της απόδοσης 5.5% (48.5% $\rightarrow$ 54.0%) σε WA και 10.7% (43.1% $\rightarrow$ 53.8%) σε UA πάνω από IS10 όταν χρησιμοποιούμε LR. Τα αποτελέσματά μας υποδηλώνουν ότι τα μέτρα του RQA διατηρούν αμετάβλητες ιδιότητες της μη γραμμικής δυναμικής της συναισθηματικής ομιλίας και ενυπαρχουν για διαφορετικούς ομιλητές.

Πίνακας 4.4: Τα αποτελέσματα SI για τις βάσεις δεδομένων SAVEE και Emo-DB

χαρακτηριστικά	Μοντέλο	SAVEE		Emo-DB	
		WA	UA	WA	UA
IS10	SVM	47.5	45.6	79.7	74.3
	LR	48.5	43.1	76.1	71.9
RQA	SVM	45.6	41.1	70.9	64.2
	LR	47.7	42.3	71.1	67.1
RQA+IS10	SVM	52.5	50.6	82.1	76.9
	LR	**54.0**	**53.8**	80.1	77.5
[49] LLDs Στατιστικά	ESR	51.5	49.3	**82.4**	**78.7**
[143] WSFHM+IS10	SVM	50.0	-	81.7	-

4.7.4 Πειράματα Άσε Μια Συνεδρεία Έξω (LOSO)

Σε αυτή την εργασία, υποθέτουμε ότι η ταυτότητα ομιλητή είναι άγνωστη, αλλά είμαστε σε θέση να εκπαιδεύσουμε το μοντέλο μας λαμβάνοντας υπόψη άλλους ομιλητές που έχουν καταγραφεί σε παρόμοιες συνθήκες. Αξιολογούμε τόσο τις μεθόδους που βασίζοντια πάνω σε όλη την ομιλία όσο και τις μεθόδους που βασίζονται σε τμήματα ομιλίας (που περιγράφονται στην Ενότητα 4.6.2) στην βάση δεδομένων IEMOCAP. Δεδομένης της παραδοχής μας, αντιμετωπίζουμε κάθε μία από τις 5 συνεδρίες ως ομάδα ομιλητών [68]. Χρησιμοποιούμε το LOSO για να δημιουργήσουμε αναδιπλώσεις εκπαίδευσης και αξιολόγησης. Σε κάθε αναδίπλωση, χρησιμοποιούμε 4 συνεδρίες για εκπαίδευση και την άλλη 1 για αξιολόγηση. Για τη δοκιμαστική συνεδρία χρησιμοποιούμε έναν ομιλητή για αξιολόγηση και τον άλλο για τον συντονισμό των υπερπαραμέτρων των μοντέλων μας. Επαναλαμβάνουμε την αξιολόγηση αντιστρέφοντας τους ρόλους των δύο ομιλητών. Στην τελική αξιολόγηση, αναφέρουμε τη μέση απόδοση που επιτυγχάνεται σε WA και UA που λαμβάνεται και από τους 2 ομιλητές [27], [28], [36]. Για να είμαστε εύκολα συγκρίσιμοι με τη βιβλιογραφία ακολουθούμε τρία διαφορετικά σενάρια κανονικοποιήσης των διανυσματικών αναπαραστάεων εισόδου. Χρησιμοποιούμε τα προαναφερθέντα συστήματα PS-N και PF-N καθώς και την καθολική z κανονικοποίηση (G-N). Στο G-N υπολογίζουμε την μέση και τυπική απόκλιση από όλα τα διαθέσιμα δείγματα στο σύνολο δεδομένων και εκτελούμε κανονικοποίηση z πάνω τους. Τα αποτελέσματα για το IEMOCAP για τα τρία διαφορετικά σχήματα κανονικοποίησης παρουσιάζονται στον Πίνακα 4.5.

Μία συνεπής βελτίωση της απόδοσης φαίνεται για όλους τους συνδυασμούς τεχνικών κανονικοποίησης και τα χρησιμοποιούμενα μοντέλα όταν χρησιμοποιείται το συντηγμένο σύνολο αντί για το IS10. Συγκεκριμένα, για το SVM το συντηγμένο σύνολο αποδίδει μια σχετική βελτίωση που κυμαίνεται από 0.3% έως 1.0% σε WA και από 0.2% σε 0.9% σε UA κάτω από όλες τις στρατηγικές κανονικοποίησης. Το ίδιο ισχύει για LR (σε WA από 0.8% σε 1.0% και σε UA από 0.3% σε 1.0%) καθώς και για A-BLSTM (σε WA από 0.1% σε 0.7% και σε UA από 0.2% σε 0.7%). Σύμφωνα με τη διαίσθησή μας [2], μια προσέγγιση που βασίζεται σε τμήματα ομιλίας χρησιμοποιώντας το A-BLSTM

ξεπερνά όλα τα μοντέλα που χρησιμοποιούνται και βασίζονται σε όλη την ομιλία σε WA από 3.4% σε 8.4% και σε UA από 3.8% σε 6.8% για όλα τα σχήματα κανονικοποίησης, όταν χρησιμοποιείται το συντηγμένο σετ. Αυτό είναι πολύ σημαντικό καθώς το εισαγόμενο σύνολο RQA παρέχει μια ευσταθή απεικόνιση χαρακτηριστικών των συναισθηματικών εκφράσεων σε μια ποικιλία προσεγγίσεων σε ποικίλες χρονικές κλίμακες.

Στα χαρακτηριστικά [27] χαμηλού επιπέδου Μελ τράπεζες φίλτρων (MFB) που τροφοδοτούνται απευθείας σε CNN. Στο [36] χρησιμοποιείται ένας στοιβαγμένος αυτόματος κωδικοποιητής για την εξαγωγή των αναπαραστάσεων χαρακτηριστικών από τα σπεκτρογράμματα των σημάτων από την γλωττίδα και στη συνέχεια ένα BLSTM χρησιμοποιείται για ταξινόμηση. Ξεπερνάμε και τα δύο αναφερόμενα αποτελέσματα με 0.2% στο UA για [27] και με ένα περιθώριο 8.7% σε WA και 8.5% σε UA για [36], αντίστοιχα ακόμη και με απλά μοντέλα. Σε σύγκριση με ένα DBN που μαθαίνει ταυτόχρονα τόσο για διακριτή ταξινόμηση συναισθημάτων όσο και για ενεργοποίηση σθένους / ενεργοποίησης στο [28], αναφέρουμε 2.0% υψηλότερη WA και 3.1% υψηλότερη UA. Αναφέρουμε επίσης 4.6% υψηλότερη UA και 1.9% χαμηλότερη WA σε σύγκριση με CNNs πάνω σε σπεκτρογράμματα [35]. Υποθέτουμε ότι αυτή η ασυνέπεια στις μετρήσεις απόδοσης συμβαίνει επειδή ακολουθείται μια ελαφρώς διαφορετική πειραματική ρύθμιση, όπου η τελική συνεδρία αποκλείεται από τον έλεγχο [35].

Πίνακας 4.5: Τα αποτελέσματα LOSO στην βάση δεδομένων IEMOCAP

Χαρακτηριστικά	Μοντέλο	PS-N		PF-N		G-N	
		WA	UA	WA	UA	WA	UA
IS10	SVM	58.3	60.9	58.9	60.1	59.2	60.5
	LR	57.5	61.2	54.6	57.9	53.5	57.5
	A-BLSTM	62.0	65.1	62.6	65.0	62.8	65.0
RQA	SVM	52.9	54.6	53.1	53.8	53.1	53.7
	LR	52.2	54.8	52.6	54.0	52.8	54.3
	A-BLSTM	55.6	59.3	56.6	58.3	56.7	58.7
RQA + IS10	SVM	59.3	61.8	59.2	60.4	59.5	60.7
	LR	58.3	62.0	55.6	58.7	54.5	58.7
	A-BLSTM	**62.7**	**65.8**	**63.0**	**65.2**	62.9	**65.5**
[27] MFB	CNN	-	61.8	-	-	-	-
[28] IS10	DBN	-	-	-	-	60.9	62.4
[35] SP	CNN	-	-	-	-	**64.8**	60.9
[36] GFS	BLSTM	-	-	50.5	51.9	-	-

Κεφάλαιο 5

Αναζήτηση προτύπων για πολυδιάστατη κλιμάκωση Pattern Search MDS

Αυτό το κεφάλαιο είναι μια εκτεταμένη έκδοση του άρθρου [3] που μπορεί να βρεθεί και διαδικτυακά στο https://arxiv.org/abs/1806.00416. Εάν ο αναγνώστης χρειάζεται να αναφέρει τμήματα αυτού του κεφαλαίου τότε θα ήταν προτιμότερο να χρησιμοποιήσει την ακόλουθη αναφορά (ή την πιο ενημερωμένη με τον ίδιο τίτλο και τους συγγραφείς που καθορίζονται):

- Giorgos Paraskevopoulos †, *Efthymios Tzinis †, Emmanuel-Vasileios Vlatakis-Gkaragkounis, and Alexandros Potamianos, "Pattern Search Multidimensional Scaling," arXiv:1806.00416, 2018.*
  †Και οι δύο συγγραφείς συμμετείχαν το ίδιο σε αυτή την δουλειά

5.1 Κίνητρο

Όπως έχουμε συζητήσει προηγουμένως στο τμήμα 1.4, τα χαρακτηριστικά που χρησιμοποιούμε για όλα τα προβλήματα ταξινόμησης θα μπορούσαν να οδηγήσουν σε τεράστιες διανυσματικές αναπαραστάσεις που καθιστούν εξαιρετικά δύσκολο να εκπαιδεύσουμε τα μοντέλα μας από την άποψη του χρόνου και της μνήμης που απαιτείται για τη συνολική διαδικασία. Το ίδιο ισχύει και για το πρόβλημα μας που είναι η συναισθηματική αναγνώριση από φωνή. Σε προηγούμενες ενότητες 4.5.1, 4.5.2 και 4.5.3 έχουμε δει ότι κάθε τμήμα ή έκφραση ομιλίας χρησιμοποιεί διανύσματα που βρίσκονται στα $\mathbb{R}^{1582}$ (σύνολο χαρακτηριστικών IS10), $\mathbb{R}^{432}$ (σύνολο χαρακτηριστικών RQA) και $\mathbb{R}^{2014}$ (RQA + IS10 σύνολο χαρακτηριστικών) αντίστοιχα. Σε αυτό το πλαίσιο, επιδιώκουμε να βρούμε παραστάσεις χαμηλής διάστασης οι οποίες είναι σε θέση να διατηρήσουν τις αρχικές αποστάσεις των αναπαραστάσεων στις υψηλές διαστάσεις. Με αυτόν τον τρόπο, επιδιώκουμε να προσεγγίσουμε μια χαμηλής διάστασης πολλαπλότητα $\mathcal{M} \in \mathbb{R}^L$ όπου $L < 20$ και είμαστε ακόμα σε θέση να πάρουμε τα ίδια αποτελέσματα όσον αφορά την απόδοση ακρίβειας. Προκειμένου να περιγράψουμε εύστοχα αυτή την αναπαράσταση χαμηλής διάστασης που μπορεί να συλλάβει τα ζεύγη συσχετίσεων μεταξύ των διαστάσεων μεγάλων διαστάσεων πρέπει να ενσωματώσουμε και τις μη γραμμικές εξαρτήσεις του χώρου χαρακτηριστικών εισόδου. Αυτό το πρόβλημα είναι γνωστό ως μη-μετρική πολυδιάστατη κλιμάκωση (MDS) ή μη γραμμικής μείωσης διαστάσεων (NLDR). Η πλειοψηφία αυτών των αλγορίθμων μείωσης διαστάσεων με μη-γραμμικό τρόπο μειώνει στην πραγματικότητα μια συνάρτηση απώλειας f. Δεδομένου αυτού του στόχου ελαχιστοποίησης, συνήθως χρησιμοποιούν μεθόδους που βασίζονται σε κλίση για να βρουν ένα ολικό ή τοπικό βέλτιστο. Σε πολλές περιπτώσεις, ωστόσο, η λειτουργία απώλειας είναι μη παραγωγίσιμη ή η εκτίμηση της κλίσης της μπορεί να είναι υπολογιστικά δαπανηρή. Επιπλέον, οι αλγόριθμοι που βασίζονται σε κλίση συνήθως αποδίδουν μια αργή σύγκλιση. απαιτούνται πολλαπλές επαναλήψεις προκειμένου να ελαχιστοποιηθεί η συνάρτηση απώλειας. Εμπνευσμένο από την πρόσφατη πρόοδο σε εργαλεία βελτιστοποίησης χωρίς παραγώγιση, προτείνουμε έναν επαναληπτικό αλγόριθμο ο οποίος αντιμετωπίζει το μη μετρικό MDS ως πρόβλημα βελτιστοποίησης χωρίς παραγώγιση.

5.2 Σχετική δουλειά

Ορισμένοι αλγόριθμοι μείωσης διαστάσεων υποθέτουν ότι οι χαμηλής διάστασης αναπαραστάσεις των δεδομένων μπορούν να ληφθούν χρησιμοποιώντας τεχνικές γραμμικής μείωσης των διαστάσεων όπως η ανάλυση σε κυρίαρχες συνιστώσες (PCA) [144], αλλά αυτό δεν ισχύει όταν αντιμετωπίζουμε δεδομένα για πραγματικές περιπτώσεις, όπως στην SER. Σε πραγματικά σενάρια, μια τέτοια παραδοχή γραμμικότητας μπορεί να είναι πολύ ισχυρή και μπορεί να οδηγήσει σε παραπλανητικές αναπαραστάσεις των δεδομένων εισόδου. Σε αυτό το πλαίσιο, έχει σημειωθεί σημαντική πρόοδος στον αλγόριθμο πολλαπλότητας μάθησης για να ληφθεί υπόψη η μη γραμμική δομή των δεδομένων εισόδου.

Οι πιό γνωστοί αλγόριθμοι μάθησης πολλαπλοτήτων περιλαμβάνουν τις κατόθι μεθόδους: ισομετρική χαρτογράφηση χαρακτηριστικών (ISOMAP) [145, 146, 147, 148, 149], σε σημείο ενδιαφέροντος ISOMAP [150, 151], Επιτόπια Γραμμική Ενσωμάτωση (LLE) [152, 153, 154, 155, 156], Τροποποιημένο LLE [157] Χεσιανό LLE [158, 159], ημιορισμένη Ενσωμάτωση [160], [161], [162], [163], Λαπλασιανοί Ιδιοχάρτες (LE) [152, 164, 165], Τοπική ευθυγράμμιση χώρου εφαπτόμενων (LTSA) [166], Σπεκτρική Συσταδοποιήση [167], t-SNE [88] κλπ. Το ISOMAP χρησιμοποιεί μια γεωδεσική απόσταση για να μετρήσει τις γεωμετρικές πληροφορίες μέσα σε μια πολλαπλότητα και έπειτα εφαρμόζει MDS. Το LLE υποθέτει ότι μια πολλαπλότητα μπορεί να προσεγγιστεί σε έναν Ευκλείδειο χώρο και οι συντελεστές ανασυγκρότησης των γειτόνων μπορούν να διατηρηθούν στο χώρο των μικρών διαστάσεων. Το LE χρησιμοποιεί ένα μη προσανατολισμένο σταθμισμένο γράφημα για να διατηρήσει τις σχέσεις τοπικών γειτόνων. Το Hessian LLE αποκτά παραστάσεις χαμηλής διάστασης εφαρμόζοντας ανάλυση με ιδιοδιανύσματα σε μια χεσιανή μήτρα συντελεστών. Το LTSA χρησιμοποιεί τοπικές εφαπτόμενες πληροφορίες για να αντιπροσωπεύει τη γεωμετρία πολλαπλότητας και επεκτείνει αυτό σε καθολικές συντεταγμένες. Τέλος, η SDE επιχειρεί να μεγιστοποιήσει την απόσταση μεταξύ σημείων που δεν ανήκουν σε μια τοπική γειτονιά. Επιπλέον, μια κοινή μη γραμμική μέθοδος για τη μείωση των διαστάσεων είναι η επέκταση του πυρήνα του PCA [168] η οποία είναι επίσης παρόμοια με την Σπεκτρική Συσταδοποιήση στην οποία χρησιμοποιείται ένας Γκαουσιανός πυρήνας ανά στοιχείο και μετά εφαρμόζεται ένας αλγόριθμος K-κέντρων για να συσταδοποιήσει τα δεδομένα σε ομάδες. Τέλος, ο t-SNE είναι κατάλληλος για την απεικόνιση των πολλαπλοτήτων (για χώρους $2D$ και $3D$) από δεδομένα μεγάλης διαστάσεως, ελαχιστοποιώντας κατά διαστήματα μια μη κυρτή συνάρτηση κόστους μέσω της μείωσης της κλίσης.

Μια ευρεία κατηγορία αλγορίθμων χωρίς παραγώγιση για τη μη γραμμική βελτιστοποίηση έχει μελετηθεί και αναλυθεί στα [169] και [170]. Οι μέθοδοι GPS αποτελούνται από ένα υποσύνολο των προαναφερθέντων αλγορίθμων που δεν απαιτούν τον ρητό υπολογισμό της κλίσης σε κάθε βήμα επανάληψης. Ορισμένοι αλγόριθμοι GPS είναι: ο αρχικός αλγόριθμος αναζήτησης μοτίβων Hooke και Jeeves [171], η εξελικτική λειτουργία χρησιμοποιώντας τον παράγοντα σχεδιασμού [172] και τον αλγόριθμο αναζήτησης πολλαπλών διευθύνσεων [173], [174]. Στο [113] παρουσιάστηκε μια ενοποιημένη θεωρητική διατύπωση αλγορίθμων GPS υπό ένα κοινό μοντέλο αναφοράς καθώς και μια εκτεταμένη ανάλυση των γενικών ιδιοτήτων σύγκλισης. Οι τοπικές ιδιότητες σύγκλισης μελετήθηκαν αργότερα στο [114]. Ειδικότερα, το θεωρητικό πλαίσιο καθώς και οι ιδιότητες σύγκλισης των μεθόδων GPS έχουν επεκταθεί σε περιπτώσεις με γραμμικούς περιορισμούς [117], περιορισμούς σε σύνορα [115] και γενικού Λανγκρατζιανού φορμαλισμού [116].

5.3 Αλγόριθμος Περιγραφή

5.3.1 Διατύπωση

Η βασική ιδέα πίσω από τον προτεινόμενο αλγόριθμο είναι να αντιμετωπίζει το MDS ως πρόβλημα χωρίς παραγώγιση, χρησιμοποιώντας μια παραλλαγή της γενικής βελτιστοποίησης αναζήτησης προτύπων για να ελαχιστοποιήσει μια συνάρτηση απώλειας. Η είσοδος στην αναζήτηση μοτίβου MDS είναι μια μήτρα $N \times N$ στόχου ανομοιότητας $\mathbf{T}$ και η διάσταση στόχου L του χώρου ενσωμάτωσης. Μια επισκόπηση του αλγορίθμου που παρουσιάζεται στον Αλγόριθμο 2 παρουσιάζεται στη

συνέχεια.

Η διαδικασία αρχικοποίησης του αλγορίθμου αποτελείται από: 1) τυχαία δειγματοληψία των N σημείων στον ενσωματωμένο χώρο και κατασκευή της μήτρας $\mathbf{X}^{(0)} = [\mathbf{x}_1^{(0)}, \mathbf{x}_2^{(0)}, ..., \mathbf{x}_N^{(0)}] \in \mathbb{R}^{N \times L}$, 2) υπολογισμό της μήτρας ανομοιότητας του ενσωματωμένου χώρου $\mathbf{D}^{(0)}$, όπου το στοιχείο $d_{ij}^{(0)}$ είναι η ευκλείδια απόσταση μεταξύ των διανυσμάτων $\mathbf{x}_i^{(0)}$ και $\mathbf{x}_j^{(0)}$ από το $\mathbf{X}^{(0)}$ και 3) τον υπολογισμό του αρχικού σφάλματος προσέγγισης $e^{(0)} = f(\mathbf{T}, \mathbf{D}^{(0)})$, όπου e είναι το μέσον τετραγωνικό σφάλμα όλων των στοιχείων (MSE) μεταξύ των δύο μητρών. Η συνάρτηση στόχος f που προσπαθούμε να ελαχιστοποιήσουμε είναι το κανονικοποιημένο τετράγωνο της Frobenius νόρμας της μήτρας $\mathbf{T} - \mathbf{D}$, δηλαδή $f(\mathbf{T}, \mathbf{D}) = (1/N^2)||\mathbf{T} - \mathbf{D}||_F^2$. Αντιστοίχως μπορεί κανείς να εκφράσει την συνάρτηση f ως εξής:

$$f(\mathbf{T}, \mathbf{D}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (t_{ij} - d_{ij})^2, \quad \text{όπου } \mathbf{T}, \mathbf{D} \in \mathbb{R}^{N \times N} \tag{5.1}$$

Algorithm 2 Αναζήτηση προτύπου MDS - Pattern Search MDS

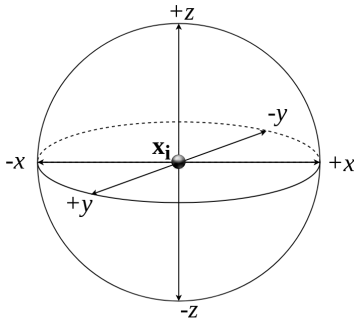
```
1:  procedure MDS(T, L, r^(0))
2:      k ← 0                                   ▷ k είναι ο δείκτης της εκάστοτε εποχής
3:      X^(k) ← UNIFORM(N × L)
4:      D^(k) ← DISTANCE_MATRIX(X^(k))
5:      e^(k) ← f(T, D^(k))
6:      e^(k−1) ← +∞
7:      r^(k) ← r^(0)
8:      while r^(k) > δ do
9:          if e^(k−1) − e^(k) ≤ ϵ · e^(k) then
10:             r^(k) ← r^(k)/2
11:         S ← SEARCH_DIRECTIONS(r^(k), L)
12:         for all x ∈ X^(k) do
13:             X*, e* ← OPTIMAL_MOVE(X^(k),x,S,e^(k))
14:             e^(k−1) ← e^(k)
15:             e^(k) ← e*
16:             X^(k) ← X*
17:         k = k + 1
```

5.3.2 Κατευθύνσεις αναζήτησης

Μετά από τα βήματα αρχικοποίησης, σε κάθε εποχή (επανάληψη), θεωρούμε την επιφάνεια μιας υπερσφαίρας ακτίνας r γύρω από κάθε σημείο $\mathbf{x}_i^{(k)}$. Οι πιθανές κατευθύνσεις αναζήτησης βρίσκονται στην επιφάνεια μιας n-διάστατης σφαίρας (αν ο χώρος αναζήτησης είναι ο $\mathbb{R}^n$) κατά μήκος της ορθογωνικής βάσης του χώρου, π.χ. στην περίπτωση διαστάσεων 3 διαστάσεων κατά μήκος των διευθύνσεων $\pm x, \pm y, \pm z$ στη σφαίρα που φαίνεται στο σχήμα 5.1. Αυτό δημιουργεί τον πίνακα των δυνατών κατευθύνσεων αναζήτησης S και συνοψίζεται στον Αλγόριθμο 3

Algorithm 3 Ορισμός κατευθύνσεων αναζήτησης

```
1:  function SEARCH_DIRECTIONS(r, L)
2:      S^+ ← r · I_L
3:      S^− ← −r · I_L
4:      S ← [S^+ / S^−]
5:      return S
```



Σχήμα 5.1: Σφαίρα ακτίνας r γύρω από το σημείο $\mathbf{x_i^{(k)}}$ και πιθανές κατευθύνσεις αναζήτησης

5.3.3 Μετακίνηση Παράλληλα με την Βέλτιστη Κατεύθυνση

Κάθε σημείο μετακινείται κατά μήκος της διάστασης που παράγει το ελάχιστο σφάλμα. Σε αυτό το στάδιο, εξετάζουμε μόνο κινήσεις που αποφέρουν μονοτονική μείωση της συνάρτησης σφάλματος. Ο αλγόριθμος 4 βρίσκει τη βέλτιστη κίνηση που ελαχιστοποιεί το $e^{(k)} = f(\mathbf{T}, \mathbf{D}^{(k)})$ για κάθε νέο σημείο $\tilde{x}$ και μετακινεί το $\mathbf{X}$ προς αυτή την κατεύθυνση. Σημειώνουμε ότι κατά την εκτίμηση του βήματος $s \in \mathbf{S}$, ο πίνακας $\mathbf{S}$ θεωρείται ότι είναι ένα σύνολο διανυσμάτων γραμμών.

Algorithm 4 Βρίσκοντας την βέλτιστη κίνηση για ένα σημείο

```
1:  function OPTIMAL_MOVE(X^(k), x, S, e)
2:      e* ← e
3:      for all s ∈ S do
4:          x̃ ← x + s
5:          X̃ ← UPDATE_POINT(X^(k), x, x̃)      ▷ Ανανέωση του σημείου x του πίνακα X^(k) με x̃
6:          D ← DISTANCE_MATRIX(X̃)
7:          ẽ ← f(T, D)
8:          if ẽ < e* then
9:              e* ← ẽ
10:             X* ← X̃
11:     return X*, e*
```

5.3.4 Υπολογισμός του σφάλματος

Το προκύπτον σφάλμα e^* υπολογίζεται αφού εκτελεστεί η βέλτιστη κίνηση για κάθε σημείο στο $\mathbf{X}^{(k)}$. Αν η μείωση του σφάλματος είναι μικρότερη από ένα κατώφλι τότε μειώνουμε στο μισό την ακτίνα αναζήτησης και συνεχίζουμε στην επόμενη εποχή. Αυτό εκφράζεται όπως φαίνεται παρακάτω:

$$e^{(k)} - e^* < \epsilon \cdot e^{(k)} \tag{5.2}$$

όπου ϵ είναι μια μικρή θετική σταθερά, δηλαδή η μείωση του σφάλματος γίνεται πολύ μικρή σε σχέση με το $e^{(k)}$. Η διαδικασία σταματά όταν η ακτίνα αναζήτησης r γίνει πολύ μικρή, δηλαδή $r < \delta$, όπου δ είναι μια μικρή σταθερά, όπως φαίνεται στον Αλγόριθμο 2.

5.3.5 Πολυπλοκότητα αλγορίθμου

Η πολυπλοκότητα του αλγόριθμου είναι $\mathcal{O}(N^2 L)$. Αυτό εξηγείται ως εξής: για κάθε εποχή αναζητούμε $2L$ διαστάσεις για τα όλα τα N σημεία. Σε κάθε αναζήτηση χρειαζόμαστε επίσης $\mathcal{O}(N)$ για την ενημέρωση της μήτρας απόστασης καθώς κινούμαστε όλα τα σημεία ανεξάρτητα για κάθε εποχή.

5.4 Προσεγγίσεις και βελτιστοποιήσεις

Στη συνέχεια παρουσιάζεται ένα σύνολο αλγοριθμικών βελτιστοποιήσεων που βελτιώνουν το χρόνο εκτέλεσης ή την ποιότητα της λύσης του αλγορίθμου 2. Είναι αξιοσημείωτο ότι, σε ορισμένες περιπτώσεις, επιτυγχάνονται βελτιστοποιήσεις στο χρόνο και απόδοση αξιολόγησης χρησιμοποιώντας αυτές τις βελτιστοποιήσεις / προσεγγίσεις. Παρουσιάζουμε επίσης τρόπους βελτίωσης του χρόνου εκτέλεσης, αναζητώντας προσεγγίσεις, καθώς και συζητήσουμε τρόπους για να χρησιμοποιήσουμε τον παράλληλο υπολογισμό για τμήματα του αλγορίθμου.

5.4.1 Ανοχή για τις κακές κινήσεις

Στον κύριο αλγόριθμο που προτείνουμε (βλέπε Ενότητα 5.3) περιορίζουμε τις αποδεκτές κινήσεις έτσι ώστε το σφάλμα να μειώνεται μονότονα. Αυτός είναι ένας λογικός περιορισμός που μας παρέχει επίσης θεωρητικές εγγυήσεις σύγκλισης. Παρ 'όλα αυτά, στο πειραματικό μας περιβάλλον παρατηρήσαμε ότι αν χαλαρώσουμε αυτόν τον περιορισμό και επιτρέπουμε σε κάθε σημείο να κάνει πάντα τη βέλτιστη κίνηση, ανεξάρτητα αν το σφάλμα (προσωρινά) αυξάνει ο αλγόριθμος συγκλίνει ταχύτερα προς καλύτερες λύσεις. Η ιδέα να επιτραπεί στους άπληστους αλγορίθμους να κάνουν κάποιες «κακές» κινήσεις με την ελπίδα να ξεπεράσουν τα τοπικά ελάχιστα μπορεί να βρεθεί σε άλλους αλγορίθμους βελτιστοποίησης, καθώς η Προσομοιωμένη ανόπτηση [175] είναι η πιο δημοφιλής. Για να το εφαρμόσουμε μπορούμε να τροποποιήσουμε την γραμμή 13 στον Αλγόριθμο 2 σε:

$\mathbf{X}^*, e^* \leftarrow$ OPTIMAL_MOVE($\mathbf{X}^{(k)}, x, S, +\infty$)

5.4.2 Ενημέρωση του πίνακα της τωρινής ανομοιότητας

Στη γραμμή 6 του Αλγορίθμου 4 παρατηρούμε ότι επανασχηματίζουμε τη μήτρα ανομοιότητας αφού κάνουμε μια κίνηση για κάθε σημείο. Αυτό μπορεί να αποφευχθεί επειδή κάθε κίνηση μεταβάλλει μόνο ένα σημείο $\mathbf{x}_i^{(k)}$, επομένως επηρεάζονται μόνο η αντίστοιχη σειρά $\mathbf{d}_{i,:}^{(k)}$ και στήλη $\mathbf{d}_{:,i}^{(k)}$ της μήτρας ανομοιότητας $\mathbf{D}^{(k)}$. Επιπλέον, μόνο μία διάσταση l του διανύσματος $\mathbf{x}_i^{(k)}$ τροποποιείται από την κίνηση, δηλ. Μόνο το στοιχείο $x_{i,l}^{(k)}$ της μήτρας $\mathbf{X}^{(k)}$. Πιο συγκεκριμένα, το στοιχείο $d_{i,j}$ που αποθηκεύει την ανομοιότητα μεταξύ των σημείων $\mathbf{x}_i$ και $\mathbf{x}_j$ θα πρέπει να ενημερωθεί ως εξής για τη μετάβαση από $x_{i,l}^{(k)}$ σε $x_{i,l}^{(k+1)}$ για $i \neq j$:

$$d_{i,j}^{(k+1)} = \sqrt{\left(d_{i,j}^{(k)}\right)^2 - \left(x_{i,l}^{(k)} - x_{j,l}^{(k)}\right)^2 + \left(x_{i,l}^{(k+1)} - x_{j,l}^{(k+1)}\right)^2} \tag{5.3}$$

5.4.3 Επιλογή κατεύθυνσης από τυχαίο υποσύνολο των διαθεσίμων κατευθύνσεων

Από την ανάγκη να αναζητηθεί η βέλτιστη κίνηση στις διαστάσεις ενσωμάτωσης L, προκύπτει ότι η πολυπλοκότητα του αλγορίθμου έχει μια γραμμική εξάρτηση από το L σε υπολογιστικό χρόνο. Μια μεγάλη τιμή του L θα επηρεάσει αρνητικά τον χρόνο εκτέλεσης του αλγορίθμου. Μια προσεγγιστική τεχνική για την λύση αυτού του προβλήματος είναι η εκτέλεση τυχαίας δειγματοληψίας των κατευθύνσεων $K < L$ σε όλες τις πιθανές κατευθύνσεις στον χώρο $\mathbb{R}^L$. Μετά από αυτό, ορίζουμε τη βέλτιστη κίνηση για κάθε σημείο με τον ίδιο τρόπο όπως πριν αλλά συγκρίνοντας τα σφάλματα στη συνάρτηση απώλειας χρησιμοποιώντας μόνο τις επιλεγμένες K κατευθύνσεις. Με αυτό τον τρόπο, μπορούμε να επιλέξουμε μια κατεύθυνση "καλή" αντί για τη βέλτιστη. Αντί των $2L$ κινήσεων ανά εποχή, θα λαμβάναμε υπόψη μόνο τις 2 κατευθύνσεις για να υπολογίσουμε τη νέα εκτίμηση του σφάλματος. Σε αυτή την περίπτωση, η συνολική πολυπλοκότητα ανά εποχή θα ήταν $\mathcal{O}(N^2K)$ αντί $\mathcal{O}(N^2L)$.

Πιθανότατα, μπορεί να υπάρχουν καλύτερες στρατηγικές για να επιλέξουμε τις K κατευθύνσεις από την αφηρημένη τυχαία δειγματοληψία όλων των δυνατών κατευθύνσεων στον χώρο αναζήτησης $\mathbb{R}^L$. Καθώς η γεωμετρία του χώρου ενσωμάτωσης αρχίζει να γίνεται φανερή, μετά από μερικές εποχές

του αλγορίθμου, έχει λογική η αυξανόμενη μεροληψία της αναζήτησης προς διανύσματα συνιστωσών της γειτονιάς του σημείου που μετακινείται. Για παράδειγμα, στον αλγόριθμο [171] του αλγορίθμου αναζήτησης μοτίβων Hooke και Jeeves γίνεται ένα αυξημένο βήμα προς την κατεύθυνση η οποία απέδωσε ικανοποιητική μείωση της συνάρτησης απώλειας από την προηγούμενη εποχή.

5.4.4 Εκτίμηση της αρχικής ακτίνας αναζήτησης

Μια σημαντική παράμετρος για τον αλγόριθμό μας είναι η ακτίνα εκκίνησης $r^{(0)}$. Αυτή η παράμετρος ελέγχει πόσο ευρεία θα είναι η αναζήτηση αρχικά και ότι έχει ένα αποτέλεσμα παρόμοιο με το ρυθμό εκμάθησης αλγόριθμων βελτιστοποίησης με βάση την μείωση της κλίσης. Μια συντηρητική επιλογή για την αρχική ακτίνα θα οδηγήσει τον αλγόριθμο να συγκλίνει αργά σε ένα τοπικό βέλτιστο. Ενώ μια υψηλή τιμή πιθανότατα θα προκαλέσει το σφάλμα να υπερβεί, κάνοντας ταυτόχρονα τον αλγόριθμο πιο δύσκολο να συγκλίνει σε ένα τοπικό ελάχιστο. Μια απλή τεχνική για την αυτόματη εύρεση μιας καλής ακτίνας εκκίνησης είναι η χρήση δυαδικής αναζήτησης μεταξύ μηδέν και μιας επαρκώς μεγάλης τιμής. Συγκεκριμένα, ρυθμίσαμε την ακτίνα έναρξης σε μια αυθαίρετη τιμή, εκτελέσαμε μια εκτέλεση του αλγορίθμου για μια εποχή χωρίς πραγματικά να παράγουμε λύση για τον χώρο των μικροτέρων διαστάσεων και παρατηρήσαμε την επίδραση στο σφάλμα. Εάν το σφάλμα αυξάνεται, μειώνουμε κατά το ήμισυ την ακτίνα. Διαφορετικά το διπλασιάζουμε και επαναλαμβάνουμε τη διαδικασία. Αυτή η διαδικασία επιτρέπεται να εκτελείται για μικρό αριθμό εποχών. Η ακτίνα εκκίνησης που βρέθηκε χρησιμοποιώντας αυτή την τεχνική είναι μια όχι πολύ απαισιόδοξη ή υπερβολικά αισιόδοξη εκτίμηση της καλύτερης τιμής παραμέτρου.

5.4.5 Υλοποιήση με παράλληλο προγραμματισμό

Ένας άλλος τρόπος να ενισχυθεί ο χρόνος εκτέλεσης είναι να χρησιμοποιηθεί παράλληλος υπολογισμός για την επιτάχυνση των τμημάτων του αλγορίθμου. Στην περίπτωσή μας μπορούμε να παραλληλιστούμε την αναζήτηση για τις βέλτιστες κινήσεις στις διαστάσεις ενσωμάτωσης χρησιμοποιώντας το μοτίβο παραλληλοποίησης του χάρτη. Συγκεκριμένα, μπορούμε να χαρτογραφήσουμε την αναζήτηση για υποψήφιες κινήσεις για εκτέλεση σε διαφορετικά νήματα και να αποθηκεύσουμε το σφάλμα για κάθε υποψήφια κίνηση σε ένα πίνακα $\mathbf{e} = [e_1, e_2.., e_{2L}]$. Με άλλα λόγια, η διαδικασία υπολογισμού της βέλτιστης κατεύθυνσης για κάθε σημείο είναι εντελώς ασυσχέτιστη με όλα τα άλλα σημεία και επομένως μπορεί να παραλληλοποιηθεί πλήρως. Μετά την ολοκλήρωση της αναζήτησης μπορούμε να εκτελέσουμε μια λειτουργία μείωσης για να βρούμε τη βέλτιστη κίνηση και το βέλτιστο σφάλμα $\mathbf{X}^*, e^*$. Για την υλοποίησή μας χρησιμοποιήσαμε το προγραμματιστικό περιβάλλον OpenMP [176]. Η παράλληλη υλοποιήση του αλγορίθμου οδήγησε σε μια σημαντική επιτάχυνση του χρόνου εκτέλεσης με τη μείωση του σε $25 - 50\%$ του αρχικού χρόνου που απαιτείται για να τρέξει (ανάλογα με τον αριθμό των νημάτων που δύναται να χρησιμοποιηθούν από τον εκάστοτε επεξεργαστή).

5.5 Περιγραφή του Pattern Search MDS ώς μέθοδος GPS

Το Pattern Search MDS ανήκει στη γενική κατηγορία των μεθόδων GPS και μπορεί να εκφραστεί χρησιμοποιώντας το ενοποιημένο σχήμα των μεθόδων GPS που εισάγεται στο τμήμα 2.8. Στη συνέχεια, εκφράζουμε τον προτεινόμενο αλγόριθμό μας και τη σχετική αντικειμενική συνάρτησή του υπό αυτόν τον φορμαλισμό.

Πρώτον, επαναλαμβάνουμε το πρόβλημα του MDS σε μορφή μόνο ενός διανύσματος. Χρησιμοποιούμε τον πίνακα με τα στοιχεία $\{\delta_{ij}\}_{1\leq i,j\leq N}$ που εκφράζουν τις ανομοιότητες μεταξύ των σημείων N στο χώρο των μεγάλων διαστάσεων. Το σύνολο των σημείων $\{\mathbf{x}_i\}_{i=1}^N$ βρίσκεται στην μικρής διαστάσεων πολλαπλότητα $\mathcal{M} \in \mathbb{R}^L$ και σχηματίζει το σετ στήλης της μήτρας $\mathbf{X}^T$. Η μήτρα $\mathbf{X} \in \mathbb{R}^{N\times L}$ θα διανυσματοποιηθεί ως ένα διάνυσμα στήλη όπως φαίνεται παρακάτω:

$$\begin{aligned} \mathbf{x}_i &= [x_{i1}, ..., x_{iL}]^T \in \mathbb{R}^L, 1 \leq i \leq N \\ \mathbf{z} &= vec(\mathbf{X}^T) = [x_{11}, ..., x_{1L}, ..., x_{N1}, ..., x_{NL}]^T \end{aligned} \tag{5.4}$$

Τώρα η νέα μας μεταβλητή $\mathbf{z}$ βρίσκεται στον χώρο αναζήτησης $\mathbb{R}^{N\cdot L}$. Η απόσταση μεταξύ οποιωνδήποτε δύο σημείων $\mathbf{x}_i$ και $\mathbf{x}_j$ της πολλαπλότητας $\mathcal{M}$ παραμένει η ίδια αλλά τώρα εκφράζεται ως συνάρτηση της μεταβλητής $\mathbf{z}$. Δηλαδή, $d_{ij}(\mathbf{X}) = ||\mathbf{x}_i - \mathbf{x}_j|| = \sqrt{\sum_{k=1}^{L}(x_{ik} - x_{jk})^2} = d_{ij}(\mathbf{z})$. Για το σκοπό αυτό, η νέα αντικειμενική μας συνάρτηση για να ελαχιστοποιήσουμε είναι η g που είναι το MSE μεταξύ των δοσμένων ανισοτήτων δ_{ij} και των ευκλειδίων αποστάσεων d_{ij} στην χαμηλής διαστάσεως πολλαπλότητα $\mathcal{M}$ όπως ορίζεται στην Εξίσωση 5.5 που φαίνεται παρακάτω:

$$g(\mathbf{z}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(d_{ij}(\mathbf{z}) - \delta_{ij})^2, \;\; \mathbf{z} \in \mathbb{R}^{N\cdot L} \tag{5.5}$$

Συνεπώς, το αρχικό MDS εκφράζεται τώρα ως ένα μη περιορισμένο μη-κυρτό πρόβλημα βελτιστοποίησης το οποίο εκφράζεται ελαχιστοποιώντας τη συνάρτηση g στον χώρο αναζήτησης του $\mathbb{R}^{N\cdot L}$ (Εξίσωση 5.6). Συγκεκριμένα, οι L συντεταγμένες για όλα τα N σημεία στην πολλαπλότητα $\mathcal{M}$ χρησιμεύουν τώρα ως βαθμοί ελευθερίας για τη λύση μας.

$$\mathbf{z}^* = \min_{\mathbf{z}\in\mathbb{R}^{N\cdot L}} g(\mathbf{z}) \tag{5.6}$$

Τώρα που έχουμε διατυπώσει το πρόβλημα και τη μεταβλητή $\mathbf{z}$ στην κατάλληλη μορφή μπορούμε να ταιριάξουμε κάθε εποχή του αρχικού μας αλγορίθμου με μια επανάληψη μιας μεθόδου GPS. Επομένως, οι κινήσεις που παράγονται από τον αλγόριθμό μας σχηματίζουν μια ακολουθία σημείων $\{\mathbf{z}^{(k)}\}$. Επιπλέον, πρόκειται να ορίσουμε τους πίνακες $\mathbf{B}, \mathbf{C}^{(k)}, \mathbf{P}^{(k)}$ για τον αλγόριθμό μας όπως στις εξισώσεις. 2.63, 2.64. Η επιλογή της βασικής μας μήτρας $\mathbf{B}$ είναι η μήτρα ταυτότητας όπως φαίνεται στην Εξίσωση 5.8.

$$\mathbf{e}_i = [0, .., \underbrace{1}_{index\;\; i}, ..., 0]^T, 1 \leq i \leq N\cdot L \tag{5.7}$$

$$\mathbf{B} = \mathbf{I}_{N\cdot L} = [\mathbf{e}_1, ..., \mathbf{e}_{N\cdot L}] \tag{5.8}$$

Ενώ η μήτρα ταυτότητας δεν είναι μοναδική και οι στήλες της επεκτείνουν θετικά στον χώρο αναζήτησης $\mathbb{R}^{N\cdot L}$, ορίζουμε επίσης το $\mathbf{M}^{(k)}$ ως μήτρα ταυτότητας. Στη μήτρα εξισώσεων 5.9 $\mathbf{\Psi}^{(k)}$ αντιπροσωπεύει την κίνηση παράλληλα με τα μοναδιαία διανύσματα ορθοκανονικής βάσης του χώρου $\mathbb{R}^{N\cdot L}$. Παρ 'όλα αυτά, η γεννήτορας μήτρα $\hat{\mathbf{C}}$ περιλαμβάνει επίσης όλες τις υπόλοιπες πιθανές διευθύνσεις που δημιουργούνται από το σύνολο $\{-1, 0, 1\}$. Συνολικά έχουμε $3^{N\cdot L} - 2\cdot N\cdot L$ επιπλέον κατευθύνσεις μέσα στον αντίστοιχο πίνακα $\mathbf{L}^{(k)}$ όπως φαίνεται στην Εξίσωση 5.10.

$$\begin{aligned} \mathbf{M}^{(k)} &= \hat{\mathbf{M}} = \mathbf{I}_{N\cdot L} \in \mathbb{Z}^{N\cdot L\times N\cdot L} \\ \mathbf{\Psi}^{(k)} &= \hat{\mathbf{\Psi}} = [\hat{\mathbf{M}} \;\; -\hat{\mathbf{M}}] \end{aligned} \tag{5.9}$$

$$\begin{aligned} \hat{S} &= \{-1, 0, 1\} \\ \mathbf{L}^{(k)} &= \hat{\mathbf{L}} \\ \hat{\mathbf{L}} &= \{\hat{v} : \hat{v} \in \underbrace{\hat{S}\times ... \times \hat{S}}_{N\cdot L} \wedge \hat{v} \notin \{\mathbf{e}_1, ..., \mathbf{e}_{N\cdot L}\}\} \end{aligned} \tag{5.10}$$

Σύμφωνα με τις εξισώσεις 5.9, 5.10, κατασκευάζουμε την πλήρη μήτρα μοτίβου $\mathbf{P}^{(k)}$ στην εξίσωση 5.11 με παρόμοιο τρόπο με την εξίσωση 2.64. Για τον αλγόριθμό μας ο πίνακας μοτίβων είναι ίσος με τον πίνακα δημιουργίας μας $\mathbf{C}^{(k)} = \hat{\mathbf{C}}$, ο οποίος επίσης είναι σταθερός για όλες τις επαναλήψεις. Εννοιολογικά, ο πίνακας δημιουργίας $\hat{\mathbf{C}}$ περιέχει όλες τις πιθανές εξερευνητικές κινήσεις ενώ χρησιμοποιείται ευρετική για την αξιολόγηση της αντικειμενικής συνάρτησης g μόνο για ένα υποσύνολο αυτών.

$$\begin{aligned} \mathbf{C}^{(k)} = \hat{\mathbf{C}} &= [\hat{\mathbf{\Psi}} \;\; \hat{\mathbf{L}}] = [\hat{\mathbf{M}} \;\; -\hat{\mathbf{M}} \;\; \hat{\mathbf{L}}] \\ \mathbf{P}^{(k)} &= \hat{\mathbf{P}} \equiv \mathbf{B}\hat{\mathbf{C}} \equiv \hat{\mathbf{C}} \end{aligned} \tag{5.11}$$

Τέλος, ρυθμίζουμε τις ενημερώσεις της παραμέτρου μήκους βήματος για κάθε κλάση επιτυχημένων και ανεπιτυχών επαναλήψεων όπως αυτές περιγράφηκαν προηγουμένως στις εξισώσεις 2.66, 2.67, αντίστοιχα. Υπενθυμίζοντας την σημείωση του τμήματος 2.8.1, το $\hat{\mathbf{s}}^{(k)}$ είναι το βήμα που επιστρέφεται από την υπορουτίνα της εξερευνητικής μας κίνησης στην kοστή επανάληψη. Για τις επιτυχείς επαναλήψεις $g(\mathbf{z}^{(k)} + \hat{\mathbf{s}}^{(k)}) < g(\mathbf{z}^{(k)})$ δεν αυξάνουμε περαιτέρω το μήκος των κινήσεών μας περιορίζοντας το $\Lambda = \{1\}$ ως εξής:

$$\Delta^{(k+1)} = \Delta^{(k)}, \quad \text{αν} \quad f(\mathbf{z}^{(k)} + \hat{\mathbf{s}}^{(k)}) < f(\mathbf{z}^{(k)}) \tag{5.12}$$

Ομοίως, για τις ανεπιτυχείς επαναλήψεις $g(\mathbf{z}^{(k)} + \hat{\mathbf{s}}^{(k)}) \geq g(\mathbf{z}^{(k)})$ μειώνουμε κατά το ήμισυ την απόσταση κατά συντελεστή 2 ρυθμίζοντας το $\theta = \frac{1}{2}$ όπως φαίνεται παρακάτω:

$$\Delta^{(k+1)} = \frac{1}{2}\Delta^{(k)}, \quad \text{αν} \quad f(\mathbf{z}^{(k)} + \hat{\mathbf{s}}^{(k)}) >= f(\mathbf{z}^{(k)}) \tag{5.13}$$

Μια σύντομη περιγραφή του αλγορίθμου μας ως μεθόδος GPS για την επίλυση του προβλήματος που αναφέρεται στην Εξίσωση 5.6, είναι η ακόλουθη: Σε κάθε επανάληψη, καθορίζουμε τη βέλτιστη κατεύθυνση συντεταγμένων για καθένα από τα σημεία που βρίσκονται στην πολλαπλότητα $\mathbf{x}_i \in \mathcal{M},\ 1 \leq i \leq N$. Για κάθε εσωτερική επανάληψη του αλγορίθμου 4, εάν η βέλτιστη κατεύθυνση παράγει μια χαμηλότερη τιμή για την αντικειμενική μας συνάρτηση g, συσσωρεύουμε αυτήν την κατεύθυνση και κινούμαστε μαζί με αυτή τη συντεταγμένη του $\mathbb{R}^{N \cdot L}$. Διαφορετικά, παραμένουμε στην ίδια θέση. Ως αποτέλεσμα, η εξερεύνηση των συντεταγμένων για το νέο σημείο $\mathbf{x}_{i+1}$ αρχίζει από αυτήν την προσωρινή θέση. Αυτή η άπληστη προσέγγιση παρέχει ένα διάνυσμα κίνησης με 1 μόνο στο διάνυσμα βάσης στο οποίο κινούμαστε (-1 αν κινούμαστε με την αντίθετη φορά) όπως περιγράφεται στην Εξίσωση 5.7. Εάν η επανάληψη είναι ανεπιτυχής το μηδενικό διάνυσμα του χώρου εξερεύνησης επιστρέφεται $\mathbf{0} \in \mathbb{R}^{N \cdot L}$. Το τελικό διάνυσμα κατεύθυνσης $\hat{\mathbf{s}}^{(k)}$ για την kοστή επανάληψη υπολογίζεται αθροίζοντας αυτά τα διανύσματα με μόνο με ένα 1 ή ένα -1 ή ένα μηδενικό στοιχείο του χώρου. Στην kοστή επανάληψη, η κίνηση θα δίδεται από τον πολλαπλασιασμό της παραμέτρου μήκους βήματος $\Delta^{(k)}$ με το τελικό διάνυσμα κατεύθυνσης με παρόμοιο τρόπο όπως ορίζεται στην Εξίσωση 2.65. Αυτό παρέχει μια απλή μείωση της αντικειμενικής συνάρτησης g ή στη χειρότερη περίπτωση αντιπροσωπεύει μηδενική κίνηση στον χώρο αναζήτησης $\mathbb{R}_{N \cdot L}$. Όσον αφορά την κίνηση σε $\hat{\mathbf{s}}^{(k)}$, είναι τετριμμένο να δείξουμε ότι αυτή η μείωση της αντικειμενικής συνάρτησης g είναι μια προσεταιριστική πράξη. Με άλλα λόγια, η συσσώρευση όλων των βέλτιστων κατευθύνσεων για όλα τα σημεία $\{\mathbf{x}_i\}_{i=1}^{N}$ και η εκτέλεση της μετακίνησης στο τέλος της kοστής επανάληψης (όπως απαιτεί ο φορμαλισμός τών μεθόδων GPS) παράγει το ίδιο αποτέλεσμα με τη λήψη καθενός από τα βήματα συντεταγμένων ξεχωριστά. Τέλος, η αναζήτηση μοτίβου MDS τερματίζεται όταν η παράμετρος μήκους βήματος $\Delta^{(k)}$ γίνεται μικρότερη από ένα προκαθορισμένο όριο.

5.6 Σύγκλιση του Pattern Search MDS

Τώρα που έχουμε εκφράσει τον αλγόριθμο Pattern Search MDS χρησιμοποιώντας το ενοποιημένο πλαίσιο GPS μπορούμε επίσης να χρησιμοποιήσουμε τα θεωρήματα που έχουν διατυπωθεί στην Ενότητα 2.8 για να αποδείξουμε τις ιδιότητες σύγκλισης του προτεινόμενου αλγορίθμου.

Πρώτα απ ’όλα, η αντικειμενική συνάρτηση g είναι πράγματι συνεχώς διαφοροποιήσιμη για όλες τις τιμές του χώρου αναζήτησης $\mathbb{R}^{N \cdot L}$ από τον ορισμό της στην Εξίσωση 5.5. Επιπλέον, η μήτρα προτύπων $\hat{\mathbf{P}}$ στην εξίσωση 5.11 περιέχει όλους τα πιθανά βήματα που παρέχονται από την ρουτίνα εύρεσης διερευνητικών κινήσεων. Έτσι, όλες οι ερευνητικές μας κινήσεις καθορίζονται από την Εξίσωση 2.65. Σε κάθε επανάληψη αξιολογούμε τα δοκιμαστικά βήματα παράλληλα με όλες τις συντεταγμένες για όλα τα σημεία $\mathbf{x}_i \in \mathcal{M},\ 1 \leq i \leq N$. Στο αναθεωρημένο ορισμό του προβλήματος (βλ. Ενότητα 5.5), αυτό μεταφράζεται σε αναζήτηση σε όλους τους πίνακες ταυτότητας $\mathbf{I}_{N \cdot L}$ και $-\mathbf{I}_{N \cdot L}$ του χώρου αναζήτησης $\mathbb{R}^{N \cdot L}$. Αλλά από τον ορισμό μας των πρώτων στηλών του γεννήτορα πίνακα μας στην Εξίσωση 5.9 αυτό αντιστοιχεί στον έλεγχο όλων των δυνατών βημάτων συντεταγμένων που παρέχονται από το $\hat{\mathbf{\Psi}} = [\mathbf{I}_{N \cdot L} \quad -\mathbf{I}_{N \cdot L}]$. Συνεπώς, εάν υπάρχει μια απλή μείωση όταν κινείται

προς οποιαδήποτε από τις κατευθύνσεις που παρέχονται από τις στήλες του $\hat{\mathbf{\Psi}}$ τότε ο αλγόριθμος μας παρέχει επίσης μια απλή μείωση. Αυτό το αποτέλεσμα επιβεβαιώνει ότι η Υπόθεση 1 ισχύει για τις εξερευνητικές κινήσεις. Συνδυάζοντας την διαφοροποίηση της αντικειμενικής μας συνάρτησης g και Hypothesis 1, το Θεώρημα 1 ισχύει για την αναζήτηση μοτίβου MDS. Ως εκ τούτου, είναι σίγουρο ότι $\lim_{k\to+\infty} \inf ||\nabla f(\mathbf{z}^{(k)})|| = 0$.

Προσπαθώντας να ενισχύσουμε περαιτέρω τις ιδιότητες σύγκλισης του προτεινόμενου αλγορίθμου, παρατηρούμε ότι πληρούνται οι περισσότερες από τις απαιτήσεις του Θεωρήματος 2 αλλά δεν πληρούμε τις προδιαγραφές της Υπόθεσης 2 για την ελάχιστη μείωση που παρέχεται από τις στήλες της γεννητόρου μήτρας $\hat{\mathbf{\Psi}}$. Ωστόσο, ο πίνακας δημιουργίας μας $\hat{\mathbf{C}} = [\hat{\mathbf{c}}_1, ..., \hat{\mathbf{c}}_{3^{N\cdot L}}]$ είναι πράγματι οριοθετημένος από τον κανόνα επειδή $||\hat{\mathbf{c}}_j||_1 \leq N \cdot L,\ \ 1 \leq j \leq 3^{N\cdot L}$. Μειώνοντας κατά το ήμισυ την παράμετρο μήκους βήματος για τις ανεπιτυχείς επαναλήψεις, διασφαλίζουμε επίσης ότι το $lim_{k\to+\infty} \Delta^{(k)}$. Προκειμένου να ικανοποιήσουμε τις προδιαγραφές του Θεωρήματος 2 θα χρειαζόμασταν μια τετραγωνική πολυπλοκότητα $\mathcal{O}((N\cdot L)^2)$ για να εξασφαλίσουμε ότι κάθε επανάληψη παρέχει την ίδια μείωση στην συνάρτηση g σαν τη μείωση που παρέχεται από τη «καλύτερη» στήλη του $\hat{\mathbf{\Psi}}$. Αυτό δηλώνεται επισήμως στο δεύτερο μέρος της Υπόθεσης 2. Αν τροποποιήσουμε τον αλγόριθμό μας για να ικανοποιήσουμε αυτές τις απαιτήσεις, δεν θα μπορούσαμε να εφαρμόσουμε όλες τις βελτιστοποιήσεις που προτάθηκαν στην Ενότητα 5.4 και ο συνολικός χρόνος εκτέλεσης θα αυξανόταν δραματικά.

5.7 Πειράματα

5.7.1 Ρύθμιση των υπερπαραμέτρων

Παρουσιάζουμε κάποιες οδηγίες σχετικά με τον τρόπο ρύθμισης των υπερπαραμέτρων για τον προτεινόμενο αλγόριθμο και την αναφορά των τιμών που χρησιμοποιούνται στα πειράματα που ακολουθούν. Πιό συγκεκριμένα:

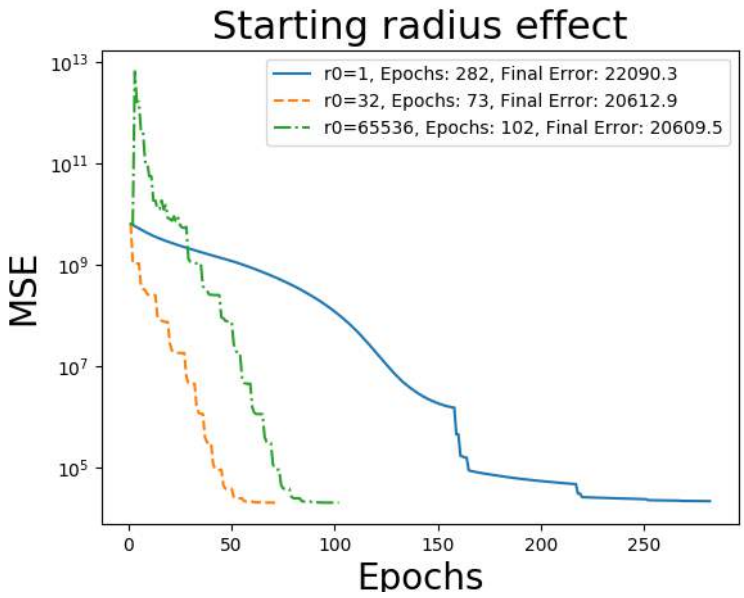
- Η σταθερή τιμή ϵ στη γραμμή 9 του Αλγορίθμου 2 καθορίζει πότε μειώνεται η ακτίνα εξερεύνησης r. Με τη ρύθμιση του ϵ σε μια τιμή πολύ κοντά στο 0, π.χ. 10^{-10}, η αναζήτηση θα πάρει περισσότερες εποχές, αλλά η λύση θα είναι πιο κοντά στο τοπικό βέλτιστο. Εάν χαλαρώσουμε το ϵ σε μια τιμή γύρω στο 10^{-2}, μπορούμε να κάνουμε μια πιό γενική εξερεύνηση του χώρου αναζήτησης που θα παράγει μια ακατέργαστη λύση σε ένα μικρότερο αριθμό εποχών. Στα πειράματά μας ορίσαμε το $\epsilon = 10^{-4}$ που παρέχει ένα καλό συνδυασμό μεταξύ της ποιότητας της λύσης και της γρήγορης σύγκλισης για τα σύνολα δεδομένων που χρησιμοποιούνται. Ωστόσο, η επιλογή αυτής της αξίας έγινε εμπειρικά.

- Πειραματικά διαπιστώσαμε ότι αν το L είναι μεγάλο, μπορούμε να ψάξουμε μόνο το 50% των διαστάσεων αναζήτησης και ακόμα να έχουμε μια καλή λύση, μειώνοντας σημαντικά τον χρόνο εκτέλεσης. Σε αυτό το πλαίσιο, δοκιμάζουμε τυχαία έναν νέο χώρο αναζήτησης για κάθε εποχή.

- Ο προτεινόμενος αλγόριθμος είναι σχετικά ισχυρός για την επιλογή του αρχικού μεγέθους της ακτίνας αναζήτησης κίνησης. Ωστόσο, η επιλογή του $r^{(0)}$ επηρεάζει την ταχύτητα σύγκλισης. Σε αυτό το πλαίσιο, Δείχνουμε τη σύγκλιση για ένα παράδειγμα διαδρομής της κλασικής Swissroll για διάφορες περιπτώσεις αρχικής ακτίνας. Προφανώς, η καλύτερη περίπτωση φαίνεται να είναι $r^{(0)} = 32$, ένας απαισιόδοξος θα είναι $r^{(0)} = 1$ και ένας αισιόδοξος: ($r^{(0)} = 65536$) αρχική ακτίνα στο σχήμα 5.2.

5.7.2 Ανακατασεκυή γεωμετρίας πολλαπλοτήτων

Η βασική υπόθεση στην πολλαπλότητα μάθηση είναι ότι τα δεδομένα εισόδου βρίσκονται σε μια χαμηλής διάστασης, μη γραμμική πολλαπλότητα, ενσωματωμένη σε ένα χώρο μεγάλης διαστάσεως. Επομένως, οι μη γραμμικές τεχνικές μείωσης των διαστάσεων αποσκοπούν στην εξαγωγή της πολλαπλότητας από τον χώρο των μεγάλων διαστάσεων. Για τα πειράματά μας χρησιμοποιούμε ποικίλα



Σχήμα 5.2: Σύγκλιση του Pattern Search MDS για διαφορετικές ακτίνες εκκίνησης

γεωμετρικά σχήματα και συγκρίνουμε τον Pattern Search MDS με άλλες καθιερωμένες τεχνικές μείωσης διαστάσεων.

Θα πρέπει να σημειωθεί ότι οι αλγόριθμοι που λύνουν το MDS με μήτρες ευκλείδιας απόστασης ως είσοδο, μπορεί να μην είναι σε θέση να συμπεράνουν γεωμετρία δεδομένων, οπότε πρέπει να παρέχουμε ως είσοδο ένα *Γεωδεσικό πίνακα αποστάσεων*. Αυτός ο πίνακας υπολογίζεται με την εκτέλεση του πιο σύντομου αλγορίθμου διαδρομής πάνω σε γράφους, τον αλγόριθμο Djikstra, αυτός ο αλγόριθμος τρέχει πανω στο γράφημα που δημιουργείται από όλα τα δεδομένα εισόδου. Για τα πειράματά μας, δοκιμάζουμε 3000 δείγματα σε 11 $3D$ σχήματα και τα μειώνουμε σε 2 διαστάσεις χρησιμοποιώντας τον Pattern Search MDS, τον SMACOF MDS [111], το περικομμένο SVD [177], το Isomap [145, 146, 147, 148, 149], το LLE [152, 153, 154, 155, 156], το Χεσιανό LLE [158, 159], το τροποποιημένο LLE [157] και το LTSA [166].

Οι μήτρες γεωδεσικής απόστασης που παρέχονται για τους αλγορίθμου που λύνουν το MDS υπολογίζονται χρησιμοποιώντας τον αλγόριθμο του Djikstra σε γράφους που απορέουν από ένα προυπολογισμό των KNN (πλησιέστερων γειτόνων) για κάθε σημείου του υψηλού σε διάσταση χώρου. Μια εκτεταμένη ανασκόπηση του αλγόριθμου KNN έχει δοθεί σε μια προηγούμενη ενότητα 2.2.4. Καταγράφουμε την ώρα που χρειάστηκε να εκτελεστεί κάθε μέθοδος. Είναι αξιοσημείωτο ότι η αναζήτηση μοτίβου MDS είναι ταχύτερη από το SMACOF MDS για όλα τα πειράματα που διεξάγουμε.

Παρουσιάζουμε τα 4 χαρακτηριστικά σχήματα που επιλέγονται από αυτά που δοκιμάσαμε και τα οποία βρίσκονται στον τρισδιάστατο χώρο $\mathbb{R}^3$ και επιδιώκουμε να βρούμε $2D$ πολλαπλότητες που διατηρούν τη γεωμετρία των δεδομένων. Στο Σχήμα 5.3 παρέχουμε τα αποτελέσματα που λαμβάνουμε από τις ανακατασκευασμένες πολλαπλότητες όλων των αλγορίθμων μείωσης διάστασιμότητας, καθώς και από το Pattern Search MDS για κάθε ένα από τα 4 επιλεγμένα σχήματα.

Στο πρώτο σχήμα, εξετάζουμε το κλασικό ελβετικό ρολό (swissroll), όπου ένα $2D$ επίπεδο είναι "περιτυλιγμένο" στον τρισδιάστατο χώρο και ο στόχος είναι να εξαχθεί το αρχικό $2D$ επίπεδο. Τα αποτελέσματα παρουσιάζονται στο Σχήμα 5.3a. Παρατηρούμε ότι οι γραμμικές τεχνικές μείωσης των διαστάσεων, όπως το περικομμένο SVD, δεν καταφέρνουν να ξεδιπλώσουν το swissroll. Επίσης το LLE εισάγει πολλή παραμόρφωση στο κατασκευασμένο επίπεδο. Ομοίως, στην Εικόνα 5.3c, το περικομένο SVD έχει επίσης ένα πρόβλημα λόγω της μη γραμμικής εμπλοκής των δεδομένων στο χώρο $3D$. Το LLE έχει επίσης πρόβλημα επειδή εισάγει καμπυλότητα στο επίπεδο λόγω του τοπικού συστήματος συντεταγμένων το οποίο είναι κατασκευασμένο για κάθε σημείο.

Στη συνέχεια, εξετάζουμε τον τρόπο με τον οποίο οι αλγόριθμοι χειρίζονται αραιωμένους πί-

νακες απόστασης. Για το σκοπό αυτό, παράγουμε ένα σύνολο δεδομένων μη συσσωρευμένων $3D$ συστάδων με μια γραμμή που συνδέει τα κεντροειδή, όπου ακολουθεί την ακεραιότητα της μήτρας απόστασης επειδή η μεγάλη πλειοψηφία των σημείων δειγματοληπτώνται πολύ στενά μέσα στις συστάδες. Μια καλή αντιστοίχιση θα πρέπει να διατηρήσει τη δομή του συμπλέγματος στις χαμηλότερες διαστάσεις. Στο Σχήμα 5.3b βλέπουμε ότι το SVD και η οικογένεια των αλγορίθμων MDS (προτεινόμενο, SMACOF, Isomap) παράγουν καλά αποτελέσματα, ενώ οι παραλλαγές LLE δεν μπορούν να χειριστούν πολύ καλά την ακεραιότητα των μήτρων απόστασης. Συγκεκριμένα, η Χεσιανό LLE και η LTSA δεν παράγουν κανένα αποτέλεσμα εξαιτίας της αριθμητικής αστάθειας. Συγκεκριμένα, στο Χεσιανό LLE, οι μήτρες που χρησιμοποιούνται για τον υπολογισμό του μηδενικού χώρου γίνονται μοναδικές, ενώ στο LTSA οι συντεταγμένες των σημείων που προκύπτουν γίνονται άπειρες καθώς επιδιώκουμε να διεξάγουμε την αποσύνθεση ιδιοτιμών. Το pattern search MDS δεν βασίζεται σε υπολογισμό ιδιοτιμών ή επίλυσης συστήματος εξισώσεων και ως εκ τούτου είναι πάντα αριθμητικά σταθερό.

Τέλος, παρουσιάζουμε πώς οι αλγόριθμοι λειτουργούν με μεταβάσεις από πυκνές σε αραιές περιοχές με σχήμα σπειροειδούς έλικας στο Σχήμα 5.3d. Μπορούμε να δούμε ότι πέντε μέθοδοι, συμπεριλαμβανομένης της αναζήτησης μοτίβου MDS, ξετυλίγουν το σχήμα στον αναμενόμενο κύκλο $2D$, ενώ το SVD παρέχει ένα σχήμα που μοιάζει με μαργαρίτα. Το Χεσιανό LLE και το LTSA προσαρμόζουν την έλικα σε πολλαπλούς κύκλους αλληλεπικάλυψης.

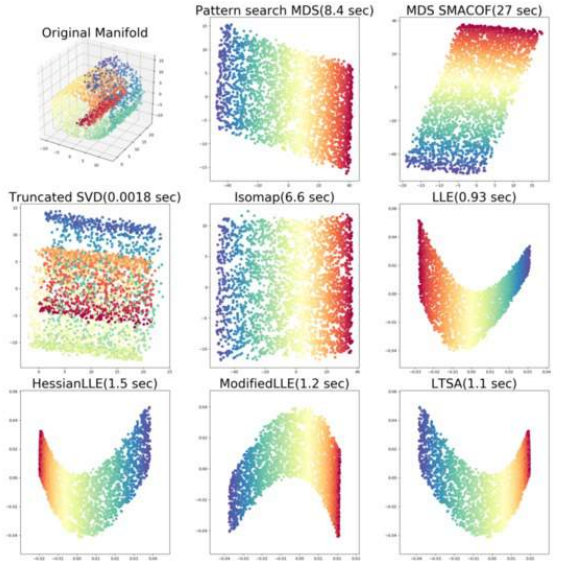
5.7.3 Σημασιολογική ομοιότητα

Η κατασκευή μοντέλων σημασιολογικών δικτύων συνίσταται στην απεικόνιση εννοιών (π.χ. λέξεων, ήχου κ.λπ.) σε ένα, ενδεχομένως, υψηλού επιπέδου, διάνυσμα του χώρου $\mathbb{R}^n$, όπου n ο αριθμός των χαρακτηριστικών κάθε δεδομένου. Οι σχέσεις μεταξύ των εννοιών προσδιορίζονται ποσοτικά ως οι αποστάσεις, ή αντιστρόφως, οι ομοιότητες συνημίτονων, μεταξύ σημασιολογικών φορέων. Το έργο σημασιολογικής ομοιότητας στοχεύει να αξιολογήσει τη συσχέτιση των ομοιοτήτων μεταξύ των εννοιών σε ένα δεδομένο σημασιολογικό χώρο έναντι ενός συνόλου τιμών ομοιότητας που παρέχονται από τους ανθρώπινους σχολιαστές.

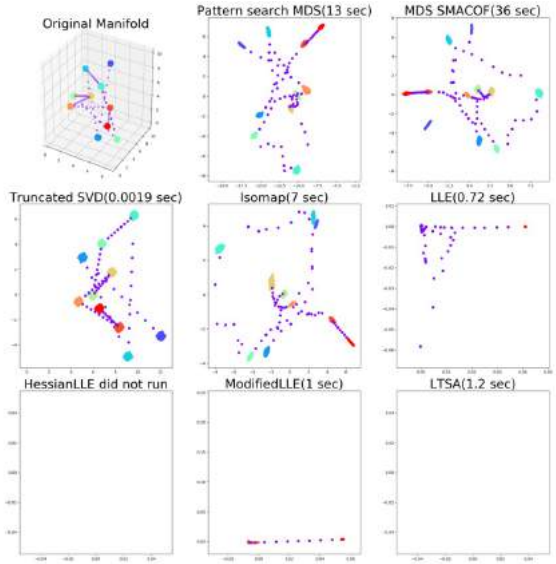
Αξιολογούμε την απόδοση των τεχνικών μείωσης διαστάσεων που διερευνήθηκαν επίσης στο τμήμα 5.7.2 και για την εργασία σημασιολογικής ομοιότητας. Χρησιμοποιούμε τα συναισθηματικά σύνολα δεδομένων MEN [178] και SimLex-999 [179] ως τις τιμές ομοιότητας που προκύπτουν από ανθρώπους και θεωρούνται αληθείς. Και τα δύο σύνολα δεδομένων παρέχονται με τη μορφή καταλόγων ζευγών λέξεων, όπου κάθε ζευγάρι συνδέεται με μια βαθμολογία ομοιότητας. Αυτό το σκορ υπολογίστηκε με το να βρούμε κατά μέσο όρο τις ομοιότητες που παρείχαν οι ανθρώπινοι σχολιαστές. Θεωρούμε αυτούς τους σχολιασμούς ως τις μόνες πραγματικές ετικέτες προκειμένου να αξιολογήσουμε και να συγκρίνουμε την αποτελεσματικότητα του αλγορίθμου μας για το πώς διατηρεί την ομοιότητα μεταξύ των ενσωματωμένων λέξεων στην ανακατασκευασμένη μαθηματική πολλαπλότητα. Οι αρχικοί μεγάλης διαστάσεως φορείς σημασιολογικής λέξης είναι 300-διαστάσεων GloVe διανύσματα που κατασκευάζονται στο [180] χρησιμοποιώντας μαι μεγάλη βάση δεδομένων από το Twitter. Μειώνουμε τη διαστατικότητα των διανυσμάτων στη διάσταση-στόχο $L < 300$ και υπολογίζουμε το συντελεστή συσχέτισης Spearman μεταξύ των ανθρωπίνων εκτιμήσεων και των αυτόματα υπολογισμένων ομοιότητων. Τα αποτελέσματα συνοψίζονται στον Πίνακα 5.1 για $L = 10$. Παρατηρούμε ότι το LLE αποδίδει τα καλύτερα αποτελέσματα για την βάση δεδομένων MEN, ενώ το Pattern Search MDS πάει καλύτερα στην βάση δεδομένων SimLex-999. Επιπλέον, παρατηρούμε ότι οι μη γραμμικές τεχνικές μείωσης των διαστάσεων μπορούν σε ορισμένες περιπτώσεις να βελτιώσουν σημαντικά την απόδοση των σημασιολογικών διανυσμάτων από ότι αν παίρναμε κατευθείαν τις αναπαραστάσεις στον μεγαλύτερο χώρο.

5.7.4 Ταξινόμηση εικόνων

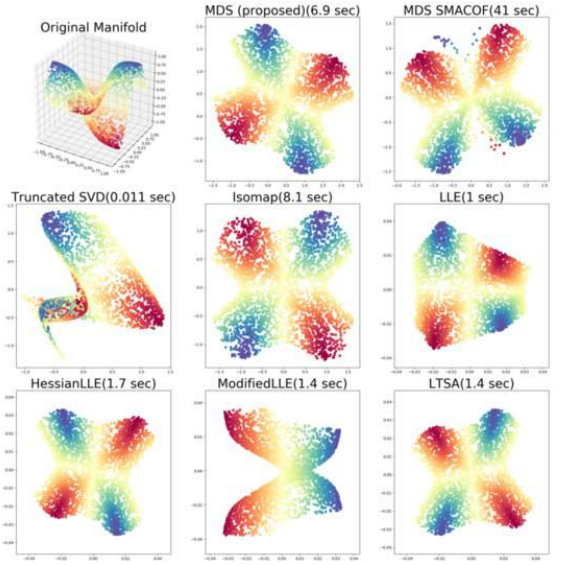
Το επόμενο σύνολο πειραμάτων στοχεύει στη σύγκριση του προτεινόμενου αλγορίθμου με άλλες μεθόδους μείωσης διαστάσεων για ταξινόμηση με KNN (βλ. Ενότητα 2.2.4 για μια εκτενή περιγραφή



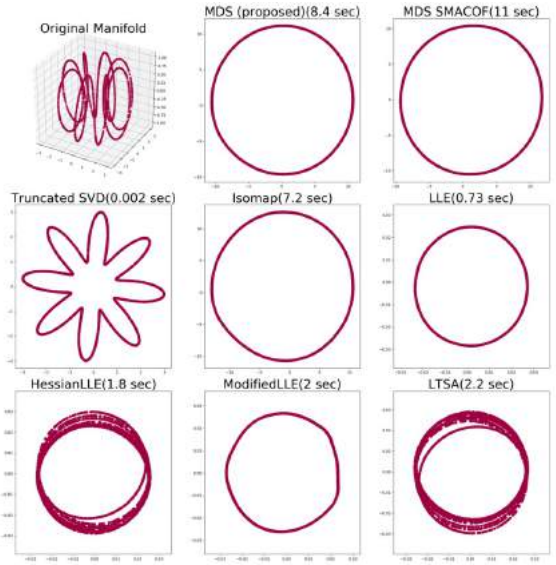
(a) Ελβετικό Ρολό (Swissroll)



(b) Συστοιχίες (Clusters) σημείων σε 3 διαστάσεις



(c) Δίδυμες κορυφές (Twin Peaks)



(d) Τοροειδής Έλικα (Toroid-Helix)

Σχήμα 5.3: Σύγκριση του Pattern Search MDS με άλλους αλγόριθμους μείωσης της διαστασικότητας για την ανακατασκευή $2D$ πολλαπλοτήτων από τεχνητά $3D$ δεδομένα εισόδου

Μέθοδος Μείωσης Διαστατικότητας	Διαστάσεις	**MEN**	**SimLex-999**
-	300	0.635	0.177
Pattern Search MDS	10	0.596	**0.242**
SMACOF	10	0.632	0.221
Isomap	10	0.625	0.132
SVD	10	0.562	0.140
LLE	10	**0.657**	0.172
Χεσσιανό LLE	10	0.157	0.004
Τροποποιημένο LLE	10	0.643	0.158
LTSA	10	0.154	0.004

Πίνακας 5.1: Σύγκριση του Pattern Search MDS με άλλους αλγόριθμους μείωσης της διαστασικότητας για πειράματα σημασιολογικής ομοιότητας με ενσωματωμένες λέξεις

της μη παραμετρικής ταξινόμησης KNN) χρησιμοποιώντας ένα σύνολο δεδομένων πραγματικής εικόνας. Επιλέγουμε να χρησιμοποιήσουμε το MNIST ως σύνολο δεδομένων αναφοράς το οποίο περιέχει $70,000$ εικόνες χειρόγραφων ψηφίων. Επιλέξαμε ένα τυχαίο υποσύνολο 1000 εικόνων και μειώσαμε τις διαστάσεις από 784 σε 20. Η απόδοση των μοντέλων αξιολογείται με ταξινόμηση $1-NN$ και τη χρήση διασταυρωμένης επικύρωσης 10-αναδιπλώσεων. Η μέτρηση αξιολόγησης είναι η μέση τιμή του F1 για κάθε κλάση ψηφίων. Ο πίνακας 5.2 συνοψίζει τα αποτελέσματα. Παρατηρήστε ότι η μείωση των διαστάσεων χρησιμοποιώντας την αναζήτηση μοτίβου MDS και SVD μπορεί να βελτιώσει την απόδοση ταξινόμησης σε σχέση με τα αρχικά δεδομένα μεγάλης διαστάσεως. Το Pattern Search MDS αποδίδει τα καλύτερα αποτελέσματα συνολικά. Το Χεσσιανό LLE, το τροποποιημένο LLE και το LTSA δεν εκτελέστηκαν λόγω της αριθμητικής αστάθειας.

Μέθοδος	διαστάσεις	**MNIST 1-NN F1 score**
MNIST	784	0.861
Pattern Search MDS	20	**0.878**
MDS SMACOF	20	0.857
Isomap	20	0.829
SVD	20	0.871
LLE	20	0.813
Χεσσιανό LLE	20	–
Τροποποιημένο LLE	20	–
LTSA	20	–

Πίνακας 5.2: Σύγκριση του Pattern Search MDS με άλλους αλγόριθμους μείωσης διαστάσεων για το σύνολο δεδομένων εικόνων MNIST

5.7.5 Αξιολόγηση της Ταχύτητας Σύγκλισης

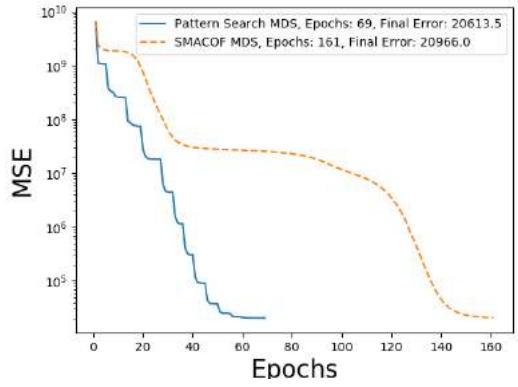
Στη συνέχεια, συγκρίνουμε την ταχύτητα σύγκλισης των Pattern Search MDS και SMACOF, με όρο τον αριθμό εποχών μέχρι να παράγουν την αντίστοιχη λύση τους. Για το σκοπό αυτό θα εξετάσουμε τα πειράματα των τμημάτων 5.7.2 και 5.7.3 και θα παρουσιάσουμε τα συγκριτικά γραφικά σύγκλισης.

Βλέπουμε τα διαγράμματα σύγκλισης για τις περιπτώσεις swissroll, $3D$ συστάδες, σπειροειδή έλικα στο σχήμα 5.4a, 5.4b και 5.4c, αντίστοιχα. Η γραφική παράσταση σύγκλισης για την εργασία σημασιολογικής ομοιότητας λέξης φαίνεται στο σχήμα 5.4d. Τα διαγράμματα παρουσιάζονται σε λογαριθμική κλίμακα του άξονα y, επειδή το σφάλμα εκκίνησης είναι πολλές τάξεις μεγέθους μεγαλύτερο από το τοπικό ελάχιστο που επιτυγχάνουν οι αλγόριθμοι.

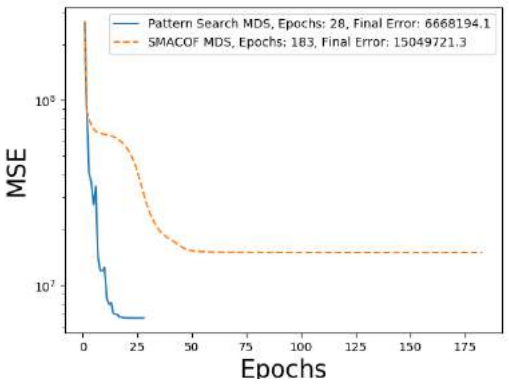
Για όλες τις περιπτώσεις, παρατηρούμε ότι το Pattern Search MDS συγκλίνει πολύ γρήγορα σε ένα παρόμοιο ή καλύτερα τοπικό βέλτιστο ενώ το SMACOF χτυπά περιοχές όπου η σύγκλιση επιβραδύνεται και στη συνέχεια ανακάμπτει. Αυτή η δομή τύπου πριονιού των διαγραμμάτων του προτεινόμενου αλγορίθμου οφείλεται στο γεγονός ότι επιτρέπουμε "κακές κινήσεις" όπως περιγράφηκε λεπτομερώς στην Ενότητα 5.4.1.

5.7.6 Ευστάθεια στον θόρυβο

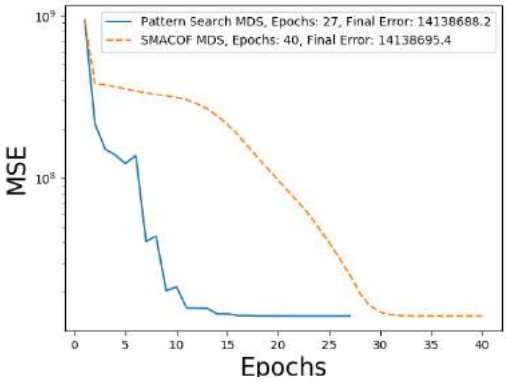
Σε αυτό το σύνολο πειραμάτων επιδιώκουμε να αποδείξουμε την ευρωστία του MDS αναζήτησης μοτίβου όταν τα δεδομένα εισόδου είναι κατεστραμμένα ή θορυβώδη. Για το σκοπό αυτό, εξετάζονται δύο περιπτώσεις καταστροφής δεδομένων: ο προσθετικός θόρυβος καθώς και τα ελλείποντα δεδομένα.



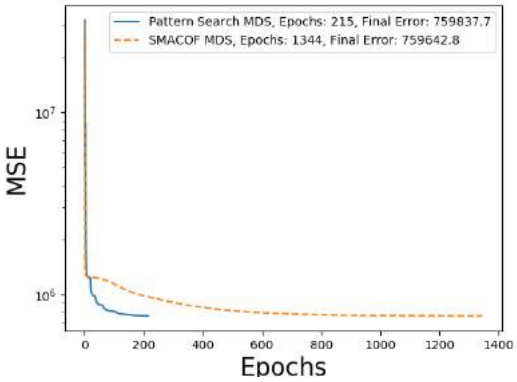
(a) Ελβετικό Ρολό (Swissroll)



(b) Συστοιχίες (Clusters) σημείων σε 3 διαστάσεις



(c) Τοροειδής Έλικα (Toroid-Helix)



(d) Πείραμα σημασιολογικής ομοιότητας

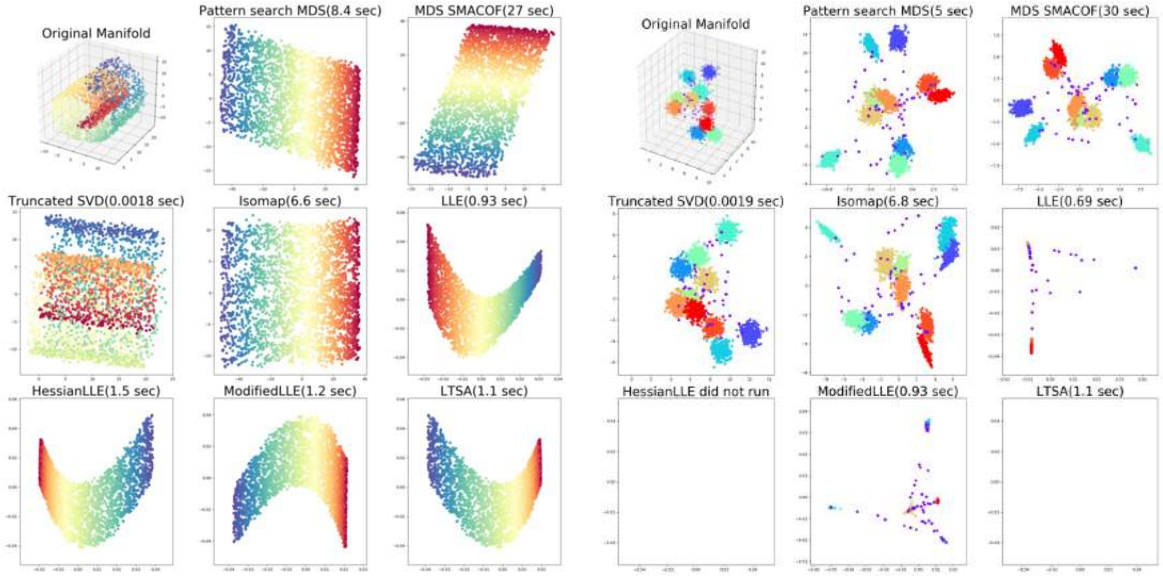
Σχήμα 5.4: Σύγκριση της ταχύτητας σύγκλισης του Pattern Search MDS με τον MDS SMACOF για την ανασυγκρότηση των γεωμετρικών σχημάτων και του πειράματος της σημασιολογικής ομοιότητας

Ευρωστία στον πρόσθετο θόρυβο

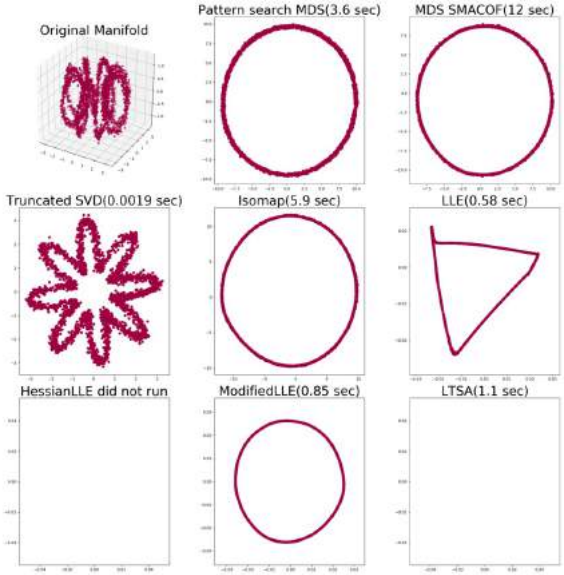
Σε αυτό το σύνολο πειραμάτων, εισάγουμε Γκαουσιανό θόρυβο τυπικής απόκλισης (σ) στα δεδομένα εισόδου και χρησιμοποιούμε τον πίνακα ανομοιότητας που υπολογίζεται στα θορυβώδη δεδομένα ως είσοδο σε κάθε έναν από τους αλγόριθμους μείωσης των διαστάσεων που αξιολογούνται.

Συγκεκριμένα, για τα συνθετικά δεδομένα του τμήματος 5.7.2, ακολουθούμε μια ποιοτική αξιολόγηση, δείχνοντας τις ανακατασκευασμένες πολλαπλότητες για υψηλά επίπεδα θορύβου. Εκτελούμε μείωση διαστασιολόγησης για συστοιχίες swissroll, τοροειδούς έλικας και $3D$ συστάδων για αυξημένα επίπεδα θορύβου. Αναφέρουμε τα αποτελέσματα για την υψηλότερη πιθανή απόκλιση θορύβου όπου μία ή περισσότερες τεχνικές παράγουν ακόμη σημαντικές πολλαπλότητες. Πέρα από αυτές τις τιμές του σ οι παραγώμενες πολλαπλότητες αλλοιώνονται και η παραγωγή όλων των μεθόδων κυριαρχείται από θόρυβο. Στις εικ. 5.5a, 5.5b, 5.5c δείχνουμε τα αποτελέσματα για θορυβώδη swissroll με $\sigma = 0.3$, $3D$ συστάδες με $\sigma = 0.4$ και τοροειδή έλικα με $\sigma = 0.07$ αντίστοιχα. Συνολικά, το Pattern Search MDS, ακολουθούμενη από τα SMACOF και Isomap, είναι πιο ανθεκτικά στον πρόσθετο θόρυβο.

Για τα πειράματα με την έγχυση του θορύβου για το πείραμα της σημασιολογικής ομοιότητας, εισάγουμε διαφορετικά επίπεδα Γκαουσιανού θορύβου στα αρχικά εμβυθισμένα διανύσματα λέξεων και αξιολογούμε την συσχέτιση τόσο σε MEN όσο και σε σύνολα δεδομένων Simlex-999 (τα ίδια με αυτά που χρησιμοποιήσαμε στα πειράματα με τα καθαρά διανύσματα ενσωματωμένων λέξεων στο τμήμα 5.7.3). Τα αποτελέσματα παρουσιάζονται στον Πίνακα 5.3. Παρατηρούμε ότι η σχετική απόδοση των αλγορίθμων διατηρείται με την πρόσθεση Γκαουσιανού θορύβου, εκτός από την μέθοδο



(a) Ελβετικό Ρολό + Γκαουσιανός Θόρυβος **(b)** Συστάδες + Γκαουσιανός Θόρυβος



(c) Τοροειδής Έλικα + Γκαουσιανός Θόρυβος

Σχήμα 5.5: Σύγκριση του Pattern Search MDS με άλλους αλγόριθμους μείωσης διαστάσεων για την ανακατασκευή $2D$ πολλαπλοτήτων από $3D$ τεχνητά δεδομένα με πρόσθεση θορύβου

LLE που δεν μπορεί να χειριστεί υψηλές ποσότητες θορύβου. Το LLE επιτυγχάνει τις καλύτερες τιμές συσχετισμού στους MEN στα $\sigma = 0.01$ και $\sigma = 0.1$, ενώ το Pattern Search MDS επιτυγχάνει την καλύτερη απόδοση στο Simlex-999.

Ευστάθεια στην απώλεια δεδομένων

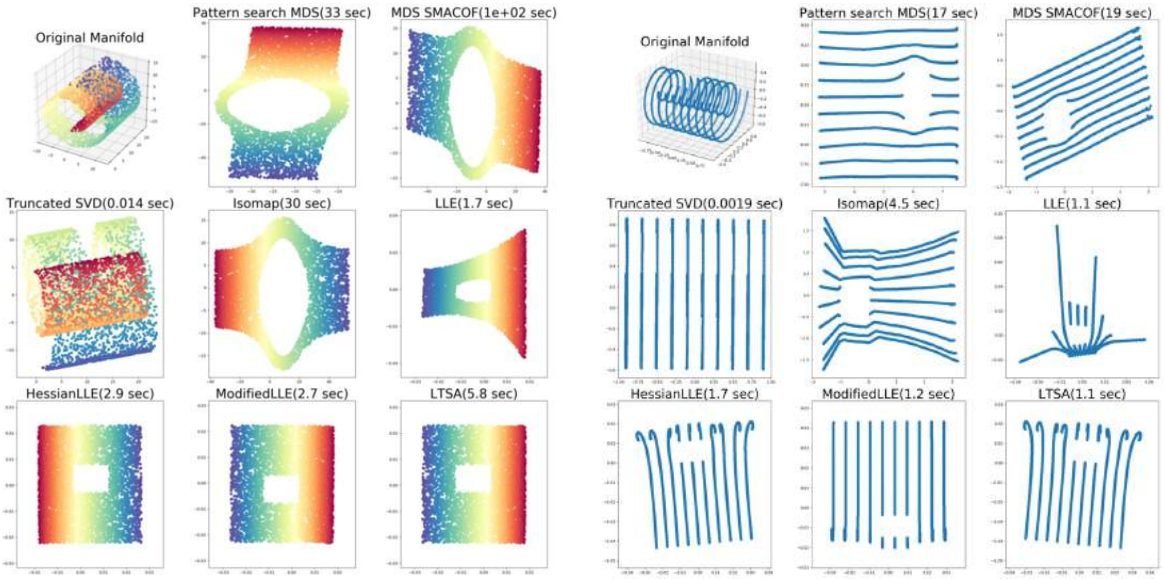
Σε πολλά πραγματικά σενάρια, ένα μέρος των δεδομένων ενδέχεται να μην είναι διαθέσιμο κατά τη διάρκεια της εκπαίδευσης και να λείπει. Πρέπει να εκτιμήσουμε την ευρωστία του MDS αναζήτησης μοτίβου όταν ένα μέρος των δεδομένων δεν είναι διαθέσιμο.

Για αυτό το πείραμα, δημιουργούμε δύο νέα συνθετικά σύνολα δεδομένων, δηλαδή ένα πυκνό και αραιό ελβετικό ρολό με μια τρύπα όπως φαίνεται στο Σχήμα 5.6. Στην εικόνα 5.6a, δείχνουμε την απόδοση των διαφόρων αλγορίθμων που εφαρμόζονται σε ένα πυκνό swissroll με μια τρύπα στη μέση. Όπως μπορούμε να δούμε μόνο το Χεσιανό LLE, τα τροποποιημένα LLE και LTSA είναι σε θέση να

Μέθοδος	Διαστάσεις	**MEN**			**SimLex-999**		
		$\sigma = 0.01$	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 0.01$	$\sigma = 0.1$	$\sigma = 0.5$
GloVe	300	0.635	0.619	0.431	0.178	0.169	0.077
Pattern Search MDS	10	0.593	0.597	0.462	**0.249**	**0.315**	**0.204**
MDS SMACOF	10	0.633	0.620	0.462	0.229	0.222	0.123
Isomap	10	0.622	0.613	**0.497**	0.134	0.124	0.079
SVD	10	0.562	0.551	0.380	0.140	0.136	0.039
LLE	10	**0.659**	**0.649**	0.369	0.175	0.166	0.052
Χεσσιανό LLE	10	0.156	0.144	0.023	0.005	0.04	0.018
Τροποποιημένο LLE	10	0.635	0.633	0.489	0.158	0.162	0.096
LTSA	10	0.155	0.141	0.020	0.06	0.04	0.002

Πίνακας 5.3: Σύγκριση της αναζήτησης MDS μοτίβου με άλλες μεθόδους μείωσης διαστάσεων για το πείραμα της σημασιολογικής ομοιότητας χρησιμοποιώντας θορυβωποιημένα διανύσματα ενσωμάτωσης λέξεων

αναδημιουργήσουν σωστά το σχήμα, ενώ οι αλγόριθμοι MDS οδηγούν σε παραμόρφωση γύρω από την τρύπα. Αυτό οφείλεται στην μη κυρτότητα που παρουσιάσαμε στο χώρο όταν προσθέσαμε την τρύπα. Αυτή η παραμόρφωση μπορεί ακόμα να παρατηρηθεί (σε μικρότερο βαθμό) στην αραιή διακύμανση που παρουσιάζεται στο Σχήμα 5.6b. Για την περίπτωση αραιών δεδομένων, παρατηρούμε ότι οι μέθοδοι LLE έχουν ως αποτέλεσμα παραμόρφωση γύρω από τις άκρες.



(a) Ελβετικό Ρολό με τρύπα **(b)** Σπειροειδές με τρύπα

Σχήμα 5.6: Σύγκριση του Pattern Search MDS με άλλους αλγόριθμους μείωσης της διαστασικότητας για σχήματα με τρύπες και μη κυρτές περιοχές

Αυτά τα προκαταρκτικά πειράματα υποδεικνύουν ότι οι παραλλαγές του LLE μπορούν να χειριστούν καλύτερα μη-κυρτότατα δεδομένα εισόδου, ενώ οι αλγόριθμοι MDS μπορούν να χειριστούν καλύτερα τα αραιά δεδομένα. Αυτό οφείλεται στο γεγονός ότι οι μέθοδοι LLE βασίζονται στο συμπέρασμα και στη συνδυασμένη γεωμετρία τοπικών δεδομένων, ενώ οι μέθοδοι MDS υπονοούν την παγκόσμια γεωμετρία.

5.7.7 Μείωση των διαστάσεων για πειράματα Ανεξαρτήτου-Ομιλητή (SI) SER

Στην ενότητα αυτή, θα συγκρίνουμε το Pattern Search MDS για τη μείωση της διάστασης των ακουστικών σύνολα δυνατοτήτων που παρουσιάζονται στα τμήματα 4.5.1 (IS10 σύνολο Χαρακτηριστικό γνώρισμα: 1582 Χαρακτηριστικά), 4.5.2 (RQA χαρακτηριστικό που 432 χαρακτηριστικά) και 4.5.3 (Συντηγμενο σύνολο Χαρακτηριστικών: 2014 Χαρακτηριστικά). Αξιολογούμε τον προτεινόμενο αλγόριθμο NLDR σε σύγκριση με άλλους αλγόριθμους για SER κάτω από πειραματική ρύθμιση Speaker Independent (SI) σε επίπεδο ολόκληρης της ομιλίας. Για μια πιο εκτενή περιγραφή αυτής της ρύθμισης αναφερόμαστε στην προηγούμενη ενότητα 4.6.1. Για τα πειράματά μας χρησιμοποιούμε την EmoDB [121] συναισθηματική βάση δεδομένων, όπως περιγράφεται στην ενότητα 4.7.1 αποτελείται από 7 συναισθήματα. Επιδιώκουμε να μειώσουμε τις διαστάσεις αυτών των παραστάσεων, χωρίς ταυτόχρονα να χάσουμε μεγάλο μέρος της περιγραφικής φύσης των χαρακτηριστικών εισόδου. Προκειμένου να προσδιοριστεί η πραγματική διακριτική ικανότητα των χαρακτηριστικών εισόδου, χρησιμοποιούμε ένα μη-παραμετρικό μοντέλο προκειμένου να αξιολογήσουμε τη μαθηματική πολυδιάστατη πολλαπλότητα $\mathcal{M}$. Συγκεκριμένα, χρησιμοποιούμε το KNN (δείτε το τμήμα 2.2.4) με διάφορους αριθμούς γειτόνων για να συμπεράνουμε την ετικέτα για ένα άγνωστο δείγμα. Στην ουσία, προκειμένου να αξιολογηθεί η ακρίβεια των παραστάσεων χαμηλών διαστάσεων πολλαπλότητα, χρησιμοποιούμε μια ποικιλία τιμών για την παράμετρο K των πλησιέστερων γειτόνων που έχουμε λάβει υπόψη κατά τη διάρκεια της αξιολόγησης. Δηλαδή επιλέγουμε το K να βρίσκεται στο σύνολο: $K \in \{1, 5, 9, 13, 17, 21\}$. Στη συνέχεια, η άγνωστη ετικέτα συνάγεται από τις αντίστοιχες ετικέτες των γειτόνων της. Αναφέρουμε τόσο τη μέτρηση επιδόσεων στις μετρικές WA και UA, όσο και την προηγούμενη πειραματική ρύθμιση που παρουσιάζεται στην ενότητα 4.7.3. Συνολικά, για κάθε πείραμα που περιγράφεται παρακάτω, τρέχουμε κάθε αλγόριθμο μείωσης των διαστάσεων χωρίς να γνωρίζουμε τις ετικέτες του συνόλου δεδομένων. Στη συνέχεια, για κάθε αναδίπλωση του σχήματος αξιολόγησης θεωρούμε έναν ομιλητή ως ομιλητή δοκιμής και όλους τους άλλους για εκπαίδευση. Το KNN εφαρμόζεται με τις ετικέτες εκπαίδευσης που είναι αυτές των ομιλητών από το σετ εκπαίδευσης. Αναφέρουμε τη μέση ακρίβεια που επιτυγχάνεται από κάθε αναδίπλωση όταν έχουν εκτιμηθεί όλες οι ομιλίες από όλους τους ομιλητές χωρίς να ληφθεί μέριμνα για τον αριθμό των δειγμάτων για κάθε ομιλητή.

Χρησιμοποιώντας το σύνολο χαρακτηριστικών IS10

Στον Πίνακα 5.4 παρουσιάζουμε τα αποτελέσματα από μια ποικιλία αλγορίθμων μείωσης διαστάσεων όταν χρησιμοποιούμε το σύνολο χαρακτηριστικών IS10 για το SI SER για μια ποικιλία μείωσης της διάστασης στόχου L και τον αριθμό των πλησιέστερων γειτόνων K. Το *Σύνολο IS10* κάτω από τη μέθοδο στήλης, σημαίνει ότι δεν πραγματοποιείται καμία άλλη μείωση διαστάσεων και το KNN εφαρμόζεται απευθείας πάνω στο χώρο των 1582 χαρακτηριστικών. Για τη σειρά του πίνακα με μηδενικά, ο αντίστοιχος αλγόριθμος μείωσης διαστάσεων δεν παρήγαγε κανένα αποτέλεσμα χρησιμοποιώντας τις προεπιλεγμένες παραμέτρους. Για κάθε διάσταση στόχου $L \in \{2, 10, 25\}$ σημειώνουμε με έντονους χαρακτήρες τα υψηλότερα αποτελέσματα που λαμβάνονται κατά τη δοκιμή όλων των μειώσεων των διαστάσεων κάτω από έναν καθορισμένο αριθμό K επιλεγμένων πλησιέστερων γειτόνων. Επισημαίνουμε, επίσης, με το **μπλέ** τα καλύτερα αποτελέσματα που προκύπτουν από όλες τις δυνατές διαμορφώσεις των μεθόδων μείωσης των διαστάσεων και από τους K επιλεγμένους πλησιέστερους γείτονες όταν ψάχνουμε για τη συμπερίληψη ενός άγνωστου δείγματος από το σετ δοκιμών.

Είναι προφανές ότι αυτό το σετ χαρακτηριστικών προσφέρει αρκετά περιγραφικές αναπαραστάσεις σε διάφορους ομιλητές για την αναγνώριση συναισθηματικών εκδηλώσεων. Χρησιμοποιώντας KNN από αυτό το χώρο χαρακτηριστικών παράγει αποτελέσματα μέχρι 69.9% σε WA και 64.1% σε UA. Ωστόσο, μπορούμε να επιτύχουμε παρόμοιες επιδόσεις με το αρχικό σύνολο χαρακτηριστικών χρησιμοποιώντας μόνο 10 διαστάσεις ή ακόμα και να ξεπεράσουμε αυτή την απόδοση και στις δύο μετρικές WA και UA χρησιμοποιώντας μόνο 25 διαστάσεις αντί να χρησιμοποιήσουμε 1582 όταν χρησιμοποιούμε τον κατάλληλο αλγόριθμο μείωσης των διαστάσεων πριν εφαρμόσουμε το KNN. Στις 2 διαστάσεις, οι αλγόριθμοι με τις καλύτερες επιδόσεις για τη μείωση της διαστασιολόγησης είναι η

Μέθοδος	L	Αριθμός Πλησιεστέρων Γειτόνων											
		WA						UA					
		1	5	9	13	17	21	1	5	9	13	17	21
Σύνολο IS10	1582	61.0	68.8	69.7	69.1	69.9	69.9	58.2	62.2	64.1	62.4	62.7	62.5
Pattern Search MDS	2	43.4	44.9	46.2	46.5	49.5	49.9	40.9	40.0	41.1	41.2	44.9	44.9
Modified LLE		40.4	44.7	46.3	48.5	49.6	49.7	40.3	41.1	42.1	44.4	44.6	42.8
Spectral Clustering		42.4	**49.2**	**51.8**	**53.1**	50.9	**52.5**	37.3	42.2	45.3	**47.3**	43.8	44.8
LTSA		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MDS SMACOF		39.7	46.7	44.7	47.0	48.7	49.5	37.4	42.5	38.9	41.5	43.0	43.8
Truncated SVD		41.7	44.6	44.8	46.8	47.1	49.3	39.7	41.0	40.6	41.8	42.1	44.1
Hessian LLE		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ISOMAP		**44.6**	48.4	49.7	50.2	48.7	49.0	**41.2**	43.5	44.9	44.0	41.3	41.6
LLE		41.3	47.6	51.1	51.2	**51.6**	52.1	38.6	**44.0**	**45.7**	46.3	**47.4**	**46.1**
Pattern Search MDS	10	55.2	64.5	64.7	63.2	65.7	64.3	50.5	58.4	**58.6**	57.7	59.9	58.0
Modified LLE		**59.6**	63.4	62.4	63.2	63.8	64.4	54.8	56.0	55.6	56.1	57.3	57.1
Spectral Clustering		59.4	64.1	**66.1**	65.4	**67.4**	66.1	**55.7**	58.5	60.8	58.9	**61.2**	59.2
LTSA		48.5	49.1	52.5	53.2	53.8	53.2	43.9	43.4	46.6	47.9	47.2	47.0
MDS SMACOF		53.6	62.9	64.2	63.7	64.7	**67.3**	49.3	56.3	57.1	57.5	58.6	**60.9**
Truncated SVD		56.3	63.0	64.1	**66.1**	64.6	64.6	52.3	56.9	57.7	**59.7**	58.8	58.1
Hessian LLE		48.8	49.3	52.5	53.2	53.8	53.2	44.1	43.6	46.6	47.9	47.2	47.0
ISOMAP		57.6	**64.9**	64.7	64.2	64.5	62.9	52.3	**59.1**	56.7	56.3	56.7	55.2
LLE		54.3	58.2	57.8	59.5	58.9	60.5	49.4	53.4	51.7	53.5	52.8	54.4
Pattern Search MDS	25	59.0	66.3	69.3	**70.4**	69.7	71.0	54.9	59.3	62.4	**63.5**	62.2	63.8
Modified LLE		55.2	58.7	65.3	67.0	66.2	66.4	51.2	53.4	62.9	61.0	63.2	62.7
Spectral Clustering		54.2	60.2	61.2	61.7	61.7	62.9	51.7	55.7	55.9	56.5	55.2	57.2
LTSA		55.2	57.0	60.2	59.1	62.2	59.2	50.6	50.3	53.1	51.8	54.9	52.3
MDS SMACOF		**62.2**	**69.9**	**69.8**	69.4	**71.4**	**71.3**	**57.0**	**62.9**	**62.5**	62.7	**65.5**	**65.5**
Truncated SVD		61.6	66.5	68.4	67.2	66.4	68.9	56.2	60.1	61.3	60.6	60.1	61.6
Hessian LLE		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ISOMAP		58.9	64.5	65.9	65.3	67.0	66.0	54.1	59.0	58.3	57.5	58.5	58.6
LLE		54.9	58.0	60.8	62.5	62.2	63.4	50.1	52.9	55.4	56.9	57.1	57.5

Πίνακας 5.4: Σύγκριση της μεθόδου αναζήτησης MDS με άλλες μεθόδους μείωσης διαστάσεων για πειράματα Ανεξαρτήτου-Ομιλητή για αναγνώριση Συναισθήματος από φωνή με τη χρήση των χαρακτηριστικών IS10

φασματική συσταδοποιήση για WA και LLE για UA στις περισσότερες περιπτώσεις. Οι πολλαπλότητες σε 2 διαστάσεις που ελήφθησαν από την προηγούμενη μέθοδο επιτυγχάνουν WA έως 53.1% όταν χρησιμοποιούμε $K = 13$ ενώ όταν ψάχνουμε $K = 17$ γείτονες πάνω από τις πολλαπλότητες της τελευταίας μεθόδου επιτυγχάνουμε ακρίβεια 47.4% στην UA. Επιπλέον, όταν αυξάνουμε τις διαστάσεις στόχου στις 10 παρατηρούμε ότι τα καλύτερα αποτελέσματα σε WA αποκτώνται και πάλι όταν χρησιμοποιούμε Φασματική Συσταδοποιήση με $K = 17$ (67.4%) αλλά και σε UA με $K = 17$ (61.2%). Αν και για διαφορετικές τιμές των K γειτόνων, τα καλύτερα αποτελέσματα λαμβάνονται με διαφορετικές μεθόδους. Όταν αυξάνουμε περαιτέρω τις διαστάσεις στόχου στις 25, έχουμε την καλύτερη απόδοση σε WA με $K = 17$ (71.4%) και σε UA με $K = 17$ (65.5%) όταν χρησιμοποιούμε SMACOF. Οι προαναφερθείσες βαθμολογίες απόδοσης ξεπερνούν την καλύτερη απόδοση του αρχικού χώρου των 1582 χαρακτηριστικών από 1.5% και 2.8% σε WA και UA, αντίστοιχα. Αυτά τα αποτελέσματα

υποδεικνύουν ότι τα χαρακτηριστικά εισόδου μπορούν να περιγραφούν κατάλληλα με πολλαπλότητες χαμηλότερων διαστάσεων ενώ εξακολουθούν να παρέχουν διακριτική ικανότητα για SI SER.

Χρησιμοποιώντας το σύνολο χαρακτηριστικών RQA

Στον Πίνακα 5.5 παρουσιάζουμε τα αποτελέσματα από μια ποικιλία αλγορίθμων μείωσης διαστάσεων όταν χρησιμοποιούμε το σύνολο χαρακτηριστικών RQA που έχει οριστεί για το SI SER για μια ποικιλία παραμετροποιήσεων της διάστασης στόχου L και τον αριθμό των πλησιέστερων γειτόνων K. Η σημείωση είναι παρόμοια με αυτή που εξηγείται προηγουμένως.

Μέθοδος	L	Number of Nearest Neighbors											
		WA						UA					
		1	5	9	13	17	21	1	5	9	13	17	21
Σύνολο RQA	432	51.8	56.5	56.2	56.5	56.9	55.4	47.3	48.4	47.8	47.1	48.1	46.9
Pattern Search MDS	2	34.8	40.6	39.4	43.1	43.5	44.1	29.4	32.4	31.3	34.1	35.7	34.7
Modified LLE		35.5	40.5	39.7	41.4	41.3	40.8	30.0	34.4	33.2	36.1	36.6	35.9
Spectral Clustering		37.6	42.1	43.6	44.9	45.6	**47.9**	35.1	36.2	**36.7**	38.4	38.5	**41.5**
LTSA		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MDS SMACOF		32.6	40.7	38.4	41.2	42.7	42.9	27.7	34.3	30.9	32.7	33.9	33.7
Truncated SVD		29.2	42.8	43.5	43.7	45.0	45.2	25.9	**36.7**	35.1	35.9	37.0	39.0
Hessian LLE		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ISOMAP		**39.8**	**43.1**	**44.8**	**48.2**	**47.5**	**47.9**	**35.4**	35.5	36.1	**40.1**	**38.7**	39.1
LLE		31.8	34.3	37.1	38.3	37.8	38.5	28.4	27.2	28.9	29.8	28.9	29.2
Pattern Search MDS	10	48.1	57.5	**58.5**	**59.9**	**60.0**	**59.4**	41.5	48.7	**51.3**	**52.7**	**52.5**	**51.4**
Modified LLE		48.6	54.0	57.0	55.5	54.4	54.2	44.5	45.6	49.5	47.1	44.9	43.6
Spectral Clustering		47.3	53.3	52.7	54.6	55.4	55.2	42.9	44.9	44.0	47.5	47.1	47.5
LTSA		34.9	41.1	43.3	43.2	45.2	43.9	28.5	33.7	35.7	34.6	36.9	35.6
MDS SMACOF		47.6	55.8	57.0	57.5	57.7	57.5	41.6	46.8	48.5	49.1	49.7	48.5
Truncated SVD		47.6	54.3	53.5	54.4	55.1	53.5	43.2	46.6	47.0	44.6	45.6	44.2
Hessian LLE		34.9	41.1	43.3	43.2	45.2	43.9	28.5	33.7	35.7	34.6	36.9	35.6
ISOMAP		48.1	51.3	52.2	52.5	51.9	52.6	43.6	42.8	42.7	44.5	43.6	44.3
LLE		**51.2**	**58.0**	57.1	57.6	58.8	58.2	**45.2**	**48.8**	47.5	49.5	50.4	48.0
Pattern Search MDS	25	**51.2**	56.0	**58.6**	**58.5**	**58.8**	58.1	45.1	48.6	**49.8**	49.3	**50.9**	48.7
Modified LLE		46.8	55.2	55.8	55.7	56.1	56.7	41.5	49.2	49.1	48.8	48.8	49.7
Spectral Clustering		49.3	55.1	54.0	57.2	56.0	56.6	42.8	47.7	45.9	48.3	48.3	49.2
LTSA		50.0	53.5	54.0	53.9	56.5	56.2	44.3	45.0	45.7	45.6	47.4	46.9
MDS SMACOF		50.5	**58.1**	57.3	57.4	56.7	57.2	**45.6**	**51.1**	48.3	48.6	46.8	47.7
Truncated SVD		49.5	57.4	57.9	57.9	57.8	58.5	42.7	49.6	49.2	48.9	48.8	51.5
Hessian LLE		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ISOMAP		50.4	54.1	53.1	52.6	53.8	54.0	44.1	46.9	46.1	44.3	45.1	46.4
LLE		49.5	52.8	55.8	56.3	56.6	**58.7**	44.5	46.1	49.4	**50.0**	49.6	**52.0**

Πίνακας 5.5: Σύγκριση της μεθόδου αναζήτησης MDS με άλλες μεθόδους μείωσης διαστάσεων για πειράματα Ανεξαρτήτου-Ομιλητή για αναγνώριση Συναισθήματος από φωνή με τη χρήση των χαρακτηριστικών RQA

Συγκρίνοντας το σύνολο χαρακτηριστικών RQA με το σύνολο χαρακτηριστικών IS10 συμπεραίνουμε ότι το πρώτο παρέχει λιγότερες πληροφορίες για τα συναισθήματα, το οποίο είναι παρόμοιο

με τα ευρήματα στο τμήμα 4.7.3 για SI SER. Χρησιμοποιώντας KNN από αυτό το χώρο χαρακτηριστικών παράγει αποτελέσματα μέχρι 56.9% σε WA και 48.4% σε UA. Ωστόσο, μπορούμε εύκολα να ξεπεράσουμε αυτή την απόδοση τόσο σε WA όσο και σε UA χρησιμοποιώντας μόνο 10 διαστάσεις, αντί να χρησιμοποιήσουμε τις αρχικές 432, όταν χρησιμοποιούμε τον κατάλληλο αλγόριθμο μείωσης των διαστάσεων πριν εφαρμόσουμε το KNN. Στις 2 διαστάσεις, ο αλγόριθμος με τις καλύτερες επιδόσεις για τη μείωση των διαστάσεων φαίνεται να είναι ο LLE για τις μετρικές WA και UA στις περισσότερες περιπτώσεις εκτός από μερικές περιπτώσεις που η φασματική συσταδοποίηση παρουσιάζει ελαφρώς καλύτερη απόδοση. Οι πολλαπλότητες σε 2 διαστάσεις που αντλήθηκαν από τη μέθοδο LLE επιτυγχάνουν WA έως 48.2% όταν χρησιμοποιούν $K = 13$. Από την άλλη πλευρά, το Spectral Clustering παρέχει τα καλύτερα αποτελέσματα για UA (41.5%) όταν ψάχνουμε $K = 21$ γείτονες πάνω στις πολλαπλότητες της τελευταίας μεθόδου. Δεν πρέπει να αγνοούμε το γεγονός ότι όταν αυξάνουμε σημαντικά τον αριθμό των γειτόνων δεν παρέχουμε χρήσιμα αποτελέσματα για τη διακριτική ικανότητα των χαρακτηριστικών, ακόμη και αν επιτύχουμε υψηλότερα αποτελέσματα. Στην ουσία, όταν αυξάνουμε δραματικά τους γείτονες που αναζητούμε τα δεδομένα, οι αναπαραστάσεις μπορεί να συρρικνωθούν μεταξύ τους και ακόμα να επιτύχουν καλύτερα αποτελέσματα. Αυτό σημαίνει ότι ένα πιο εξελιγμένο παραμετρικό μοντέλο θα μπορούσε ακόμα να μην είναι σε θέση να συμπεράνει τις γεωμετρίες για να ταξινομήσει κατάλληλα δείγματα από το σετ δοκιμών.

Επιπλέον, όταν αυξάνουμε τις διαστάσεις στόχου στις 10, παρατηρούμε ότι επιτυγχάνουμε τα καλύτερα αποτελέσματα τόσο σε WA όσο και σε UA όταν χρησιμοποιούμε τον προτεινόμενο αλγόριθμο Patter Search MDS. Συγκεκριμένα, έχουμε WA έως 60.0% με $K = 17$ καθώς και 52.7% σε UA με $K = 13$. Για τιμές $K = 1$ και $K = 5$ το LLE φαίνεται να καταγράφει σωστά τη γεωμετρία των δεδομένων. Αυτές οι βαθμολογίες απόδοσης ξεπερνούν σημαντικά τις καλύτερες βαθμολογίες απόδοσης του αρχικού χώρου χαρακτηριστικών των 432 διαστάσεων από 3.1% και 2.3% σε WA και UA, αντίστοιχα. Σε αυτό το πλαίσιο, παρατηρούμε το αντίθετο φαινόμενο των αναπαραστάσεων δεδομένων που καταρρέουν, το οποίο μπορεί να είναι ότι οι μικρού μεγέθους πολλαπλότητες μπορούν να αποτελούνται από μικρά συνδεδεμένα στοιχεία που ανήκουν στην ίδια συναισθηματική τάξη χωρίς να είναι σε θέση να διατηρήσουν την γενική δομή των δεδομένων. Αυτό θα μπορούσε να είναι η περίπτωση που το LLE μπορεί να εκτελέσει ελαφρώς καλύτερα από το Pattern Search MDS, αλλά συνολικά οι παραστάσεις του τελευταίου μπορεί να αποδειχθούν πιο ανθεκτικές και κλιμακούμενες για την κατάρτιση χρησιμοποιώντας πολλαπλά δεδομένα.

Όταν αυξάνουμε περαιτέρω τις διαστάσεις στόχου σε 25, παρατηρούμε μικρή μείωση της απόδοσης σε μετρήσεις WA και UA. Η καλύτερη επίδοση για αυτές τις διαστάσεις επιτυγχάνεται με το $K = 17$ (58.8%) χρησιμοποιώντας την αναζήτηση μοτίβου MDS και με $K = 21$ (52.0%) χρησιμοποιώντας LLE, για WA και UA, αντίστοιχα. Οι προαναφερθείσες βαθμολογίες απόδοσης εξακολουθούν να ξεπερνούν την καλύτερη απόδοση του αρχικού χώρου χαρακτηριστικών των χαρακτηριστικών των 432 διαστάσεων από 1.9% και 1.6% σε WA και UA, αντίστοιχα. Αυτά τα αποτελέσματα υποδεικνύουν ότι τα χαρακτηριστικά εισόδου μπορούν να περιγραφούν κατάλληλα από πολλαπλότητες σε κατώτερες διαστάσεις και εξακολουθούν να παρέχουν διακριτική ικανότητα για SI SER.

Χρησιμοποιώντας το συντηγμένο σύνολο ακουστικών χαρακτηριστικών (RQA + IS10)

Στον Πίνακα 5.6 παρουσιάζουμε τα αποτελέσματα από μια ποικιλία αλγορίθμων μείωσης διαστάσεων όταν χρησιμοποιούμε το συντηγμένο σύνολο χαρακτηριστικών (RQA + IS10) για SI SER κάτω από μια ποικιλία παραμετροποιήσεων της διάστασης στόχου L και τον αριθμό των πλησιέστερων γειτόνων K. Η σημείωση είναι παρόμοια με αυτή που εξηγείται προηγουμένως.

Σύμφωνα με τα ευρήματά μας στην προηγούμενη ενότητα 4.7.3 για το SI SER, παρατηρούμε ότι έχουμε ακόμα καλύτερη απόδοση στις μετρήσεις ακρίβειας των WA και UA σε σύγκριση με τα προηγούμενα σύνολα χαρακτηριστικών όταν χρησιμοποιούμε το σύνολο (RQA + IS10). Η εφαρμογή του KNN απευθείας στις παραστάσεις του συντηγμένου σετ παράγει αποτελέσματα έως και 72.4% σε WA και 65.9% στην UA. Ωστόσο, μπορούμε εύκολα να ξεπεράσουμε αυτήν την απόδοση σε WA και UA χρησιμοποιώντας μόνο 25 διαστάσεις, αντί να χρησιμοποιήσουμε τις αρχικές 2014, όταν χρησιμοποιούμε το Pattern Seach MDS πριν εφαρμόσουμε το KNN. Στις 2 διαστάσεις, οι αλγόριθμοι με

Μέθοδος	L	Αριθμός Πλησιεστέρων Γειτόνω											
		Weighted Accuracy (WA)						Unweighted Accuracy (UA)					
		1	5	9	13	17	21	1	5	9	13	17	21
Σύνολο RQA + IS10	2014	62.4	69.3	72.3	70.9	72.4	72.0	58.4	63.7	65.9	63.8	65.1	65.0
Pattern Search MDS	2	**45.7**	**51.7**	50.3	**51.9**	51.5	50.3	39.5	**44.7**	42.8	44.9	44.4	42.9
Modified LLE		42.2	46.6	46.8	49.5	51.1	48.7	39.6	40.4	39.9	44.1	47.9	41.6
Spectral Clustering		43.0	48.0	52.0	51.8	52.5	52.2	39.5	41.3	45.7	45.1	45.1	45.0
LTSA		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MDS SMACOF		43.5	50.1	50.1	51.2	50.5	50.9	39.4	44.5	43.3	44.5	44.2	44.1
Truncated SVD		41.5	41.7	47.2	48.6	49.6	52.0	36.4	36.3	41.9	44.5	43.2	45.2
Hessian LLE		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ISOMAP		42.5	46.4	45.6	46.9	47.4	47.8	38.0	39.5	38.8	39.3	39.8	39.7
LLE		43.5	47.3	**53.1**	51.3	**54.2**	**53.5**	**40.6**	44.0	**48.1**	**46.9**	**49.7**	**48.6**
Pattern Search MDS	10	56.7	66.1	**69.1**	**68.8**	**69.7**	**69.9**	51.5	59.6	**63.1**	**61.6**	**62.6**	**62.9**
Modified LLE		59.3	66.9	66.5	68.0	66.4	66.1	53.5	59.3	58.5	60.5	58.7	58.5
Spectral Clustering		55.6	64.5	65.2	67.6	66.9	65.9	51.5	59.2	59.1	61.1	60.3	59.0
LTSA		39.2	44.4	47.1	47.5	48.9	48.9	34.0	37.3	39.5	40.2	42.4	41.2
MDS SMACOF		58.2	66.1	65.9	66.6	67.6	68.5	52.6	58.4	57.8	58.8	59.4	61.4
Truncated SVD		57.9	65.9	66.2	66.9	66.8	66.8	52.3	58.6	58.8	59.2	59.4	59.1
Hessian LLE		39.2	44.4	47.1	47.5	48.9	48.9	34.0	37.3	39.5	40.2	42.4	41.2
ISOMAP		**62.5**	**68.0**	67.5	67.3	67.8	66.7	**57.7**	**61.1**	57.7	58.1	58.7	57.6
LLE		60.9	62.4	62.2	64.1	62.7	63.0	55.7	54.5	53.9	56.5	55.1	54.2
Pattern Search MDS	25	62.1	68.3	71.4	73.0	**74.4**	**74.1**	57.6	61.3	65.4	66.5	**68.8**	**68.6**
Modified LLE		59.7	62.4	66.5	66.6	66.7	66.7	55.4	57.5	61.0	60.9	59.4	59.6
Spectral Clustering		59.3	65.4	67.7	66.3	67.4	67.3	53.7	59.1	60.5	58.6	61.2	59.3
LTSA		57.4	61.3	63.8	62.9	61.8	61.0	54.6	52.9	55.9	54.3	53.7	52.4
MDS SMACOF		60.8	70.7	**73.6**	**73.9**	74.0	72.9	57.2	64.0	**66.7**	**66.7**	66.2	65.0
Truncated SVD		61.4	**70.9**	72.5	73.0	71.9	71.4	55.5	**64.4**	64.3	65.3	64.3	63.5
Hessian LLE		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ISOMAP		**64.4**	69.8	68.3	68.5	68.1	67.7	**58.5**	62.5	60.3	58.9	59.6	58.7
LLE		59.6	62.0	61.6	62.5	63.3	64.5	55.0	56.0	56.1	57.9	57.2	59.4

Πίνακας 5.6: Σύγκριση της μεθόδου αναζήτησης MDS με άλλες μεθόδους μείωσης διαστάσεων για πειράματα Ανεξαρτήτου-Ομιλητή για αναγνώριση Συναισθήματος από φωνή με τη χρήση των χαρακτηριστικών (RQA + IS10)

τις καλύτερες επιδόσεις για τη μείωσης διάστασης φαίνεται να είναι οι Pattern Seach MDS και LLE για τις μετρικές WA και UA. Οι πολλαπλότητες σε 2 διαστάσεις που έχουν μάθει από τη μέθοδο LLE επιτυγχάνουν WA έως 54.2% όταν χρησιμοποιούνται $K = 17$ γείτονες και UA από 49.7% για ίσο αριθμό γειτόνων. Αυτή η τεράστια πτώση στις μετρικές απόδοσης είναι ενδεικτική ότι οι 2 διαστάσεις δεν μπορούν να καταγράψουν την πλήρη διακριτική ικανότητα των χαρακτηριστικών σε οποιαδήποτε διαμόρφωση των πλησιέστερων γειτόνων και του αλγόριθμου μείωσης των διαστάσεων.

Επιπλέον, όταν αυξάνουμε τις διαστάσεις στόχου σε 10 παρατηρούμε ότι επιτυγχάνουμε πολύ καλύτερα αποτελέσματα σε WA και UA όταν χρησιμοποιούμε τον προτεινόμενο αλγόριθμο Pattern Search MDS για τις περισσότερες από τις τιμές του K, παρόμοιες με αυτές των προηγούμενων αποτελεσμάτων για το RQA Σύνολο (βλέπε πίνακα 5.5). Συγκεκριμένα, επιτυγχάνουμε WA έως 69.9% με $K = 21$ καθώς και 63.1% σε UA με $K = 9$ χρησιμοποιώντας Pattern Seach MDS. Για τιμές $K = 1$

και $K = 5$ το LLE φαίνεται να καταγράφει καλύτερα τη γεωμετρία των δεδομένων από τον προτεινόμενο αλγόριθμο. Αυτό υποδεικνύει ότι η αναζήτηση μοτίβου MDS είναι μια ανθεκτική μέθοδος για τη διατήρηση των συναισθηματικών πληροφοριών από τις αναπαραστάσεις επιπέδου ομιλίας, ακόμα και αν ο χώρος στόχος είναι αρκετά χαμηλός στις διαστάσεις.

Όταν αυξάνουμε περαιτέρω τις διαστάσεις στόχου σε 25, αποκτούμε την καλύτερη απόδοση στην WA με $K = 17$ (74.4%) και στην UA με $K = 17$ (68.8%) όταν χρησιμοποιούμε Pattern Search MDS πριν εφαρμόσουμε το KNN. Αυτές οι βαθμολογίες απόδοσης υπερβαίνουν τις καλύτερες βαθμολογίες απόδοσης του αρχικού χώρου χαρακτηριστικών των 2014 διαστάσεων από 2.0% και 2.9% σε WA και UA, αντίστοιχα. Υπενθυμίζοντας την καλύτερη απόδοση που αποκτήσαμε κάτω από την ίδια πειραματική ρύθμιση των πειραμάτων SI SER χρησιμοποιώντας βάση δεδομένων EmoDB για αξιολόγηση, όπως παρουσιάζεται στον Πίνακα 4.4, αποκτήσαμε 82.1% WA χρησιμοποιώντας SVM και 77.5% χρησιμοποιώντας LR. Συγκρίνοντας τα αποτελέσματα που επιτεύχθηκαν από τις καλύτερες επιδόσεις μας εδώ, ο συνδυασμός ενός αλγόριθμου NLDR και του KNN αποδίδει μια χαμηλότερη απόδοση με 7.7% σε WA και 8.7% στην UA. Ωστόσο, αυτό μπορεί να αντισταθμιστεί όταν εξετάζουμε τον αριθμό των χαρακτηριστικών όπου εκπαιδεύτηκε το σύστημα SER (25 αντί 2014) και ότι το KNN είναι ένα από τα απλούστερα μοντέλα σε σύγκριση με πολύ πιο πολύπλοκους ταξινομητές όπως SVM και LR. Ως εκ τούτου, αυτά τα αποτελέσματα υποδεικνύουν ότι ακόμη και ένα από τα απλούστερα μη παραμετρικά μοντέλα που είναι KNN (βλέπε Ενότητα 2.2.4 για επεξήγηση) μπορεί να ανταγωνιστεί έναντι παραμετρικών μοντέλων όπως SVM και LR (βλέπε ενότητες 2.2.2 και 2.2.3 για την περιγραφή αυτών των μοντέλων) χρησιμοποιώντας αλγόριθμους NLDR πριν από την ταξινόμηση. Είναι αξιοσημείωτο ότι αυτά τα ευρήματα υποδηλώνουν ότι ακόμη και στα πιο εξελιγμένα μοντέλα SER θα μπορούσαν να χρησιμοποιηθούν αλγόριθμοι NLDR για την προετοιμασία των παραμέτρων τους, χωρίς να χάσουν την εκφραστικότητα των χαρακτηριστικών εισόδου και να εστιάσουν μόνο στα κυριότερα τμήματα αυτών των αναπαραστάσεων.

5.7.8 Συγκρίνοντας τον συνδυασμό (Pattern Search MDS + KNN) με παραμετρικά μοντέλα του Κεφαλαίου 4

Σε αυτή την ενότητα παρουσιάζουμε τα καλύτερα αποτελέσματα που προκύπτουν από το συνδυασμό Pattern Search MDS με ταξινομητή KNN για SER και τα συγκρίνουμε με τα αποτελέσματα που παρουσιάστηκαν στις προηγούμενες ενότητες 4.7.3 και 4.7.4. Τα αποτελέσματα που παρουσιάζονται περιλαμβάνουν όλους τους συνδυασμούς συνόλων χαρακτηριστικών των RQA και IS10 όπως παρουσιάζονται σε προηγούμενες ενότητες. Εκτός από τα πειράματα που παρουσιάστηκαν προηγουμένως σχετικά με το SI SER χρησιμοποιώντας το EmoDB, εκτελούμε παρόμοια πειράματα με το LOSO με τη βάση δεδομένων IEMOCAP χρησιμοποιώντας τους ίδιους τρόπους εξαγωγής χαρακτηριστικών (δείτε το τμήμα 4.7.4). Πραγματοποιούμε Γενική Κανονικοποίηση (GN) πριν εφαρμόσουμε Pattern Search MDS για τη μείωση των διαστάσεων.

Παρουσιάζουμε τα αποτελέσματα όλων των πειραμάτων χρησιμοποιώντας το Pattern Search MDS σε συνδυασμό με το KNN για το SER καθώς επίσης μεταφέρουμε τα αποτελέσματα στις ίδιες πειραματικές ρυθμίσεις στο επίπεδο ολόκληρης της ομιλίας από τα τμήματα 4.7.3 και 4.7.4 στον πίνακα 5.7. Τα σύνολα χαρακτηριστικών είναι τα ίδια με αυτά που χρησιμοποιήθηκαν σε προηγούμενες ενότητες: 4.7.2, 4.7.3 και 4.7.4, και συγκεκριμένα: IS10, RQA και το συντηγμένο σύνολο (RQA + IS10). Οι δύο πρώτες σειρές που αντιστοιχούν σε κάθε ομάδα χαρακτηριστικών αφορούν τα καλύτερα αποτελέσματα που έχουν επιτευχθεί για τις δοκιμαστικές ρυθμίσεις σε EmoDB και IEMOCAP με πειραματικές ρυθμίσεις στα πειράματα SI και LOSO που παρουσιάζονται σε προηγούμενες ενότητες. Στις τελευταίες δύο σειρές που αντιστοιχούν σε κάθε σύνολο χαρακτηριστικών, τα καλύτερα αποτελέσματα που επιτυγχάνονται από ένα KNN με $K \in \{k \mid k \leq 40 and\ k\ mod\ 4 = 1\}$ μετά την εφαρμογή του Pattern Search MDS προκειμένου να μειωθούν οι διαστάσεις από το αρχικό χαρακτηριστικό που έχει οριστεί σε 10 ή 25. Το L αντιστοιχεί στη διάσταση προορισμού του MDS αναζήτησης μοτίβων, ενώ για τις δύο πρώτες σειρές, όπου δεν εφαρμόζεται μείωση διαστάσεων (αυτό σημειώνεται από τη στήλη "-" στη στήλη "Μέθοδος Μείωσης"), αυτή η στήλη καθορίζει τις διαστάσεις κάθε αναπαράστασης σε επίπεδο ολόκληρης της ομιλίας.

Πίνακας 5.7: Συγκρίνοντας τον συνδυασμό (Pattern Search MDS + KNN) με παραμετρικά μοντέλα του Κεφαλαίου 4

Χαρακτηριστικά	Διαστάσεις Μέθοδος Μείωσης	L	Ταξινομητής	EmoDB		IEMOCAP	
				WA	UA	WA	UA
IS10	-	1582	SVM	79.7	74.3	59.2	60.5
	-	1582	LR	76.1	71.9	53.5	57.5
	-	1582	KNN	69.9	64.1	53.1	55.7
	Pattern Search MDS	10	KNN	65.7	59.9	53.8	55.2
	Pattern Search MDS	25	KNN	70.4	63.5	**54.5**	56.8
RQA	-	432	SVM	70.9	64.2	53.1	53.7
	-	432	LR	71.1	67.1	52.8	54.3
	-	432	KNN	56.9	48.4	46.9	48.8
	Pattern Search MDS	10	KNN	60.0	52.7	46.4	47.2
	Pattern Search MDS	25	KNN	58.8	50.9	47.6	49.3
RQA+IS10	-	2014	SVM	82.1	76.9	59.5	60.7
	-	2014	LR	80.1	77.5	54.5	58.7
	-	2014	KNN	72.4	65.9	52.6	55.1
	Pattern Search MDS	10	KNN	69.9	63.1	52.9	54.4
	Pattern Search MDS	25	KNN	74.4	68.8	**54.9**	57.2

Τα αποτελέσματα που εμφανίζονται στον Πίνακα 5.7 υποδηλώνουν τη σημασία της εφαρμογής της σωστής μείωσης των διαστάσεων πριν από την εφαρμογή ενός ταξινομητή επιπέδου ολόκληρης της ομιλίας για SER. Γενικά, ο συνδυασμός Pattern Search MDS και ο μη παραμετρικός ταξινομητής KNN δεν αποφέρει καλύτερα αποτελέσματα από την εφαρμογή ενός ταξινομητή SVM ή LR σε σχέση με τους αρχικά διανύσματα χαρακτηριστικών. Για όλα τα χαρακτηριστικά γνωρίσματα παρατηρούμε ότι ένας ταξινομητής SVM πάνω από τους αρχικούς φορείς χαρακτηριστικών ξεπερνά τον καλύτερο συνδυασμό Pattern Search MDS και KNN με περιθώρια $7.7\% - 10.9\%$ σε WA και $8.1 - 11.5\%$ σε UA στην EmoDB για SI πειράματα καθώς και με $4.7 - 5.5\%$ σε WA και $3.5 - 4.4\%$ σε UA για την IEMOCAP. Ομοίως, όταν ένα μοντέλο LR εφαρμόζεται πάνω από στα αρχικά διανύσματα χαρακτηριστικών, τότε έχουμε παρόμοια περιθώρια που η τελευταία προσέγγιση ξεπερνά τον καλύτερο συνδυασμό Pattern Search MDS και KNN. Παρόλο που υπάρχουν κάποιες περιπτώσεις όπου ο προτεινόμενος αλγόριθμος σε συνδυασμό με KNN παρέχει καλύτερα αποτελέσματα για το πείραμα IEMOCAP σε σύγκριση με το LR. Δηλαδή, στο WA $53.5\% \rightarrow 54.5\%$ όταν χρησιμοποιειούμε το IS10 Σύνολο χαρακτηριστικών και πάλι στο WA $54.5\% \rightarrow 54.9\%$ όταν χρησιμοποιειούμε το συντηγμένο σύνολο χαρακτηριστικών. Αυτό είναι πολύ ενδιαφέρον αν λάβουμε υπόψη ότι με 10 ή 25 διαστάσεις χαρακτηριστικών εισόδου και χρησιμοποιώντας μόνο ένα μη παραμετρικό KNN για SER, χωρίς να εκτελεσθούν κόλπα πυρήνα όπως τα μοντέλα SVM και LR προκειμένου να δημιουργήσουν διαχωριστικά υπερεπίπεδα. Επομένως, η μείωση της απόδοσης είναι αποδεκτή όταν μειώσουμε το χώρο των χαρακτηριστικών στο $\approx 0.5 - 1\%$ της αρχικής της διαστάσεων.

Επιπλέον, η μείωση τόσο των μετρήσεων απόδοσης των WA και UA κατά τη χρήση του Pattern Search MDS με KNN αντί των πιο εξελιγμένων μοντέλων επίπεδο έκφρασης όπως SVM και LR είναι πολύ πιο εμφανής στο πείραμα EmoDB, ενώ στο IEMOCAP η απόδοση που επιτυγχάνεται είναι πολύ πιο συγκρίσιμη με την καλύτερη ένας. Πιθανότατα, αυτό μπορεί να εξηγηθεί επειδή στο πείραμα EMoDB έχουμε μόνο 535 δείγματα προκειμένου να κατασκευάσουμε την χαμηλής διαστάσεως πολλαπλότητα με 7 συναισθηματικές τάξεις αντί για το πείραμα IEMOCAP όπου 5531 ομιλίες είναι διαθέσιμες μόνο για 4 συναισθηματικές τάξεις. Σε αυτό το πλαίσιο, η κατανομή των δεδομένων εισόδου που προσπαθούμε να προσεγγίσουμε σε χαμηλές διαστάσεις με τη χρήση MDS αναζήτησης προτύπων θα είναι αραιά δειγματοληπτημένη και επομένως η ανακατασκευή της πολλαπλότητας θα ήταν πιθανότατα δυσκολότερη στην εκτέλεση. Σε όλες τις περιπτώσεις, το απλό KNN πάνω από τις

μειωμένες παραστάσεις μεγέθους αποδίδει καλύτερη απόδοση και στις δύο μετρήσεις σε σύγκριση με την περίπτωση όπου εφαρμόζουμε KNN απευθείας πάνω από τις αρχικές αναπαραστάσεις χαρακτηριστικών.

5.8 Οπτικοποίηση συναισθηματικών πολλαπλοτήτων από ακουστικά χαρακτηριστικά γνωρίσματα

Σε αυτή την ενότητα παρουσιάζουμε μερικές $2D$ και $3D$ πολλαπλότητες οι οποίες δημιουργήθηκαν μετά την εφαρμογή αλγορίθμων NLDR πάνω στα σύνολα ακουστικών χαρακτηριστικών που χρησιμοποιήθηκαν σε πειράματα SI SER στην προηγούμενη ενότητα 5.7.7. Θα θέλαμε να παρουσιάσουμε μια εικόνα του τρόπου εμφάνισης αυτών των παραστάσεων, προκειμένου να αντλήσουμε ορισμένες ποιοτικές διαβεβαιώσεις των αποτελεσμάτων που παρουσιάζονται στους Πίνακες 5.4, 5.5, 5.6 και 5.7 καθώς και τον τρόπο με τον οποίο λειτουργεί κάθε μέθοδος μείωσης των διαστάσεων, ειδικά για την προτεινόμενη. Για κάθε απεικόνιση επιλέγουμε τον αλγόριθμο που απέδωσε τις καλύτερες επιδόσεις για τα προηγούμενα πειράματα SI SER στο EmoDB (βλ. Πίνακες 5.4, 5.5, 5.6). Συγκεκριμένα, αναλύουμε τις πολλαπλότητες που παράγονται από τα *Pattern Search MDS*, *MDS SMACOF*, *Spectral Clustering, LLE*, *ISOMAP* και *Truncated SVD* από αριστερά προς τα δεξιά και από πάνω προς τα κάτω.

5.8.1 Δισδιάστατες πολλαπλότητες από το IS10 σύνολο χαρακτηριστικών για την βάση δεδομένων EmoDB

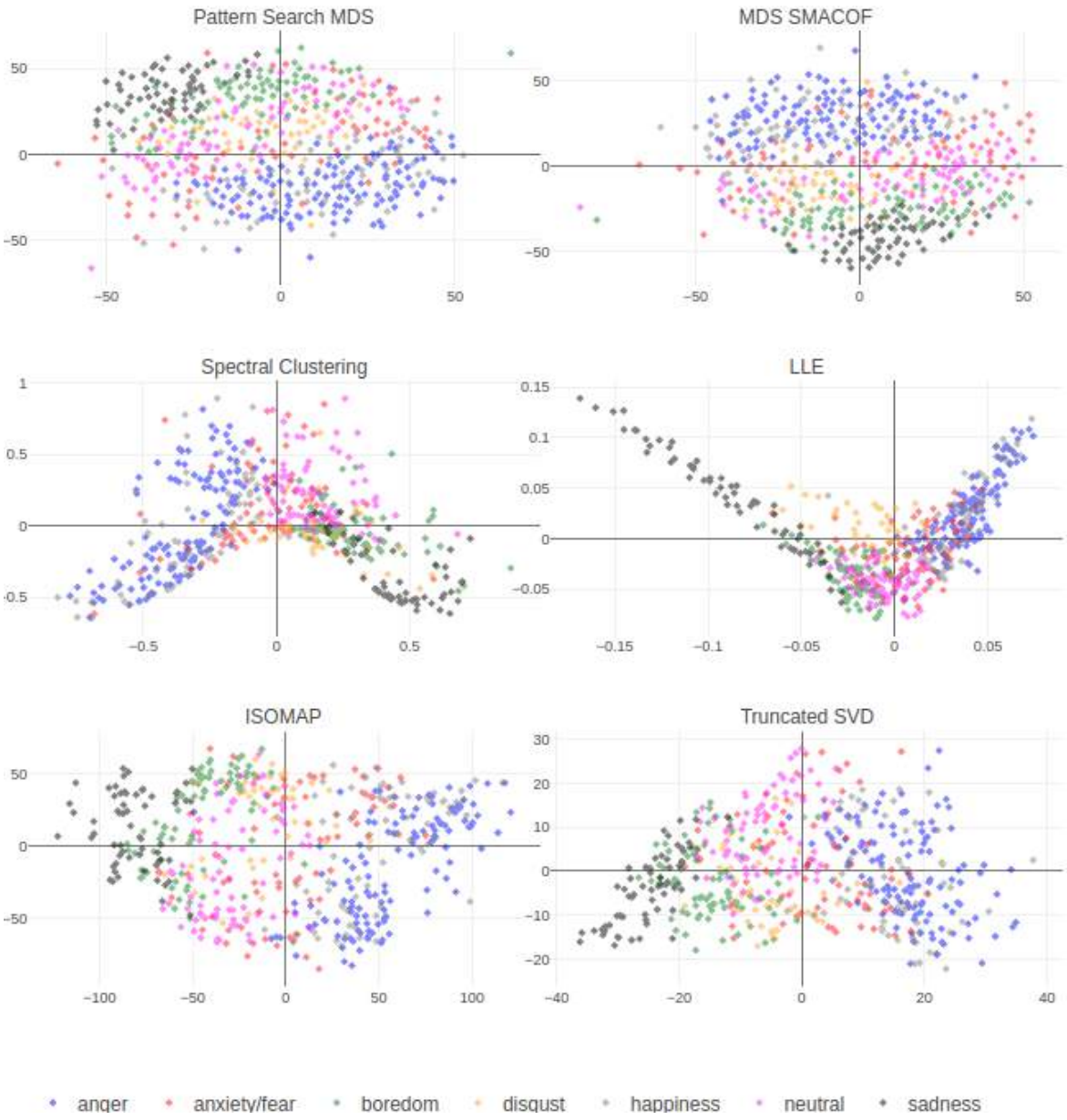
Στο Σχήμα 5.7, παρουσιάζουμε τις πολλαπλότητες 2 διαστάσεων που έχουν ληφθεί χρησιμοποιώντας τα χαρακτηριστικά IS10 που έχουν εξαχθεί σε χρονική κλίμακα ολόκληρης της ομιλίας της βάσης δεδομένων EmoDB (συμπεριλαμβάνονται όλα τα 7 συναισθήματα).

Όλες οι εμφανιζόμενες μέθοδοι παράγουν πολλαπλότητες όπου πολλές κλάσεις καταρρέουν η μία πάνω στην άλλη. Για παράδειγμα, οι περιπτώσεις *happiness* και *anger* φαίνεται να επικαλύπτονται σε όλες τις πολλαπλότητες. Επιπλέον, μερικές κλάσεις όπως το *anxiety/fear* φαίνεται να είναι διασκορπισμένες πάνω από την πολλαπλότητα, υποδεικνύοντας ότι η δυναμική αυτής της κλάσης δεν έχει απεμπλακεί χρησιμοποιώντας αυτό το σύνολο χαρακτηριστικών ώστε να αντανακλά αυτό σε έναν χώρο 2 διαστάσεων. Το LLE και το Spectral Clustering φαίνεται να παρέχουν ορθογώνιες και καμπύλες πολλαπλότητες, αντίστοιχα, που περιλαμβάνονται σε περιοχές διακριτών συναισθημάτων που είναι αρκετά διακριτές ανάμεσα στις διαφορετικές συναισθηματικές τάξεις, χωρίς να αποφεύγεται η επικάλυψη μεταξύ ορισμένων τάξεων. Αυτό απεικονίζεται επίσης στις επιδόσεις που λαμβάνονται με αυτές τις μεθόδους χρησιμοποιώντας το KNN στον Πίνακα 5.7. Επιπλέον, το Pattern Search MDS παράγει μια πολλαπλότητα πολύ παρόμοια με την MDS SMACOF, από την άποψη της τοπολογίας και του τρόπου με τον οποίο εκπροσωπούνται όλες οι συναισθηματικές τάξεις.

5.8.2 Δισδιάστατες πολλαπλότητες από το RQA σύνολο χαρακτηριστικών για την βάση δεδομένων EmoDB

Στο Σχήμα 5.8, παρουσιάζουμε τις πολλαπλότητες 2 διαστάσεων που έχουν ληφθεί χρησιμοποιώντας τα χαρακτηριστικά RQA που έχουν εξαχθεί σε χρονική κλίμακα ολόκληρης της ομιλίας της βάσης δεδομένων EmoDB (συμπεριλαμβάνονται όλα τα 7 συναισθήματα).

Γενικά, αυτές οι πολλαπλότητες αποτελούνται από πολύ πιο επικαλυπτόμενες συναισθηματικές περιοχές από αυτές που αποκτήθηκαν κατά τη χρήση του IS10. Αν και αυτό σχετίζεται εξ ολοκλήρου με την εκφραστικότητα κάθε τύπου χαρακτηριστικού και σύμφωνα με τα ευρήματά μας σε προηγούμενες ενότητες της παρούσας διπλωματικής εργασίας έχουμε καταλήξει στο συμπέρασμα ότι το RQA είναι λιγότερο περιγραφικό από το σύνολο των χαρακτηριστικών IS10 όταν αξιολογούνται ξεχωριστά (βλ. Πίνακες 4.4, 4.5 και 5.7). Υπάρχει πλήρης κατάρρευση της κατανομής των δειγμάτων για τη μέθοδο LLE, η οποία είναι επίσης εμφανής από την απόδοση που επιτυγχάνεται χρησιμοποιώντας το

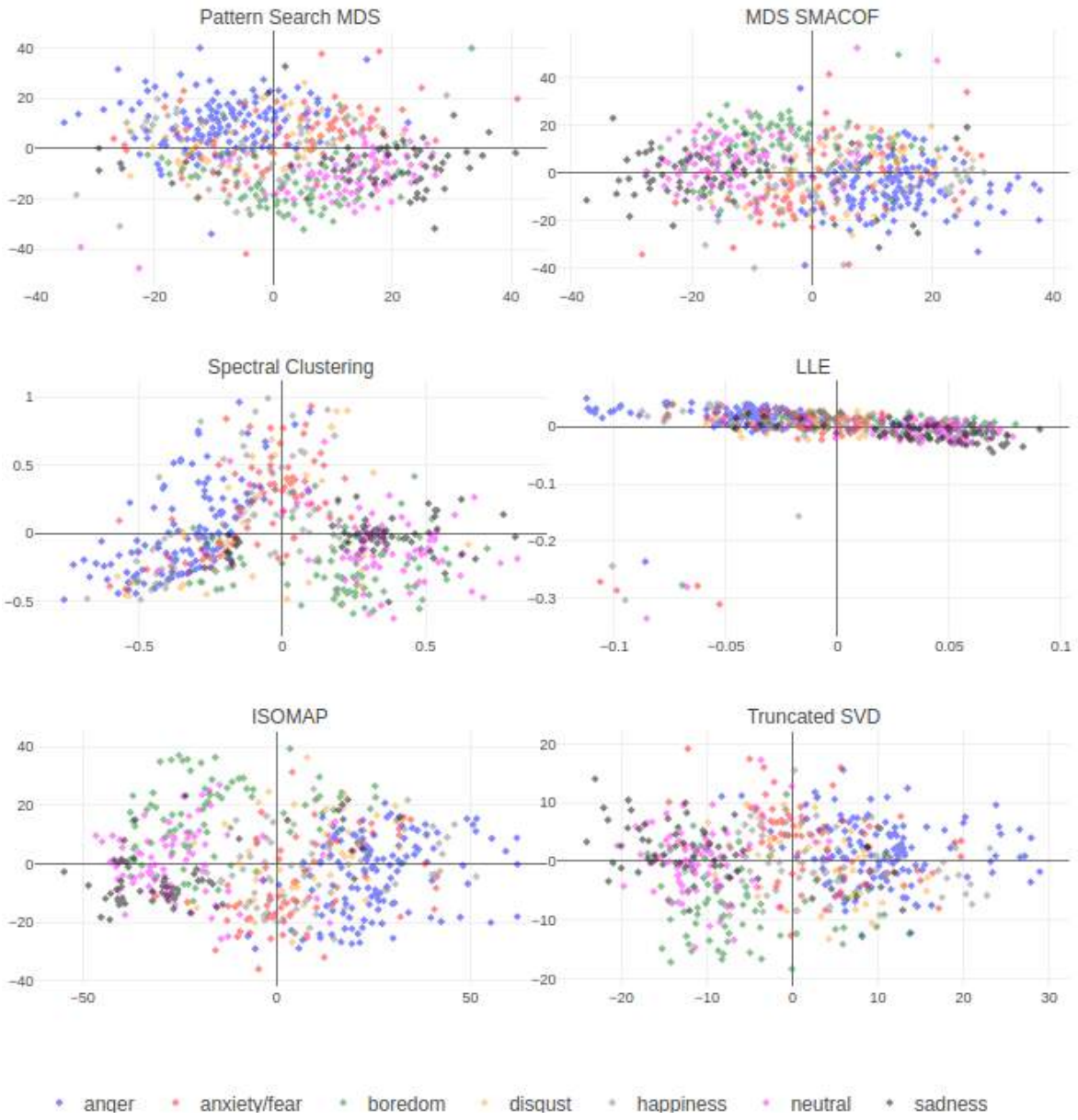


Σχήμα 5.7: Σύγκριση των παραγόμενων $2D$ πολλαπλοτήτων από την εφαρμογή μεθόδων μείωσης των διαστάσεων από τις αναπαραστάσεις ακουστικών χαρακτηριστικών IS10 για την βάση δεδομένων EmoDB

KNN (βλέπε πίνακα 5.5). Είναι αρκετά προφανές ότι για τις αναπαραστάσεις του RQA, το ISOMAP σημειώνει τα καλύτερα αποτελέσματα στη δημιουργία περιοχών με μόνο ένα ή δύο συναισθήματα και αυτό απεικονίζεται και στις μετρήσεις απόδοσης WA και UA στον Πίνακα 5.5. Οι περιοχές των ίδιων συναισθηματικών δειγμάτων δεν είναι τόσο ομοιογενείς για το Pattern Search MDS, το Truncated SVD και το MDS SMACOF.

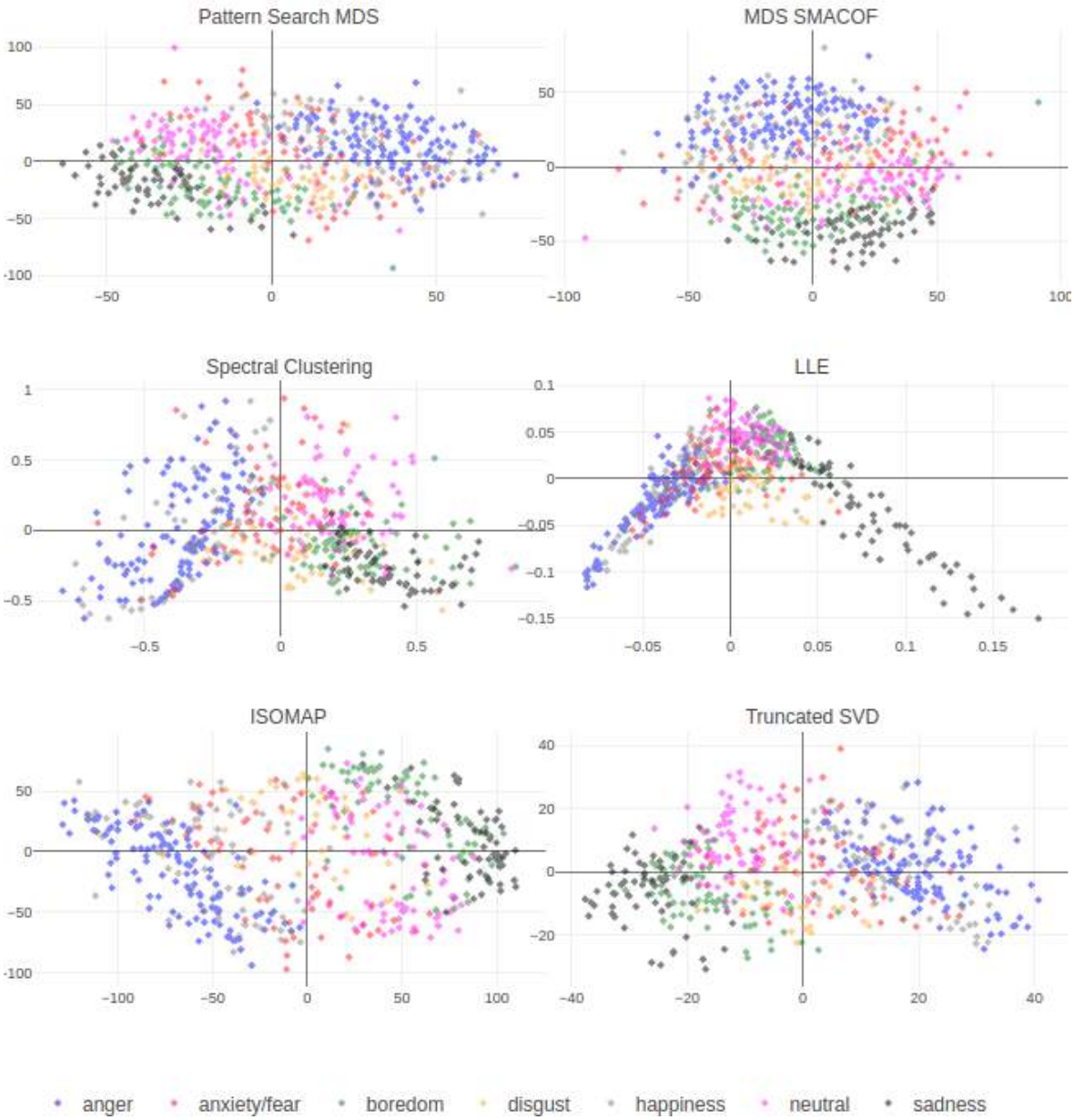
5.8.3 Δισδιάστατες πολλαπλότητες από το συντηγμένο σύνολο χαρακτηριστικών (RQA + IS10) για την βάση δεδομένων EmoDB

Στο Σχήμα 5.9, παρουσιάζουμε τις πολλαπλότητες 2 διαστάσεων που έχουν ληφθεί χρησιμοποιώντας τα χαρακτηριστικά (RQA + IS10) που έχουν εξαχθεί σε χρονική κλίμακα ολόκληρης της ομιλίας της βάσης δεδομένων EmoDB (συμπεριλαμβάνονται όλα τα 7 συναισθήματα).



Σχήμα 5.8: Σύγκριση των παραγόμενων $2D$ πολλαπλοτήτων από την εφαρμογή μεθόδων μείωσης των διαστάσεων από τις αναπαραστάσεις ακουστικών χαρακτηριστικών RQA για την βάση δεδομένων EmoDB

Για το σετ χαρακτηριστικών (RQA + IS10) παρατηρούμε ότι όλες οι μέθοδοι παράγουν δισδιάστατες πολλαπλότητες που αποτελούνται από αρκετά γραμμικά διαχωρίσιμες περιοχές συναισθηματικών περιπτώσεων. Ωστόσο, το Pattern Search MDS, LLE και Spectral Clustering φαίνεται να παρέχουν τις καλύτερες αναπαραστάσεις λόγω της γειτνίασης των δειγμάτων που ανήκουν σε κάθε κατηγορία. Αυτό επιβεβαιώνεται επίσης από την απόδοση της ταξινόμησης KNN που παρουσιάζεται στον πίνακα 5.6. Παρόλο που το ISOMAP παράγει καλά διασυνδεδεμένες περιοχές συναισθηματικών τάξεων, γενικά ο παραγόμενος συναισθηματικός χάρτης περιέχει μεγάλες τρύπες και αραιές περιοχές που απεικονίζεται επίσης στις επιδόσεις του. Προφανώς, αυτό οφείλεται στις λανθασμένες γεωδεσικές αποστάσεις που υπολογίζονται απευθείας από τις αποστάσεις των σημείων στον Ευκλείδειο Χώρο. Τέλος, το SVD παράγει επίσης έναν χάρτη στον οποίο οι περιοχές για κάθε συναισθηματική τάξη είναι αρκετά πυκνή μέσα στις τάξεις αλλά χωρίς να αποφεύγονται οι αλληλεπικαλυπτόμενες περιοχές μεταξύ παρόμοιων εκφρασμένων συναισθημάτων.



Σχήμα 5.9: Σύγκριση των παραγόμενων $2D$ πολλαπλοτήτων από την εφαρμογή μεθόδων μείωσης των διαστάσεων από τις αναπαραστάσεις ακουστικών χαρακτηριστικών (RQA + IS10) για την βάση δεδομένων EmoDB

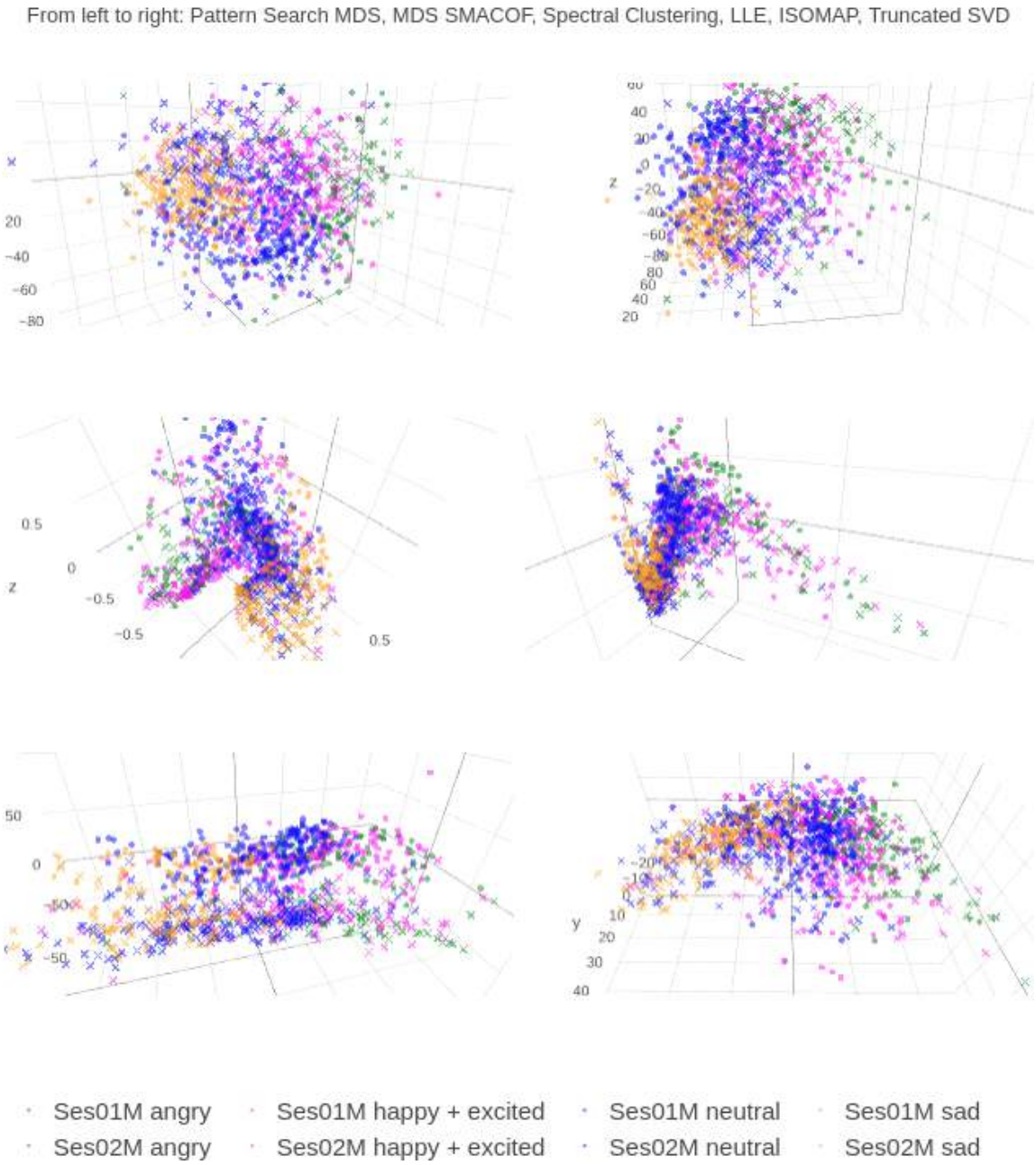
5.8.4 Τρισδιάστατες πολλαπλότητες από το (RQA + IS10) σύνολο χαρακτηριστικών για δύο άντρες ομιλητές της βάσης δεδομένων IEMOCAP

Με βάση τα ποιοτικά συμπεράσματα σχετικά με τον τρόπο κατασκευής αυτών των πολλαπλοτήτων για κάθε αλγόριθμο μείωσης των διαστάσεων, θα προσθέσουμε μια ακόμη διάσταση στη διάσταση-στόχο L. Θα θέλαμε επίσης να δούμε τη συμπεριφορά αυτών των αλγορίθμων με μεγαλύτερο αριθμό δειγμάτων και με διαφορετικούς ομιλητές. Απομονούμε δύο άντρες ομιλητές από τη βάση δεδομένων IEMOCAP, δηλαδή τα τους άντρες ομιλητές από τις δύο πρώτες συνεδρίες και υπολογίζουμε τις μαθηματικές πολλαπλότητες σε 3 διαστάσεις για όλες τις προαναφερθείσες μεθόδους μείωσης των διαστάσεων στο σχήμα 5.10. Ακολουθούμε την ίδια πειραματική ρύθμιση όπως περιγράψαμε στην προηγούμενη ενότητα 4.7.1 για τη λήψη ενός υποσυνόλου του συνόλου δεδομένων IEMOCAP, λαμβάνοντας υπόψη μόνο 5531 εκφράσεις, συμπεριλαμβανομένων των 4 συναισθημάτων (1103 θυμός, 1636 ευτυχία, 1708 ουδετερότητα και 1084 λύπη), όπου συγχωνεύουμε τον ενθουσιασμό και την

ευτυχία στην τελευταία ομάδα συναισθημάτων. Επιπλέον, θεωρούμε μόνο το συντηγμένο σύνολο χαρακτηριστικών (RQA + IS10) λόγω του περιγραφικού του χαρακτήρα για το SER. Με αυτό τον τρόπο, μπορούμε να επικεντρωθούμε σε αυτές τις αναπαραστάσεις λαμβάνοντας υπόψη πολλά περισσότερα δείγματα από τη διανομή για κάθε συναισθηματική τάξη.

Σε αυτή τη ρύθμιση σημειώνουμε με ξεχωριστά σύμβολα "x" και "σφαίρες" τις συναισθηματικές αναπαραστάσεις που αντιστοιχούν σε κάθε άντρα ομιλιτή από την πρώτη και την δεύτερη συνεδρία της IEMOCAP, αντίστοιχα. Μπορούμε να διαπιστώσουμε ότι για παράδειγμα, το ISOMAP δεν καταγράφει τις κατανομές των ίδιων συναισθηματικών τάξεων και για τους δύο ομιλητές. Υπάρχουν δύο αποσυνδεδεμένες περιοχές που αντιστοιχούν σε διαφορετικούς ομιλητές αν κοιτάξουμε πιο κοντά στο σχήμα 5.10 (κάτω και αριστερά). Πιθανώς, αυτές οι αποσυνδεδεμένες περιοχές δημιουργούνται εξαιτίας κάποιων κυρίαρχων χαρακτηριστικών όπως την θεμελιώδη συχνότητα και την ενέργεια που κυριαρχούν πανω από άλλα χαρακτηριστικά, καθώς και σε χώρους με μεγάλες διαστάσεις. Σε αυτό το πλαίσιο, όταν υπολογίζουμε τις γεωδεσικές αποστάσεις σε όλα τα σημεία, οι δηλώσεις του ίδιου ομιλητή περιέχουν πληροφορίες στενά συνδεδεμένες με την ταυτότητα του ομιλητή, οι οποίες αποδίδουν μεγαλύτερη σύνδεση μεταξύ δειγμάτων του ίδιου ομιλητή παρά της συναισθηματικής κλάσης αυτής κάθε αυτής. Αυτό θα μπορούσε επίσης να οδηγήσει σε σημαντική μείωση της απόδοσης του συστήματος SER για το οποίο, η δυνατότητα γενίκευσης των χαρακτηριστικών εισόδου είναι μια ανάγκη για ακριβή συμπεράσματα.

Επιπλέον, τα MDS SMACOF και Pattern Search MDS παράγουν και πάλι πολύ παρόμοια αποτελέσματα με ελλειψοειδή τα οποία είναι σύμφωνα με τα προηγούμενα ευρήματα σχετικά με τις μαθηματικές πολλαπλότητες 2 διαστάσεων όταν χρησιμοποιούμε το σύνολο συντηγμένων χαρακτηριστικών (βλέπε σχήμα 5.9). Σε αυτό το πλαίσιο, το LLE φαίνεται επίσης να κατασκευάζει την πολλαπλότητα χαμηλών διαστάσεων χρησιμοποιώντας μια μεταφορά, μία περιστροφή και καθρεπτισμό των τοπικών δεδομένων που είναι ακριβώς οι πράξεις στις οποίες τα βάρη που ορίζει το LLE για κάθε σημείο δεδομένων είναι αμετάβλητα. Η φασματική συσταδοποιήση ακολουθεί παρόμοιο μοτίβο όταν ανακατασκευάζεται την πολλαπλότητα στο $3D$ που μοιάζει αρκετά με την πολλαπλότητα που έχει δειχθεί προηγουμένως στο $2D$ (βλ. Εικόνα 5.9). Αυτό σχετίζεται άμεσα με τον χειροποίητο πυρήνα που χρησιμοποιεί το Spectral Clustering για να ομαλοποιήσει τη μήτρα καθολικής συσχέτισης μεταξύ των σημείων του χώρου μεγάλης διαστάσεως.



Σχήμα 5.10: Σύγκριση των παραγόμενων $3D$ πολλαπλοτήτων κατά την εφαρμογή μεθόδων μείωσης των διαστάσεων στις αναπαραστάσεις ακουστικών χαρακτηριστικών (RQA + IS10) για δύο άντρες ομιλητές της βάσης δεδομένων IEMOCAP

Κεφάλαιο 6

Επίλογος

6.1 Συμπεράσματα

Από αυτό το έργο εξάγουμε ορισμένα συμπεράσματα τα οποία μπορούν να χωριστούν σε τρεις κύριες κατηγορίες που αντιστοιχούν στα Κεφάλαια 3, 4 και 5, αντίστοιχα. Ως εκ τούτου, θα απαριθμήσουμε τα πιο σημαντικά συμπεράσματα που αντλήσαμε από το έργο που παρουσιάζεται σε αυτή τη διπλωματική εργασία σε σχέση με τα κεφάλαια που αναφέρονται παραπάνω. Τα παρακάτω συμπεράσματα σχετίζονται τελικά με τα ερωτήματα που τίθενται στα αντίστοιχα τμήματα 1.4.1, 1.4.2, 1.4.3. Ειδικότερα, τα παρακάτω συμπεράσματα παρουσιάζονται με τέτοιο τρόπο ώστε να μπορούν να δοθούν ορισμένες απαντήσεις στα προαναφερθέντα ερωτήματα, λαμβάνοντας υπόψη το έργο που έχει γίνει για τη διεξαγωγή αυτής της διπλωματικής εργασίας.

6.1.1 Συμπεράσματα από το κεφάλαιο 3

Έχουμε δείξει ότι η χρονική κλίμακα στην οποία εξάγουμε τα χαρακτηριστικά για το SER, έχει μεγάλη επίδραση στην απόδοση RNNs. Τα LLDs λειτουργούν καλά σε χρονική κλίμακα που αντιστοιχεί περίπου στο επίπεδο φωνημάτων. Αντιστρόφως, τα στατιστικά χαρακτηριστικά κωδικοποιούν αξιοσημείωτα το συναισθηματικό πλαίσιο σε χρονική κλίμακα που αντιστοιχεί στο επίπεδο λέξεων. Στο μοντέλο LSTM που προτείνουμε, εξάγουμε στατιστικά χαρακτηριστικά πάνω σε τμήματα λόγου που αντιστοιχούν σε μερικές λέξεις και αναφέρουμε τα καλύτερα αποτελέσματα στη βάση δεδομένων IEMOCAP, με πολύ χαμηλότερη υπολογιστική πολυπλοκότητα και χωρίς να χρησιμοποιούμε μηχανισμούς προσοχής ή πιο πολύπλοκα δίκτυα.

6.1.2 Συμπεράσματα από το κεφάλαιο 4

Ερευνήσαμε τη χρήση μη γραμμικών μέτρων (Ποσοτικοποιημένης ανάλυσης της επαναληψιμότητας) RQA που εξάγονται από (Γραφήματα επαναληψιμότητας) RPs για SER. Η αποτελεσματικότητα αυτών των χαρακτηριστικών δοκιμάστηκε τόσο σε βάσεις γνώσεων όσο και σε τμηματικές προσεγγίσεις σε τρεις βάσεις δεδομένων συναισθημάτων. Η σύντηξη μη γραμμικών και συμβατικών συνόλων χαρακτηριστικών παρέχει σημαντική βελτίωση της απόδοσης έναντι των παραδοσιακών συνόλων χαρακτηριστικών για όλα τα πειράματα SER. η βελτίωση της απόδοσης είναι ιδιαίτερα μεγάλη όταν η ταυτότητα του ομιλητή είναι άγνωστη. Το συντηγμενο σύνολο ακουστικών χαρακτηριστικών βελτιώνει την απόδοση στην SER υπό τις περισσότερες συνθήκες δοκιμής, μεθόδους ταξινόμησης και σύνολα δεδομένων. Είναι αξιοσημείωτο ότι τα πειράματα δείχνουν ότι το προτεινόμενο σύνολο χαρακτηριστικών RQA μπορεί να χρησιμοποιηθεί επαρκώς ακόμη και μόνο του για SER χωρίς οποιαδήποτε χρήση ενός συνδυασμού επιλογής χαρακτηριστικών ή αλγόριθμου μείωσης διαστάσεων. Ως αποτέλεσμα, πιστεύουμε ότι η ανάλυση της υποτροπιάζουσας δυναμικής των σημάτων ομιλίας αποτελεί μια πολλά υποσχόμενη κατεύθυνση για την έρευνα στην αναγνώριση συναισθημάτων από φωνή.

6.1.3 Συμπεράσματα από το κεφάλαιο 5

Προτείνουμε τον αλγόριθμο Pattern Search MDS, ένα νέο αλγόριθμο για τη μείωση διαστάσεων, εμπνευσμένο από μεθόδους βελτιστοποίησης χωρίς την ανάγκη υπολογισμού της παραγώγου.

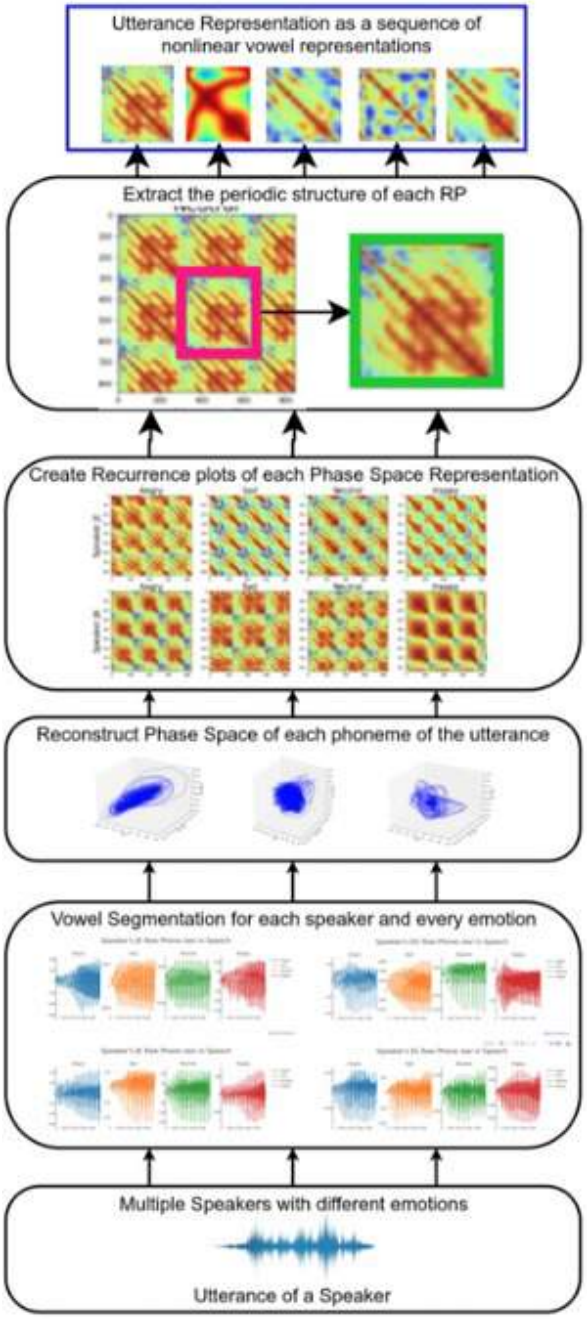
Ο Pattern Search MDS διαμορφώνεται ως παράδειγμα της ευρύτερης οικογένειας μεθόδων GPS, παρέχοντας έτσι θεωρητικές εγγυήσεις σύγκλισης μέχρι ένα σταθερό σημείο. Επιπλέον παρουσιάζουμε και κάποιες περαιτέρω βελτιστοποιήσεις στην απόδοση του αλγορίθμου μας όσον αφορά στην υπολογιστική αποδοτικότητα, στην ευρωστία και στην ποιότητα της λύσης. Η ποιοτική αξιολόγηση ενάντια σε άλλες δημοφιλείς τεχνικές μείωσης των διαστάσεων για καθαρά και θορυβώδη σχήματα γεωμετρίας πολλαπλότητας υποδεικνύει ότι ο Pattern Search MDS μπορεί να συμπεράνει με ακρίβεια την εγγενή γεωμετρία πολλαπλοτήτων που είναι ενσωματωμένες σε χώρους μεγάλης διαστάσεως. Επιπλέον, η σύγκριση των χαρακτηριστικών σύγκλισης με το SMACOF MDS δείχνει ότι το Pattern Search MDS συγκλίνει σε λιγότερες εποχές σε παρόμοιες ή καλύτερες λύσεις. Τα πειράματα για την απόδοση πραγματικών δεδομένων είναι συγκρίσιμα με τα αποτελέσματα τελευταίας τεχνολογίας τόσο για μια λεξικολογική εργασία σημασιολογικής ομοιότητας όσο και για την ταξινόμηση MNIST για KNN.

Όταν εξετάζουμε μια πειραματική ρύθμιση για το SI SER, τα ευρήματά μας υποδηλώνουν ότι ο προτεινόμενος αλγόριθμος μπορεί καταλλήλως να συλλάβει τις διαφορές του χώρου χαρακτηριστικών εισόδου και να την ενσωματώσει σε μια πολλαπλότητα. Σε σύγκριση με άλλους αλγόριθμους μείωσης διαστάσεων παρατηρούμε ότι στις περισσότερες περιπτώσεις το Pattern Search MDS παρέχει τα καλύτερα αποτελέσματα μεταξύ όλων των άλλων μεθόδων για SI SER. Πιθανώς, οι αναπαραστάσεις RQA έχουν μια εξαιρετικά μη γραμμική δομή μεταξύ διαφορετικών τάξεων συναισθημάτων και αυτός είναι ο λόγος για τον οποίο οι γραμμικοί αλγόριθμοι μείωσης της διαστασιολόγησης αποτυγχάνουν να διατηρήσουν εκείνη την πληροφορία που φέρουν σε παραστάσεις μικρού διαστάσεων. Οι αναπαραστάσεις που λαμβάνονται από τις μεθόδους NLDR μπορούν ακόμα να διατηρήσουν την διακριτική ικανότητα του χώρου μεγάλης διαστάσεως για όλες τις διαμορφώσεις και τα σύνολα χαρακτηριστικών και ακόμη και να ξεπεράσουν τις μετρήσεις απόδοσης WA και UA χρησιμοποιώντας το KNN. Το ίδιο ισχύει όταν αξιολογούμε τον αλγόριθμό μας προς τα πειράματα LOSO χρησιμοποιώντας το σύνολο δεδομένων IEMOCAP. Όλα αυτά τα πειράματα έχουν εκτελεστεί για όλους τους συνδυασμούς συνόλων χαρακτηριστικών που προτείνονται στις Ενότητες 4.5.1 (Χαρακτηριστικά IS10: Χαρακτηριστικά 1582), 4.5.2 (Χαρακτηριστικά RQA Χαρακτηριστικά 432 Χαρακτηριστικά) και 4.5.3 (Συνδυασμένη Λειτουργία Χαρακτηριστικών: Χαρακτηριστικά 2014). Διαπιστώνουμε με συνέπεια ότι ο ταξινομητής KNN τα πηγαίνει καλύτερα όταν χρησιμοποιούμε τον προτεινόμενο αλγόριθμο σε σύγκριση με τις αρχικές, υψηλού επιπέδου παραστάσεις συναισθηματικών ομιλιών. Τέλος, παρέχουμε ορισμένες απεικονίσεις από τις πολλαπλότητες που μαθαίνονται για τις συναισθηματικές εκφράσεις σε ευκλείδιους χώρους $2D$ και $3D$ προκειμένου να αποκτήσουμε μια ποιοτική άποψη των πλεονεκτημάτων και των μειονεκτημάτων της εφαρμογής κάθε αλγόριθμου μείωσης των διαστάσεων.

6.2 Μελλοντική δουλειά

Στο πλαίσιο εξαγωγής καλύτερων μη γραμμικών αναπαραστάσεων από RPs προκειμένου να βελτιωθεί η απόδοση των συστημάτων SER, θα μπορούσαμε να προσπαθήσουμε να απομονώσουμε τα αμετάβλητα χαρακτηριστικά από αυτές τις αναπαραστάσεις, που είναι η μικρή περιοδική υποεικόνα των εικόνων. Όπως αναφέρθηκε προηγουμένως, μία από τις κύριες προκλήσεις είναι να συλλάβουμε αμετάβλητες δυναμικές υπογραφές από κάθε πλαίσιο ομιλίας (βλ. Το τμήμα 1.4.2). Τα φωνήεντα κυριαρχούνται από τη θεμελιώδη συχνότητα του ομιλητή και, ως εκ τούτου, θα θέλαμε να απαλλαγούμε από αυτά τα χαρακτηριστικά που σχετίζονται με τους ομιλητές στις μη γραμμικές παραστάσεις μας από RPs. Μια περίληψη του προτεινόμενου σχεδίου εξαγωγής χαρακτηριστικών παρουσιάζεται στο σχήμα 6.1. Πρώτα, εκτελούμε τμηματοποίηση σε φωνήεντα, προκειμένου να επικεντρωθούμε μόνο στις περιοχές ομιλίας που κυριαρχούνται από την θεμελιώδη συχνότητα του σήματος ομιλίας. Η υπόλοιπη διαδικασία εξαγωγής χαρακτηριστικών είναι πανομοιότυπη με αυτήν που παρουσιάστηκε στο προηγούμενο τμήμα 4.5.2 μέχρι το τμήμα της εξαγωγής του περιοδικού υπογράφου από κάθε RP. Μετά από αυτό, κάθε κομμάτι ομιλίας ή μάλιστα η έκφραση συναισθηματικής ομιλίας μπορεί να αναπαρασταθεί από μια ακολουθία αυτών των υπογράφων RP που ελπίζουμε να διατηρήσουν δομές ενδεικτικές της φύσης κάθε συναισθηματικής εκδήλωσης. Τέλος, οι συνεχείς αναπαραστάσεις

RP είναι ήδη κανονικοποιημένες με ενέργεια και υπό τον προτεινόμενο αλγόριθμο εξαγωγής χαρακτηριστικών θα μπορούσαμε να εξαγάγουμε πολύ πιο ανθεκτικά μη γραμμικά χαρακτηριστικά για το SER χωρίς να κυριαρχούν στις αναπαραστάσεις χαρακτηριστικών μας στοιχεία που σχετίζονται με τον ομιλητή και όχι τόσο με την υποκείμενη δυναμική των συναισθηματικών εκφράσεων ομιλίας.



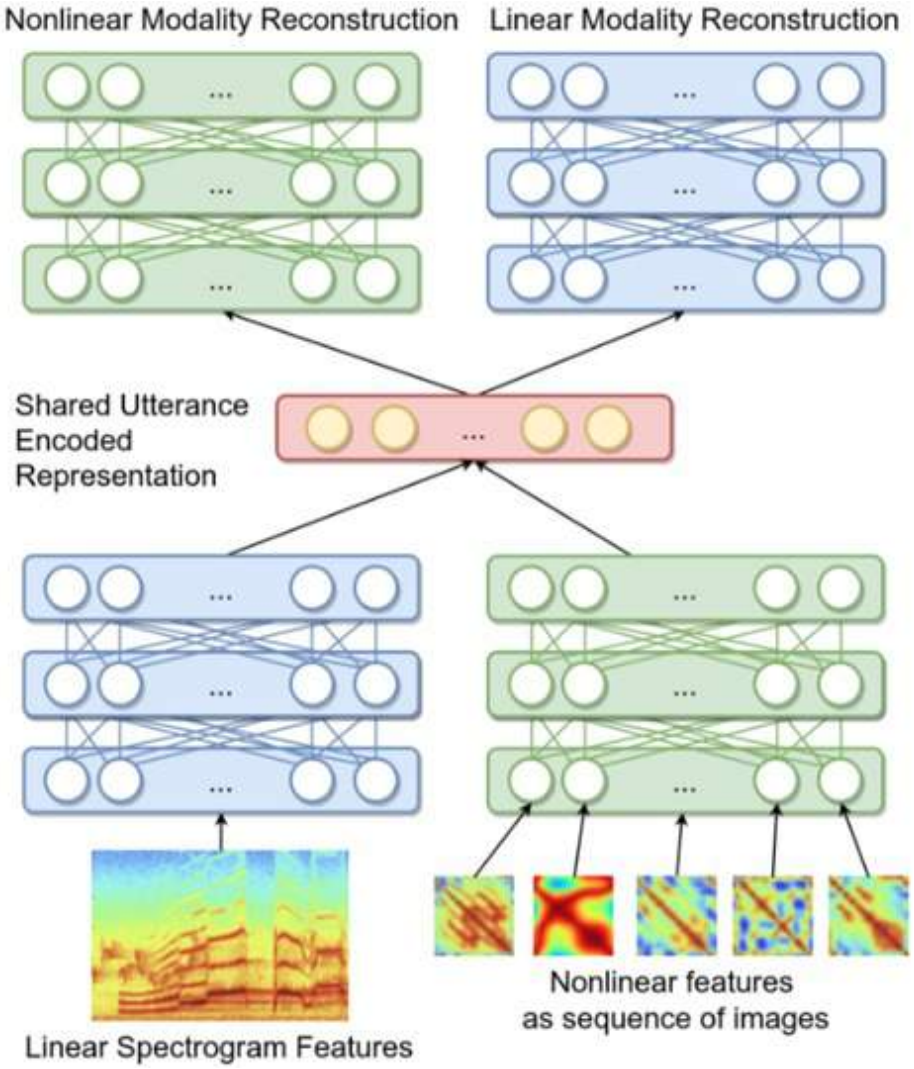
Σχήμα 6.1: Απομόνωση περιοδικών υπογραφών από RPs προκειμένου να εκχυλίσουμε αναλλοίωτες αναπαραστάσεις για SER

Επιπλέον, μπορούμε επίσης να προσπαθήσουμε να συγχωνεύσουμε καλύτερα τις πληροφορίες από διαφορετικά σύνολα ακουστικών χαρακτηριστικών για την κατασκευή πιο ισχυρών συστημάτων SER. Όπως έχουμε δει στα τμήματα 4.7.2, 4.7.3, 4.7.4 και 5.7.7, το σύνολο χαρακτηριστικών Fused (RQA + IS10) απέδωσε σημαντική βελτίωση σε σχέση με τα σύνολα χαρακτηριστικών IS10 και RQA. Ωστόσο, η απλή συνένωση των δύο χαρακτηριστικών παραστάσεων είναι ο πιο αφελής τρόπος συνδυασμού πληροφοριών που εξάγονται από διαφορετικές πηγές. Σε αυτό το πλαίσιο, μπορούμε να

βασιστούμε στην προηγούμενη ιδέα της εξαγωγής μιας ακολουθίας υπογραφών RP χωρίς περιοδικότητα από τα φωνήεντα κάθε τμήματος ομιλίας ή έκφρασης και να τα συνδυάσουμε με φασματικές αναπαραστάσεις (ουσιαστικά μια ακολουθία μετασχηματισμών Fourier από όλα τα πλαίσια ομιλίας που περιλαμβάνονται στο το τμήμα ομιλίας). Ο συνδυασμός και των δύο καναλιών πληροφοριών (γραμμικών και μη γραμμικών παραστάσεων) θα μπορούσε να γίνει αποτελεσματικά με τη χρήση ενός διπολικού αυτοκωδικοποιητή όπως αυτός που παρουσιάζεται στην Εικόνα 6.2. Εκπαιδεύουμε έναν αυτόνομο κωδικοποιητή, ο οποίος βασίζεται στην ανακατασκευή της πληροφορίας και των δύο καναλιών όταν μόνο μία μόνη δίοδος πληροφορίας είναι διαθέσιμη κάθε φορά. Με αυτή την τεχνική, το προτεινόμενο πλαίσιο θα ήταν ικανό όχι μόνο να ανακατασκευάσει την ελλείπουσα πληροφορία αλλά και να αποκτήσει μια ικανοποιητική αναπαράσταση μέσα στο ενδιάμεσο στρώμα που ονομάζεται ‘’Κοινή-Shared αναπαράσταση-representation”, στην οποία θα έχει συμπυκνωθεί σημαντική πληροφορία και από τα δύο κανάλια. Κάθε ένα από τα δύο μέρη του αυτοκωδικοποιητή θα μπορούσε να υλοποιήσει οποιαδήποτε αρχιτεκτονική DNN ή RNN, αλλά θα πρέπει να είναι ταυτόσημη μεταξύ των αντικατοπτρισμένων μπλοκ. Συγκεκριμένα, το τμήμα *μπλέ* του autoencoder (κάτω αριστερά και κάτω δεξιά) θα τιμωρείται όχι μόνο από το σφάλμα της ανακατασκευής (“Ανακατασκευή Γραμμικής Πληροφορίας”) αλλά και από την ανακατασκευή του διπλού καναλιού πληροφοριών (‘ Μη γραμμική ανακατασκευή πληροφορίας”). Παρόμοια πράγματα θα ισχύουν για την μη γραμμική δυική αντιστοίχηση του δικτύου (*πράσινο*). Ελπίζουμε ότι η ενδιάμεση αναπαράσταση θα κωδικοποιούσε σημαντικές συναισθηματικές πληροφορίες στο τέλος της εκπαιδευτικής διαδικασίας και θα μπορούσε να χρησιμοποιηθεί ως διάνυσμα εισόδου για τα επόμενα στρώματα βαθύτερων συστημάτων SER.

Λαμβάνοντας υπόψη κάποια επόμενα βήματα για την προτεινόμενη Pattern Search MDS 5.3, η μελλοντική εργασία θα επικεντρωθεί επίσης στη βελτίωση της απόδοσης εκτέλεσης και της κλιμάκωσης του προτεινόμενου αλγορίθμου. Συγκεκριμένα, μια προσέγγιση για τη μείωση της υπολογιστικής πολυπλοκότητας ανά εποχή θα ήταν να περιοριστεί ο χώρος αναζήτησης πιθανών κινήσεων καθώς η γεωμετρία του χώρου ενσωμάτωσης καθίσταται περισσότερο εμφανής με την απομάκρυνση των κινήσεων προς τα κύρια διανύσματα της γειτονιάς του σημείου που μετακινήθηκε στην εκάστωτε επανάληψη. Αυτό μπορεί να θεωρηθεί ως ένας συνδυασμός αναζήτησης προτύπων και μείωσης της κλίσης, όπου ο χώρος αναζήτησης των κινήσεων είναι ευρύς στην αρχή και στη συνέχεια γίνεται ολοένα και πιο μικρό, με την κατεύθυνση της κλίσης να μοιάζει η βέλτιστη επιλογή. Ο αλγόριθμός μας μπορεί να κλιμακωθεί σε μεγάλους αριθμούς σημείων χρησιμοποιώντας σημεία Φάρους (Landmarks) [150] ή γρήγορες προσεγγίσεις στο MDS [181]. Αυτές οι προσεγγίσεις αποσκοπούν στην μείωση του κόστους υπολογισμών και μνήμης για τον υπολογισμό της μήτρας πλήρους απόστασης, προσεγγίζοντας τη γεωμετρία των δεδομένων χρησιμοποιώντας μικρότερα υποτμήματα του αρχικού πίνακα ομοιοτήτων. Επιπλέον, οι στοχαστικές προσεγγίσεις όπως το στοχαστικό SMACOF [182] μπορούν να προσαρμοστούν σε MDS αναζήτησης προτύπων.

Σκοπεύουμε επίσης να παρέχουμε περισσότερες σε βάθος θεωρητικές ιδέες και τρόπους για να επιτρέψουμε στο Pattern Search MDS να καταγράψει πολύπλοκες γεωμετρικές ιδιότητες των δεδομένων εισόδου. Σκοπός μας είναι να εκτελέσουμε μια λεπτομερή ανάλυση σχετικά με το πώς τα θεωρητικά και ιδιαίτερα τις “κακές κινήσεις” που επηρεάζουν την απόδοση του Pattern Search MDS. Επιπλέον, στις Ενότητες 5.7.2 και 5.7.6 παρουσιάσαμε ότι το MDS μπορεί να χειριστεί καλύτερα τα αραιά δεδομένα και το LLE μπορεί να χειριστεί καλύτερα τη μη κυρτότητα και την έλλειψη δεδομένων. Αυτό έχει νόημα, καθώς το MDS λαμβάνει υπόψη την παγκόσμια γεωμετρία του χώρου ενσωμάτωσης, ενώ το LLE επικεντρώνεται στη γεωμετρία των τοπικών γειτονιών. Σχεδιάζουμε να συνδυάσουμε τις λειτουργίες κόστους αυτών των προσεγγίσεων για να εξάγουμε τόσο την παγκόσμια όσο και την τοπική γεωμετρία της πολλαπλότητας πάνω στην οποία κινούνται τα δεδομένα και είναι μικρής διαστασιμότητας. Ένας άλλος τρόπος για να αυξηθεί η εκφραστικότητα του αλγορίθμου είναι η διερεύνηση μιας ευρύτερης ποικιλίας μετρήσεων απόστασης, και συγκεκριμένα μη συμμετρικών μετρήσεων απόστασης, με κίνητρο τις επιστήμες που μελετούν αυτά τα μοντέλα σε αντιστοίχηση με το πώς δουλεύει ο εγκέφαλος [183].



Σχήμα 6.2: Διπολικός Αυτοκωδικοποιητής για συνδυασμό φασματικών και μη γραμμικών αναπαραστάσεων σημάτων ομιλίας

Βιβλιογραφία

[1] E. Tzinis, G. Paraskevopoulos, C. Baziotis, and A. Potamianos, "Integrating recurrence dynamics for speech emotion recognition," in *Proceedings of INTERSPEECH, (in press)*, 2018.

[2] E. Tzinis and A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 2017, pp. 190–195.

[3] G. Paraskevopoulos, E. Tzinis, E.-V. Vlatakis-Gkaragkounis, and A. Potamianos, "Pattern search multidimensional scaling," *arXiv:1806.00416*, 2018.

[4] A. Chorianopoulou, E. Tzinis, E. Iosif, A. Papoulidi, C. Papailiou, and A. Potamianos, "Engagement detection for children with autism spectrum disorder," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5055–5059.

[5] K. Nesbitt, K. Blackmore, G. Hookham, F. Kay-Lambkin, and P. Walla, "Using the startle eye-blink to measure affect in players," in *Serious Games Analytics*. Springer, 2015, pp. 401–434.

[6] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

[7] A. Jafari, F. Almasganj, and M. N. Bidhendi, "Statistical modeling of speech poincaré sections in combination of frequency analysis to improve speech recognition performance," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 20, no. 3, p. 033106, 2010.

[8] P. M. Niedenthal, L. W. Barsalou, P. Winkielman, S. Krauth-Gruber, and F. Ric, "Embodiment in attitudes, social perception, and emotion," *Personality and social psychology review*, vol. 9, no. 3, pp. 184–211, 2005.

[9] R. R. Cornelius, *The science of emotion: Research and tradition in the psychology of emotions.* Prentice-Hall, Inc, 1996.

[10] T. C. Schneirla, "An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal." 1959.

[11] H. Schlosberg, "Three dimensions of emotion." *Psychological review*, vol. 61, no. 2, p. 81, 1954.

[12] K. R. Scherer, "Vocal affect expression: A review and a model for future research." *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[13] I. Fónagy and K. Magdics, "Emotional patterns in intonation and music," *STUF-Language Typology and Universals*, vol. 16, no. 1-4, pp. 293–326, 1963.

[14] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[15] C. E. Izard, *Human emotions*. Springer Science & Business Media, 2013.

[16] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proceedings of Artificial Neural Networks in Engineering*, vol. 710, 1999.

[17] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[18] R. W. Picard *et al.*, "Affective computing," 1995.

[19] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[20] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[21] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.

[22] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. Newman, "Stress and emotion classification using jitter and shimmer features," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1081.

[23] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proceedings of INTERSPEECH*, 2010, pp. 2794–2797.

[24] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[25] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[26] S. Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*. University of Erlangen-Nuremberg Erlangen, Germany, 2009.

[27] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *Proceedings of ICASSP*, 2017, pp. 2741–2745.

[28] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, 2017.

[29] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[30] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2011.

[31] B. Schuller, R. Müller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[32] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[33] C. M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 240–243.

[34] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," *Information processing & management*, vol. 45, no. 3, pp. 315–328, 2009.

[35] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[36] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition." in *Proceedings of INTERSPEECH*, 2016, pp. 3603–3607.

[37] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[38] J. C. Vasquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, J. Vargas-Bonilla, and E. Noeth, "Wavelet-based time-frequency representations for automatic recognition of emotions from speech," in *Proceedings of the 12. ITG Symposium on Speech Communication*, 2016, pp. 1–5.

[39] F. Shah *et al.*, "Wavelet packets for speech emotion recognition," in *Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017 Third International Conference on*. IEEE, 2017, pp. 479–481.

[40] R. Sun and E. Moore, "Investigating glottal parameters and teager energy operators in emotion recognition," in *Affective Computing and Intelligent Interaction*, 2011, pp. 425–434.

[41] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory teager energy cepstrum coefficients for robust speech recognition," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[42] A. Shahzadi, A. Ahmadyfard, A. Harimi, and K. Yaghmaie, "Speech emotion recognition using nonlinear dynamics features," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 23, no. Sup. 1, pp. 2056–2073, 2015.

[43] T. Chaspari, D. Dimitriadis, and P. Maragos, "Emotion classification of speech using modulation features," in *Proceedings of Signal Processing Conference (EUSIPCO)*, 2014, pp. 1552–1556.

[44] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5200–5204.

[45] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[46] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 582–596, 2009.

[47] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 205–211.

[48] R. Xia, J. Deng, B. Schuller, and Y. Liu, "Modeling gender information for emotion recognition using denoising autoencoder," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 990–994.

[49] Y. Sun and G. Wen, "Ensemble softmax regression model for speech emotion recognition," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 8305–8328, 2017.

[50] B. Schuller and G. Rigoll, "Timing levels in segment-based speech emotion recognition," in *Proceedings of INTERSPEECH*, 2006.

[51] J. H. Jeon, R. Xia, and Y. Liu, "Sentence level emotion recognition based on decisions from subsentence segments," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4940–4943.

[52] X. Mao, L. Chen, and L. Fu, "Multi-level speech emotion recognition based on hmm and ann," in *Computer Science and Information Engineering, 2009 WRI World Congress on*, vol. 7. IEEE, 2009, pp. 225–229.

[53] E. M. Provost, "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3682–3686.

[54] C.-H. Wu, W.-B. Liang, K.-C. Cheng, and J.-C. Lin, "Hierarchical modeling of temporal course in emotional expression for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 810–814.

[55] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[56] A. Chorianopoulou, P. Koutsakis, and A. Potamianos, "Speech emotion recognition using affective saliency." in *INTERSPEECH*, 2016, pp. 500–504.

[57] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden markov models," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[58] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[59] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[60] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[61] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," 2015.

[62] Z.-Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5150–5154.

[63] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[64] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[65] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proceedings of ICASSP*, 2017, pp. 2227–2231.

[66] C.-W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition." in *Proceedings of INTERSPEECH*, 2016, pp. 1387–1391.

[67] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[68] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[69] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Transactions on Affective Computing*, 2017.

[70] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," *arXiv preprint arXiv:1706.03256*, 2017.

[71] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[72] Y. Bengio and P. Frasconi, "Credit assignment through time: Alternatives to backpropagation," in *Advances in Neural Information Processing Systems*, 1994, pp. 75–82.

[73] N. van Noord and E. Postma, "Learning scale-variant and scale-invariant features for deep image classification," *Pattern Recognition*, vol. 61, pp. 583–592, 2017.

[74] K. Honda, "Physiological processes of speech production," in *Springer handbook of speech processing*. Springer, 2008, pp. 7–26.

[75] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering OnLine*, vol. 6, no. 1, p. 23, 2007.

[76] H. Herzel, "Bifurcations and chaos in voice signals," *Applied Mechanics Reviews*, vol. 46, no. 7, pp. 399–413, 1993.

[77] W. T. Fitch, J. Neubauer, and H. Herzel, "Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production," *Animal behaviour*, vol. 63, no. 3, pp. 407–418, 2002.

[78] V. Pitsikalis and P. Maragos, "Analysis and classification of speech signals by generalized fractal dimension features," *Speech Communication*, vol. 51, no. 12, pp. 1206–1223, 2009.

[79] C. L. Webber Jr and J. P. Zbilut, "Recurrence quantification analysis of nonlinear dynamical systems," *Tutorials in contemporary nonlinear methods for the behavioral sciences*, pp. 26–94, 2005.

[80] H. Herzel, D. Berry, I. Titze, and I. Steinecke, "Nonlinear dynamics of the voice: signal analysis and biomechanical modeling," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 5, no. 1, pp. 30–34, 1995.

[81] S. S. Narayanan and A. A. Alwan, "A nonlinear dynamical systems analysis of fricative consonants," *The Journal of the Acoustical Society of America*, vol. 97, no. 4, pp. 2511–2524, 1995.

[82] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of statistical Physics*, vol. 65, no. 3-4, pp. 579–616, 1991.

[83] G. Boeing, "Visual analysis of nonlinear dynamical systems: Chaos, fractals, self-similarity and the limits of prediction," *Systems*, vol. 4, no. 4, 2016.

[84] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," *Physics reports*, vol. 438, no. 5-6, pp. 237–329, 2007.

[85] R. Bellman, "Adaptative control processes," 1961.

[86] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: a comparative," *J Mach Learn Res*, vol. 10, pp. 66–71, 2009.

[87] L. Cayton, "Algorithms for manifold learning," *Univ. of California at San Diego Tech. Rep*, vol. 12, no. 1-17, p. 1, 2005.

[88] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[89] Y. LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[90] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural networks for perception*. Elsevier, 1992, pp. 65–93.

[91] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[92] M. Aizerman, E. M. Braverman, and L. Rozonoer, "Theoretical foundations of potential function method in pattern recognition," *Automation and Remote Control*, vol. 25, pp. 917–936, 1964.

[93] M. B. Christopher, *PATTERN RECOGNITION AND MACHINE LEARNING*. Springer-Verlag New York, 2016.

[94] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical review A*, vol. 33, no. 2, p. 1134, 1986.

[95] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Phys. Rev. A*, vol. 45, pp. 3403–3411, 1992.

[96] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *EPL (Europhysics Letters)*, vol. 4, no. 9, p. 973, 1987.

[97] S. Schinkel, O. Dimigen, and N. Marwan, "Selection of recurrence threshold for signal detection," *The european physical journal special topics*, vol. 164, no. 1, pp. 45–53, 2008.

[98] M. Thiel, M. C. Romano, J. Kurths, R. Meucci, E. Allaria, and F. T. Arecchi, "Influence of observational noise on the recurrence quantification analysis," *Physica D: Nonlinear Phenomena*, vol. 171, no. 3, pp. 138–152, 2002.

[99] W. S. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, Dec 1952.

[100] I. Borg and P. J. F. Groenen, *Modern multidimensional scaling: theory and applications*. Springer, 2005.

[101] J. C. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis," *Biometrika*, vol. 53, no. 3-4, pp. 325–338, 1966.

[102] M. A. A. Cox and T. F. Cox, "Multidimensional scaling on the sphere," in *Compstat*, D. Edwards and N. E. Raun, Eds. Heidelberg: Physica-Verlag HD, 1988, pp. 323–328.

[103] A. Cvetkovski and M. Crovella, "Low-stress data embedding in the hyperbolic plane using multidimensional scaling," *Appl. Math*, vol. 11, no. 1, pp. 5–12, 2017.

[104] H. Lindman and T. Caelli, "Constant curvature riemannian scaling," *Journal of Mathematical Psychology*, vol. 17, no. 2, pp. 89–109, 1978.

[105] R. N. Shepard, "The analysis of proximities: Multidimensional scaling with an unknown distance function. i." *Psychometrika*, vol. 27, no. 2, pp. 125–140, Jun 1962.

[106] ——, "The analysis of proximities: Multidimensional scaling with an unknown distance function. ii," *Psychometrika*, vol. 27, no. 3, pp. 219–246, Sep 1962.

[107] P. Groenen and I. Borg, "Past, present, and future of multidimensional scaling," *Visualization and Verbalization of Data*, pp. 95–117, 2014.

[108] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, Mar 1964.

[109] ——, "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, Jun 1964.

[110] S. L. France and J. D. Carroll, "Two-way multidimensional scaling: A review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 5, pp. 644–661, 2011.

[111] J. D. Leeuw, I. J. R. Barra, F. Brodeau, G. Romier, and B. V. C. (eds, "Applications of convex analysis to multidimensional scaling," in *Recent Developments in Statistics*. North Holland Publishing Company, 1977, pp. 133–146.

[112] J. de Leeuw, "Convergence of the majorization method for multidimensional scaling," *Journal of Classification*, vol. 5, no. 2, pp. 163–180, Sep 1988.

[113] V. Torczon, "On the convergence of pattern search algorithms," *SIAM Journal on optimization*, vol. 7, no. 1, pp. 1–25, 1997.

[114] E. D. Dolan, R. M. Lewis, and V. Torczon, "On the local convergence of pattern search," *SIAM Journal on Optimization*, vol. 14, no. 2, pp. 567–583, 2003.

[115] R. M. Lewis and V. Torczon, "Pattern search algorithms for bound constrained minimization," *SIAM Journal on Optimization*, vol. 9, no. 4, pp. 1082–1099, 1999.

[116] A. R. Conn, N. I. M. Gould, and P. Toint, "A globally convergent augmented lagrangian algorithm for optimization with general constraints and simple bounds," *SIAM Journal on Numerical Analysis*, vol. 28, no. 2, pp. 545–572, 1991.

[117] R. M. Lewis and V. Torczon, "Pattern search methods for linearly constrained minimization," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 917–941, 2000.

[118] C. Audet, "Convergence results for generalized pattern search algorithms are tight," *Optimization and Engineering*, vol. 5, no. 2, pp. 101–122, Jun 2004.

[119] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013, pp. 835–838.

[120] L. R. Rabiner, R. W. Schafer *et al.*, "Introduction to digital speech processing," *Foundations and Trends® in Signal Processing*, vol. 1, no. 1–2, pp. 1–194, 2007.

[121] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[122] D. M. Hawkins, "The problem of overfitting," *Journal of chemical information and computer sciences*, vol. 44, no. 1, pp. 1–12, 2004.

[123] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[124] J. Bayer, C. Osendorfer, D. Korhammer, N. Chen, S. Urban, and P. van der Smagt, "On fast dropout and its applicability to recurrent networks," *arXiv preprint arXiv:1311.0701*, 2013.

[125] T. Dozat, "Incorporating nesterov momentum into adam," 2016.

[126] C. Laurent, G. Pereyra, P. Brakel, Y. Zhang, and Y. Bengio, "Batch normalized recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2657–2661.

[127] F. Chollet *et al.*, "Keras," 2015.

[128] J. Bergstra, F. Bastien, O. Breuleux, P. Lamblin, R. Pascanu, O. Delalleau, G. Desjardins, D. Warde-Farley, I. Goodfellow, A. Bergeron *et al.*, "Theano: Deep learning on gpus with python," in *NIPS 2011, BigLearning Workshop, Granada, Spain*, vol. 3. Citeseer, 2011.

[129] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[130] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.

[131] P. Henríquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, and J. R. Orozco-Arroyave, "Nonlinear dynamics characterization of emotional speech," *Neurocomputing*, vol. 132, pp. 126–135, 2014.

[132] A. Lombardi, P. Guccione, and C. Guaragnella, "Exploring recurrence properties of vowels for analysis of emotions in speech," *Sensors & Transducers*, vol. 204, no. 9, p. 45, 2016.

[133] R. V. Donner, M. Small, J. F. Donges, N. Marwan, Y. Zou, R. Xiang, and J. Kurths, "Recurrence-based time series analysis by means of complex network methods," *International Journal of Bifurcation and Chaos*, vol. 21, no. 04, pp. 1019–1046, 2011.

[134] F. Orsucci, R. Petrosino, G. Paoloni, L. Canestri, E. Conte, M. A. Reda, and M. Fulcheri, "Prosody and synchronization in cognitive neuroscience," *EPJ Nonlinear Biomedical Physics*, vol. 1, no. 1, p. 6, 2013.

[135] N. D. Duran, R. Dale, C. T. Kello, C. N. Street, and D. C. Richardson, "Exploring the movement dynamics of deception," *Frontiers in psychology*, vol. 4, p. 140, 2013.

[136] A. Goshvarpour, A. Abbasi, and A. Goshvarpour, "Dynamical analysis of emotional states from electroencephalogram signals," *Biomedical Engineering: Applications, Basis and Communications*, vol. 28, no. 02, p. 1650015, 2016.

[137] T. Tošić, K. K. Sellers, F. Fröhlich, M. Fedotenkova, A. Hutt *et al.*, "Statistical frequency-dependent analysis of trial-to-trial variability in single time series by recurrence plots," *Frontiers in systems neuroscience*, vol. 9, p. 184, 2016.

[138] N. Hatami, Y. Gavet, and J. Debayle, "Classification of time-series images using deep convolutional neural networks," in *Tenth International Conference on Machine Vision (ICMV 2017)*, vol. 10696. International Society for Optics and Photonics, 2018, p. 106960Y.

[139] S. Haq and P. Jackson, "Speaker-dependent audio-visual emotion recognition," in *Proceedings Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Norwich, UK*, Sept. 2009.

[140] N. Marwan, J. Kurths, and S. Foerster, "Analysing spatially extended high-dimensional dynamics by recurrence plots," *Physics Letters A*, vol. 379, no. 10, pp. 894 – 900, 2015.

[141] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[142] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[143] Y. Sun, G. Wen, and J. Wang, "Weighted spectral features based on local hu moments for speech emotion recognition," *Biomedical signal processing and control*, vol. 18, pp. 80–90, 2015.

[144] K. F. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[145] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[146] M. Bernstein, V. D. Silva, J. C. Langford, and J. B. Tenenbaum, "Graph approximations to geodesics on embedded manifolds," 2000.

[147] H. Zha and Z. Zhang, "Isometric embedding and continuum isomap," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 2003, pp. 864–871.

[148] D. L. Donoho and C. Grimes, "Image manifolds which are isometric to euclidean space," *Journal of Mathematical Imaging and Vision*, vol. 23, no. 1, pp. 5–24, Jul 2005.

[149] R. Pless, "Image spaces and video trajectories: Using isomap to explore video sequences." in *ICCV*, vol. 3, 2003, pp. 1433–1440.

[150] V. Silva and J. B. Tenenbaum, "Sparse multidimensional scaling using landmark points," 01 2004.

[151] V. D. Silva and J. B. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction," in *Advances in Neural Information Processing Systems 15*. MIT Press, 2003, pp. 705–712.

[152] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003.

[153] L. Cayton and S. Dasgupta, "Robust euclidean embedding," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 169–176.

[154] L. K. Saul and S. T. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

[155] F. Sha and L. K. Saul, "Analysis and extension of spectral methods for nonlinear dimensionality reduction," in *Proceedings of the 22Nd International Conference on Machine Learning*, ser. ICML '05. New York, NY, USA: ACM, 2005, pp. 784–791.

[156] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002, pp. 585–591.

[157] Z. Zhang and J. Wang, "Mlle: Modified locally linear embedding using multiple weights," in *Advances in neural information processing systems*, 2007, pp. 1593–1600.

[158] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM journal on scientific computing*, vol. 26, no. 1, pp. 313–338, 2004.

[159] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5591–5596, 2003.

[160] K. Q. Weinberger and L. K. Saul, "Unsupervised learning of image manifolds by semidefinite programming," *International journal of computer vision*, vol. 70, no. 1, pp. 77–90, 2006.

[161] K. Q. Weinberger, B. Packer, and L. K. Saul, "Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization." in *AISTATS*. Citeseer, 2005.

[162] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Rev.*, vol. 38, no. 1, pp. 49–95, Mar. 1996.

[163] D. P. Bertsekas, *Nonlinear programming*. Athena scientific Belmont, 1999.

[164] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in neural information processing systems*, 2002, pp. 585–591.

[165] ——, "Convergence of laplacian eigenmaps," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 129–136.

[166] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM J. Sci. Comput.*, vol. 26, no. 1, pp. 313–338, Jan. 2005.

[167] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[168] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[169] L. M. Rios and N. V. Sahinidis, “Derivative-free optimization: a review of algorithms and comparison of software implementations,” *Journal of Global Optimization*, vol. 56, no. 3, pp. 1247–1293, Jul 2013.

[170] M. Avriel, *Nonlinear programming: analysis and methods*. Courier Corporation, 2003.

[171] R. Hooke and T. A. Jeeves, ““ direct search” solution of numerical and statistical problems,” *J. ACM*, vol. 8, no. 2, pp. 212–229, Apr. 1961.

[172] G. E. Box, “Evolutionary operation: A method for increasing industrial productivity,” *Applied statistics*, pp. 81–101, 1957.

[173] V. J. Torczon, “Multidirectional search: a direct search algorithm for parallel machines,” Ph.D. dissertation, Rice University, 1989.

[174] J. J. E. Dennis and V. Torczon, “Direct search methods on parallel machines,” *SIAM Journal on Optimization*, vol. 1, no. 4, pp. 448–474, 1991.

[175] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[176] L. Dagum and R. Menon, “Openmp: an industry standard api for shared-memory programming,” *IEEE computational science and engineering*, vol. 5, no. 1, pp. 46–55, 1998.

[177] G. Golub and W. Kahan, “Calculating the singular values and pseudo-inverse of a matrix,” *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, vol. 2, no. 2, pp. 205–224, 1965.

[178] E. Bruni, N. K. Tran, and M. Baroni, “Multimodal distributional semantics,” *J. Artif. Int. Res.*, vol. 49, no. 1, pp. 1–47, Jan. 2014.

[179] F. Hill, R. Reichart, and A. Korhonen, “Simlex-999: Evaluating semantic models with (genuine) similarity estimation,” *Computational Linguistics*, 2015.

[180] C. Baziotis, N. Pelekis, and C. Doulkeridis, “Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis,” in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, August 2017, pp. 747–754.

[181] T. Yang, J. Liu, L. Mcmillan, and W. Wang, “A fast approximation to multidimensional scaling,” in *In Proc. of the IEEE Workshop on Computation Intensive Methods for Computer Vision*, 2006.

[182] K. Rajawat and S. Kumar, “Stochastic multidimensional scaling,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 2, pp. 360–375, 2017.

[183] E. Pothos and J. Busemeyer, “A quantum probability explanation for violations of symmetry in similarity judgments,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 33, no. 33, 2011.

Παράρτημα Α

Συντομογραφίες

(ACF): Λειτουργία αυτοσυσχέτισης
(AI): Τεχνητή Νοημοσύνη
(AMI): Μέση Αμοιβαία Πληροφορία
(ANN): Τεχνητό νευρωνικό δίκτυο
(ASD): Διαταραχή Αυτιστικού Φάσματος
(ASR): Αυτόματη αναγνώριση ομιλίας
(AWVL): Μέσο Μήκος Λευκών Κάθετων Γραμμών, μέτρηση ποσοτικής ανάλυσης υποτροπής
(BLSTMs): Μονάδες αμφίδρομης μνήμης μεγάλης διάρκειας μικρού χρόνου
(CNNs): Συνελικτικά Νευρωνικά Δίκτυα
(CTC): Συνδεσμιακή χρονική ταξινόμηση
(DAE): Αυτόματοι κωδικοποιητές με αποθορυβοποίηση
(DBN): Δίκτυο βαθιάς πίστης
(DDP): Διαφορά της διαφοράς των περιόδων
(DENTR): Διαγωνική Εντροπία, μέτρηση ποσοτικής ανάλυσης υποτροπής
(DET): Ντετερμινισμός, μέτρηση ποσοτικής ανάλυσης υποτροπής
(DNN): Βαθιά Νευρωνικά Δίκτυα
(EEG): Σήματα Ηλεκτροεγκεφαλογράμματος
(ELM): Μηχανές ακραίας μάθησης
(EM): Αλγόριθμος μεγιστοποίησης προσδοκίας
(EP): Προφίλ Συναισθημάτων
(ESR): Μοντέλο Softmax για Κατηγοριοποίηση Συναισθήματος
(FNN): Λανθάνοντες Πλησιέστεροι Γείτονες
(GFS): Σπεκτρόγραμμα Φάσματος Γλωττίδας
(ΓΤΜ): Μοντέλα Μείγματος Γκαουσιανών
(GN): Παγκόσμια Κανονικοποίηση
(GPS): Γενική αναζήτηση προτύπων
(GPU): Μονάδα Γραφικής Επεξεργασίας
(HMI): Διεπαφές ανθρώπου-μηχανής
(HMM): Κρυφό μοντέλο Markov
(HNR): Ακουστικό Χαρακτηριστικό: κλάσμα αρμονικών σε θόρυβο
(IEMOCAP): βάση δεδομένων Interactive Emotional Dyadic Motion Capture
(ISOMAP): Ισομετρική χαρτογράφηση χαρακτηριστικών, μη γραμμικός αλγόριθμος μείωσης διαστάσεων
(KNNs): μη-παραμετρικός ταξινομητής Κ-Πλησιέστεροι γείτονες
(L): Μέσο μήκος διαγωνίου, μέτρο ανάλυσης ποσοτικοποίησης επαναληψιμότητας
(LAM): Λαμιναριότητα, μέτρηση ποσοτικής ανάλυσης υποτροπής
(LE): Λαπλασιανοί Ιδιοχάρτες, μη γραμμικός αλγόριθμος μείωσης διαστάσεων
(LLDs): Περιγραφείς χαμηλού επιπέδου που εξάγονται από παράθυρα ομιλίας
(LLE): Τοπική γραμμική ενσωμάτωση, μη γραμμικός αλγόριθμος μείωσης διαστάσεων
(LOSO): Σχήμα αξιολόγησης: Άσε μια Συνεδρία Έξω
(LR): Ταξινομητής λογιστικής παλινδρόμησης

(LSP): Φασματικά ζεύγη γραμμών
(LSTM): Μονάδα μνήμης μεγάλης διάρκειας μικρού χρόνου
(LTSA): Τοπική ευθυγράμμιση χώρου εφαπτόμενων
(MDS): Πολυδιάστατος αλγόριθμος κλιμάκωσης για μη γραμμική μείωση διαστάσεων
(MFB): Χαμηλού επιπέδου περιγραφητής: Μελ Τράπεζα Φίλτρων
(MFCCs): Χαμηλού επιπέδου περιγραφητής: Μελ Κέπστρουμ Συντελεστές
(ML): Μηχανική Μάθηση
(MSE): μέσο τετραγωνικό σφάλμα
(NLDR): Λειτουργία μη γραμμικής μείωσης διαστάσεων
(NN): Νευρωνικό Δίκτυο
(PCA): Ανάλυση βασικών συνιστωσών Αλγόριθμος μείωσης διαστάσεων
(PS): Χώρος Φάσης ή χώρος κατάστασης ενός υπό ανάλυση συστήματος
(RBF): πυρήνας ακτινικής βάσης
(RMS): ρίζα μέσης τετραγωνικής τιμής
(RNNs): Ανατροφοδοτούμενα Νευρωνικά Δίκτυα
(RP): Σχέδιο-Γράφημα επαναληψιμότητας
(RQA): Ποσοτική ανάλυση επαναληψιμότητας, μέτρα πολυπλοκότητας που εξάγονται από ένα γράφημα επανάληψιμότητας
(RR): Ποσοστό επαναληψιμότητας
(SAE): Αραιά αυτόματοι κωδικοποιητές
(SAVEE): Βάση δεδομένων συναισθηματικών εκφράσεων Surrey Audio-Visual Expressed Emotion
(SD): Σχήμα αξιολόγησης: Εξαρτημένου-ομιλητή
(SER): Αναγνώριση συναισθημάτων από φωνή
(SHS): Υπο-αρμονικό άθροισμα
(SI): Σχήμα αξιολόγησης: Ανεξαρτήτως-ομιλητή
(SP): Σπεκτρόγραμμα
(USA): Υπο-προτασιακή Προσοχή για ένα ανατροφοδοτούμενο νευρωνικό δίκτυο
(SVD): Αποσύνθεση Singular τιμών
(SVM): Ταξινομητής με υποστηριξιακά διανύσματα
(TD): Τυπικά ανεπτυσσόμενα παιδιά
(TEO): Ενεργειακός χειριστής Teager
(TT): Χρόνος παγίδευσης, μέτρο ανάλυσης ποσοτικοποίησης επανάληψιμότητας
(UA): Μη σταθμισμένη μέτρηση επιτυχίας
(VAD): Μηχανισμός ανίχνευσης φωνητικής δραστηριότητας
(VENTR): Εντροπία κατακόρυφων γραμμών, μέτρο ανάλυσης ποσοτικοποίησης επανάληψιμότητας
(WA): Σταθμισμένη μέτρηση επιτυχίας
(WENTR): Λευκή κατακόρυφη εντροπία, μέτρο ανάλυσης ποσοτικοποίησης επανάληψιμότητας
(WPA): Μηχανισμός σταθμισμένης συγκέντρωσης προσοχής για ανατροφοδοτούμενα νευρωνικά δίκτυα
(WSFHM): Σταθμισμένα φασματικά χαρακτηριστικά βασισμένα σε Hu Moments
(ZCR): Ποσοστό μηδενικής διέλευσης