Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Σηματων Ελεγχου και Ρομποτικης

Πολυτροπικές Σημασιολογικές Αναπαραστάσεις Λέξεων με Βάση την Ανθρώπινη Αντίληψη

Διπλωματικη Εργασια

του

ΙΩΑΝΝΗ ΚΑΡΑΜΑΝΩΛΑΚΗ

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

Εργαστηριο Ορασης Υπολογιστων, Επεξεργασιας Λογου και Επεξεργασιας Σηματων
Αθήνα, Ιούνιος 2017

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

Πολυτροπικές Σημασιολογικές Αναπαραστάσεις Λέξεων με Βάση την Ανθρώπινη Αντίληψη

Διπλωματικη Εργασια

του

ΙΩΑΝΝΗ ΚΑΡΑΜΑΝΩΛΑΚΗ

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 12 Ιουνίου 2017.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Αλέξανδρος Ποταμιάνος	Πέτρος Μαραγκός	Άγγελος Πικράκης
Αναπληρωτής Καθηγητής	Καθηγητής	Επίκουρος Καθηγητής
Ε.Μ.Π.	Ε.Μ.Π.	Πανεπιστήμιο Πειραιώς

Αθήνα, Ιούνιος 2017

.....

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

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

Ευχαριστίες

Με την παρούσα Διπλωματική εργασία κλείνει ο κύκλος των προπτυχιακών μου σπουδών στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Κατά τη διάρκεια αυτών των χρόνων ήρθα σε επαφή με πολλούς αξιόλογους ανθρώπους, τους οποίους θα ήθελα να ευχαριστήσω θερμά για όσα μου προσέφεραν.

Πρώτα από όλα, θα ήθελα να ευχαριστήσω τον καθηγητή κ. Αλέξανδρο Ποταμιάνο, με τον οποίο είχα την τύχη να συνεργαστώ στα δύο τελευταία έτη των σπουδών μου. Οι διαλέξεις του στα μαθήματα 'Επεξεργασία Φωνής και Φυσικής Γλώσσας' και 'Αναγνώριση Προτύπων' αλλά και η συχνή από την πλευρά του αναφορά στη σύνδεση των μαθηματικών εννοιών με την ανθρώπινη νόηση και τη μουσική αρμονία έπαιξαν καθοριστικό ρόλο στην απόφασή μου να εμβαθύνω στους τομείς της Μηχανικής Μάθησης, της Υπολογιστικής Γλωσσολογίας και της Ανάκτησης Μουσικής Πληροφορίας. Επιπλέον, οι γνώσεις και η διαίσθησή του σε συνδυασμό με τις ατελείωτες συζητήσεις αναφορικά με τους παραπάνω τομείς συνέβαλαν καταλυτικά στην περάτωση αυτής της εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω ιδιαίτερα τους καθηγητές κ. Πέτρο Μαραγκό και κ. Άγγελο Πικράκη για τις εποικοδομητικές συζητήσεις που είχαμε αλλά και για τις πολύ ενδιαφέρουσες ερευνητικές ιδέες που μου πρότειναν. Ακόμη, ευχαριστώ τους μεταδιδακτορικούς ερευνητές Ηλία Ιωσήφ και Νάνσυ Ζλατίντση για τις πολύτιμες συμβουλές και τη βοήθεια που μου προσέφεραν σε όλα τα στάδια αυτής της εργασίας αλλά και τον Θεόδωρο Γιαννακόπουλο για την ευκαιρία που μου έδωσε να συνεργαστώ μαζί του στα πλαίσια πρακτικής άσκησης στο ερευνητικό κέντρο Ε.Κ.Ε.Φ.Ε 'Δημόκριτος'.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στην οικογένεια και τους φίλους μου που στέκονται στο πλευρό μου, με βοηθούν να ξεπεράσω κάθε δυσκολία και μου δίνουν δύναμη για να συνεχίσω την ερευνητική μου πορεία.

Γιάννης Καραμανωλάκης
Ιούνιος 2017

Η Διπλωματική αυτή εργασία αφιερώνεται στον παππού μου Νίκο, χωρίς τη συμβολή του οποίου δε θα είχα εκτιμήσει τη χαρά του να μαθαίνεις.

Περίληψη

Τα Κατανεμημένα Σημασιολογικά Μοντέλα είναι από τις πιο γνωστές υπολογιστικές μεθόδους για την αναπαράσταση της έννοιας των λέξεων, μοντελοποιώντας τα πρότυπα συνεμφάνισης λέξεων σε πηγές κειμένου. Λόγω της δυνατότητας υπολογισμού της σημασιολογικής ομοιότητας μεταξύ λέξεων, εφαρμόζονται με επιτυχία σε πολλά προβλήματα σχετικά με την Επεξεργασία Φυσικής Γλώσσας και την Εξαγωγή και Ανάκτηση Πληροφορίας. Ωστόσο, τα μοντέλα αυτά αδυνατούν να αναπαραστήσουν την έννοια των λέξεων σύμφωνα με την ανθρώπινη αντίληψη, καθώς ο άνθρωπος συνδυάζει την έννοια των λέξεων με τα ερεθίσματα που δέχεται μέσω των αισθήσεών του, όπως για παράδειγμα μέσω της όρασης ή της ακοής.

Στόχος αυτής της Διπλωματικής εργασίας είναι η αντιμετώπιση του παραπάνω προβλήματος μέσω της δημιουργίας Πολυτροπικών Κατανεμημένων Σημασιολογικών Μοντέλων. Αρχικά, γίνεται περιγραφή της διαδικασίας για τη δημιουργία Κατανεμημένων Σημασιολογικών Μοντέλων και στη συνέχεια παρουσιάζεται η επέκτασή τους σε πολυτροπικά μοντέλα. Δίνεται έμφαση σε πολυτροπικά μοντέλα ήχου για την αναπαράσταση των λέξεων με βάση τις ακουστικές τους ιδιότητες, ωστόσο πραγματοποιείται συνοπτική περιγραφή και των πολυτροπικών μοντέλων εικόνας. Έπειτα, γίνεται αναφορά σε μεθόδους για την Πολυτροπική Σύμπτυξη, δηλαδή τη σύμπτυξη των μοντέλων που βασίζονται σε χαρακτηριστικά κειμένων, ήχων και εικόνων με σκοπό τη δημιουργία μίας ενιαίας πολυτροπικής αναπαράστασης λέξεων.

Σύμφωνα με όσα γνωρίζουμε, η παρούσα εργασία αποτελεί την πρώτη προσπάθεια για τη δημιουργία αναπαραστάσεων λέξεων χρησιμοποιώντας ταυτόχρονα τα κειμενικά, ακουστικά και οπτικά τους χαρακτηριστικά. Μάλιστα, διαπιστώνεται ότι η απόδοση των πολυτροπικών μοντέλων ξεπερνά την απόδοση των παραδοσιακών κατανεμημένων μοντέλων στο πρόβλημα αξιολόγησης της σημασιολογικής ομοιότητας των λέξεων. Επίσης, προτείνονται δύο επεκτάσεις των πολυτροπικών μοντέλων ήχου, η μία εκ των οποίων σχετίζεται με τη χρήση και σύμπτυξη πολλαπλών χώρων χαρακτηριστικών ανάλογα με τη φύση του ήχου. Αυτή η επέκταση οδηγεί σε βελτίωση της απόδοσης στο ίδιο πρόβλημα σε σύγκριση με τα πολυτροπικά μοντέλα ήχου που αναφέρονται στη βιβλιογραφία.

Επιπλέον, οι πολυτροπικές αναπαραστάσεις λέξεων αποδεικνύονται χρήσιμες όχι μόνο για την εκτίμηση της ομοιότητας λέξεων αλλά και για την ομοιότητα μεταξύ λέξεων και ήχων, λέξεων και εικόνων κ.ο.κ. Στα πλαίσια αυτής της εργασίας, τα Κατανεμημένα Σημασιολογικά Μοντέλα Ήχου εφαρμόζονται για πρώτη φορά σε δύο προβλήματα του τομέα Ανάκτησης Μουσικής Πληροφορίας: τον αυτόματο χαρακτηρισμό ηχητικών αποσπασμάτων και την αξιολόγηση μουσικής ομοιότητας. Διαπιστώνεται ότι ο προτεινόμενος αλγόριθμος για τον χαρακτηρισμό

ήχων προβλέπει αντιπροσωπευτικές λέξεις που περιγράφουν το ηχητικό περιεχόμενο και θα μπορούσαν να αξιοποιηθούν σε περιπτώσεις έλλειψης μεταδεδομένων. Ακόμη, η ταυτόχρονη αξιοποίηση των ακουστικών χαρακτηριστικών και των περιγραφικών λέξεων για την αξιολόγηση της μουσικής ομοιότητας, καθιστά τον προτεινόμενο αλγορίθμο (μη επιβλεπόμενης μάθησης) συγκρίσιμο με τους αλγορίθμους που προτείνονται στη βιβλιογραφία. Η ικανοποιητική απόδοση των πολυτροπικών μοντέλων σε καθένα από τα παραπάνω προβλήματα αποτελεί κίνητρο για την εφαρμογή περισσότερων ιδεών στα πλαίσια μελλοντικής έρευνας.

Λέξεις Κλειδιά

Πολυτροπικά Κατανεμημένα Σημασιολογικά Μοντέλα, Πολυτροπική Σύμπτυξη, Σημασιολογική Ομοιότητα Λέξεων, Μουσική Ομοιότητα, Μάθηση Αναπαραστάσεων

Abstract

Distributional Semantic Models (DSMs) is a popular method for the representation of word meaning by modeling the patterns of word co-occurrence in text corpora. The ability of DSMs to estimate word similarity justifies their successful application for various problems related to Natural Language Processing and Information Extraction. However, DSMs have been criticized as "disembodied", since they rely solely on linguistic information without being grounded in human perception and action.

The goal of this diploma thesis is to alleviate this problem via the creation of Multimodal DSMs. First, we describe the procedure for the creation of traditional DSMs and then we present their extension to Multimodal DSMs. We focus on models based on the audio modality but we also describe models based on the visual modality. Moreover, we present various methods for the Multimodal Fusion, i.e., the fusion of multimodal text, audio and image models, in order to obtain joint word representations.

Up to our knowledge, this is the first attempt to create multimodal word representations using text, acoustic and visual features at the same time. The proposed multimodal DSMs outperform the traditional DSMs at the estimation of word semantic similarity. Also, an extension of the baseline audio-based DSM enables the fusion of multiple feature spaces depending on the nature of sound and yields significant improvement of accuracy compared to the state-of-the-art acoustic DSMs that are proposed in the literature.

In addition to semantic word similarity, multimodal word representations can be used for the estimation of similarity between words and sounds, words and images etc. In this work, the audio-based DSMs (ADSMs) are applied for two problems in the field of Music Information Retrieval, namely Audio Auto-tagging and Music Similarity Estimation. It appears that the proposed auto-tagging algorithm predicts words that accurately describe the audio content and could be used in the case of missing metadata. Moreover, due to the consideration of both acoustic features and tags for music similarity, the proposed unsupervised algorithm is comparable with state-of-the-art algorithms, which provides as a motivation for the examination of more research ideas in the future.

Keywords

Multimodal Distributional Semantic Models, Multimodal Fusion, Lexical Semantic Similarity, Auto-tagging, Music Similarity, Representation Learning

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Η Σημασιολογία των Λέξεων

Πώς αντιλαμβάνεται ο άνθρωπος τη σημασιολογία των λέξεων; Σκεφτείτε για παράδειγμα τη λέξη 'μπανάνα'. Κάποιος θα την αντιστοίχιζε στη λέξη 'φρούτο' λόγω της γνώσης ότι η μπανάνα ταξινομείται στην κατηγορία των φρούτων, αλλά και στη λέξη 'κίτρινο' λόγω του χρώματος της μπανάνας. Βέβαια, δεν έχει αναφερθεί κάπου ρητά ότι υπάρχουν μόνο κίτρινες μπανάνες, ωστόσο η επαναλαμβανόμενη οπτική επαφή του ανθρώπου με κίτρινες μπανάνες τον οδηγεί στην, ενδεχομένως αυθόρμητη, αντιστοίχιση της μπανάνας με το κίτρινο χρώμα. Αντίστοιχα, η ανθρώπινη αντίληψη της έννοιας της λέξης 'κιθάρα' συνδέεται άμεσα με τον ήχο που παράγεται από το μουσικό όργανο, ιδιότητα που το διαχωρίζει από άλλα μουσικά όργανα που έχουν διαφορετική ακουστική χροιά. Γενικότερα, υπάρχουν ενδείξεις ότι ο άνθρωπος μαθαίνει τη σημασιολογία των λέξεων βασίζοντας τις λέξεις στα ερεθίσματα που λαμβάνει μέσω των αισθήσεών του (όραση, ακοή, όσφρηση, γεύση, αφή) αλλά και στην αισθησεο-κινητική του εμπειρία [7, 8]. Χαρακτηριστικά είναι τα αποτελέσματα μίας σειράς πειραμάτων [9], όπου διαπιστώθηκε ότι η προφορά των ρημάτων 'κλωτσάω', 'πιάνω' και 'γλείφω' προκαλούσε την ενεργοποίηση των περιοχών του ανθρώπινου εγκεφάλου που σχετίζονται με την κίνηση του ποδιού, του χεριού και της γλώσσας αντίστοιχα. Εξάλλου, ο άνθρωπος χρησιμοποιεί λέξεις για να περιγράψει ό,τι ακούει, βλέπει, μυρίζει, γεύεται και νιώθει, γεγονός που επιβεβαιώνει τη σύνδεση της σημασιολογίας των λέξεων με τις ανθρώπινες αισθήσεις και εμπειρίες.

1.2 Τα Κατανεμημένα Σημασιολογικά Μοντέλα

Τα Κατανεμημένα Σημασιολογικά Μοντέλα (Distributional Semantic Models - DSMs), είναι ιδιαίτερα διαδεδομένα μοντέλα για τον υπολογισμό των σημασιολογικών αναπαραστάσεων των λέξεων. Σύμφωνα με την υπόθεση της κατανεμημένης έννοιας των λέξεων (distributional hypothesis) [10, 11], λέξεις που εμφανίζονται σε παρόμοια συμφραζόμενα τείνουν να έχουν παρόμοιες έννοιες. Στηριζόμενα στην παραπάνω υπόθεση, τα Κατανεμημένα Σημασιολογικά Μοντέλα επικεντρώνονται στη χρήση στατιστικών μεθόδων για την μοντελοποίηση των προτύπων συνεμφάνισης των λέξεων σε πηγές κειμένου. Οι αναπαραστάσεις λέξεων που

υπολογίζονται με τα μοντέλα αυτά αποδεικνύονται χρήσιμες για την αξιολόγηση της σημασιολογικής ομοιότητας λέξεων και ως αποτέλεσμα τα μοντέλα αυτά χρησιμοποιούνται κατά κόρον σε ποικίλα προβλήματα που εντάσσονται στον τομέα της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing) και της Εξαγωγής και Ανάκτησης Πληροφορίας (Information Extraction and Retrieval). Ορισμένες μόνο από τις εφαρμογές των μοντέλων αυτών είναι η αποσαφήνιση της σημασίας των λέξεων (word sense disambiguation) [12], η απόδοση σημασιολογικών ρόλων (semantic role labeling) [13], η ανίχνευση μεταφορικού λόγου (metaphor detection) [14], η συνθετικότητα φράσεων (phrasal compositionality) [15, 16] κλπ. Παρόλο που η υπόθεση κατανεμημένης έννοιας των λέξεων πηγάζει από τον τομέα της Γλωσσολογίας, τα Κατανεμημένα Σημασιολογικά Μοντέλα έχουν κινήσει το ενδιαφέρον και από την πλευρά της Γνωσιακής Επιστήμης (Cognitive Science), όπου έχει γίνει διερεύνηση των ομοιοτήτων μεταξύ της λειτουργίας των μοντέλων αυτών και της γνωστικής λειτουργίας του ανθρώπου [17, 18, 19, 20, 21].

1.3 Το Πρόβλημα Βασισμού των Συμβόλων

Παρόλο που τα Κατανεμημένα Σημασιολογικά Μοντέλα έχουν εφαρμοστεί με επιτυχία σε πληθώρα εφαρμογών, έχουν δεχτεί κριτική από ορισμένες μελέτες [22, 23, 24]. Υποστηρίζεται ότι η υπολογιστική αναπαράσταση της σημασιολογίας των λέξεων λαμβάνοντας υπόψη μόνο χαρακτηριστικά κειμένου δεν συνάδει με την ανθρώπινη αντίληψη, καθώς όπως αναφέρθηκε προηγουμένως, ο άνθρωπος συνδέει τη σημασιολογία των λέξεων με τις αισθήσεις του αλλά και με την αισθησεο-κινητική του εμπειρία. Η αδυναμία των σημασιολογικών μοντέλων στην μοντελοποίηση αυτής της σύνδεσης αναφέρεται ως 'Πρόβλημα Βασισμού των Συμβόλων' (Symbol Grounding Problem) [25]. Το πρόβλημα αυτό δεν αναφέρεται αποκλειστικά στα Κατανεμημένα Σημασιολογικά Μοντέλα αλλά είναι γενικότερου φιλοσοφικού ενδιαφέροντος και έχει απασχολήσει πολλούς ερευνητές στον τομέα της Τεχνητής Νοημοσύνης.

1.4 Στόχος και Συνεισφορά της Εργασίας

Στόχος αυτής της εργασίας είναι η αντιμετώπιση του προβλήματος 'Βασισμού των Συμβόλων' μέσω της δημιουργίας Πολυτροπικών Κατανεμημένων Σημασιολογικών Μοντέλων (Multimodal DSMs). Τα μοντέλα αυτά χαρακτηρίζονται ως 'πολυτροπικά' (multimodal), διότι εκτός της πληροφορίας που παρέχεται μέσω των κειμένων κωδικοποιούν και την πληροφορία από τις αισθήσεις (sense modalities) του ανθρώπου.

Πιο συγκεκριμένα, επικεντρωνόμαστε στα πολυτροπικά μοντέλα ήχου και εικόνας για την αναπαράσταση των λέξεων με βάση τις ακουστικές και οπτικές τους ιδιότητες αντίστοιχα. Δίνεται έμφαση στα πολυτροπικά μοντέλα ήχου, τα οποία έχουν μελετηθεί σε μικρότερο βαθμό στη βιβλιογραφία. Οι ήχοι που μπορεί να σχετίζονται με μία λέξη δεν είναι απαραίτητα σήματα μουσικής ή φωνής αλλά είναι πιθανό να αντλούνται από οποιοδήποτε ακουστικό γεγονός της καθημερινότητας. Εφόσον λοιπόν οι ήχοι χαρακτηρίζονται από μεγάλη ποικιλομορφία και δοθέντος ότι συγκεκριμένα ακουστικά χαρακτηριστικά μοντελοποιούν αποδοτικά συγκε-

κριμένους τύπους ήχων, προτείνεται η επέκταση των πολυτροπικών μοντέλων ήχου, ώστε κατά την αναπαράσταση ενός ηχητικού αποσπάσματος να λαμβάνονται υπόψη συγκεκριμένα ακουστικά χαρακτηριστικά ανάλογα με τη φύση του αποσπάσματος. Επίσης, προτείνεται μία σειρά από μεθόδους για την Πολυτροπική Σύμπτυξη με βάρη, ώστε να γίνει συνδυασμός των μοντέλων κειμένου, ήχου και εικόνας για τη δημιουργία μίας κοινής αναπαράστασης λέξεων. Σε αυτό το σημείο να αναφερθεί ότι, σύμφωνα με όσα γνωρίζουμε, δεν υπάρχουν άλλες εργασίες σχετικές με την πολυτροπική αναπαράσταση των λέξεων χρησιμοποιώντας ταυτόχρονα κειμενικά, ακουστικά και οπτικά χαρακτηριστικά. Στα πλαίσια αυτής της διπλωματικής ασχολούμαστε με τρία γνωστά προβλήματα:

1. **Αξιολόγηση της σημασιολογικής ομοιότητας λέξεων**: Το πρόβλημα αυτό είναι μείζονος σημασίας για την αξιολόγηση των σημασιολογικών μοντέλων, καθώς εγγενής τους λειτουργία είναι η δημιουργία αξιόπιστων αναπαραστάσεων λέξεων. Βέβαια, χρησιμοποιώντας πολυτροπικά μοντέλα ήχου και εικόνας, η αξιολόγηση της ομοιότητας λέξεων επεκτείνεται και σε ιδιότητες πέραν των σημασιολογικών, δηλαδή τις ακουστικές και οπτικές ιδιότητες των λέξεων.

2. **Αυτόματος χαρακτηρισμός ηχητικών αποσπασμάτων (auto-tagging)**: Σημαντική ιδιότητα των πολυτροπικών μοντέλων ήχου είναι η αναπαράσταση λέξεων και ήχων στον ίδιο διανυσματικό χώρο. Έτσι, τα μοντέλα αυτά μπορούν να χρησιμοποιηθούν για την αξιολόγηση της ομοιότητας μεταξύ ήχων και λέξεων. Εδώ, γίνεται αξιοποίηση αυτής της ιδιότητας για τον αυτόματο χαρακτηρισμό των ηχητικών αποσπασμάτων με λέξεις (tags) που περιγράφουν το περιεχόμενό τους

3. **Αξιολόγηση της μουσικής ομοιότητας**: Το πρόβλημα αυτό αποτελεί την καρδιά του τομέα Ανάκτησης Μουσικής Πληροφορίας (Music Information Retrieval). Τα πολυτροπικά μοντέλα ήχου παρέχουν αναπαραστάσεις λέξεων στον ακουστικό χώρο. Εκμεταλλευόμενοι αυτή την ιδιότητα προτείνουμε έναν αλγόριθμο για την παράλληλη αξιοποίηση του ηχητικού σήματος και των tags με σκοπό την αξιολόγηση της ομοιότητας μεταξύ μουσικών κομματιών. Μάλιστα, ακόμα και στην περίπτωση απουσίας των tags, ο προτεινόμενος αλγόριθμος είναι δυνατό να χρησιμοποιηθεί σε συνδυασμό με τον αλγόριθμο auto-tagging, δίνοντας εξίσου καλές εκτιμήσεις.

Τέλος, καθώς ο χρόνος για πειραματισμό σε επίπεδο διπλωματικής ήταν περιορισμένος, γίνεται πρόταση ορισμένων ιδεών για την επέκταση των πολυτροπικών μοντέλων στα πλαίσια μελλοντικής έρευνας.

1.5 Διάρθρωση της Εργασίας

Στο Κεφάλαιο 2 περιγράφονται τα παραδοσιακά σημασιολογικά μοντέλα, δηλαδή τα υπολογιστικά μοντέλα για εξαγωγή σημασιολογίας από πηγές κειμένου. Παρουσιάζονται σταδιακά τα βήματα για τη δημιουργία των Κατανεμημένων Σημασιολογικών Μοντέλων μέσω της εφαρμογής στατιστικών μεθόδων σε συλλογές κειμένων. Επίσης, περιγράφεται η μέθοδος για την

αξιολόγηση της σημασιολογικής ομοιότητας λέξεων με βάση τις αναπαραστάσεις σύμφωνα με τα παραπάνω μοντέλα. Ακόμη, γίνεται σύγκριση μεταξύ των Κατανεμημένων Σημασιολογικών Μοντέλων και άλλων σημασιολογικών μοντέλων που μελετώνται στη βιβλιογραφία.

Το Κεφάλαιο 3 αποτελεί το βασικότερο κεφάλαιο αυτής της διπλωματικής, καθώς περιγράφει τη διαδικασία δημιουργίας Πολυτροπικών Σημασιολογικών Μοντέλων (Multimodal Semantic Models). Έχοντας ως κίνητρο την επίλυση του προβλήματος βασισμού των συμβόλων (symbol grounding problem), γίνεται περιγραφή των πολυτροπικών μοντέλων για την αναπαράσταση λέξεων με χρήση των ακουστικών και οπτικών τους ιδιοτήτων. Στην Ενότητα 3.1.2 γίνεται μία συνοπτική παρουσίαση σε μεθόδους για την εξαγωγή ακουστικών χαρακτηριστικών από το σήμα του ήχου και παρουσιάζονται τα πιο διαδεδομένα ακουστικά χαρακτηριστικά. Παρόλο που δίνεται έμφαση στα ακουστικά-σημασιολογικά (ADSM) μοντέλα, γίνεται σύντομη περιγραφή και των οπτικών-σημασιολογικών (VDSM) μοντέλων. Επίσης, γίνεται αναφορά στις μεθόδους για την Πολυτροπική Σύμπτυξη, δηλαδή τη σύμπτυξη των μοντέλων που βασίζονται σε χαρακτηριστικά κειμένου, ήχου και εικόνας με τελικό σκοπό τη δημιουργία μίας κοινής αναπαράστασης λέξεων, εδραιωμένης (grounded) στις αισθήσεις του ανθρώπου.

Στο Κεφάλαιο 4 παρουσιάζονται τρεις διαφορετικές εφαρμογές των Πολυτροπικών Σημασιολογικών Μοντέλων. Στην Ενότητα 4.1, το ακουστικό-σημασιολογικό μοντέλο χρησιμοποιείται για την αξιολόγηση της σημασιολογικής ομοιότητας λέξεων με βάση τις ακουστικές τους ιδιότητες. Γίνεται διερεύνηση της επίδρασης διαφορετικών παραμέτρων στην απόδοση του μοντέλου και εφαρμογή των μεθόδων για τη Πολυτροπική Σύμπτυξη με σκοπό την συμπερίληψη κειμενικών, ακουστικών και οπτικών ιδιοτήτων των λέξεων για την αξιολόγηση της σημασιολογικής ομοιότητας. Στην Ενότητα 4.2, το ακουστικό-σημασιολογικό μοντέλο χρησιμοποιείται για την αυτόματο χαρακτηρισμό μουσικών αποσπασμάτων με περιγραφικές λέξεις που είναι σχετικές με το περιεχόμενό τους (auto-tagging). Τέλος, στην Ενότητα 4.3, το ακουστικό-σημασιολογικό μοντέλο χρησιμοποιείται για την αξιολόγηση της μουσικής ομοιότητας.

Κεφάλαιο 2

Κατανεμημένα Σημασιολογικά Μοντέλα

Σε ταινίες επιστημονικής φαντασίας παρουσιάζονται υπολογιστές που κατανοούν την ανθρώπινη γλώσσα και συνομιλούν με ανθρώπους. Η αλληλεπίδραση μεταξύ ανθρώπου και υπολογιστή, στην πραγματικότητα είναι ιδιαίτερα πολύπλοκη, καθώς ο υπολογιστής έχει ως ρόλο την εξαγωγή νοήματος από μία αλληλουχία χαρακτήρων κειμένου. Επαιτείται, λοιπόν κατάλληλη επεξεργασία και ανάλυση της αλληλουχίας χαρακτήρων ώστε να πάρει τη μορφή της ανθρώπινης (φυσικής) γλώσσας. Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing) είναι ένας επιστημονικός κλάδος, στενά συνδεδεμένος με την επιστήμη της πληροφορικής, της τεχνητής νοημοσύνης και της υπολογιστικής γλωσσολογίας και ασχολείται με την κατανόηση και την παραγωγή φυσικής γλώσσας. Συμπεριλαμβάνει την επεξεργασία της γλώσσας σε μορφολογικό, συντακτικό, σημασιολογικό και νοηματικό επίπεδο, ενώ ορισμένες από τις εφαρμογές της είναι η επισήμανση μερών του λόγου, συντακτική ανάλυση, αυτόματη μετάφραση και σύνθεση ομιλίας. Η παρούσα εργασία επικεντρώνεται στο σημασιολογικό επίπεδο της γλώσσας. Ο όρος 'σημασιολογία' αφορά την κατανόηση του νοήματος που προκύπτει από μία λέξη, φράση, πρόταση ή ακόμη και από ολόκληρο κείμενο. Με τον όρο 'σημασιολογικό μοντέλο' υποδηλώνεται ο αλγόριθμος εξαγωγής σημασιολογίας. Εδώ θα δοθεί έμφαση στη δημιουργία σημασιολογικών μοντέλων στο επίπεδο λέξεων, το οποίο είναι συνήθως προαπαιτούμενο στάδιο για την εξαγωγή σημασιολογίας από πιο πολύπλοκες δομές λόγου όπως οι φράσεις και οι προτάσεις.

Υπάρχουν πολλά διαφορετικά μοντέλα για την αναπαράσταση λέξεων. Πολλά από αυτά ασχολούνται με την αναπαράσταση λέξεων ως διανύσματα (vectors) σε έναν διανυσματικό χώρο (Vector Space), ο οποίος συχνά αναφέρεται ως σημασιολογικός χώρος (Semantic Space). Η επίδοση των μοντέλων αυτών είναι εντυπωσιακή σε ορισμένες εφαρμογές. Χαρακτηριστικό παράδειγμα είναι η επίτευξη απόδοσης 92.5% σε ερωτήσεις πολλαπλής επιλογής που αντλήθηκαν από το γνωστό Test of English as a Foreign Language (TOEFL) [26]. Σημειώνεται ότι η μέση απόδοση ανθρώπων στο συγκεκριμένο τεστ είχε υπολογιστεί ως 64.5%, δηλαδή σημαντικά χαμηλότερη από την προαναφερθείσα. Άλλες ενδιαφέρουσες εφαρμογές των σημασιολογικών μοντέλων θα αναφερθούν για κάθε μοντέλο ξεχωριστά στη συνέχεια.

Τα Κατανεμημένα Σημασιολογικά Μοντέλα (Distributional Semantic Models - DSMs), βασίζονται στην υπόθεση της κατανεμημένης έννοιας των λέξεων (distributional hypothesis), σύμφωνα με την οποία λέξεις που εμφανίζονται σε παρόμοια συμφραζόμενα τείνουν να έχουν παρόμοιες έννοιες. Έτσι τα μοντέλα αυτά επικεντρώνονται στη χρήση στατιστικών μεθόδων για τη μοντελοποίηση των προτύπων συνεμφάνισης των λέξεων σε πηγές κειμένου. Τα Κατανεμημένα Σημασιολογικά Μοντέλα εντάσσονται στην κατηγορία των μοντέλων Vector Space καθώς οι λέξεις αναπαρίστανται ως διανύσματα στον σημασιολογικό χώρο. Επίσης, λειτουργούν με βάση την υπόθεση bag-of-words, σύμφωνα με την οποία δε λαμβάνεται υπόψη η σειρά των λέξεων ενός κειμένου. Παρόλο που αυτή η υπόθεση μπορεί να ακούγεται παράλογη αρχικά, στην πράξη έχει φανεί ότι ανταποκρίνεται με επιτυχία σε πολλά προβλήματα που υπάγονται στους τομείς της Επεξεργασίας Φυσικής Γλώσσας αλλά και της Εξαγωγής και Ανάκτησης Πληροφορίας. Εξάλλου, υπάρχουν ενδείξεις ότι το ανθρώπινο μυαλό δεν επεξεργάζεται σειριακά τους χαρακτήρες που διαβάζει [18]. Στη συνέχεια θα γίνει περιγραφή των βασικών σταδίων για τη δημιουργία των DSMs, όπως αυτά περιγράφονται στη βιβλιογραφία [27, 28, 29, 30].

2.1 Υπολογισμός του Πίνακα Συνεμφάνισης Λέξεων

Ο Πίνακας Λέξεων-Συμφραζομένων (Word-Context Matrix)

Έστω ένα σύνολο λέξεων (λεξικό) μεγέθους K. Μία λέξη w αναπαριστάταται από ένα διάνυσμα διάστασης K, όπου η τιμή του i-οστού στοιχείου αντιπροσωπεύει το βαθμό συνεμφάνισης της λέξης w με την i-οστή από τις K λέξεις του λεξικού. Οι λέξεις για τις οποίες επιθυμούμε να εξάγουμε αναπαραστάσεις ονομάζονται λέξεις-στόχοι (target words). Αν θεωρήσουμε N λέξεις στόχους και για κάθε λέξη-στόχο γίνει υπολογισμός ενός διανύσματος διάστασης K, καταλήγουμε στη δημιουργία ενός πίνακα διάστασης $N \times K$, ο οποίος αναφέρεται ως πίνακας λέξεων-συμφραζομένων (word-context matrix). Για να γίνει κατανοητή η παραπάνω διαδικασία, ας θεωρήσουμε το ακόλουθο παράδειγμα κειμένου (πηγή: [2]):

An **automobile** is a wheeled **motor** vehicle used for **transporting passengers**. A **car** is a form of **transport**, usually with four **wheels** and the capacity to carry around five **passengers**. **Transport** for the **London** games is limited, with spectators strongly advised to avoid the use of **cars**. The **London** 2012 **soccer tournament** began yesterday, with plenty of **goals** in the opening **matches**. Giggs **scored** the first **goal** of the **football tournament** at Wembley, North **London**. Bellamy was largely a **passenger** in the **football match**, playing no part in either **goal**.

Ο πίνακας λέξεων-συμφραζομένων (Πίνακας 2.1) προκύπτει μέσω του υπολογισμού της συνεμφάνισης των λέξεων-στόχων με τις λέξεις που ορίζονται στο λεξικό. Ως στήλες-χαρακτηριστικά απεικονίζονται κάποιες από τις λέξεις του κειμένου, ενώ τέσσερις λέξεις έχουν επιλεγεί ως λέξεις-στόχοι (γραμμές του πίνακα). Η τιμή κάθε κελιού του πίνακα λέξεων-συμφραζομένων προκύπτει ως το πλήθος συνεμφανίσεων των αντίστοιχων λέξεων. Στο συγκεκριμένο πα-

	wheel	transport	passenger	tournament	London	goal	match
automobile	1	1	1	0	0	0	0
car	1	2	1	0	1	0	0
soccer	0	0	0	1	1	1	1
football	0	0	1	1	1	2	1

Πίνακας 2.1: Παράδειγμα ενός πίνακα λέξεων-συμφραζομένων. Πηγή: [2].

ράδειγμα, η εμβέλεια συμφραζομένων έχει οριστεί σε επίπεδο πρότασης, δηλαδή θεωρείται ότι υπάρχει συνεμφάνιση δύο λέξεων, μόνο αν βρίσκονται στην ίδια πρόταση. Ωστόσο, υπάρχουν πολλοί εναλλακτικοί τρόποι ορισμού της εμβέλειας συμφραζομένων [2]. Συχνά γίνεται χρήση ενός παραθύρου σταθερού μεγέθους (σε λέξεις), το οποίο έχει ως κέντρο τη λέξη-στόχο και μόνο οι λέξεις που βρίσκονται εντός του παραθύρου προσμετρώνται στην αναπαράσταση της λέξης-στόχου [31, 32].

Ο Πίνακας Όρων-Κειμένων (Term-Document Matrix)

Ένα βασικό ερώτημα που συναντάται συχνά στον τομέα της Ανάκτησης Πληροφορίας και της Επεξεργασίας Φυσικής Γλώσσας είναι το ακόλουθο: 'Δεδομένου ενός ερωτήματος (απαρτιζόμενο από μία σειρά λέξεων) και ενός συνόλου κειμένων, βαθμολόγησε τα κείμενα από το πιο σχετικό στο λιγότερο σχετικό ως προς το ερώτημα'. Η υπόθεση του πολυσυνόλου-από-λέξεις (bag-of-words hypothesis) αποτελεί τη βάση για την εφαρμογή των Κατανεμημένων Σημασιολογικών Μοντέλων στο παραπάνω πρόβλημα [33]. Πιο συγκεκριμένα, η υπόθεση bag-of-words υποδεικνύει ότι ο υπολογισμός της σχετικότητας των κειμένων ως προς το δοθέν ερώτημα μπορεί να πραγματοποιηθεί με την αναπαράσταση των κειμένων και του ερωτήματος ως πολυσύνολα από λέξεις. Ένα πολυσύνολο (bag ή multiset) έχει τις ίδιες ιδιότητες με ένα σύνολο (set) με εξαίρεση την επίτρεψη διπλοτύπων. Για παράδειγμα, το ακόλουθο πολυσύνολο-από-λέξεις: {'dog', 'cat', 'dog', 'dog', 'pig', 'cat'} μπορεί να αναπαρασταθεί με το διάνυσμα $x =< 3, 1, 2 >$, όπου η πρώτη διάσταση αντιστοιχεί στις εμφανίσεις της λέξης `dog', η δεύτερη στις εμφανίσεις της λέξης `cat' και η τρίτη στις εμφανίσεις της λέξης 'pig'. Η παραπάνω μέθοδος συμβολισμού συχνά αναφέρεται και ως bag-of-words μοντέλο.

Έστω ένα λεξιλόγιο από T όρους (λέξεις) και ένα σύνολο από D κείμενα. Ο πίνακας όρων-κειμένων έχει διάσταση $T \times D$ και το στοιχείο (i, j) προκύπτει ως το πλήθος εμφανίσεων του i-οστού όρου στο j-οστό κείμενο. Επομένως, η j-οστή στήλη του πίνακα είναι η αναπαράσταση του j-οστού κειμένου με βάση την υπόθεση bag-of-words. Στην γενική περίπτωση, ο πίνακας αυτός είναι αραιός (sparse), καθώς σε κάθε κείμενο εμφανίζεται μόνο ένα υποσύνολο των διαθέσιμων όρων.

Τα Κατανεμημένα Σημασιολογικά Μοντέλα έχουν εφαρμοστεί με επιτυχία στην εξαγωγή πληροφορίας από κείμενα καθώς, όπως είναι αναμενόμενο, κείμενα που είναι σχετικά με ένα ερώτημα συνήθως περιλαμβάνουν σε μεγάλο βαθμό τις λέξεις του ερωτήματος. Έστω για παράδειγμα ένα ερώτημα q (query) που διατυπώνεται ως ένα σύνολο από λέξεις, ένα σύνολο κειμένων και επιθυμούμε να επιστρέψουμε ως απάντηση στο ερώτημα ένα υποσύνολο K κειμένων τα οποία θεωρούνται πιο σχετικά με το ερώτημα. Τότε ο υπολογισμός του πίνακα

όρων-κειμένων μπορεί να δώσει λύση στο πρόβλημα, καθώς οι στήλες (κείμενα) που αντιστοιχούν στα στοιχεία του πίνακα με τη μεγαλύτερη τιμή θεωρούνται πιο σχετικά προς το δοθέν ερώτημα. Επομένως, μία ενδεχόμενη λύση στο πρόβλημα είναι ο υπολογισμός του αθροίσματος κατά στήλες και η επιστροφή των K κειμένων που αντιστοιχούν στις στήλες με το μεγαλύτερο άθροισμα. Βέβαια, οι λύσεις που χρησιμοποιούνται στην πράξη είναι πιο πολύπλοκες [34, 35] και περιλαμβάνουν την επεξεργασία των αναπαραστάσεων με κλασσικές μεθόδους της Θεωρίας Πληροφορίας.

2.2 Τεχνικές Στάθμισης των Αναπαραστάσεων

Το πλήθος συνεμφανίσεων λέξεων δεν είναι απαραίτητα αποδοτικός τρόπος για τη δημιουργία αναπαραστάσεων λέξεων. Πιο συγκεκριμένα, έχει διαπιστωθεί [36] ότι διατηρώντας απλώς το πλήθος συνεμφανίσεων λέξεων σε έναν πίνακα δημιουργεί πολλά προβλήματα. Χαρακτηριστικό είναι το πρόβλημα των διαφορετικών συχνοτήτων λέξεων. Για παράδειγμα, το αγγλικό οριστικό άρθρο 'the¨, εμφανίζεται πολύ συχνά σε ένα κείμενο, ωστόσο δεν παρέχει χρήσιμη πληροφορία για τις λέξεις που συνοδεύει. Από την άλλη, η ύπαρξη μιας σπάνιας λέξης μπορεί να αλλάξει σημαντικά το περιεχόμενο των λέξεων με τις οποίες συνεμφανίζεται. Εξ΄ άλλου είναι γνωστό από τον τομέα της Θεωρίας Πληροφορίας ότι τα σπάνια γεγονότα προσδίδουν περισσότερη πληροφορία από τα συχνά και αναμενόμενα γεγονότα [37]. Επιπλέον, έχει παρατηρηθεί ότι οι λέξεις μίας γλώσσας τείνουν να κατανέμονται σύμφωνα με το νόμο του Zipf [38], δηλαδή σχετικά λίγες λέξεις χρησιμοποιούνται πάρα πολύ συχνά ενώ οι υπόλοιπες χρησιμοποιούνται πολύ σπάνια. Το φαινόμενο αυτό οδηγεί σε αραιές αναπαραστάσεις στον πίνακα συνεμφάνισης. Με βάση τα παραπάνω, Είναι επιθυμητό να δίνεται έμφαση στα σπάνια στοιχεία αλλά και να αντιμετωπιστεί το πρόβλημα των αραιών αναπαραστάσεων. Αυτό επιτυγχάνεται με συγκεκριμένους μεθόδων για τη στάθμιση (απόδοση τιμών βάρους - weighting) των πινάκων συχνοτήτων.

Η στάθμιση tf-idf (Term Frequency - Inverse Document Frequency)

Η πιο γνωστή τεχνική στάθμισης ονομάζεται **tf-idf** (Term Frequency - Inverse Document Frequency) [39]. Ο όρος tf-idf προκύπτει ως το γινόμενο δύο στατιστικών όρων: του term frequency και inverse document frequency. Έστω ότι έχει υπολογιστεί ο πίνακας όρων-κειμένων (βλ. Ενότητα 2.1) και $f_{t,d}$ το στοιχείο (t, d) του πίνακα, δηλαδή η απόλυτη συχνότητα συνεμφάνισης του όρου t με το κείμενο d. Τότε ο πιο απλός ορισμός του όρου term frequency είναι ο ακόλουθος:

$$tf(t, d) = f_{t,d} \tag{2.1}$$

Ένας άλλος ορισμός του όρου είναι ο εξής:

$$tf(t, d) = 1 + log f_{t,d} \tag{2.2}$$

ενώ πιο συχνά χρησιμοποιείται ο ακόλουθος:

$$tf(t, d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{max\{f_{t',d} : t' \in d\}} \tag{2.3}$$

ο οποίος προκύπτει ως η κανονικοποιημένη και αυξημένη τιμή του πρώτου. Έστω το παράδειγμα του ερωτήματος (query) που αναφέρθηκε στην προηγούμενη ενότητα. Υπενθυμίζεται ότι το ερώτημα διατυπώνεται ως ένα σύνολο από λέξεις, οι οποίες αποτελούν τους όρους (γραμμές) του πίνακα. Με τη στάθμιση term frequency όλοι οι όροι του ερωτήματος θεωρούνται εξίσου σημαντικοί κατά την αξιολόγηση της ¨σχετικότητας' των κειμένων με το ερώτημα. Ωστόσο, ορισμένοι όροι έχουν στην πραγματικότητα ελάχιστη η μηδενική επίδραση στην αξιολόγηση. Για παράδειγμα, σε μία συλλογή κειμένων σχετική με μία βιομηχανίες αυτοκινήτων, η λέξη ¨αυτοκίνητο' θα εμφανίζεται σε κάθε κείμενο. Για τη μείωση της επίδρασης των όρων που εμφανίζονται στα περισσότερα κείμενα γίνεται χρήση του όρου inverse document frequency. Αν D είναι η συλλογή των κειμένων, τότε ορίζεται:

$$idf(t, D) = log \frac{|D|}{|d \in D : t \in d|} \tag{2.4}$$

Με άλλα λόγια, όσο μεγαλύτερος είναι ο αριθμός των κειμένων στα οποία εμφανίζεται ο όρος t, τόσο μικρότερη είναι η τιμή $idf(t, D)$. Έτσι, ορίζουμε:

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \tag{2.5}$$

Ο σταθμισμένος πίνακας όρων-κειμένων προκύπτει με την εφαρμογή του βάρους σε κάθε στοιχείο του ξεχωριστά. Έτσι, ένα στοιχείο (t, d) του πίνακα λαμβάνει μεγάλο βάρος όταν ο όρος t εμφανίζεται συχνά στο κείμενο d αλλά εμφανίζεται σπάνια στα άλλα κείμενα της συλλογής D.

Η στάθμιση κατά PMI (Pointwise Mutual Information)

Μία εναλλακτική μέθοδος στάθμισης βασίζεται στην κατά σημείο αμοιβαία πληροφορία (pointwise mutual information) [40]. Η κατά σημείο αμοιβαία πληροφορία μεταξύ δύο ενδεχομένων x, y είναι ένα μέτρο του κατά πόσο η πιθανότητα της συνεμφάνισης των δύο ενδεχομένων διαφέρει από την αντίστοιχη πιθανότητα στην περίπτωση που τα ενδεχόμενα ήταν ανεξάρτητα.

$$PMI(x, y) = log \frac{p(x, y)}{p(x)p(y)} \tag{2.6}$$

Σημειώνεται ότι αν τα γεγονότα x και y είναι ανεξάρτητα, τότε $p(x, y) = p(x)p(y)$, οπότε η κατά σημείο κοινή πληροφορία λαμβάνει μηδενική τιμή, διαφορετικά μπορεί να πάρει είτε θετικές είτε αρνητικές τιμές.

Έστω οι λέξεις-στόχοι $x_1, x_2, ..., x_N$, οι λέξεις του λεξικού $y_1, y_2, ..., y_K$ και ο πίνακας λέξεων-συμφραζομένων (βλ. Ενότητα 2.1), διάστασης $N \times K$ και $f(x_n, y_k)$ η συχνότητα συνεμφάνισης της λέξης x_n με τη λέξη y_k, όπου $n \in \{1, ..., N\}$ και $k \in \{1, ..., K\}$. Η πιθανότητα συνεμφάνισης της λέξης x_n με τη λέξη y_k ορίζεται ως:

$$p(x_n, y_k) = \frac{f(x_n, y_j)}{\sum_{n=1}^{N} \sum_{k=1}^{K} f(x_n, y_k)} \tag{2.7}$$

Η πιθανότητα εμφάνισης της λέξης x_n ορίζεται ως:

$$p(x_n) = \frac{\sum_{k=1}^{K} f(x_n, y_j)}{\sum_{n=1}^{N} \sum_{k=1}^{K} f(x_n, y_k)} \tag{2.8}$$

ενώ πιθανότητα εμφάνισης της λέξης y_j ορίζεται ως:

$$p(y_j) = \frac{\sum_{n=1}^{N} f(x_n, y_k)}{\sum_{n=1}^{N}\sum_{k=1}^{K} f(x_n, y_k)} \tag{2.9}$$

Έτσι, κατά τη στάθμιση του πίνακα, κάθε στοιχείο $f(i, j)$ του πίνακα λέξεων-συμφραζομένων αντικαθίσταται με την κατά σημείο αμοιβαία πληροφορία των λέξεων x_n και y_k:

$$f'(x_n, y_k) = PMI(x_n, y_k) = log\frac{p(x_n, y_k)}{p(x_n)p(y_k)} \tag{2.10}$$

Αν υπάρχει σημασιολογική συσχέτιση μεταξύ των λέξεων x_n και y_k, αναμένεται ότι η από κοινού πιθανότητα θα είναι μεγαλύτερη από το γινόμενο των επί μέρους πιθανοτήτων, επομένως η τιμή $f'(x_n, y_k)$ θα είναι θετική, γεγονός το οποίο είναι σύμφωνο με την υπόθεση κατανεμημένης έννοιας λέξεων (distributional hypothesis) (βλ. Εισαγωγή της Ενότητας 2). Μία εναλλακτική μορφή στάθμισης είναι η θετική κατά σημείο αμοιβαία πληροφορία (Positive PMI ή PPMI), σύμφωνα με την οποία κατά τη στάθμιση του πίνακα, οι αρνητικές τιμές $pmi(x_n, y_k)$ αντικαθίστανται με τη μηδενική τιμή.

$$PPMI(x, y) = max(0, PMI(x, y)) \tag{2.11}$$

Η μέθοδος PPMI έχει εφαρμοστεί με επιτυχία στην αξιολόγηση της σημασιολογικής ομοιότητας μεταξύ λέξεων [41, 42, 43], γι΄ αυτό και θα χρησιμοποιηθεί στο πειραματικό στάδιο. Υπάρχουν πολλές ακόμη μέθοδοι για στάθμιση, οι οποίες δε θα παρουσιαστούν στα πλαίσια αυτής της διπλωματικής αλλά αξίζει να μελετηθούν στη βιβλιογραφία [44, 45, 46].

2.3 Τεχνικές Μείωσης Διαστασιμότητας

Συχνά, οι σημασιολογικές αναπαραστάσεις λέξεων (π.χ. γραμμές του πίνακα λέξεων-συμφραζομένων) χαρακτηρίζονται ως αραιές και με μεγάλη διαστασιμότητα. Με τον όρο 'αραιή' εννοούμε την αναπαράσταση που αποτελείται κυρίως από μηδενικά στοιχεία. Συνήθως το μέγεθος του λεξικού K, δηλαδή το πλήθος των στηλών του πίνακα λέξεων-συμφραζομένων, είναι ιδιαίτερα μεγάλο, επομένως υπάρχουν πολλά ζευγάρια λέξεων για τα οποία η συχνότητα συνεμφάνισης είναι μηδενική. Σκοπός της μείωσης της διαστασιμότητας των αναπαραστάσεων (dimensionality reduction) είναι η ανίχνευση των 'κρυφών' διαστάσεων του σημασιολογικού χώρου, οι οποίες διατηρούν τα σημαντικότερα χαρακτηριστικά των αναπαραστάσεων και ταυτόχρονα δίνουν στα σημασιολογικά μοντέλα καλύτερη ικανότητα γενίκευσης σε σύγκριση με τις αρχικές διαστάσεις. Επιπλέον, με τη μείωση του πλήθους των διαστάσεων επιτυγχάνεται αύξηση της επίδοσης των υπολογιστικών μεθόδων που εφαρμόζονται στις παραγόμενες αναπαραστάσεις (π.χ. ταξινόμηση/ομαδοποίηση/σύγκριση των αναπαραστάσεων). Τέλος, η μείωση στη μία, δύο ή τρεις διαστάσεις δίνει τη δυνατότητα οπτικοποίησης των αναπαραστάσεων, γεγονός το οποίο βοηθά στην κατανόηση σημαντικών ιδιοτήτων του σημασιολογικού μοντέλου.

Η έννοια των 'κρυφών' διαστάσεων συναντάται και στην περίπτωση της ανθρώπινης αντίληψης. Πιο συγκεκριμένα, θεωρείται ότι η νοητική αναπαράσταση του κόσμου βασίζεται

σε έναν μικρό αριθμό κρυφών χαρακτηριστικών. Αυτά δημιουργούνται μέσω των μεγαλύτερης διάστασης χαρακτηριστικών που προέρχονται από τα ερεθίσματα, όπως για παράδειγμα την όραση και την ακοή του ανθρώπου.

Πριν αναφερθούμε στη γνωστότερη μέθοδο μείωσης της διαστασιμότητας, αξίζει να γίνει αναφορά σε κάποιους βασικούς ορισμούς της Γραμμικής Άλγεβρας. Έστω V ένας διανυσματικός χώρος διάστασης k. Ένα πεπερασμένο σύνολο διανυσμάτων $\{v_1, ..., v_n\}$ του χώρου V, λέγεται **γραμμικώς ανεξάρτητο**, αν όλοι οι συντελεστές οποιουδήποτε γραμμικού συνδυασμού των διανυσμάτων, ο οποίος ισούται με 0, οφείλουν να είναι ίσοι με το 0. Δηλαδή, όταν από οποιαδήποτε σχέση της μορφής

$$\sum_{i=1}^{n} \lambda_i v_i = 0, \quad \lambda_1, ..., \lambda_n \in k \tag{2.12}$$

προκύπτει ότι $\lambda_1 = ... = \lambda_n = 0$. Ένα υποσύνολο $B \subseteq V$ καλείται **βάση** του V όταν είναι γραμμικώς ανεξάρτητο και παράγει ολόκληρον τον V. Κάθε διανυσματικός χώρος διαθέτει τουλάχιστον μία βάση, ενώ όλες οι βάσεις του V είναι ισοπληθείς. Ως **διάσταση** του χώρου V ορίζεται ο πληθικός αριθμός της βάσης B.

Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA)

Η Ανάλυση Κύριων Συνιστωσών (PCA) [47] είναι πιθανότατα η πιο γνωστή μέθοδος μείωσης της διαστασιμότητας. Έστω X ο αρχικός πίνακας διάστασης $k \times n$. Στόχος της PCA είναι η εύρεση ενός ορθογώνιου γραμμικού μετασχηματισμού P διάστασης $k' \times k$, όπου $k' \leq k$, ώστε ο πίνακας $Y = P \cdot X$ που θα προκύψει να διατηρεί το μεγαλύτερο μέρος της πληροφορίας του X. Το πρόβλημα μεγιστοποίησης της πληροφορίας μπορεί να θεωρηθεί ισοδύναμο με το πρόβλημα μεγιστοποίησης της διακύμανσης των δεδομένων. Επομένως, ο ρόλος της PCA είναι η απεικόνιση (projection) των δεδομένων από τον αρχικό χώρο διάστασης k, σε έναν χώρο μικρότερης διάστασης, k', με τρόπο τέτοιο ώστε να μεγιστοποιηθεί η συνολική διακύμανση των δεδομένων στον νέο χώρο. Ο πίνακας συνδιακύμανσης του Y υπολογίζεται ως:

$$\begin{aligned} C_Y &= \frac{1}{n} Y Y^T \\ &= \frac{1}{n} P X X^T P \\ &= P C_X P^T \end{aligned} \tag{2.13}$$

όπου $C_X = \frac{1}{n} X X^T$ ο πίνακας συνδιακύμανσης του X. Εδώ πρέπει να αναφερθεί ότι η μέθοδος προαπαιτεί τη στάθμιση του C_X ώστε οι γραμμές του να έχουν μηδενική μέση τιμή. Με επίλυση των εξισώσεων Lagrange, αποδεικνύεται ότι η βάση του μετασχηματισμού P είναι το σύνολο των ιδιοδιανυσμάτων του πίνακα συνδιακύμανσης C_X. Άρα, το πρόβλημα της PCA περιορίζεται στο πρόβλημα υπολογισμού των ιδιοδιανυσμάτων του C_X. Τα ιδιοδιανύσματα του C_X ονομάζονται 'κύριες διευθύνσεις', ενώ οι προβολές των αρχικών δεδομένων στις κύριες διευθύνσεις ονομάζονται 'κύριες συνιστώσες'.

Ένας αποδοτικός τρόπος υπολογισμού των ιδιοδιανυσμάτων του πίνακα συνδιακύμανσης C_X είναι η εφαρμογή της μεθόδου SVD (Singular Value Decomposition) στον πίνακα X.

Σύμφωνα με τη μέθοδο αυτή, ο πίνακας X γράφεται ως:

$$X = U\Sigma V^T \tag{2.14}$$

όπου οι πίνακες U, V είναι ορθοκανονικοί και περιλαμβάνουν τα ιδιοδιανύσματα του χώρου των στηλών και των γραμμών του X αντίστοιχα, ενώ ο Σ είναι διαγώνιος και η διαγώνιος του αποτελείται από τις $\sigma_1, ..., \sigma_k$, που ονομάζονται μοναδιαίες τιμές (singular values). Έτσι, ο πίνακας συνδιακύμανσης υπολογίζεται:

$$\begin{aligned} C_X &= (U\Sigma V^T)(U\Sigma V^T)^T \\ &= U\Sigma V^T V\Sigma^T U^T \\ &= U\Sigma\Sigma^T U^T \\ &= U\Sigma^2 U^T \end{aligned} \tag{2.15}$$

καθώς ο V είναι ορθοκανονικός, άρα $VV^T = I$, και ο Σ είναι διαγώνιος, άρα $\Sigma\Sigma^T = \Sigma^2$. Επομένως, οι ιδιοτιμές του C_X προκύπτουν ως $\lambda_i = \sigma_i^2$. Η συνολική διακύμανση υπολογίζεται ως το άθροισμα των διαγωνίων στοιχείων του C_Y, το οποίο άθροισμα, ισούται με το άθροισμα των ιδιοτιμών λ_i με βάση τη σχέση 2.15. Επομένως, για μείωση της διαστασιμότητας στις k' διαστάσεις, τα ιδιοδιανύσματα επιλέγονται σε φθίνουσα σειρά (με κριτήριο την ιδιοτιμή) και επιλέγονται τα k' πρώτα ιδιοδιανύσματα. Τα ιδιοδιανύσματα με τη σειρά που ορίστηκε αποτελούν τις γραμμές του πίνακα μετασχηματισμού P.

Αποδεικνύεται ότι αν ως κριτήριο κόστους οριστεί το Ελάχιστο Τετραγωνικό Σφάλμα (Mean Squared Error), η PCA αποτελεί βέλτιστη μέθοδο για μείωση της διαστασιμότητας. Επίσης, η PCA κατατάσσεται στις μεθόδους μάθησης χωρίς επίβλεψη, καθώς δε λαμβάνονται υπόψη πιθανές ετικέτες (labels) των δεδομένων ως προς κάποια κατηγορία. Μία εναλλακτική μέθοδος για μείωση της διαστασιμότητας, η οποία λαμβάνει υπόψη τις ετικέτες των δεδομένων είναι η Linear Discriminant Analysis. Ένα επίσης βασικό χαρακτηριστικό της PCA, είναι η υπόθεση ότι ο μεταχηματισμός βάσης είναι ορθογώνιος και γραμμικός. Συχνά, η μέθοδος αυτή δέχεται κριτικές, γιατί οι συσχετίσεις των δεδομένων τις οποίες επιλύει είναι μέχρι δευτέρου βαθμού. Ωστόσο, υπάρχουν περιπτώσεις πραγματικών δεδομένων με συσχετίσεις μεγαλύτερου βαθμού τις οποίες αδυνατεί να επιλύσει η PCA και ως αποτέλεσμα η μείωση των διαστάσεων δεν αποκαλύπτει την πραγματική δομή των δεδομένων. Για να αντιμετωπιστεί το παραπάνω πρόβλημα, γίνεται χρήση μεθόδων που χρησιμοποιούν μη-γραμμικούς μετασχηματισμούς βάσης. Ένα παράδειγμα είναι η PCA με χρήση μη-γραμμικών πυρήνων (kernel PCA) η οποία εφαρμόζεται με την προϋπόθεση ότι έχουμε εκ των προτέρων κάποια πληροφορία για τη φύση των δεδομένων. Αλλες μέθοδοι που υπάγονται στον γενικότερο τομέα του Manifold Learning είναι η μέθοδος Locality Linear Embedding (LLE), η Isomap, οι Autoencoders κλπ. Ωστόσο, η περιγραφή των τελευταίων μεθόδων ξεφεύγει από τα όρια της διπλωματικής.

Η τεχνική Truncated SVD

Η τεχνική Truncated SVD είναι μία εναλλακτική τεχνική για τη μείωση διαστασιμότητας, ιδιαίτερα διαδεδομένη για τη δημιουργία σημασιολογικών αναπαραστάσεων λέξεων. Συνήθης

είναι η εφαρμογή της σε πίνακες που έχουν δομή αντίστοιχη με αυτή του Πίνακα Όρων-Κειμένων και έχουν σταθμιστεί με την τεχνική tf-idf. Σε αυτά τα πλαίσια εφαρμογής, η τεχνική Truncated SVD αναφέρεται και ως Latent Semantic Analysis - LSA) ή Latent Semantic Indexing - LSI [48]. Σε αντίθεση με την PCA η οποία εφαρμόζεται στον πίνακα συνδιακύμανσης, η LSA εφαρμόζεται απευθείας στον πίνακα Όρων-Κειμένων και δεν απαιτείται κανονικοποίηση των δεδομένων ως προς τη μέση τιμή. Ως αποτέλεσμα, η LSA λειτουργεί αποδοτικά στην περίπτωση αραιών (sparse) πινάκων. Κύρια εφαρμογή της τεχνικής LSA είναι η θεματική μοντελοποίηση (Topic Modeling), δηλαδή η αντιστοίχιση κειμένων σε συγκεκριμένους θεματικούς τομείς (topics).

2.4 Σημασιολογική Ομοιότητα Λέξεων

Η υπόθεση κατανεμημένης έννοιας των λέξεων (distributional hypothesis) υποδεικνύει ότι όσο πιο κοινή είναι η σημασιολογία δύο λέξεων, τόσο πιο συχνά θα συνεμφανίζονται σε κείμενα ως κοινά συμφραζόμενα και ως αποτέλεσμα τόσο πιο 'κοντινές' θα είναι οι κατανεμημένες αναπαραστάσεις τους. Η υπόθεση αυτή, λοιπόν, παρακινεί τη χρήση των κατανεμημένων αναπαραστάσεων των λέξεων για την αξιολόγηση της σημασιολογικής τους ομοιότητας [49]. Έστω $x =< x_1, x_2, ..., x_k >$ και $y =< y_1, y_2, ..., y_k >$ οι διανυσματικές αναπαραστάσεις δύο λέξεων στο σημασιολογικό χώρο. Αναμένουμε ότι όσο πιο κοντά βρίσκονται οι αναπαραστάσεις x και y στο σημασιολογικό χώρο, τόσο πιο σημασιολογικά συγγενείς θα είναι οι λέξεις.

Για την αξιολόγηση της απόστασης δύο διανυσματικών αναπαραστάσεων γίνεται χρήση γνωστών μετρικών ομοιότητας η απόστασης. Μία ιδιαίτερα χρησιμοποιούμενη μετρική είναι η ομοιότητα συνημιτόνου. Ορίζεται ως το συνημίτονο της γωνίας που σχηματίζεται μεταξύ των διανυσμάτων x και y, δηλαδή:

$$\begin{aligned} cos(x, y) &= \frac{x \cdot y}{||x|| \cdot ||y||} \\ &= \frac{\sum_{i=1}^{k} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{k} x_i^2} \cdot \sqrt{\sum_{i=1}^{k} y_i^2}} \end{aligned} \tag{2.16}$$

Η μετρική αυτή λαμβάνει συνεχείς τιμές στο διάστημα $[-1, 1]$, όπου η τιμή 1 υποδηλώνει ίδια κατεύθυνση (γωνία 0 μοιρών) ενώ η τιμή -1 υποδηλώνει αντίθετη κατεύθυνση (γωνία 180 μοιρών). Χαρακτηριστικό πλεονέκτημα της ομοιότητας συνημιτόνου είναι η κανονικοποίηση ως προς το μήκος των διανυσμάτων ($L2$ νόρμα) με αποτέλεσμα να μην επηρεάζεται η αξιολόγηση από το μήκος. Η ιδιότητα αυτή είναι πολύ σημαντική, ειδικά για τις σημασιολογικές αναπαραστάσεις λέξεων, διότι λέξεις που παρουσιάζονται συχνότερα σε ένα κείμενο οδηγούν σε διανύσματα με μεγαλύτερο μήκος στο διανυσματικό χώρο. Ωστόσο είναι πιθανό να υπάρχουν συνώνυμες λέξεις ή γενικότερα λέξεις με παρόμοια σημασιολογία των οποίων η συχνότητα εμφάνισης είναι σημαντικά διαφορετική. Με χρήση, λοιπόν, της ομοιότητας συνημιτόνου μπορεί να ανιχνευτεί τέτοιου είδους ομοιότητα λέξεων. Αν τα δύο διανύσματα που συγκρίνονται έχουν μόνο θετικές τιμές σε κάθε διάσταση, τότε η ομοιότητα συνημιτόνου λαμβάνει τιμές στο εύρος $[0, 1]$. Έτσι, στις περιπτώσεις στάθμισης όπως tf-idf ή ppmi, όπου

οι πίνακες συχνοτήτων αποτελούνται αποκλειστικά από θετικά στοιχεία, οπότε η ομοιότητα συνημιτόνου λαμβάνει θετικές τιμές. Δεν ισχύει όμως το ίδιο στην περίπτωση της στάθμισης pmi, ούτε σε οποιαδήποτε περίπτωση εφαρμογής της μεθόδου PCA (με χρήση SVD) για τη μείωση της διαστασιμότητας των αναπαραστάσεων.

Αντί των μετρικών ομοιότητας, είναι συνηθέστερη η χρήση μετρικών απόστασης για τη σύγκριση διανυσματικών αναπαραστάσεων. Βέβαια, η μετατροπή από μετρική απόστασης σε μετρική ομοιότητας και το αντίστροφο είναι πολύ απλή και συνήθως γίνεται με αντιστροφή:

$$sim(x, y) = \frac{1}{dist(x, y)} \tag{2.17}$$

ή με αφαίρεση:

$$sim(x, y) = 1 - dist(x, y) \tag{2.18}$$

Η συνηθέστερη μετρική απόστασης είναι η Ευκλείδια απόσταση:

$$euclidean\_dist(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.19}$$

ενώ συχνά συναντώνται και άλλες μετρικές όπως η απόσταση Manhattan (η ποία ορίζεται στην περίπτωση πλέγματος), η Bhattacharaya και η Kullback-Leibler. Στα πλαίσια της διπλωματικής αυτής θα γίνει χρήση της ομοιότητας συνημιτόνου καθώς έχει διαπιστωθεί [41, 50] ότι η συγκεκριμένη μετρική είναι καλύτερη από τις προαναφερθείσες για τη σύγκριση διανυσμάτων λέξεων.

2.5 Σύγκριση με Άλλα Σημασιολογικά Μοντέλα

Στις προηγούμενες ενότητες έγινε περιγραφή της διαδικασίας δημιουργίας Κατανεμημένων Σημασιολογικών Μοντέλων (Distributional Semantic Models) τα οποία βασίζονται στην υπόθεση της κατανεμημένης έννοιας των λέξεων (distributional hypothesis). Επειδή τα Κατανεμημένα Σημασιολογικά Μοντέλα χρησιμοποιούνται ευρέως στους τομείς της Εξαγωγής Πληροφορίας και της Επεξεργασίας Φυσικής Γλώσσας, παρουσιάζονται πολλές τεχνικές με διαφορετικά ονόματα λόγω των διαφορετικών εφαρμογών, οι οποίες όμως έχουν ως κεντρικό άξονα την υπόθεση της κατανεμημένης έννοια των λέξεων. Για παράδειγμα οι τεχνικές Latent Semantic Indexing (LSI), Latent Semantic Analysis (LSA) έχουν την ίδια λειτουργία και προκύπτουν αν αντί της τεχνικής PCA γίνει χρήση της τεχνικής Truncated SVD για τη μείωση διαστασιμότητας. Αξίζει εδώ να αναφερθεί ότι στην περίπτωση του τομέα Κειμενικής Μοντελοποίησης (Topic Modeling), βρέθηκε ότι η τεχνική LSA (ή LSI) χαρακτηρίζεται από συγκεκριμένα προβλήματα λόγω του ασταθούς στατιστικού υποβάθρου της. Πιο συγκεκριμένα, με τη μέθοδο LSA υποτίθεται ότι οι λέξεις και τα κείμενα σχηματίζουν ένα κοινό γκαουσιανό μοντέλο ενώ έχει παρατηρηθεί ότι το μοντέλο ακολουθεί την κατανομή Poisson. Ως αποτέλεσμα, η μέθοδος τροποποιήθηκε στην probabilistic Latent Semantic Analysis (pLSA) ή αντίστοιχα (pLSI) [51]. Και πάλι, η τεχνική pLSA αντιμετωπίζει σημαντικά μειονεκτήματα λόγω του γεγονότος ότι δεν παρέχει κάποιο πιθανοτικό μοντέλο στο επίπεδο μοντελοποίησης κειμένων. Μία μέθοδος που επιλύει το συγκεκριμένο πρόβλημα και χρησιμοποιείται με επιτυχία εδώ και χρόνια στον τομέα του Topic Modeling είναι η Latent Dirichlet Allocation (LDA) [52].

Επίσης, υπάρχουν πολλά σημασιολογικά μονέλα που ξεφεύγουν από την φιλοσοφία των Κατανεμημένων Σημασιολογικών Μοντέλων. Πολύ διαδεδομένα είναι τα Νευρωνικά Γλωσσικά Μοντέλα (Neural Language Models) [53] τα οποία διαφοροποιοούνται από τα κατανεμημένα μοντέλα λόγω του ότι αντί να μετρούν τις συνεμφανίσεις μεταξύ λέξεων, αντιμετωπίζουν το πρόβλημα αναπαράστασης των λέξεων ως ένα πρόβλεψης των αναπαραστάσεων [36]. Το παραπάνω πρόβλημα προσεγγίζεται με τη χρήση τεχνητών νευρωνικών δικτύων. Συνοπτικά, ένα Νευρωνικό Γλωσσικό Μοντέλο αποτελείται από ένα feed-forward νευρωνικό δίκτυο το οποίο χρησιμοποιείται για την πρόβλεψη της επόμενης λέξης σε μία ακολουθία λέξεων πλήθους n. Στόχος του δικτύου είναι η μεγιστοποίηση της ακόλουθης συνάρτησης:

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} \log f(w_t, w_{t-1}, ..., w_{t-n+1}), \tag{2.20}$$

όπου:

$$f(w_t, w_{t-1}, ..., w_{t-n+1}) = p(w_t | w_{t-1}, ..., w_{t-n+1}). \tag{2.21}$$

Η τελευταία τιμή υποδηλώνει την έξοδο του νευρωνικού δικτύου μετά από την προσθήκη ενός επιπέδου softmax το οποίο χρησιμοποιείται για τη διασφάλιση ότι οι έξοδοι υποδηλώνουν τιμές πιθανότητας και αθροίζονται στη μονάδα. Συχνά, οι αναπαραστάσεις λέξεων που μαθαίνονται με τις παραπάνω μεθόδους ονομάζονται embeddings για να διαχωριστούν από τις κατανεμημένες αναπαραστάσεις λέξεων. Βέβαια χαρακτηρίζονται και με τον όρο distributed

representations, ο οποίος διαφέρει από τον όρο distributional representation, η διαφοροποίηση των οποίων δεν είναι σαφής με χρήση ελληνικών ορολογιών. Στη συνέχεια θα γίνει περιγραφή δύο ιδιαίτερα γνωστών distributed σημασιολογικών μοντέλων:

Το μοντέλο Word2Vec

Το μοντέλο Word2Vec [54, 55, 56] είναι ίσως το ποιο γνωστό μοντέλο για τη δημιουργία διανυσμάτων λέξεων (word embeddings). Το μοντέλο αυτό αποτελείται από δύο επιμέρους μοντέλα: το Continuous bag of words (CBOW) [54] και το Skip-gram [56].

Το μοντέλο Continuous bag of words (CBOW) αποσκοπεί στην πρόβλεψη μία λέξης δοθέντων των συμφραζομένων της. Το πρόβλημα αυτό μεταφράζεται στην ακόλουθη συνάρτηση κόστους:

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} \log p(w_t|w_{t-n}, ..., w_{t-1}, w_{t+1}, ..., w_{t+n}). \tag{2.22}$$

Η παραπάνω συνάρτηση κόστους μοιάζει αρκετά με τη συνάρτηση της Εξίσωσης2.20, με τη διαφορά ότι η πρόβλεψη της λέξης w_t δε γίνεται με βάση τις n προηγούμενες λέξεις αλλά με βάση τις n προηγούμενες και τις n επόμενες λέξεις.

Το μοντέλο skip-gram [56] αποσκοπεί στην πρόβλεψη των συμφραζομένων μίας λέξης, ως ένα σύνολο από λέξεις που είναι πιο πιθανό να τη συνοδεύουν. Ως αποτέλεσμα προκύπτει η ακόλουθη συνάρτηση κόστους:

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T}\sum -n \leq j \leq n, \neq 0 \log p(w_{t+j}|w_t), \tag{2.23}$$

όπου αν V είναι το διαθέσιμο λεξικό, τότε:α

$$p(w_{t+j}|w_t) = \frac{exp(v_{w_t}^T v'_{w_{t+j}})}{\sum_{w_i \in V} exp(v_{w_t}^T v'_{w_i})} \tag{2.24}$$

Τόσο το μοντέλο CBOW όσο και το μοντέλο Skip-gram επικεντρώνονται στη μάθηση ενός πίνακα βαρών W, μέσω του ορισμού δύο κριτηρίων βελτιστοποίησης. Το word2vec μοντέλο έχει κινήσει το ενδιαφέρον στο πεδίο μάθησης αναπαραστάσεων, καθώς συλλαμβάνει ενδιαφέρουσες λεξιλογικές ιδιότητες. Για παράδειγμα η διαφορά των διανυσμάτων λέξεων μοντελοποιεί την αντιπαράθεση των λέξεων. Γνωστότερο παράδειγμα είναι το ακόλουθο: $king - man + woman \approx queen$. Διαδεδομένη είναι και η επέκταση του παραπάνω μοντέλου, η οποία αναφέρεται ως GloVe [57] ενώ έχουν υπάρξει και άλλες προσεγγίσεις με τεχνητά νευρωνικά δίκτυα [58, 59].

Σύγκριση μεταξύ distributed και distributional μοντέλων

Τα distributional μοντέλα χαρακτηρίσονται ως "count" μοντέλα, καθώς 'μετράνε' το πλήθος των συνεμφανίσεων μεταξύ των λέξεων. Αντίθετα, τα distributed μοντέλα χαρακτηρίζονται ως "predict" μοντέλα, καθώς στόχος τους είναι η πρόβλεψη των λέξεων δεδομένων

των συμφραζομένων τους. Σε πρόσφατη δημοσίευση [36], διαπιστώθηκε ότι τα distributed μοντέλα ξεπερνούν σε απόδοση τα distributional μοντέλα σε μία σειρά προβλημάτων. Ωστόσο, δεν είναι σαφής ο διαχωρισμός τους από την θεωρητική πλευρά. Πιο συγκεκριμένα, έχει βρεθεί ότι τα distributed μοντέλα αποτελούν στην πράξη προσεγγίσεις της παραγοντοποίησης των σταθμισμένων με τη διαδικασία PMI πινάκων, όπως αυτοί υπολογίζονται με τα distributional μοντέλα [57, 60]. Φαίνεται λοιπόν ότι και οι δύο κατηγορίες μοντέλων βασίζονται στην υπόθεση κατανεμημένης έννοιας των λέξεων (distributional hypothesis) ενώ έχει υποστηριχτεί ότι βασικός λόγος για την καλύτερη απόδοση των distributed μοντέλων είναι η διαδικασία βέλτιστης επιλογής των αντίστοιχων υπερπαραμέτρων [61]. Βέβαια ένα βασικό πλεονέκτημα των distributed μοντέλων είναι ότι μπορούν να υπολογιστούν με σημαντικά ταχύτερο τρόπο σε σχέση με τα distributional μοντέλα.

Σχόλια για την υπόθεση bag-of-words

Ένα βασικό χαρακτηριστικό των σημασιολογικών μοντέλων που λειτουργούν με την υπόθεση bag-of-words είναι το γεγονός ότι δε λαμβάνουν υπόψη την αλληλουχία των λέξεων ενός κειμένου [62]. Για παράδειγμα, οι φράσεις 'Η Μαρία είναι πιο γρήγορη από τη Χριστίνα' και 'Η Χριστίνα είναι πιο γρήγορη από τη Μαρία', αποτελούνται από το ίδιο σύνολο λέξεων αλλά με διαφορετική αλληλουχία. Ενώ είναι ξεκάθαρο ότι η διαφορετική αλληλουχία προσδίδει διαφορετικό νόημα στις φράσεις, οι αναπαραστάσεις τους στο διανυσματικό χώρο με τις μεθόδους που περιγράφηκαν προηγούμενα είναι πανομοιότυπες, γιάυτό και τα συγκεκριμένα σημασιολογικά μοντέλα έχουν κριθεί αρνητικά στο παρελθόν [50, 63]. Ωστόσο, το μειονέκτημα των μοντέλων αυτών έχει αντιμετωπιστεί με επεκτάσεις τους, οι οποίες λαμβάνουν υπόψη την αλληλουχία των λέξεων [15, 64, 65]. Ακόμη, έχει διαπιστωθεί τα μοντέλα που λειτουργούν με την υπόθεση bag-of-words ότι ανταποκρίνεται με επιτυχία σε πληθώρα προβλημάτων ενώ υπάρχουν ενδείξεις η επεξεργασία των χαρακτήρων από το ανθρώπινο μυαλό δεν είναι σειριακή [18], επιβεβαιώνοντας τη φιλοσοφία των μοντέλων αυτών. Επίσης, η απλότητα των bag-of-words σημασιολογικών μοντέλων τα καθιστά ιδιαίτερα χρήσιμα σε περιπτώσεις όπου παίζει ρόλο υπολογιστική ταχύτητα. Πιο πολύπλοκα μοντέλα που λειτουργούν με χρήση συντακτικών αναλυτών και βαθιών νευρωνικών δικτύων είναι πολύ πιο απαιτητικά ως προς το χρόνο. Για το λόγο, σε πολύ πρόσφατη έρευνα [66] προτάθηκε ο συνδυασμός bag-of-words μοντέλων και πιο πολύπλοκων μοντέλων, ώστε τα τελευταία να χρησιμοποιούνται μόνο σε προτάσεις με πολύπλοκη δομή, στις οποίες αναμένεται ότι θα έχουν καλύτερη απόδοση από τα bag-of-words μοντέλα. Πάντως, το κατά πόσο οι στατιστικές ιδιότητες των λέξεων είναι ικανοποιητικές για τη μοντελοποίηση εννοιών είναι ένα ανοιχτό ερώτημα, το οποίο έχει απασχολήσει και συνεχίζει να απασχολεί την ερευνητική κοινότητα.

Κεφάλαιο 3

Πολυτροπικά Κατανεμημένα Σημασιολογικά Μοντέλα

Τα Πολυτροπικά Κατανεμημένα Σημασιολογικά Μοντέλα (Multimodal DSMs) επικεντρώνονται στις πολυτροπικές αναπαραστάσεις λέξεων, δηλαδή αναπαραστάσεις λέξεων με βάση χαρακτηριστικά που προέρχονται από τις ανθρώπινες αισθήσεις, επιδιώκοντας με αυτόν τον τρόπο την επίλυση του προβλήματος 'Βασισμού των συμβόλων' (Symbol Grounding Problem) (βλ Ενότητα 1.3). Πώς όμως θα μπορούσε να αναπαρασταθεί σε ένα σημασιολογικό μοντέλο πληροφορία σχετική με τις αισθήσεις του ανθρώπου; Στο κεφάλαιο αυτό θα γίνει περιγραφή των πολυτροπικών μοντέλων ήχων (Ενότητα 3.1) και εικόνων (Ενότητα 3.2.4), ώστε στην αναπαράσταση των λέξεων να κωδικοποιηθεί η πληροφορία που αντιστοιχεί στις ακουστικές και οπτικές τους ιδιότητες. Στη συνέχεια (Ενότητα 3.3) θα δούμε πώς μπορούν να συνδυαστούν αποδοτικά τα κλασσικά σημασιολογικά μοντέλα με τα δύο παραπάνω μοντέλα για τη δημιουργία πολυτροπικών σημασιολογικών μοντέλων.

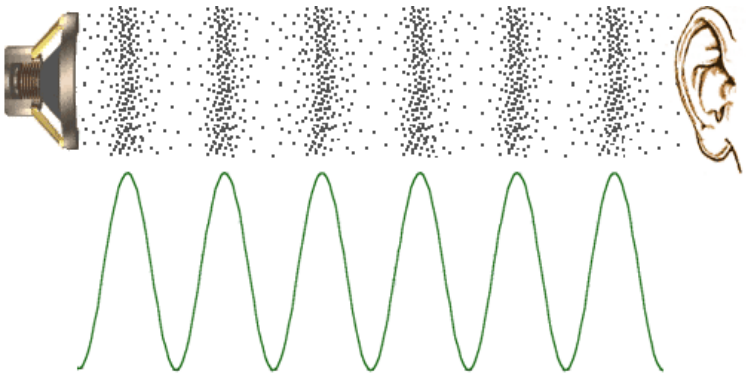
3.1 Πολυτροπικά Μοντέλα Ήχου

Σκοπός αυτής της ενότητας είναι η δημιουργία σημασιολογικών μοντέλων για αναπαραστάσεις λέξεων με βάση τις ακουστικές τους ιδιότητες. Για παράδειγμα, η λέξη 'γάτα' αναπαρίσταται ως προς τις ακουστικές ιδιότητες του ζώου. Μία μέθοδος για να επιτευχθεί το παραπάνω ονομάζεται bag-of-audio-words και είναι εμπνευσμένη από την παραδοσιακή μέθοδο δημιουργίας των μοντέλων bag-of-words (βλ. Ενότητα 2.1). Προτού όμως προχωρήσουμε στην περιγραφή του μοντέλου bag-of-audio-words, θα γίνει γενικότερη περιγραφή του ήχου ως σήμα αλλά και αναφορά στα πιο γνωστά ακουστικά χαρακτηριστικών και στη διαδικασία εξαγωγής τους.

3.1.1 Ο Ήχος ως Σήμα

Παραγωγή και Διάδοση του Ήχου

Ο αέρας είναι ένα σύνολο από μόρια τα οποία βρίσκονται σε μεγαλύτερη απόσταση το ένα από το άλλο σε σχέση με τα στερεά και τα υγρά. Όταν ένα αντικείμενο πάλλεται, προκαλεί διαταραχή της πίεσης του αέρα. Η διαταραχή αυτή οδηγεί στην ταλάντωση των μορίων του και η μετάδοσή της στον αέρα με τη μορφή κύματος πίεσης γίνεται αντιληπτή από το ανθρώπινο αυτί ως ήχος. Μία απλή σχηματική αναπαράσταση του φαινομένου δίνεται στο Σχήμα 3.1.



Σχήμα 3.1: Σχηματική αναπαράσταση της διάδοσης του ήχου από την πηγή μέχρι το ανθρώπινο αυτί ως κύμα πίεσης.

Ο ήχος χαρακτηρίζεται ως διαμήκες σφαιρικό κύμα. Διαμήκες, γιατί η ταλάντωση των μορίων του αέρα γίνεται κατά τη διεύθυνση μετάδοσης της της κυματικής κίνησης (σε αντίθεση με τα εγκάρσια κύματα, των οποίων η διεύθυνση είναι κάθετη στη διεύθυνση μετάδοσης). Σφαιρικό, γιατί διαδίδεται από την πηγή προς όλες τις κατευθύνσεις.

Ιδιότητες του Ήχου

Τέσσερις είναι οι βασικές ιδιότητες του ήχου, οι οποίες σχετίζονται με τον τρόπο που τον αντιλαμβάνεται ο άνθρωπος. Πρώτη ιδιότητα είναι η **ένταση** (loudness) που σχετίζεται με το **πλάτος** (amplitude) του ηχητικού κύματος. Σύνηθες μέτρο έντασης είναι τα decibel (dB). Για παράδειγμα, το όριο πόνου στο ανθρώπινο αυτί είναι 120 dB. Δεύτερη ιδιότητα είναι το **τονικό ύψος** (pitch) το οποίο θα μπορούσε να θεωρηθεί ότι αντιστοιχεί στη συχνότητα του ηχητικού κύματος. Μονάδα μέτρησης του τονικού ύψους είναι τα Hertz (Hz). Κάθε νότα του πενταγράμμου αντιστοιχεί σε μία συγκεκριμένη συχνότητα. Το αυτί ενός ενήλικα είναι ικανό να διακρίνει ήχους με εύρος συχνοτήτων από 20 έως 22000 Hz. Ωστόσο, δεν ισχύει πάντα η αντιστοίχιση του τονικού ύψους με τη συχνότητα του κύματος, όπως για παράδειγμα στο φαινόμενο της έλλειψης θεμελιώδους συχνότητας (missing fundamendal). Τρίτη ιδιότητα είναι η **χροιά** (timbre) ή αλλιώς **ηχόχρωμα** το οποίο σχετίζεται με το φάσμα (spectrum) του ήχου και είναι το χαρακτηριστικό που διαφοροποιεί δύο ήχους με ίδια ένταση και ίδιο τονικό ύψος. Η χροιά είναι ίσως το πιο δυσνόητο και ενδιαφέρον χαρακτηριστικό καθώς είναι καθοριστικό για το διαχωρισμό των μουσικών οργάνων. Τέταρτη ιδιότητα είναι η **διάρκεια**

(duration) και αφορά την χρονική διάρκεια για την οποία ένας ήχος είναι αντιληπτός. Μονάδα μέτρησης της διάρκειας είναι τα δευτερόλεπτα (seconds).

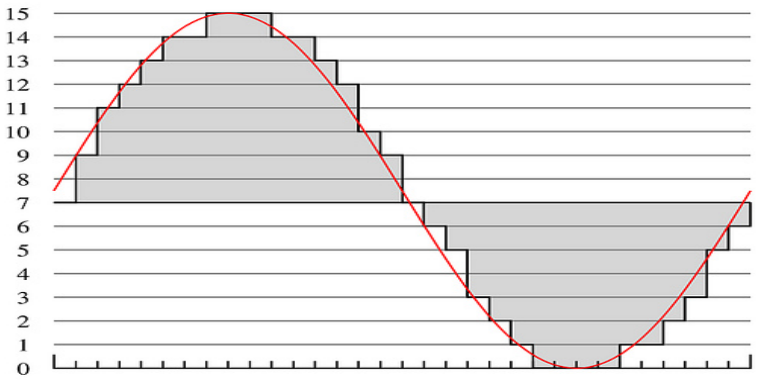
Παρόλο που ο ορισμός των παραπάνω χαρακτηριστικών φαίνεται απλός, στην πραγματικότητα η αντίληψη του ήχου είναι ιδιαίτερα πολύπλοκη [67]. Η Ψυχοακουστική είναι ένας κλάδος της ακουστικής ο οποίος μελετά τον υποκειμενικό τρόπο με τον οποίο ο άνθρωπος αντιλαβάνεται τον ήχο, με τον οποίο παράλληλα εμπλέκονται και άλλες επιστήμες όπως η ψυχολογία και οι νευροεπιστήμες. Για περισσότερες λεπτομέρειες αναφορικά με τις ιδιότητες του ήχου από μαθηματικής πλευράς αξίζει η ανάγνωση του βιβλίου 'Music: A Mathematical Offering' [68].

Ο Ήχος ως Ψηφιακό Σήμα

Παρόλο που ο ήχος χαρακτηρίζεται ως ακουστικό σήμα, είναι ιδιαίτερα βολική η μετατροπή του σε ηλεκτρική μορφή καθώς επιτρέπει πληθώρα μετατροπών στο σήμα, όπως ενίσχυση, ηχογράφηση ή μίξη [69]. Η μετατροπή του ήχου σε ηλεκτρικό σήμα πραγματοποιείται από το μικρόφωνο, όπου αν θεωρήσουμε ότι είναι ιδανικό (κάτι που δε συμβαίνει στην πραγματικότητα), το ηλεκτρικό σήμα που θα προέκυπτε θα είχε ακριβώς τις ίδιες ιδιότητες με το ακουστικό σήμα. Το παραγόμενο ηλεκτρικό σήμα χαρακτηρίζεται ως αναλογικό, δηλαδή είναι ορισμένο σε ένα συνεχές χρονικό διάστημα και μπορεί να πάρει οποιαδήποτε τιμή στο σύνολο των πραγματικών αριθμών. Ωστόσο η αναπαράσταση του ηχητικού σήματος στον ηλεκτρονικό υπολογιστή απαιτεί την ψηφιοποίησή του. Η μετατροπή του αναλογικού σε ψηφιακό σήμα γίνεται σε δύο στάδια: τη δειγματοληψία και τον κβαντισμό [70].

Η μετατροπή του σήματος συνεχούς χρόνου σε σήμα διακριτού χρόνου επιτυγχάνεται μέσω της **δειγματοληψίας** (sampling), δηλαδή της επιλογής ορισμένων τιμών του συνεχούς σήματος ανά καθορισμένα χρονικά διαστήματα. Το σταθερό χρονικό διάστημα που μεσολαβεί ανάμεσα σε δύο δείγματα του σήματος ονομάζεται **περίοδος**, T της δειγματοληψίας. Η συχνότητα (ρυθμός) δειγματοληψίας, f_s, ορίζεται ως ο αριθμός δειγμάτων του σήματος που λαμβάνονται στο χρονικό διάστημα του ενός δευτερολέπτου και ισχύει η σχέση $f_s = \frac{1}{T}$. Συνήθεις ρυθμοί δειγματοληψίας για σήματα φωνής είναι 8kHz και 16kHz ενώ για σήματα μουσικής τα 22.05 kHz και 42.1kHz. Οι τιμές αυτές είναι βασισμένες σε ένα από τα σημαντικότερα θεωρήματα στον τομέα της Θεωρίας Πληροφορίας: το θεώρημα των Nyquist-Shannon σύμφωνα με το οποίο για να πραγματοποιηθεί ψηφιοποίηση χωρίς απώλεια πληροφορίας, ο ρυθμός δειγματοληψίας πρέπει να είναι τουλάχιστον διπλάσιος από την μέγιστη συχνότητα του σήματος.

Το δεύτερο στάδιο για την ψηφιοποίηση του ήχου είναι ο **κβαντισμός** (quantization) του δειγματοληπτημένου σήματος, δηλαδή η απεικόνιση των τιμών του (που είναι πραγματικοί αριθμοί) σε ένα μικρότερο σύνολο αριθμών. Τυπικές μορφές αναπαράστασης των δειγμάτων είναι αυτές των 8-bit, 16-bit και 20-bit. Γενικά, αν χρησιμοποιούνται N bits για τον κβαντισμό του σήματος, η απεικόνιση γίνεται σε 2^N διαφορετικά επίπεδα. Ένα παράδειγμα μετατροπής του αναλογικού σήματος σε ψηφιακό μέσω της δειγματοληψίας και του κβαντισμού παρατίθεται στο Σχήμα 3.4.



Σχήμα 3.2: Μετατροπή αναλογικού σήματος σε ψηφιακό.
Πηγή: https://www.videomaker.com/article/c4/14524-digital-audio-sampling.

Αναπαράσταση του ήχου στο πεδίο της συχνότητας

Με την ολοκλήρωση της ψηφιοποίησης έχουμε την ψηφιακή αναπαράσταση ενός ηλεκτρικού σήματος που αντιπροσωπεύει το ακουστικό σήμα. Η αναπαράσταση του σήματος έχει υπολογιστεί στο πεδίο του χρόνου, δηλαδή διαδοχικά δείγματα του σήματος αντιστοιχούν σε διαδοχικές χρονικές στιγμές με διάστημα μίας περιόδου (T) δειγματοληψίας. Ωστόσο, η αναπαράσταση στο πεδίο του χρόνου δεν είναι βολική για τη μελέτη του συχνοτικού περιεχομένου του ήχου, επομένως είναι ουσιώδης ο ρόλος του μετασχηματισμού του σήματος στο πεδίο της συχνότητας. Η μετάβαση στο πεδίο της συχνότητας επιτυγχάνεται με χρήση του Διακριτού Μετασχηματισμού Fourier (DFT), όπως επεξηγείται στη συνέχεια.

Έστω ένα σήμα διακριτού χρόνου το οποίο αποτελείται από N ισαπέχοντα δείγματα: $x[0], ..., x[N-1]$. Ο DFT [71] του σήματος ορίζεται ως:

$$X[n] = \sum_{k=0}^{N-1} x[k] e^{-j\frac{2\pi}{N}nk}, \forall n \in [0, N-1] \tag{3.1}$$

Η ακολουθία $X[n]$ ονομάζεται φασματογράφημα (spectrum) της ακολουθίας $x[k]$ και παρέχει πληροφορία για την κατανομή των συχνοτήτων του ηχητικού σήματος. Συμβολίζουμε:

$$X[n] = \mathscr{F}\{x[k]\}$$

Ο αντίστροφος μετασχηματισμός Fourier ορίζεται ως:

$$x[k] = \frac{1}{N} \sum_{n=0}^{N-1} X[n] e^{+j\frac{2\pi}{N}nk}, \forall k \in [0, N-1] \tag{3.2}$$

και μετασχηματίζει το σήμα από το πεδίο της συχνότητας πίσω στο πεδίο του χρόνου.

Ο Τμηματικός Μετασχηματισμός (STFT) λειτουργεί διαχωρίζοντας το αρχικό σήματος σε μικρά κομμάτια ίσης χρονικής διάρκειας και εφαρμόζοντας τον προηγούμενο μετασχηματισμό σε κάθε κομμάτι. Έτσι, με χρήση του Τμηματικού Σχηματισμού, δίνεται η δυνατότητα μελέτης των συχνοτικών χαρακτηριστικών καθώς το σήμα εξελίσσεται στο χρόνο.

3.1.2 Διαδεδομένα Ακουστικά Χαρακτηριστικά

Δεδομένης της ποικιλομορφίας των ήχων που ακούμε καθημερινά, κρίνεται μείζονος σημασίας η εξαγωγή ακουστικών χαρακτηριστικών, κατάλληλων να μοντελοποιήσουν κάθε μορφή ήχου. Εφόσον μία από τις εφαρμογές τις παρούσας διπλωματικής είναι η αξιολόγηση ηχητικής ομοιότητας, τα ακουστικά χαρακτηριστικά πρέπει να ομαδοποιούν τους ήχους που ακούγονται ως 'κοντινοί' στο ανθρώπινο αυτί ενώ ταυτόχρονα να τους διαχωρίζουν από πιο 'μακρινούς' ήχους. Η δυσκολία της διαδικασίας αυτής έγκειται στο γεγονός ότι η αντίληψη της απόστασης μεταξύ ήχων είναι υποκειμενική. Ωστόσο, πληθώρα χαρακτηριστικών έχου μελετηθεί και αποδειχτεί χρήσιμα για μεμονωμένες ή και πιο γενικές εφαρμογές. Αρχικά, θα γίνει γενικότερη περιγραφή του ήχου ως σήμα και στη συνέχεια θα γίνει παρουσίαση των πιο γνωστών ακουστικών χαρακτηριστικών και της διαδικασίας εξαγωγής τους. Η εξαγωγή ακουστικών χαρακτηριστικών είναι ένα από τα βασικότερα στάδια των εφαρμογών ανάλυσης ήχου όπως για παράδειγμα η Αυτόματη Αναγνώριση Φωνής (Automatic Speech Recognition - ASR) και η Εξαγωγή Μουσικής Πληροφορίας (Music Information Retrieval - MIR). Υπάρχουν ορισμένα ακουστικά χαρακτηριστικά, τα οποία είναι ιδιαίτερα διαδεδομένα λόγω της αποδοτικής αξιοποίησής τους σε πολλαπλές εφαρμογές. Ορισμένα από αυτά θα παρουσιαστούν στη συνέχεια.

Σε κάθε περίπτωση εξαγωγής χαρακτηριστικών, το ακουστικό σήμα περνάει από διαδικασία παραθύρωσης, δηλαδή χωρισμού σε (συνήθως) επικαλυπτόμενα παράθυρα. Η διάρκεια των παραθύρων εξαρτάται κυρίως από τη φύση του σήματος και συνήθως είναι ικανοποιητικά μικρή, ώστε οι στατιστικές ιδιότητες του σήματος να παραμένουν αμετάβλητες σε κάθε παράθυρο. Συνήθως, για την αποφυγή ασυνεχειών στα άκρα των παραθύρων, εφαρμόζεται παραθύρωση Hamming σε κάθε παράθυρο, δηλαδή συνέλιξη στο πεδίο του χρόνου (αντίστοιχα πολλαπλασιασμός στο πεδίο της συχνότητας) με το παράθυρο Hamming, το οποίο ορίζεται ως:

$$w(n) = 0.54 - 0.46cos(2\pi\frac{n}{N}) \tag{3.3}$$

Σημειώνεται ότι σε ορισμένες εφαρμογές το ακουστικό σήμα υφίσταται πιο πολύπλοκη προεπεξεργασία, όπως για παράδειγμα πέρασμα από εξειδικευμένα φίλτρα για την ενίσχυση συγκεκριμένου εύρους συχνοτήτων.

Mel Frequency Cepstral Coefficients (MFCCs)

Οι MFCCs (Συντελεστές Αναφάσματος στις Mel Συχνότητες) [72] είναι τα πιο διαδεδομένα σύνολα χαρακτηριστικών που χρησιμοποιούνται σε εφαρμογές όπως η αναγνώριση φωνής [73], ωστόσο έχουν αποδειχτεί χρήσιμα και για την αναπαράσταση μουσικών κομματιών ή ήχων γενικότερα [74]. Κεντρική ιδέα στην εξαγωγή των MFCCs είναι η κατασκευή μίας συστοιχίας φίλτρων (filterbank) σύμφωνα με την κλίμακα mel, η οποία αναφέρεται στη συνέχεια.

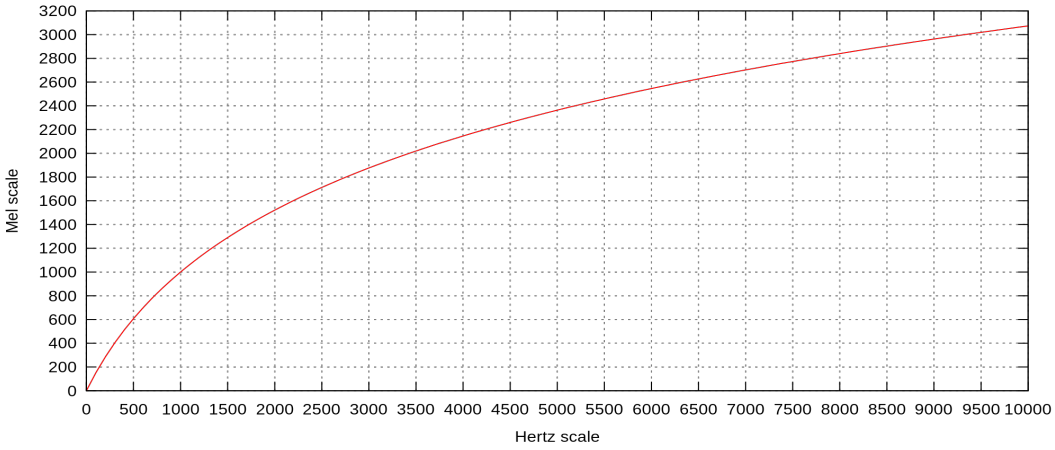
Η κλίμακα mel χρησιμοποιείται έναντι της γραμμικής κλίμακας (κλίμακα Hz) για να προσωμειώσει την ανθρώπινη αντίληψη της συχνότητας [75]. Πιο συγκεκριμένα, ο άνθρωπος έχει μεγαλύτερη διακριτική ικανότητα στις χαμηλές συχνότητες σε σύγκριση με τις υψηλές. Έτσι,

με χρήση της κλίμακας mel, ζεύγη ήχων που απέχουν το ίδιο μεταξύ τους συχνοτικά σύμφωνα με το ανθρώπινο αυτί, διαχωρίζονται από τον ίδιο αριθμό από mels. Η πιο διαδεδομένη (αλλά όχι μοναδική) σχέση αντιστοίχισης είναι η ακόλουθη:

$$M(f) = 2595 \log_{10}(1 + \frac{1}{700}) \tag{3.4}$$

όπως απεικονίζεται στο Σχήμα 3.3, ενώ η αντίστροφη αντιστοίχιση δίνεται από τη σχέση:

$$M^{-1}(m) = 700(10^{m/2595} - 1) \tag{3.5}$$



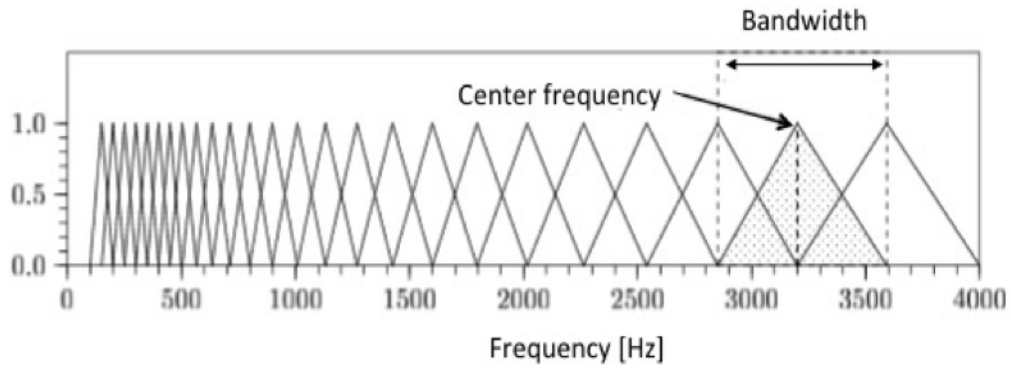
Σχήμα 3.3: Αντιστοίχηση της κλίμακας Hz στην κλίμακα Mel.

Η διαδικασία εξαγωγής των MFCCs στηρίζεται στην παραγωγή μίας συστοιχίας Q τριγωνικών φίλτρων H^j των οποίων οι κεντρικές συχνότητες (center frequencies) είναι ισοκατανεμημένες στην κλίμακα mel. Οι συχνότητες αποκοπής κάθε φίλτρου ορίζονται ως οι κεντρικές συχνότητες των δύο γειτονικών του φίλτρων. Συμβολίζοντας ως f_c^j την κεντρική συχνότητα του φίλτρου j και θεωρώντας πως $H^j(f_c^j) = 1$ καταλήγουμε στην ακόλουθη εξίσωση για το φίλτρο j:

$$H^j[k] = \begin{cases} 0 & , k < f_c^{j-1} \\ \frac{k - f_c^{j-1}}{f_c^j - f_c^{j-1}} & , f_c^{j-1} \leq k \leq f_c^j \\ \frac{f_c^{j+1} - k}{f_c^{j+1} - f_c^j} & , f_c^j \leq k \leq f_c^{j+1} \\ 0 & , k > f_c^{j+1} \end{cases} \tag{3.6}$$

Ο αριθμός των φίλτρων, Q, συνήθως επιλέγεται μεταξύ 20 και 40. Στο Σχήμα 3.1.2 παρουσιάζεται στην κλίμακα Hz μία συστοιχία 25 φίλτρων με ισαπέχουσες συχνότητες στην κλίμακα mel. Είναι εμφανές ότι τα φίλτρα είναι αραιότερα στην περιοχή των υψηλών συχνοτήτων το οποίο συμφωνεί με το γεγονός ότι ο άνθρωπος έχει μικρότερη διακριτική ικανότητα στις υψηλές συχνότητες.

Στη συνέχεια, υπολογίζεται η απόκριση κάθε φίλτρου με είσοδο το παραθυρωμένο σήμα, ακολουθούμενη από τη λογαρίθμησή της και έπειτα εφαρμόζεται ο Διακριτός Μετασχηματισμός



Σχήμα 3.4: Συστοιχία 25 τριγωνικών φίλτρων με κεντρικές συχνότητες στην κλίμακα mel. Πηγή: http://recognize-speech.com/feature-extraction/mfcc

Συνημιτόνου (DCT). Οι συντελεστές MFCC προκύπτουν ως οι πρώτοι N_c συντελεστές του σήματος που προέκυψε από τον DCT. Συνήθως $N_c = 13$ ενώ σε ορισμένες περιπτώσεις, τη θέση του πρώτου συντελεστή λαμβάνει η λογαριθμημένη τετραγωνική ενέργεια του σήματος.

Παράγωγοι των MFCCs

Οι MFCCs περιγράφουν μία εικόνα του σπεκτρογράμματος ενός μοναδικού παραθύρου του ακουστικού σήματος. Ωστόσο, δε δίνουν καμία πληροφορία για τη σχέση του με τα γειτονικά παράθυρα, το οποίο είναι ιδιαίτερα χρήσιμο για την αναγνώριση φωνής ή ηχητικών γεγονότων. Για το λόγο αυτό γίνεται υπολογισμός των πρώτων παραγώγων (Deltas) των MFCCs, οι οποίοι αναφέρονται και ως συντελεστές ταχύτητας:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}, \tag{3.7}$$

όπου d_t είναι ο προσδιοριστέος συντελεστής που αντιστοιχεί στο παράθυρο t του σήματος, ενώ η σταθερά N αντιπροσωπεύει το πλήθος γειτονικών πλαισίων που λαμβάνονται υπόψη για τον υπολογισμό των παραγώγων.

Αν η αντίστοιχη διαδικασία επαναληφθεί χρησιμοποιώντας τους συντελεστές ταχύτητας στη θέση των MFCCs λαμβάνονται οι δεύτεροι παράγωγοι Delta-Deltas ή συντελεστές επιτάχυνσης. Σε εφαρμογές λοιπόν όπου εξάγονται 13 MFCCs και οι συντελεστές ταχύτητας και επιτάχυνσης, για κάθε παράθυρο του σήματος προκύπτει ένα διάνυσμα χαρακτηριστικών με 39 συντελεστές.

Ρυθμός Μηδενισμού (Zero Crossing Rate)

Ο ρυθμός μηδενισμού (Zero Crossing Rate) είναι χρονικό χαρακτηριστικό, δηλαδή υπολογίζεται απευθείας από την αναπαράσταση του ήχου στο πεδίο του χρόνου. Ορίζεται ως ο μέσος αριθμός των μηδενισμών του σήματος σε ένα χρονικό διάστημα μήκους T:

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{\mathbb{R}_{<0}}(s_t s_{t-1}), \tag{3.8}$$

όπου s_t, η τιμή του σήματος τη χρονική στιγμή t και η συνάρτηση $1_{\mathbb{R}_{<0}}(x)$ λαμβάνει μοναδιαία τιμή αν $x < 0$, διαφορετικά λαμβάνει μηδενική τιμή (indicator function).

Χρωμόγραμμα (Chromagram)

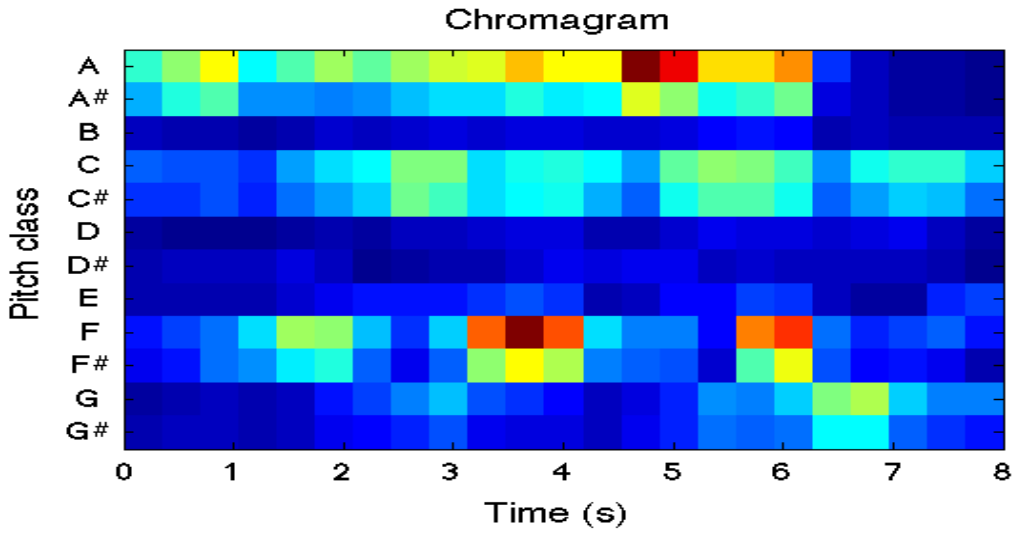
Το Χρωμόγραμμα (Chromagram ή Pitch Class Profile) έχει άμεση εφαρμογή στην ανάλυση μουσικών κομματιών λόγω του συσχετισμού του με τη μουσική αντίληψη του ανθρώπου. Όπως είναι γνωστό από τη Μουσική Θεωρία, μία οκτάβα αποτελείται από 12 ημιτόνια (semitones): Ντο, Ντο#, Ρε, Ρε#, Μι, Φα, Φα#, Σολ, Σολ#, Λα, Λα#, Σι (C, C#, D, D#, E, F, F#, G, G#, A, A#, B). Μία νότα προσδιορίζεται μέσω δύο χαρακτηριστικών: το τονικό ύψος (height) και το τονικό χρώμα (chroma). Το ύψος σχετίζεται με την κατάταξη της νότας σε διαφορετικές οκτάβες. Για παράδειγμα, η Ντο^4 (C^4) αντιστοιχεί στην 4η οκτάβα του πιάνου ενώ η Ντο^5 (C^5) αντιστοιχεί στην 5η οκτάβα του πιάνου. Το χρώμα αντιπροσωπεύει τη σχετική θέση της νότας σε μία οκτάβα. Ο όρος τονική κλάση (pitch class ή chroma class) μπορεί να επεξηγηθεί ως το σύνολο όλων των νοτών με ίδιο χρώμα. Για παράδειγμα, η τονική κλάση της νότας Ντο είναι: {..., Ντο^{-2}, Ντο^{-1}, Ντο^0, Ντο^1, Ντο^2, ...}. Το χρωμόγραμμα λοιπόν είναι η ομαδοποίηση των νοτών ενός μουσικού κομματιού στις 12 τονικές κλάσεις. Στην περίπτωση που το μουσικό δείγμα αναπαρίσταται με την MIDI σημειογραφία, τότε η διαδικασία δημιουργίας του χρωμογράματος είναι απλή. Ωστόσο, στην περίπτωση που δεν υπάρχει η MIDI πληροφορία, τότε υπάρχουν πολλοί διαφορετικοί αλγόριθμοι για τον υπολογισμό του χρωμογράμματος. Ένας εξ' αυτών είναι η χρήση του Τμηματικού Μετασχηματισμού (STFT) (βλ. Ενότητα 3.1.1) σε συνδυασμό με ειδικές τεχνικές binning [76]. Έτοιμη υλοποίηση για εξαγωγή χρωμογράμματος παρέχεται από τη συνάρτηση *mirchromagram()* του εργαλείου MIR Toolbox. Ένα παράδειγμα χρωμογράμματος απεικονίζεται στο Σχήμα 3.5. Το χρωμόγραμμα έχει χρησιμοποιηθεί ευρέως σε εφαρμογές όπως η ανίχνευση μουσικών διασκευών [77], αυτόματη ανίχνευση συγχορδιών (chord recognition), συγχρονισμός MIDI σημειογραφίας με το μουσικό απόσπασμα [78] κλπ.

Θεμελιώδης Συχνότητα (F0)

Η θεμελιώδης συχνότητα (F0) είναι ένα ιδιαίτερα συνηθισμένο χαρακτηριστικό που λαμβάνεται από σήματα φωνής. Ορίζεται ως η χαμηλότερη συχνότητα που εμφανίζεται σε μία περιοδική κυματομορφή και υποδηλώνει το ύψος (pitch) της φωνής. Η θεμελιώδης συχνότητα έχει οριστεί και σε σήματα μουσικής ως μία εκτίμηση του μουσικού ύψους, δηλαδή της νότας που ηχεί μία συγκεκριμένη χρονική στιγμή. Σε σήματα που δεν είναι περιοδικά, η θεμελιώδης συχνότητα υπολογίζεται από τη θέση του μεγίστου της αυτοσυσχέτισης του σήματος [79], ωστόσο η περιγραφή του αλγορίθμου παραλείπεται καθώς ξεφεύγει από τα όρια αυτής της εργασίας.

Ηχητική Τραχύτητα (Auditory Rouchness)

Το χαρακτηριστικό Auditory Rouchness [80] ή απλά Roughness αποτελεί μία εκτίμηση της 'τραχύτητας' του ήχου. Με τον όρο 'τραχύ' εννοούμε έναν ήχο που ακούγεται ως μη αρμονικός



Σχήμα 3.5: Παράδειγμα των χαρακτηριστικών χρωμογράμματος (Chromagram features) για μουσικό κομμάτι 8 δευτερολέπτων.

στο αυτί και παραπέμπει σε αίσθηση δυσαρέσκειας. Η τραχύτητα είναι άμεσα συνδεδεμένη με την έννοια του διακροτήματος, δηλαδή του φαινομένου συνήχησης δύο ημιτονικών ήχων με συχνότητες που διαφέρουν ελάχιστα. Η σύνθεση τέτοιου ήδους ήχων ακούγεται δυσάρεστη εξαιτίας της αδυναμίας του ανθρώπινου αυτιού στο να διαχωρίσει τις πολύ κοντινές συχνότητες. Ένας τρόπος εκτίμησης της τραχύτητας του ήχου σχετίζεται με το λόγο συχνοτήτων σε ζεύγη από ημίτονα [80] ενώ μία άλλη μέθοδος [81] αφορά τον υπολογισμό κορυφών στο φάσμα του ήχου. Η τελευταία μέθοδος υλοποιείται από τη συνάρτηση *mirroughness()* του εργαλείου MIR Toolbox.

Φασματική Ροή (Spectral Flux)

Το χαρακτηριστικό Spectral Flux (φασματική ροή) είναι ένα μέτρο του πόσο γρήγορα μεταβάλλεται με το χρόνο το φάσμα του σήματος. Ο υπολογισμός της γίνεται μέσω της Ευκλείδιας απόστασης μεταξύ των κανονικοποιημένων φασμάτων διαδοχικών παραθύρων [82]. Έτοιμη υλοποίηση για υπολογισμό της φασματικής ροής, παρέχεται από τη συνάρτηση *mirflux()* του MIR Toolbox.

3.1.3 Δημιουργία του Ακουστικού Χώρου

Παραπάνω έγινε περιγραφή της γενικότερης διαδικασίας για την εξαγωγή ακουστικών χαρακτηριστικών από ένα ηχητικό αποσπάσματα. Έστω τώρα ένα σύνολο από M ηχητικά αποσπάσματα. Κάθε ηχητικό απόσπασμα διαχωρίζεται σε ένα σύνολο από επικαλυπτόμενα χρονικά παράθυρα o_t σταθερής διάρκειας. Το πλήθος των παραθύρων διαφέρει από απόσπασμα σε απόσπασμα, καθώς αποσπάσματα μεγαλύτερης διάρκειας αποτελούνται από περισσότερα χρονικά παράθυρα. Στη συνέχεια, για κάθε χρονικό παράθυρο o_t υπολογίζεται ένα διάνυσμα χαρακτηριστικών $x_t \in \mathbb{R}^{\mathrm{d}}$, όπου d το πλήθος των ακουστικών χαρακτηριστικών. Η ίδια διαδι-

κασία για την εξαγωγή χαρακτηριστικών ακολουθείται για κάθε ένα από τα M αποσπάσματα. Έτσι, ως ακουστικό χώρο ονομάζουμε τον χώρο διάστασης d στον οποίο αναπαρίστανται όλα τα διανύσματα χαρακτηριστικών.

3.1.4 Το Μοντέλο Bag-of-Audio-Words (BoAW)

Υπενθυμίζεται ότι στο Bag-of-Words μοντέλο, ένα κείμενο αναπαρίσταται ως ένα πολυσύνολο (bag ή multiset) από τις λέξεις που εμφανίζονται σε αυτό. Αν αντί για κείμενο θεωρήσουμε ένα ηχητικό απόσπασμα και αντί για λέξεις θεωρήσουμε ηχητικά παράθυρα του αποσπάσματος, τότε με αντίστοιχο τρόπο, ένα ηχητικό απόσπασμα θα μπορούσε να αναπαρασταθεί ως ένα πολυσύνολο από τα χρονικά παράθυρα (παραθυρωμένα ακουστικά σήματα) που εμφανίζονται σε αυτό. Μια σημαντική διαφορά όμως ανάμεσα στις λέξεις και στα χρονικά παράθυρα είναι ότι οι πρώτες επαναλαμβάνονται αυτούσιες μέσα σε ένα κείμενο. Αντίθετα, δύο παράθυρα δεν επαναλαμβάνονται ποτέ αυτούσια, παρόλο που μπορεί να ηχούν παρόμοια στο ανθρώπινο αυτί. Έτσι, θεωρείται ότι χρονικά παράθυρα που έχουν παρόμοια χαρακτηριστικά αντιστοιχούν στην ίδια 'ακουστική λέξη' (audio-word - από εκεί προκύπτει και το όνομα του bag-of-audio-words μοντέλου). Για να γίνει κατανοητός ο όρος 'ακουστική λέξη', ας θεωρήσουμε το παράδειγμα ενός μουσικού κομματιού. Αυτό αντιμετωπίζεται ως κείμενο και κάθε νότα που ακούγεται θα μπορούσε να εκφραστεί ως ακουστική λέξη. Τότε, μία αλληλουχία απο ακουστικές λέξεις (νότες) αντιστοιχεί σε μία 'φράση'. Άλλωστε, ο όρος 'μουσικές φράσεις' είναι συχνά χρησιμοποιούμενος στο χώρο της μουσικής. Βέβαια, στα πλαίσια αυτής της διπλωματικής οι ακουστικές λέξεις πρέπει να αναφέρονται, όχι σε μουσικές νότες, αλλά γενικότερα σε οποιουδήποτε τύπου ηχητικό σήμα.

Η έννοια των bag-of-audio-words μοντέλων έχει ήδη χρησιμοποιηθεί στο παρελθόν για εφαρμογές όπως η ανάκτηση ηχητικών αποσπασμάτων από ερωτήματα κειμένου (text queries) [83], ανίχνευση περιστατικών κλοπής βίντεο [84] και ανίχνευση σημαντικών γεγονότων σε βίντεο [85, 86]. Στη συνέχεια παρουσιάζεται ο αλγόριθμος για τη δημιουργία του μοντέλου Bag-of-Audio-Words (BoAW).

3.1.5 Υπολογισμός των Ακουστικών Λέξεων

Όπως αναφέρθηκε προηγουμένως, μία ακουστική λέξη αντιστοιχεί σε ένα σύνολο χρονικών παραθύρων τα οποία έχουν παρόμοιο άκουσμα στο ανθρώπινο αυτί. Αν θεωρήσουμε ότι τα εξαγόμενα ακουστικά χαρακτηριστικά είναι αντιπροσωπευτικά, τότε παρόμοια χρονικά παράθυρα βρίσκονται σε κοντινές αποστάσεις στον ακουστικό χώρο. Αφού πρώτα γίνει εξαγωγή ακουστικών χαρακτηριστικών από έναν ικανοποιητικό αριθμό από ηχητικά αποσπάσματα, ο υπολογισμός των ακουστικών λέξεων πραγματοποιείται μέσω της ομαδοποίησης (clustering) των ακουστικών χαρακτηριστικών στον ακουστικό χώρο.

Ο πιο γνωστός αλγόριθμος ομαδοποίησης είναι ο αλγόριθμος k-μέσων (k-means). Έστω $x_1, ..., x_n \in \mathbb{R}^d$ τα n ακουστικά χαρακτηριστικά προς ομαδοποίηση. Δοθούσης της τιμής $k \leq n$, ο αλγόριθμος k-means ομαδοποιεί τα n δείγματα σε k συστάδες (clusters) C με κεντρίδια (centroids) $\mu_1, ..., \mu_k \in \mathbb{R}^d$, με κριτήριο την ελαχιστοποίηση του τετραγώνου της

απόστασης κάθε δείγματος από το κοντινότερό του κεντρίδιο:

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_j - \mu_i||^2) \tag{3.9}$$

Η παραπάνω ποσότητα αναφέρεται ως (inertia) και συμβολίζει τη συνεκτικότητα των συστάδων. Συνήθως, για τον υπολογισμό των κεντριδίων γίνεται χρήση του προσεγγιστικού αλγορίθμου Lloyd, ο οποίος λειτουργεί σε τρία βήματα. Κατά το πρώτο βήμα γίνεται αρχικοποίηση των k κεντριδίων, δίνοντάς τους συνήθως τις τιμές υπαρχόντων δειγμάτων. Δεύτερο βήμα είναι η ανάθεση καθενός από τα δείγματα στην κοντινότερη συστάδα με κριτήριο την Ευκλείδια απόσταση από το αντίστοιχο κεντρίδιο. Τρίτο βήμα είναι ο επανυπολογισμός των κεντριδίων κάθε συστάδας ως η μέση τιμή των δειγμάτων που ομαδοποιήθηκαν σε καθένα από αυτά στο προηγούμενο βήμα. Το δεύτερο και τρίτο βήμα επαναλαμβάνονται έως ότου να μην υπάρχει σημαντική μεταβολή στις τιμές των κεντριδίων. Συνήθως γίνεται χρήση κάποιου κατωφλίου για τον καθορισμό της συνθήκης τερματισμού. Εδώ πρέπει να τονιστεί ότι η αρχικοποίηση των κεντριδίων παίζει πολύ σημαντικό ρόλο για τη σύγκλιση του αλγορίθμου και ακατάλληλες τιμές αρχικοποίησης είναι πιθανό να οδηγήσουν στη σύγκλιση σε τοπικά ελάχιστα και όχι σε ολικά ελάχιστα της συνάρτησης ελαχιστοποίησης. Γι΄ αυτό το λόγο συνήθως πραγματοποιείται επανάληψη του αλγορίθμου με διαφορετικές τυχαίες αρχικοποιήσεις και γίνεται επιλογή της βέλτιστης λύσης. Επίσης, συχνά γίνεται αρχικοποίηση των κεντριδίων με χρήση του αλγορίθμου k-means++ [87].

Η πολυπλοκότητα του αλγορίθμου k-means με χρήση του αλγορίθμου *Lloyd* είναι $O(nkdi)$, όπου n ο αριθμός των δειγμάτων, k ο αριθμός των συστάδων, d η διάσταση των δειγμάτων και i το πλήθος επαναλήψεων του αλγορίθμου. Μία παραλλαγή του αλγορίθμου k-means είναι ο Mini Batch K-means. Σε αντίθεση με τον k-means, κατά τον οποίο λαμβάνονται υπόψη όλα τα δείγματα σε κάθε επανάληψη, στον Mini Batch k-means, λαμβάνεται υπόψη ένα τυχαίο υποσύνολο των δειγμάτων σε κάθε επανάληψη. Ως αποτέλεσμα, με χρήση του τελευταίου αλγορίθμου μειώνεται σημαντικά το χρονικό κόστος υπολογισμού των κεντριδίων και η ακρίβεια των προβλέψεων είναι ελάχιστα χειρότερη σε σύγκριση με αυτή των προβλέψεων του αλγορίθμου k-means.

Οι k ακουστικές λέξεις προκύπτουν ως τα k κεντροειδή (centroids) των συστάδων (clusters) που επιστρέφει ο αλγόριθμος k-means. Επομένως το σύνολο των ακουστικών λέξεων, το οποίο αναφέρεται και ως Λεξικό Ακουστικών Λέξεων (Audio-word Vocabulary) είναι το $C = \{c_1, c_2, ..., c_k\}$. Το πλήθος των ακουστικών λέξεων, k, είναι μία ιδιαίτερα σημαντική παράμετρος που επηρεάζει την απόδοση του μοντέλου και θα μελετηθεί αναλυτικά στην Ενότητα 4.1.

3.1.6 Αναπαραστάσεις Ηχητικών Αποσπασμάτων

Σε αυτή την ενότητα περιγράφονται δύο μέθοδοι για την αναπαράσταση ηχητικών αποσπασμάτων με χρήση του bag-of-audio-words μοντέλου. Από εδώ και στο εξής, τέτοιου τύπου αναπαραστάσεις θα αναφέρονται ως bag-of-audio-words (BoAW) αναπαραστάσεις. Έστω ένα

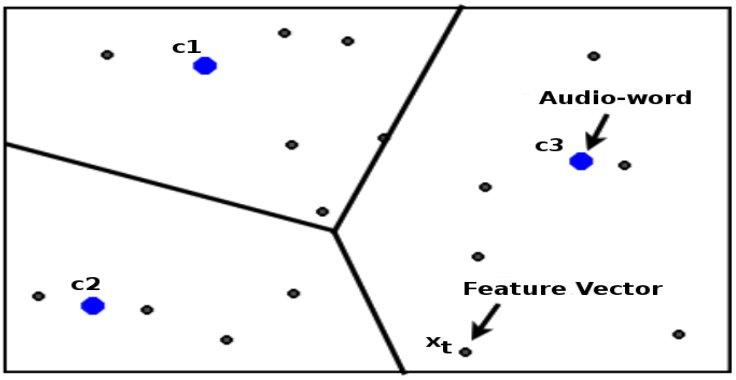
νέο ηχητικό απόσπασμα για το οποίο θέλουμε να υπολογίσουμε την bag-of-audio-words αναπαράσταση. Ακολουθώντας τη διαδικασία που περιγράφεται στην Ενότητα 3.1.2, για κάθε χρονικό παράθυρο $o_t, \quad t = 1, 2, ..., T$ του αποσπάσματος υπολογίζεται ένα διάνυσμα χαρακτηριστικών $x_t \in \mathbb{R}^d$.

Hard Encoding

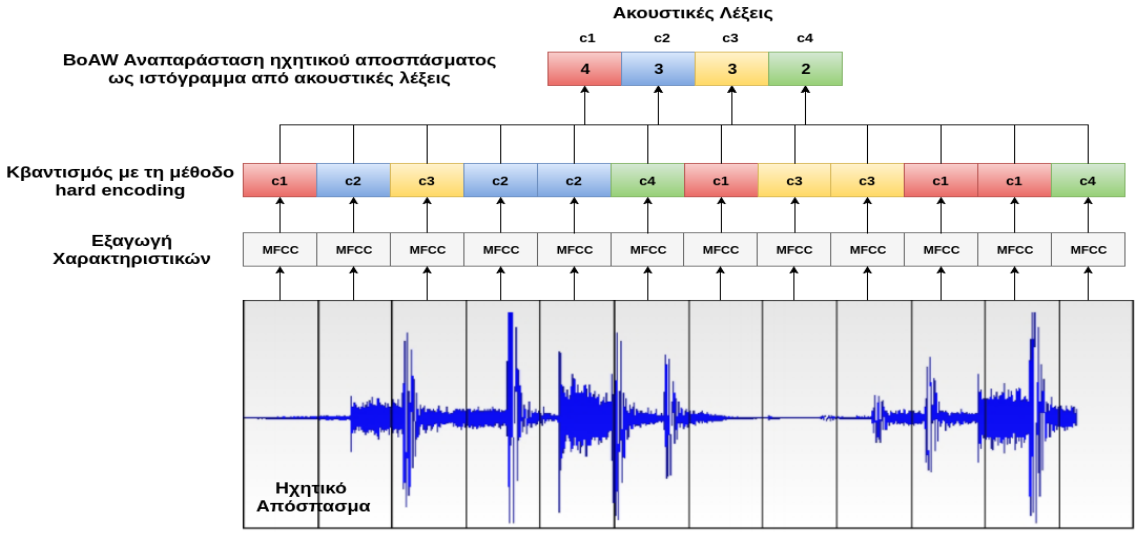
Σύμφωνα με την πρώτη μέθοδο (hard-encoding), κάθε διάνυσμα χαρακτηριστικών x_t αντιστοιχίζεται στην κοντινότερη ακουστική λέξη c_i, με κριτήριο την Ευκλείδια απόσταση. Η μέθοδος αυτή αναφέρεται και σε γενικότερες εφαρμογές ως vector quantization. Έτσι, η bag-of-audio-words αναπαράσταση του διανύσματος x_t συμβολίζεται ως e_t και αναπαρίσταται ως ένα διάνυσμα μήκους k, όπου ένα στοιχείο έχει μοναδιαία τιμή ενώ τα υπόλοιπα $k-1$ στοιχεία έχουν μηδενική τιμή:

$$e_t = (0, ..., 1, 0, ..., 0). \tag{3.10}$$

Το μη-μηδενικό στοιχείο αντιστοιχεί στην κοντινότερη ακουστική λέξη. Τέτοιου είδους αναπαραστάσεις, όπου μόνο ένα στοιχείο είναι μη μηδενικό αναφέρονται συχνά ως one-hot αναπαραστάσεις. Η BoAW αναπαράσταση του ηχητικού αποσπάσματος προκύπτει ως το άθροισμα των BoAW αναπαραστάσεων των επιμέρους χρονικών παραθύρων. Επομένως, η τελική αναπαράσταση μπορεί να θεωρηθεί ως ένα ιστόγραμμα των ακουστικών λέξεων και υποδηλώνει το βαθμό συμμετοχής κάθε ακουστικής λέξης στο συγκεκριμένο ηχητικό απόσπασμα.



Σχήμα 3.6: Σχηματική αναπαράσταση της μεθόδου hard encoding σε μορφή διαγράμματος Voronoi, όπου $k = 3$ και $d = 2$ (διδιάστατος χώρος χαρακτηριστικών).



Σχήμα 3.7: Η διαδικασία αναπαράστασης ενός ηχητικού αποσπάσματος με τη μέθοδο Bag-of-Audio-Words (BoAW). Πλήθος ακουστικών λέξεων: $k = 4$. Μέθοδος κβαντισμού: 'hard encoding'.

Soft Encoding

Με χρήση της μεθόδου hard encoding, κάθε χρονικό παράθυρο του ηχητικού αποσπάσματος αντιστοιχίζεται με μία μόνο ακουστική λέξη. Μία εναλλακτική μέθοδος για τον υπολογισμό των BoAW αναπαραστάσεων ονομάζεται soft encoding και αναμένεται ότι οδηγεί σε εύρωστες αναπαραστάσεις, ακόμα και στην περίπτωση ύπαρξης θορύβου στο σήμα. Η βασική ιδέα της μεθόδου είναι η συνεισφορά περισσότερων από μία ακουστικών λέξεων στην αναπαράσταση e_t ενός χρονικού παραθύρου. Αντί λοιπόν να συνεισφέρει μόνο μία ακουστική λέξη με βάρος 1 και όλες οι υπόλοιπες με βάρος 0, εδώ κάθε ακουστική λέξη c_i συνεισφέρει με βάρος w_i με την προϋπόθεση ότι $\sum_{i=1}^{k} w_i = 1$. Το ζήτημα που προκύπτει σε αυτή την περίπτωση είναι ο υπολογισμός των βαρών w_i. Η απλούστερη λύση είναι η ακόλουθη:

$$w_i = \frac{\frac{1}{dist(c_i, x_t)}}{\sum_{j=1}^{k} \frac{1}{dist(c_j, x_t)}}, \tag{3.11}$$

όπου $dist(c_i, x_t)$ η Ευκλείδια απόσταση μεταξύ των αναπαραστάσεων c_t και x_t ενώ ο όρος του αθροίσματος λειτουργεί ως όρος κανονικοποίησης ώστε να ικανοποιείται ο περιορισμός $\sum_{i=1}^{k} w_i = 1$. Ωστόσο με χρήση της Εξίσωσης 3.11 δε συνυπολογίζεται το γεγονός ότι κάθε ακουστική λέξη έχει διαφορετική κατανομή στον ακουστικό χώρο.

Σε αυτή την εργασία προτείνεται μία εναλλακτική μέθοδος για τον υπολογισμό του βαθμού συνεισφοράς των ακουστικών λέξεων στην αναπαράσταση ενός ηχητικού αποσπάσματος, μοντελοποιώντας κάθε ακουστική λέξη με διαφορετική κατανομή στον ακουστικό χώρο. Όπως αναφέρθηκε στην Ενότητα 3.1.5 οι ακουστικές λέξεις προκύπτουν ως τα κεντροειδή των clusters που επιστρέφει ο αλγόριθμος k-means. Αν θεωρηθεί ότι τα διανύσματα χαρακτηριστικών που αντιστοιχούν στις ακουστικές λέξεις ακολουθούν κανονική (ή γκαουσιανή)

κατανομή, τότε οι παράμετροι κάθε κατανομής (μέση τιμή $\mu \in \mathbb{R}^{\mathrm{d}}$ και ο πίνακας συνδιακύμανσης $\Sigma \in \mathbb{R}^{\mathrm{d}\times\mathrm{d}}$) υπολογίζονται πάνω στο αποτέλεσμα του αλγορίθμου k-means μέσω της Εκτίμησης Μέγιστης Πιθανοφάνειας (Maximum Likelihood Estimation). Εδώ θεωρούμε για απλότητα ότι ο πίνακας συνδιακύμανσης είναι διαγώνιος, δηλαδή τα d χαρακτηριστικά είναι γραμμικώς ανεξάρτητα (και καθώς πρόκειται για κανονικές κατανομές είναι πλήρως ανεξάρτητα). Έτσι, για κάθε ακουστική λέξη c_i, η μέση τιμή $\mu_i = (\mu_{i1}, ..., \mu_{id})$ και η διασπορά $\sigma_i^2 = (\sigma_{i1}^2, ..., \sigma_{id}^2)$ (διαγώνια στοιχεία του πίνακα συνδιακύμανσης) προκύπτουν ως η μέση τιμή και η διασπορά των διανυσμάτων χαρακτηριστικών τα οποία κατηγοριοποιήθηκαν στο αντίστοιχο cluster. Επομένως, το βάρος w_i υπολογίζεται ως:

$$w_i = \frac{p(c_i|x_t)}{\sum_{j=1}^{k} p(c_j|x_t)}, \tag{3.12}$$

όπου, με τον όρο $p(.)$ υποδηλώνεται η έννοια της πιθανότητας, οπότε $p(c_i)$ είναι η εκ των προτέρων (a priori) πιθανότητα ενώ $p(c_i|x_t)$ είναι η εκ των υστέρων (a posteriori) της ακουστικής λέξης c_i. Σύμφωνα με τον Μπεϋζιανό κανόνα:

$$p(c_j|x_t) = \frac{p(x_t|c_j)p(c_j)}{p(x_t)}. \tag{3.13}$$

Καθώς όμως έχουμε θεωρήσει ότι οι ακουστικές λέξεις ακολουθούν κανονική κατανομή, προκύπτει:

$$p(x_t|c_j) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}\cdot h_{tj}^2}, \tag{3.14}$$

όπου h_{tj} είναι η απόσταση Mahalanobis μεταξύ των x_t και c_j και Σ είναι ο πίνακας συνδιακύμανσης. Σημειώνεται ότι η Εξίσωση 3.14 ισχύει λόγω της υπόθεσης ότι ο πίνακας συνδιακύμανσης είναι διαγώνιος. Συνδυάζοντας τις Εξισώσεις 3.12, 3.13 και 3.14 καταλήγουμε στην τελική σχέση υπολογισμού των βαρών:

$$w_i = \frac{p(c_i)|\Sigma_i|^{-\frac{1}{2}} e^{-h_{ti}^2}}{\sum_{j=1}^{k} p(c_j)|\Sigma_j|^{-\frac{1}{2}} e^{-h_{tj}^2}}. \tag{3.15}$$

Επομένως, η BoAW αναπαράσταση του διανύσματος x_t υπολογίζεται ως:

$$e_t^{'} = (w_1, w_2, ..., w_k). \tag{3.16}$$

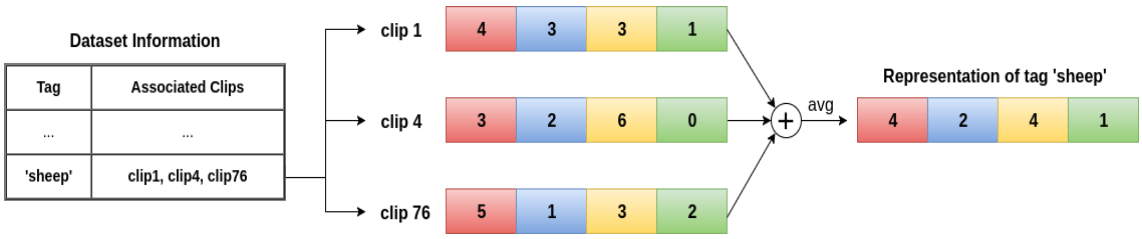
Σε περίπτωση που είναι επιθυμητή η συνεισφορά συγκεκριμένου πλήθους $N < k$ ακουστικών λέξεων στην παραπάνω αναπαράσταση, τότε διατηρούνται τα N μεγαλύτερα βάρη, όλα τα υπόλοιπα μηδενίζονται και γίνεται ξανά κανονικοποίηση ώστε οι εναπομείνασες τιμές να αθροίζονται στην μονάδα.

Αναφέρεται ότι εκτός της παραπάνω μεθόδου έχουν χρησιμοποιηθεί και άλλες μέθοδοι για soft encoding στη βιβλιογραφία [88], με χρήση των οποίων διαπιστώθηκε βελτίωση της απόδοσης του bag-of-audio-words μοντέλου. Στην Ενότητα 4 θα παρουσιαστούν αντίστοιχα πειράματα σύγκρισης των μεθόδων hard encoding και soft encoding.

3.1.7 Το Ακουστικό-Σημασιολογικό (ADSM) Μοντέλο

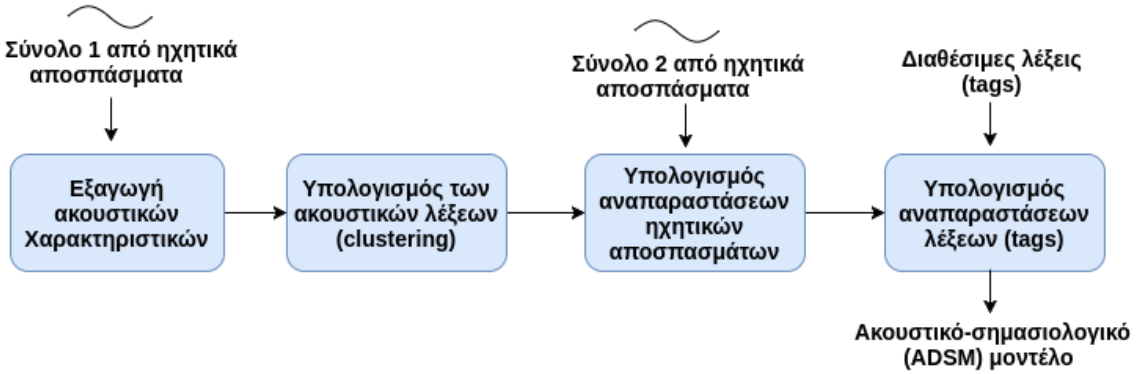
Συχνά, τα ηχητικά αποσπάσματα συνοδεύονται από μία ή περισσότερες λέξεις-ετικέτες (tags) οι οποίες περιγράφουν το περιεχόμενό τους. Για παράδειγμα ένα μουσικό απόσπασμα θα μπορούσε να συνοδεύεται από τις λέξεις: "rock", "singing", "guitar", "happy" κλπ. ενώ ένα ηχητικό απόσπασμα άσχετο με μουσική θα μπορούσε να περιγράφεται από τις λέξεις: "car", "birds", "loud", "silence", "horror", "ring" πολλές από τις οποίες υποδηλώνουν κάποιο συγκεκριμένο ακουστικό γεγονός ενώ άλλες περιγράφουν το συναίσθημα που προκαλεί ο ήχος στον άνθρωπο. Η πληροφορία που παρέχεται μέσω των tags αποδεικνύεται ιδιαίτερα χρήσιμη σε πολλές εφαρμογές σχετικές με την Εξαγωγή Πληροφορίας κυρίως από μουσικά αποσπάσματα αλλά και γενικότερα [89, 90, 91, 92, 93]. Στα πλαίσια αυτής της διπλωματικής, η σύνδεση των ηχητικών αποσπασμάτων με tags αξιοποιείται για τη δημιουργία σημασιολογικών αναπαραστάσεων λέξεων σύμφωνα με τις ακουστικές τους ιδιότητες. Επειδή η ιδέα για τη δημιουργία αυτών των αναπαραστάσεων αντλείται από τη μέθοδο για την κατασκευή των παραδοσιακών σημασιολογικών μοντέλων (βλ. Ενότητα 2), συχνά γίνεται αναφορά στις παραπάνω αναπαραστάσεις ως ακουστικό-σημασιολογικό μοντέλο (Audio-based Distributional Semantic Model - ADSM). Στη συνέχεια περιγράφεται η διαδικασία για την κατασκευή του ADSM μοντέλου.

Έστω ένα σύνολο από M ηχητικά αποσπάσματα, καθένα από τα οποία συνοδεύεται από ένα ή περισσότερα tags. Με χρήση των μεθόδων soft encoding και hard encoding της προηγούμενης υποενότητας προκύπτουν bag-of-audio-words (BoAW) αναπαραστάσεις ηχητικών αποσπασμάτων. Η BoAW αναπαράσταση ενός tag προκύπτει ως ο μέσος όρος (ανά στοιχείο) των αναπαραστάσεων των ηχητικών αποσπασμάτων που περιλαμβάνουν το συγκεκριμένο tag στην περιγραφή τους. Έτσι, οι BoAW αναπαραστάσεις υπολογίζονται ως προς k ακουστικές



Σχήμα 3.8: Αναπαραστάσεις λέξεων με το ADSM μοντέλο

λέξεις και συνολικά υπάρχουν T μοναδικά tags, τότε η επανάληψη της παραπάνω διαδικασίας για κάθε tag ξεχωριστά οδηγεί στη δημιουργία ενός πίνακα διάστασης $T \times k$, όπου οι γραμμές υποδηλώνουν τις αναπαραστάσεις διαφορετικών tags ενώ οι στήλες υποδηλώνουν το βαθμό εμφάνισης συγκεκριμένων ακουστικών λέξεων. Ο πίνακας αυτός έχει ίδια δομή με τον πίνακα λέξεων-συμφραζομένων (βλ. Ενότητα 2.1), επομένως είναι αναμενόμενο ότι η στάθμιση κατά PPMI (βλ. Ενότητα 2.2) θα οδηγήσει σε αναπαραστάσεις κατάλληλες για την αξιολόγηση της σημασιολογικής ομοιότητας λέξεων (στην περίπτωσή μας tags). Επιπλέον, είναι δυνατό να εφαρμοστεί μείωση της διαστασιμότητας του πίνακα με χρήση της τεχνικής Principal Component Analysis (PCA) (βλ. Ενότητα 2.3).



Σχήμα 3.9: Σχηματική αναπαράσταση των βημάτων για τη δημιουργία του ακουστικού - σημασιολογικού μοντέλου (ADSM).

Συνοψίζοντας, ο χαρακτηρισμός των ηχητικών αποσπασμάτων με περιγραφικές λέξεις (tags) δίνει τη δυνατότητα δημιουργίας αναπαραστάσεων των λέξεων αυτών με χρήση των ακουστικών χαρακτηριστικών που προέκυψαν από τα αποσπάσματα. Τα βασικά βήματα της παραπάνω διαδικασίας αναπαριστώνται στο Σχήμα 3.9. Σημειώνεται ότι για τον υπολογισμό των ακουστικών-λέξεων μπορεί να χρησιμοποιηθεί ένα σύνολο από ηχητικά αποσπάσματα (Σύνολο 1), το οποίο είναι ανεξάρτητο από το σύνολο αποσπασμάτων (Σύνολο 2) που συνοδεύονται από περιγραφικές λέξεις (tags). Ένα μειονέκτημα της μεθόδου είναι το γεγονός ότι είναι εφικτή η δημιουργία αναπαράστασης μόνο για λέξεις οι οποίες έχουν το ρόλο tags και συνοδεύονται από ηχητικά αποσπάσματα. Βέβαια, είναι αναμενόμενο ότι ορισμένες λέξεις, για παράδειγμα η λέξη 'δημοκρατία', δε σχετίζονται με ακουστικά ερεθίσματα, επομένως είναι λογικό να μην προκύπτουν αναπαραστάσεις για αυτές. Παρόλα αυτά, είναι δυνατή η επέκταση του ADSM μοντέλου, ώστε η ηχητική αναπαράσταση μίας 'άγνωστης' λέξης να προκύπτει μέσω του συνδυασμού των αναπαραστάσεων από διαφορετικά μοντέλα. Η μέθοδος αυτή θα περιγραφεί στην Ενότητα 3.3. Αξίζει επίσης να αναφερθεί ότι μέσω του ADSM μοντέλου, οι αναπαραστάσεις των λέξεων υπολογίζονται ως διανύσματα στον ίδιο διανυσματικό χώρο όπου αναπαρίστανται ηχητικά αποσπάσματα. Το γεγονός αυτό προσφέρει τη δυνατότητα για άμεση χρήση του ADSM μοντέλου σε πληθώρα εφαρμογών, οι οποίες περιγράφονται αναλυτικά στο Κεφάλαιο 4.

3.1.8 Επέκταση: Σύμπτυξη Πολλαπλών Ακουστικών Χώρων

Για τη δημιουργία του ADSM μοντέλου, πραγματοποιείται εξαγωγή συγκεκριμένου τύπου ακουστικών χαρακτηριστικών, με βάση τα οποία γίνεται υπολογισμός των ακουστικών λέξεων. Στη βιβλιογραφία αναφέρεται ότι με χρήση των MFCC συντελεστών επιτυγχάνεται ικανοποιητική απόδοση σε πληθώρα εφαρμογών [74, 94, 95, 96, 97, 98]. Ωστόσο, ορισμένα ακουστικά χαρακτηριστικά μοντελοποιούν αποδοτικά ηχητικά αποσπάσματα σηγκεκριμένου τύπου, ενώ αδυνατούν να μοντελοποιήσουν αποσπάσματα διαφορετικού τύπου [99, 100]. Για παράδειγμα, όταν πρόκειται για την ανίχνευση πλαισίων φωνής [101, 102, 103, 104], γίνεται χρήση χαρακτηριστικών όπως οι Linear Prediction Coefficients (LPC), ο Σύντομος Μετασχηματι-

σμός Φουριέ (Short-Time Fourier Transform) και τα AM-FM χαρακτηριστικά, ενώ για την αξιολόγηση της ομοιότητας μουσικών κομματιών [77, 105] είναι πολύ συχνή η χρήση του Χρωμογράμματος (Chromagram). Με βάση τα παραπάνω και δεδομένου ότι το ADSM μοντέλο θα χρησιμοποιηθεί σε ηχητικά αποσπάσματα με μεγάλη ποικιλομορφία (σήματα μουσικής, φωνής ή γενικότερου τύπου), κρίνεται σημαντική η δημιουργία εύρωστων αναπαραστάσεων για κάθε είδους σήμα. Στα πλαίσια αυτής της διπλωματικής, γίνεται πρόταση μίας μεθόδου για την επέκταση του ADSM μοντέλου η οποία περιλαμβάνει τη χρήση διαφορετικών ακουστικών χαρακτηριστικών με βάση τη 'φύση' του ήχου και στη συνέχεια την σύμπτυξη των διαφορετικών αναπαραστάσεων για δημιουργία μίας ενιαίας αναπαράστασης.

Ταξινόμηση των ηχητικών αποσπασμάτων σε τρεις κατηγορίες

Η 'φύση' ενός ηχητικού αποσπάσματος προσδιορίζεται μέσω της ταξινόμησής του σε μία από τις ακόλουθες κλάσεις 1) 'μουσική', 2) 'φωνή' και 3) 'γενικού τύπου απόσπασμα'. Φυσικά, ο τύπος και το πλήθος των κλάσεων θα μπορούσε να είναι διαφορετικός, ωστόσο στα πλαίσια αυτής τις διπλωματικής θα γίνει χρήση των προηγούμενων κλάσεων. Και αυτό, πρώτον γιατί καλύπτουν το σύνολο των διαφορετικών ηχητικών αποσπασμάτων και δεύτερον γιατί έχουν προηγηθεί πολλές δημοσιεύσεις αναφορικά με το ποια χαρακτηριστικά μοντελοποιούν καλύτερα τα σήματα κάθε κλάσης ξεχωριστά [77, 101, 102, 103, 105].

Στην παρούσα διπλωματική το πρόβλημα της ταξινόμησης στις κλάσεις ΄μουσική΄, ΄φωνή΄ και ΄γενικού τύπου απόσπασμα΄ επιλύεται με χρήση Μηχανών Διανυσματικής Υποστήριξης (Support Vector Machines - SVMs) [106]. Σε προβλήματα ταξινόμησης σε δύο κλάσεις (binary classification) όπου τα δεδομένα $x_i \in \mathbb{R}^{\mathrm{d}}$ των δύο κλάσεων είναι γραμμικώς διαχωρίσιμα, υπάρχουν άπειρα υπερεπίπεδα διαχωρισμού των δεδομένων:

$$w^T x + b = 0, \quad w \in \mathbb{R}^{\mathrm{d}}, \mathrm{b} \in \mathbb{R}. \tag{3.17}$$

Με χρήση του ταξινομητή SVM γίνεται εύρεση του βέλτιστου υπερεπίπεδου διαχωρισμού στον d-διάστατο χώρο των δεδομένων, το οποίο προκύπτει ως το υπερεπίπεδο με το μεγαλύτερο περιθώριο διαχωρισμού (margin). Περιθώριο διαχωρισμού είναι η ελάχιστη απόσταση των δεδομένων των δύο κλάσεων από το επίπεδο διαχωρισμού και τα δεδομένα με τη μικρότερη απόσταση από το υπερεπίπεδο ονομάζονται διανύσματα υποστήριξης (support vectors). Αποδεικνύεται ότι το πρόβλημα εύρεσης του βέλτιστου υπερεπιπέδου διαχωρισμού μετατρέπεται σε πρόβλημα δευτεροβάθμιου προγραμματισμού (quadratic programming):

$$\min \frac{1}{2}||w||^2 \tag{3.18}$$

$$y_k(w^T x + b) \geq 1, \quad k = 1, ..., n \tag{3.19}$$

όπου μεταβλητή $y_k \in \{-1, +1\}$ υποδηλώνει την κλάση του δεδομένου x_k ($y_k = 1$ αν το δεδομένο x_k ανήκει στην πρώτη κλάση ενώ $y_k = -1$ αν το δεδομένο ανήκει στην δεύτερη κλάση). Το παραπάνω πρόβλημα επιλύεται με χρήση των εξισώσεων Lagrange, ωστόσο η διαδικασία επίλυσης ξεφεύγει από τα όρια της διπλωματικής. Στην περίπτωση που τα δεδομένα είναι γραμμικώς μη διαχωρίσιμα, δηλαδή δεν είναι δυνατή η εύρεση υπερεπιπέδου διαχωρισμού,

προτείνεται η μέθοδος χαλαρού περιθωρίου (soft margin), σύμφωνα με την οποία προστίθενται κάποιες σταθερές 'χαλάρωσης' των περιορισμών της εξίσωσης 3.19:

$$\xi_k = \max\{0, 1 - y_k(w^T x_k + b)\}. \tag{3.20}$$

Έτσι, επιτρέπονται 'λάθη' κατά το διαχωρισμό των δεδομένων και προκύπτει το νέο πρόβλημα βελτιστοποίησης:

$$\min_{w,\xi} \left\{ \frac{1}{2}||w||^2 + C\sum_{k=1}^{n} \xi_k \right\} \tag{3.21}$$

$$y_k(w^T x + b) \geq 1 - \xi_k, \quad \xi_k \geq 0, k = 1, ..., n \tag{3.22}$$

Η σταθερά C λειτουργεί ως όρος κανονικοποίησης (regularization term) και όσο μικρότερη τιμή λαμβάνει, τόσο πιο 'χαλαροί' είναι η περιορισμοί κατηγοριοποίησης στη σωστή κλάση.

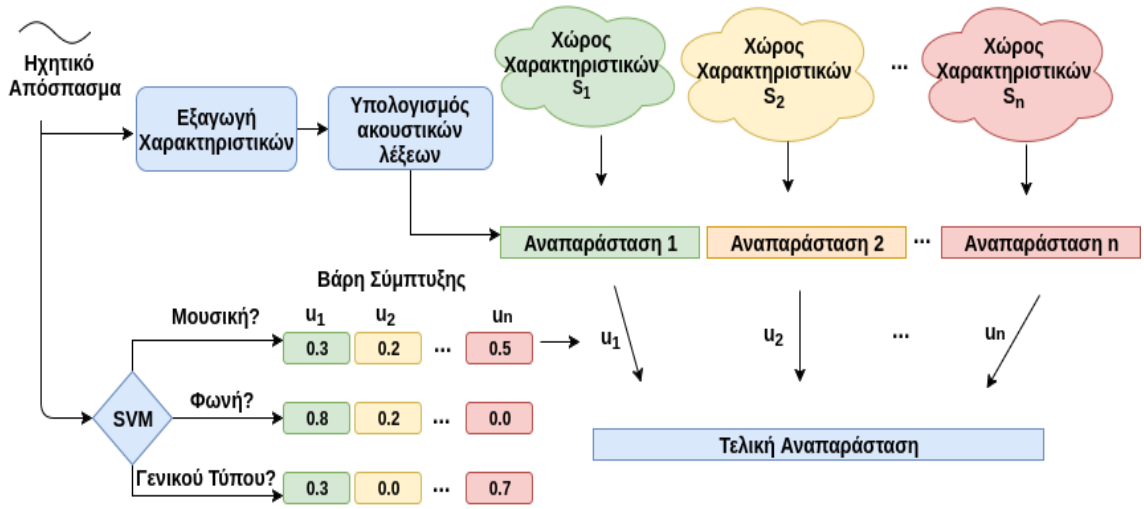
Μία ακόμη επέκταση του ταξινομητή SVM είναι η χρήση συναρτήσεων πυρήνα (kernel functions). Η γραμμική μη-διαχωρισιμότητα είναι δυνατό να επιλυθεί μέσω της εφαρμογής μίας συνάρτησης $\phi(x)$ στα δεδομένα, ώστε να λάβουν αναπαραστάσεις σε χώρο μεγαλύτερης διάστασης, στον οποίο θα είναι γραμμικώς διαχωρίσιμα. Το πρόβλημα που προκύπτει είναι το ίδιο με το προηγούμενο αν τα δεδομένα x_i αντικατασταθούν από τα δεδομένα $\phi(x_i)$ Ωστόσο, για την επίλυση του προβλήματος δεν είναι απαραίτητο να γνωρίζουμε τη συνάρτηση $\phi(x)$ αλλά τα εσωτερικά γινόμενα $\phi(x_i)^T\phi(x_j)$. Έτσι, ορίζεται η συνάρτηση πυρήνα $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$. Οι πιο γνωστοί τύποι συναρτήσεων πυρήνα είναι οι ακόλουθοι:

- Γραμμική: $K(x_i, x_j) = x_i^T x_j$
- Πολυωνυμική βαθμού p: $K(x_i, x_j) = (x_i^T x_j + c)^p$
- Γκαουσιανή (rbf): $K(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{\sigma})$

Στην περίπτωση κατηγοριοποίησης σε N κλάσεις (multiclass classification), γίνεται χρήση της μεθόδου 'ένα-προς-πολλά' (one-versus-all), σύμφωνα με την οποία γίνεται εκπαίδευση N διαφορετικών ταξινομητών. Πιο συγκεκριμένα, για κάθε κλάση εκπαιδεύεται ένας ταξινομητής για το διαχωρισμό δεδομένων της συγκεκριμένης κλάσης από το σύνολο των άλλων κλάσεων. Έτσι, στα πλαίσια αυτής της διπλωματικής όπου υπάρχουν τρεις διαφορετικές κλάσεις, γίνεται εκπαίδευση τριών ταξινομητών SVM από τους οποίους ο πρώτος ταξινομεί κάθε απόσπασμα στις κλάσεις 1) 'μουσική', 2) 'όχι μουσική', ο δεύτερος στις κλάσεις 1) 'φωνή', 2) 'όχι φωνή' κ.ο.κ. Η τελική κλάση για κάθε δείγμα προκύπτει ως η κλάση στην οποία έγινε κατηγοριοποίηση με μεγαλύτερη αξιοπιστία (confidence score), δηλαδή στην περίπτωση των SVMs η περίπτωση στην οποία υπήρχε μεγαλύτερη απόσταση του δείγματος από το υπερεπίπεδο διαχωρισμού.

Υπολογισμός και σύμπτυξη των ηχητικών αναπαραστάσεων

Όπως αναφέρθηκε προηγουμένως, σκοπός της επέκτασης του ADSM μοντέλου είναι η χρήση διαφορετικών ακουστικών χαρακτηριστικών και η σύμπτυξή τους σε μία ενιαία σημασιολογική αναπαράσταση. Έστω ότι για κάθε σήμα εξάγονται n διαφορετικά σύνολα ακουστικών χαρακτηριστικών. Έτσι, δημιουργούνται οι χώροι χαρακτηριστικών $S_1, S_2, ..., S_n$ σε



Σχήμα 3.10: Σχηματική αναπαράσταση του επεκτεταμένου συστήματος για τη δημιουργία αναπαραστάσεων με χρήση πολλαπλών χώρων χαρακτηριστικών.

κάθε έναν από τους οποίους γίνεται υπολογισμός των ακουστικών λέξεων με τη μέθοδο που περιγράφεται στην Ενότητα 3.1.5. Καθώς δεν είναι απαραίτητο να υπολογιστεί ίδιο πλήθος ακουστικών λέξεων σε κάθε χώρο χαρακτηριστικών, ορίζουμε με αντίστοιχο τρόπο τα πλήθη των ακουστικών λέξεων ως $k_1, k_2, ..., k_n$. Έστω τώρα ένα ηχητικό απόσπασμα q για το οποίο επιθυμούμε να υπολογίσουμε την bag-of-audio-words (BoAW) αναπαράσταση. Αρχικά, γίνεται υπολογισμός της BoAW αναπαράστασης σε κάθε έναν από τους χώρους χαρακτηριστικών $S_1, S_2, ..., S_n$ ξεχωριστά, με χρήση της μεθόδου που περιγράφεται στην Ενότητα 3.1.6.

Η τελική αναπαράσταση του αποσπάσματος q υπολογίζεται ως το επαυξημένο διάνυσμα διάστασης $k = k_1 + k_2 + ... + k_n$ που προκύπτει ως η αλληλουχία των n επιμέρους BoAW αναπαραστάσεων, αφού αυτές πρώτα σταθμιστούν με συγκεκριμένες τιμές βάρους $u_1, u_2..., u_n$ αντίστοιχα:

$$e_q = (u_1 e_q^1, u_2 e_q^2, ..., u_n e_q^n), \tag{3.23}$$

όπου $\sum_{i=1}^{n} u_i = 1$. Τα βάρη u_i λαμβάνουν διαφορετικές τιμές για κάθε απόσπασμα με βάση την ταξινόμησή του στις κατηγορίες 'μουσική', 'φωνή' και 'γενικού τύπου απόσπασμα'. Πιο συγκεκριμένα, για κάθε μία από τις τρεις κλάσεις προσδιορίζεται ένα διαφορετικό σύνολο από βάρη. Έτσι, αν για παράδειγμα ένα απόσπασμα ταξινομείται στην κατηγορία 'μουσική', τα βάρη ορίζουν τη συνεισφορά συγκεκριμένων χώρους χαρακτηριστικών ενώ αν ταξινομείται στην κατηγορία 'φωνή', γίνεται συνεισφορά διαφορετικών χώρων χαρακτηριστικών.

Το ζήτημα που απομένει είναι ο καθορισμός των τιμών βάρους $u_1, ..., u_n$ για κάθε κατηγορία ξεχωριστά. Μία λύση θα ήταν να δοθούν αυθαίρετες τιμές στα βάρη με κριτήριο το πόσο αποδοτικά περιγράφει κάθε χώρος χαρακτηριστικών το σήμα. Για παράδειγμα, όπως αναφέρθηκε νωρίτερα, ο χώρος των χαρακτηριστικών Χρωμογράμματος (Chromagram feautures) περιγράφει καλύτερα τα σήματα μουσικής, επομένως θα δινόταν μεγαλύτερο βάρος στον χώρο αυτό στην περίπτωση της κατηγορίας 'μουσική'. Ωστόσο, η λύση αυτή είναι καθαρά υποκειμενική και επιπλέον δεν είναι εύκολο να ποσοτικοποιηθεί η 'ποιότητα' των χαρακτηριστικών

για κάθε είδους σήμα. Γι΄ αυτό το λόγο, τα βάρη $u_1, ..., u_n$ αφήνονται ως παράμετροι του μοντέλου και θα προσδιοριστούν σε κάθε εφαρμογή ξεχωριστά με χρήση γνωστών μεθόδων ρύθμισης των παραμέτρων (parameter tuning). Αφού γίνει υπολογισμός των BoAW αναπαραστάσεων για ένα σύνολο αποσπασμάτων, οι BoAW αναπαραστάσεις λέξεων προκύπτουν με χρήση του ADSM μοντέλου που περιγράφεται στην Ενότητα 3.1.7. Σημειώνεται ότι με τη συγκεκριμένη μέθοδο, η κατηγοριοποίηση ενός ηχητικού αποσπάσματος σε μία από τις τρεις κατηγορίες καθορίζει την τιμή των βαρών που θα χρησιμοποιηθούν για όλα τα χρονικά παράθυρα του αποσπάσματος. Η μέθοδος αυτή μπορεί να επεκταθεί ώστε να πραγματοποιείται κατηγοριοποίηση κάθε παραθύρου ξεχωριστά. Έτσι, αν για παράδειγμα παρουσιάζεται μέσα σε ένα ηχητικό απόσπασμα εμφάνιση παραθύρων φωνής αλλά και παραθύρων μουσικής, τότε για κάθε παράθυρο θα γίνει χρήση διαφορετικού συνδυασμού βαρών.

Συνοψίζοντας, σε αυτή την ενότητα έγινε περιγραφή της μεθοδολογίας για τη δημιουργία του ADSM μοντέλου, σύμφωνα με το οποίο κάθε λέξη αποκτά μία διανυσματική αναπαράσταση βασισμένη στα ακουστικά της χαρακτηριστικά. Επιπλέον έγινε πρόταση της μεθόδου soft encoding, σύμφωνα με την οποία περισσότερες από μία ακουστικές λέξεις συνεισφέρουν στην κωδικοποίηση ενός διανύσματος χαρακτηριστικών με βάρη που προκύπτουν μέσω της θεώρησης ότι κάθε ακουστική λέξη ακολουθεί κανονική κατανομή διαφορετικών παραμέτρων (μέση τιμή και πίνακα συνδιακύμανσης). Τέλος, έγινε πρόταση μίας μεθόδου για τη δημιουργία πολλαπλών χώρων χαρακτηριστικών και την αποδοτική συνεισφορά τους στη δημιουργία της τελικής BoAW αναπαράστασης με κριτήριο την ταξινόμηση των αποσπασμάτων στις κατηγορίες ‘μουσική’, ‘φωνή’ και ‘γενικού τύπου απόσπασμα’.

3.2 Πολυτροπικά Μοντέλα Εικόνας

Πολύ συχνά γίνεται αναφορά στη φράση ¨Μία εικόνα χίλιες λέξεις¨, ώστε να τονιστεί ότι το περιεχόμενο μίας εικόνας παρέχει ιδιαίτερα χρήσιμη πληροφορία σημασιολογικού ενδιαφέροντος. Επίσης, όπως αναφέρθηκε στην εισαγωγή αυτής της ενότητας, ορισμένα από τα περιγραφικά χαρακτηριστικά των λέξεων (όπως για παράδειγμα ότι η μπανάνα έχει κίτρινο χρώμα) θεωρούνται αυτονόητα από τον άνθρωπο με αποτέλεσμα να μη γίνεται αναφορά σε αυτά σε πηγές κειμένου. Αναμένεται, λοιπόν, ότι όπως η σημασιολογική αναπαράσταση λέξεων με χρήση των ακουστικών τους χαρακτηριστικών (βλ. Ενότητα 3.1) έτσι και η αναπαράσταση λέξεων με χρήση των οπτικών τους χαρακτηριστικών θα συμβάλλει στην θεμελίωση (grounding) των σημασιολογικών αναπαραστάσεων σύμφωνα με την αισθητήρια αντίληψη του ανθρώπου.

Η διαδικασία για την αναπαράσταση του περιεχομένου μίας εικόνας είναι παρόμοια με αυτήν που ακολουθείται στην περίπτωση ηχητικών αποσπασμάτων και βασίζεται στη μέθοδο bag-of-visual-words. Κάθε εικόνα δηλαδή αναπαρίσταται ως ένα πολυσύνολο από οπτικές λέξεις. Η μέθοδος αυτή έχει μελετηθεί εκτεταμένα στο παρελθόν και έχει εφαρμοστεί σε πληθώρα εφαρμογών [1, 107, 108, 109, 110, 111]. Εφόσον στην παρούσα διπλωματική δόθηκε έμφαση στη δημιουργία και τη βελτίωση των ηχητικών αναπαραστάσεων, θα γίνει αφορά στη δημιουργία οπτικών αναπαραστάσεων όπως περιγράφεται σε πρόσφατη δημοσίευση [112].

3.2.1 Εξαγωγή Οπτικών Χαρακτηριστικών

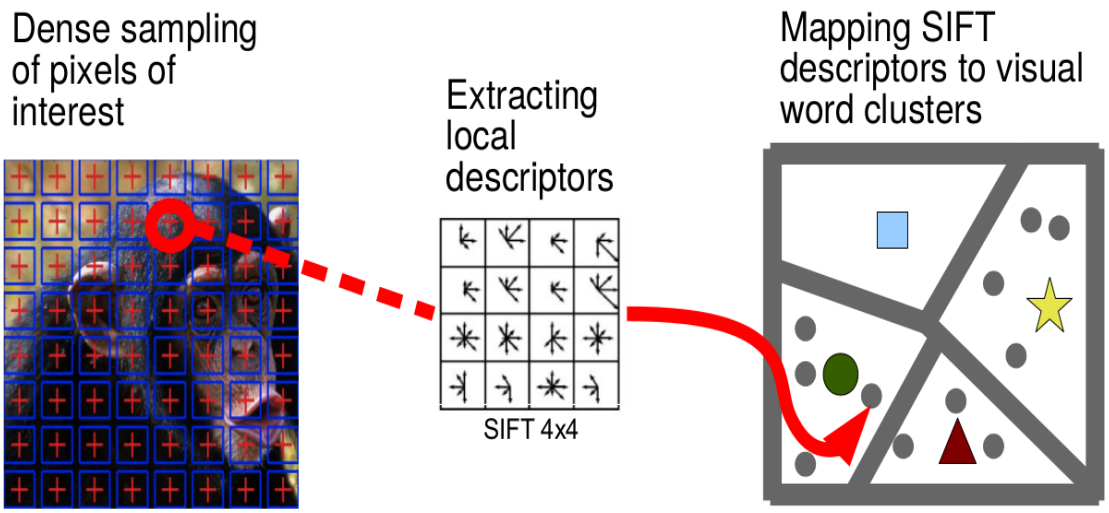
Η διαδικασία εξαγωγής χαρακτηριστικών από εικόνες έχει μελετηθεί διεξοδικά στο παρελθόν και έχουν βρεθεί χαρακτηριστικά που περιγράφουν συγκεκριμένες ιδιότητες του περιεχομένου των εικόνων, όπως για παράδειγμα οι ακμές [113], το σχήμα [114], το χρώμα [115] και η υφή [116]. Πολύ γνωστός αλγόριθμος εξαγωγής χαρακτηριστικών είναι ο SIFT (Scale Invariant Feature Transform) [117, 118] με χρήση του οποίου είναι δυνατή η ανίχνευση και περιγραφή τοπικών χαρακτηριστικών σε εικόνες ενώ σημαντικό πλεονέκτημα του αλγορίθμου είναι η ανεξαρτησία στην κλίμακα, περιστροφή αλλά και στο φωτισμό. Γενικότερα, ο SIFT αλγόριθμος είναι ανεξάρτητος από οποιουδήποτε τύπου affine μετασχηματισμούς. Λόγω της διακριτικής του ικανότητας ανταποκρίνεται με επιτυχία στον εντοπισμό σημαντικών σημείων αλλά και στην αντιστοίχηση σημείων ανάμεσα σε δύο εικόνες και έχει χρησιμοποιηθεί σε εφαρμογές όπως η αναγνώριση αντικειμένων σε εικόνες [117]. Συνήθως, η εξαγωγή χαρακτηριστικών γίνεται στον τριδιάστατο χώρο RGB (Red, Green, Blue), όπου κάθε διάσταση ενός pixel συμβολίζει την ένταση των χρωμάτων κόκκινο, πράσσινο και μπλε αντίστοιχα. Ωστόσο συχνά πραγματοποιείται η απεικόνιση των pixels σε διαφορετικούς χώρους όπως ο HSV (Hue, Saturation Value) στον οποίο η πληροφορία του χρώματος κωδικοποιείται με τρόπο που προσομοιώνει την κωδικοποίηση των χρωμάτων από τον άνθρωπο. Αν το πλήθος των χαρακτηριστικών που εξάγονται είναι d, τότε η εξαγωγή χαρακτηριστικών από έναν τριδιάστατο χρωματικό χώρο οδηγεί σε ένα διάνυσμα χαρακτηριστικών στο χώρο $\mathbb{R}^{3\times d}$. Το τελικό διάνυσμα χαρακτηριστικών προκύπτει ως ο μέσος όρος των τριών τιμών σε κάθε μία από τις d διαστάσεις, επομένως αναπαριστάται στο χώρο $\mathbb{R}^{d}$.

3.2.2 Υπολογισμός των Οπτικών Λέξεων

Ο υπολογισμός των οπτικών λέξεων πραγματοποιείται με την ίδια ακριβώς διαδικασία όπως ο υπολογισμός των ακουστικών λέξεων. Δηλαδή, έχοντας εξάγει οπτικά χαρακτηριστικά από έναν ικανοποιητικό πλήθος εικόνων, πραγματοποιείται ομαδοποίηση (clustering) των ακουστικών με χρήση του αλγορίθμου k-means. Οι k οπτικές λέξεις προκύπτουν ως τα κεντροειδή των συμπλεγμάτων (clusters) που επιστρέφει ο αλγόριθμος k-means. Έχοντας προκαθορίσει το πλήθος των ακουστικών λέξεων, καταλήγουμε στη δημιουργία του Λεξικού Οπτικών Λέξεων (Visual Word Vocabulary): $C_v = \{c_{v1}, c_{v2}, ..., c_{vk}\}$.

3.2.3 Αναπαραστάσεις Εικόνων

Οι αναπαραστάσεις εικόνων υπολογίζονται με χρήση του bag-of-visual-words μοντέλου, δηλαδή ως ιστογράμματα οπτικών λέξεων. Αν για παράδειγμα η i-οστή διάσταση της αναπαράστασης μίας εικόνας έχει τιμή 4 (η αναπαράσταση ως ιστόγραμμα περιλαμβάνει θετικές ακέραιες τιμές), υποδηλώνεται ότι η i-οστή οπτική λέξη εμφανίζεται 4 φορές στην εικόνα. Βέβαια, αναπαριστώντας την εικόνα απλά ως ιστόγραμμα οπτικών λέξεων δε γνωρίζουμε σε ποιο σημείο της εικόνας εμφανίζεται κάθε ακουστική λέξη. Επομένως, για να γίνει εισαγωγή της γεωμετρίας της εικόνας στην αναπαράστασή της, η μέθοδος επεκτείνεται εμπερικλείοντας τον όρο των χωρικών ιστογραμμάτων (spatial histograms) [119, 120]. Πιο συγκεκριμένα,



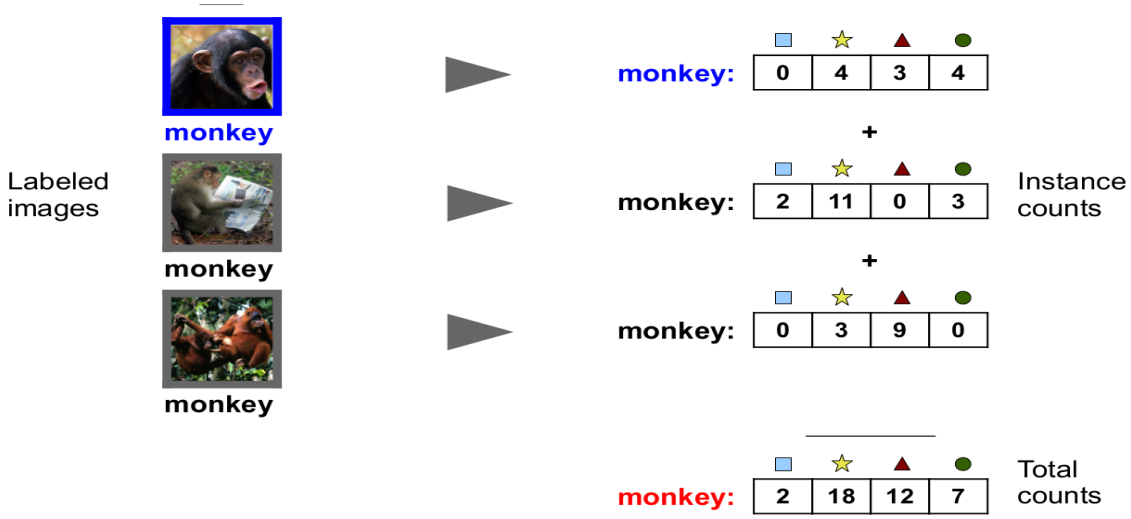
Σχήμα 3.11: Αναπαράσταση εικόνων με τη μέθοδο Bag of Visual Words. Πηγή: [1]

κάθε εικόνα χωρίζεται σε $P \times Q$ περιοχές και η εξαγωγή χαρακτηριστικών πραγματοποιείται σε κάθε μία περιοχή ανεξάρτητα. Έτσι, προκύπτουν $P \times Q$ διανύσματα χαρακτηριστικών για μία εικόνα και κάθε ένα από αυτά κωδικοποιείται με μία bag-of-visual-words αναπαράσταση, με κριτήριο την απόστασή του από τις οπτικές λέξεις. Η τελική baf-of-visual-words αναπαράσταση της εικόνας υπολογίζεται ως το επαυξημένο διάνυσμα διάστασης $k_v = P \cdot Q \cdot k$ που προκύπτει ως η αλληλουχία των $P \cdot Q$ αναπαραστάσεων των επιμέρους περιοχών.

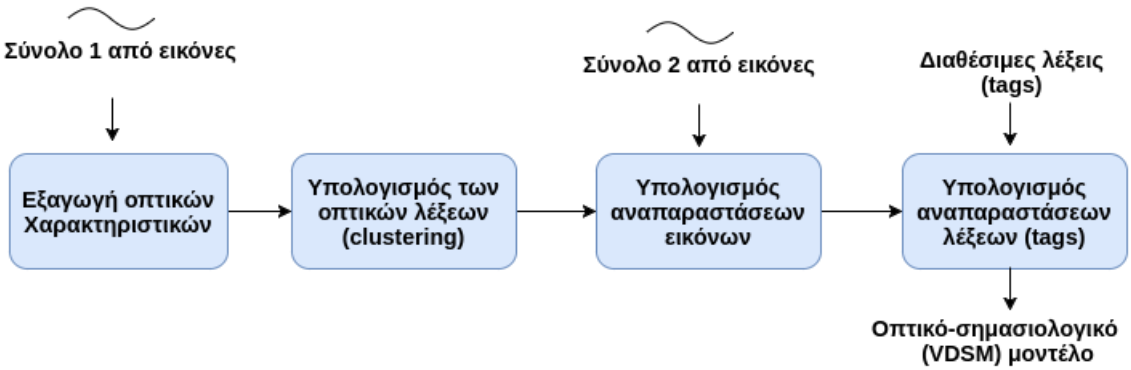
3.2.4 Το Οπτικό-Σημασιολογικό (VDSM) Μοντέλο

Στην ενότητα 3.1.7 έγινε περιγραφή της δημιουργίας του ακουστικού-σημασιολογικού ADSM μοντέλου με χρήση ενός συνόλου ηχητικών αποσπασμάτων τα οποία συνοδεύονται από περιγραφικές λέξεις (tags). Με αντίστοιχο τρόπο, για τη δημιουργία αναπαραστάσεων λέξεων με χρήση των οπτικών τους χαρακτηριστικών γίνεται χρήση του οπτικού-σημασιολογικού (VDSM) μοντέλου. Έστω ένα σύνολο από M_v εικόνες, κάθε μία από τις οποίες χαρακτηρίζεται από ένα ή περισσότερα tags. Τότε, η bag-of-visual-words αναπαράσταση ενός tag υπολογίζεται ως ο μέσος όρος των bag-of-visual-words αναπαραστάσεων των εικόνων που περιλαμβάνουν το συγκεκριμένο tag στην περιγραφή τους. Ένα παράδειγμα της διαδικασίας δημιουργίας αναπαραστάσεων λέξεων με το μοντέλο VDSM απεικονίζεται στο Σχήμα 3.12. Έτσι, αν συνολικά υπάρχουν T_v μοναδικά tags, υπολογίζοντας την bag-of-visual-words αναπαράσταση κάθε εικόνας ξεχωριστά καταλήγουμε σε έναν πίνακα διάστασης $T_v \times k_v$. Οι γραμμές του πίνακα συμβολίζουν τις αναπαραστάσεις των διαφορετικών tags ενώ οι στήλες συμβολίζουν το βαθμό εμφάνισης των οπτικών λέξεων σε κάθε αναπαράσταση. Ο παραπάνω πίνακας έχει ίδια δομή με τον πίνακα λέξεων-συμφραζομένων (βλ. Ενότητα 2.1), επομένως αναμένεται ότι η στάθμιση κατά PPMI (βλ. Ενότητα 2.2) θα δημιουργήσει πιο κατάλληλες αναπαραστάσεις λέξεων. Ακόμη, εφόσον είναι επιθυμητό, μπορεί να γίνει εφαρμογή κάποιας μεθόδου μείωσης διαστασιμότητας του πίνακα, όπως για παράδειγμα η Principal Component Analysis (βλ. Ενότητα 2.3). Η συνολική διαδικασία για τη δημιουργία του μοντέλου VDSM με

τη μέθοδο Bag of Visual Words απεικονίζεται στο Σχήμα 3.13. Συγκρίνοντας αυτό το σχήμα με το Σχήμα 3.9 παρατηρείται ότι η δημιουργία των μοντέλων ADSM και VDSM παρουσιάζει εμφανείς ομοιότητες στα περισσότερα στάδια.



Σχήμα 3.12: Παράδειγμα αναπαράστασης της λέξης 'monkey' με βάση το οπτικό-σημασιολογικό μοντέλο VDSM. Πηγή: [1].



Σχήμα 3.13: Σχηματική αναπαράσταση των βημάτων για τη δημιουργία του οπτικού - σημασιολογικού μοντέλου (VDSM). Πηγή: [1].

3.3 Πολυτροπική Σύμπτυξη

Με χρήση του ακουστικού-σημασιολογικού (ADSM) μοντέλου προκύπτουν διανυσματικές αναπαραστάσεις λέξεων με βάση τα ακουστικά τους χαρακτηριστικά οι οποίες αναφέρονται ως bag-of-audio-words αναπαραστάσεις. Αντίστοιχα, το οπτικό-σημασιολογικό (VDSM) μοντέλο περιλαμβάνει τις bag-of-visual-words αναπαραστάσεις λέξεων, δηλαδή αναπαραστάσεις με βάση τα οπτικά τους χαρακτηριστικά. Εύλογα προκύπτει το ερώτημα για το πώς είναι δυνατό να συνδυαστούν οι αναπαραστάσεις λέξεων που προκύπτουν από το κλασσικό (DSM) μοντέλο με τις αναπαραστάσεις των ADSM και VDSM μοντέλων για τη δημιουργία μίας κοινής αναπαράστασης. Η σύμπτυξη των αναπαραστάσεων αναφέρεται ώς Πολυτροπική Σύμπτυξη (Multimodal Fusion), καθώς συνδυάζεται η πληροφορία από διαφορετικού τύπου αισθητήριες πηγές (sense modalities). Στη βιβλιογραφία έχουν προταθεί πολλές μέθοδοι για την Πολυτροπική Σύμπτηξη [1, 121, 122, 123, 124, 125, 126]. Σε ορισμένες από αυτές έχει γίνει προσπάθεια μοντελοποίησης της σχέσης ανάμεσα στις πηγές εικόνας, ήχου και κειμένου. Στα πλαίσια αυτής της διπλωματικής θα δοθεί έμφαση στη σύμπτυξη αναπαραστάσεων λέξεων που προκύπτουν από τα DSM, ADSM και VDSM μοντέλα οπότε δε θα ασχοληθούμε με θέματα μοντελοποίησης της άμεσης σύνδεσης μεταξύ ήχων και εικόνων μέσα από τα οποία προκύπτουν ζητήματα όπως ο συγχρονισμός των πολυτροπικών πηγών. Κεντρικός στόχος λοιπόν είναι η αξιολόγηση της σημασιολογικής ομοιότητας λέξεων με χρήση των παραπάνω σημασιολογικών μοντέλων.

Η πολυτροπική σύμπτυξη είναι δυνατό να πραγματοποιηθεί σε διαφορετικά επίπεδα. Πρώτον, στο επίπεδο των αναπαραστάσεων (Feature Level Fusion ή Early Fusion) όπου μία λέξη λαμβάνει μία αναπαράσταση που προκύπτει ως ο συνδυασμός των επιμέρους αναπαραστάσεων των DSM, ADSM και VDSM μοντέλων. Πιο συγκεκριμένα, μπορούν να χρησιμοποιηθούν διαφορετικοί τύποι συνδυασμού των επιμέρους αναπαραστάσεων. Εδώ, αν για μία λέξη q υπολογιστούν οι αναπαραστάσεις e_q^T, e_q^A και e_q^V από τα DSM, ADSM και VDSM μοντέλα αντίστοιχα, τότε η πολυτροπική αναπαράσταση προκύπτει ως η σταθμισμένη αλληλουχία των τριών αναπαραστάσεων, δηλαδή:

$$e_q = (w_q^T e_q^T, w_q^A e_q^A, w_q^V e_q^V), \tag{3.24}$$

όπου w_q^T,w_q^A,w_q^V βάρη τα οποία αθροίζονται στη μονάδα. Τα βάρη αυτά είναι δυνατό να λαμβάνουν διαφορετική τιμή για κάθε λέξη. Βέβαια, η αναπαράσταση e_q περιλαμβάνει διαχωρισμένες τις αναπαραστάσεις που προέρχονται από τα τρία διαφορετικά μοντέλα. Η τελική αναπαράσταση της λέξης q προκύπτει μέσω της 'ανάμειξης' των διαστάσεων. Για την ανάμειξη των διαστάσεων γίνεται χρήση της μεθόδου Principal Component Analysis (με χρήση της τεχνικής SVD) καθώς έχει αναφερθεί (βλ. Ενότητα 2.3) ότι με χρήση αυτής της μεθόδου γίνεται αποσυσχέτιση των γραμμικά εξαρτημένων χαρακτηριστικών αλλά και ο υπολογισμός των 'κρυφών' διαστάσεων γίνει μεγιστοποίηση της διακύμανσης των δεδομένων στον χώρο που θα προκύψει.

Εκτός από το επίπεδο των αναπαραστάσεων, η πολυτροπική σύμπτυξη είναι δυνατό να πραγματοποιηθεί στο επίπεδο της αξιολόγησης (Scoring Level Fusion ή Late Fusion) όπου

γίνεται αξιολόγηση της σημασιολογικής ομοιότητας λέξεων για κάθε μοντέλο ξεχωριστά και η τελική τιμή προκύπτει ως ο σταθμισμένος γραμμικός συνδυασμός των επιμέρους τιμών. Δηλαδή, αν γίνει υπολογισμός της σημασιολογικής ομοιότητας για δύο λέξεις p, q και προκύψουν οι τιμές ομοιότητας s_{pq}^T, s_{pq}^A, s_{pq}^V από τα μοντέλα DSM, ADSM και VDSM αντίστοιχα, τότε η τελική τιμή προκύπτει ως:

$$s_{pq} = (w_{pq}^T s_{pq}^T, w_{pq}^A s_{pq}^A, w_{pq}^V s_{pq}^V), \tag{3.25}$$

όπου $w_{pq}^T + w_{pq}^A + w_{pq}^V = 1$. Με χρήση του συγκεκριμένου συμβολισμού για τις τιμές των βαρών υποννοείται ότι είναι δυνατό να λάβουν διαφορετικές τιμές για κάθε ζεύγος λέξεων. Παρόλο που έγινε αναφορά στη της τελευταίας μεθόδου για την αξιολόγηση της σημασιολογικής ομοιότητας, η μέθοδος μπορεί να γενικευτεί ώστε να εφαρμόζεται και σε διαφορετικά ζητήματα, τα οποία όμως ξεφεύγουν από τα όρια της διπλωματικής.

Τέλος, αξίζει να αναφερθεί ότι τελευταία έχουν προταθεί μέθοδοι για την πολυτροπική σύμπτυξη με χρήση βαθιών νευρωνικών δικτύων (deep neural networks) [127, 128, 129, 130]. Οι μέθοδοι αυτές παρουσιάζουν ενδιαφέρον καθώς δε χρειάζεται να προκαθοριστούν οι τιμές των βαρών για την σύμπτυξη ούτε να οριστεί συγκεκριμένη μέθοδος για την 'ανάμειξη' των τριών αναπαραστάσεων. Αντίθετα, οι τελικές αναπαραστάσεις προκύπτουν μέσα από την εκπαίδευση του νευρωνικού δικτύου ώστε να βελτιστοποιηθεί κάποια συνάρτηση κόστους με πιθανή προσθήκη ενός ή περισσότερων όρων κανονικοποίησης.

Κεφάλαιο 4

Εφαρμογές των Πολυτροπικών Σημασιολογικών Μοντέλων

4.1 Σημασιολογική Ομοιότητα/Σχετικότητα Λέξεων

Στον τομέα της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing) εμφανίζεται συχνά η ανάγκη για την αξιολόγηση της σημασιολογικής ομοιότητας (semantic similarity) και της σημασιολογικής σχετικότητας (semantic relatedness) μεταξύ λέξεων ή κειμένων. Η σημασιολογική ομοιότητα δύο λέξεων υποδηλώνει ότι οι λέξεις συνδέονται σημασιολογικά μέσω της σχέσης 'είναι'. Παραδείγματα σημασιολογικά όμοιων λέξεων είναι τα ζεύγη (καθηγητής - διδάκτωρ), (αυτοκίνητο - λεωφορείο) κλπ. Ωστόσο, δύο λέξεις μπορεί να σχετίζονται χωρίς να είναι απαραίτητα όμοιες. Για παράδειγμα οι λέξεις (σούπα - κουτάλι) σχετίζονται καθώς η σούπα τρώγεται με χρήση κουταλιού και οι λέξεις (αυτοκίνητο - ρόδα) σχετίζονται γιατί η ρόδα είναι εξάρτημα ενός αυτοκινήτου. Επομένως, η σημασιολογική σχετικότητα είναι πιο γενικός όρος από την σημασιολογική ομοιότητα και υποδηλώνει οποιαδήποτε σχέση μεταξύ λέξεων [131]. Όπως αναφέρθηκε στην Ενότητα 2, για τον υπολογισμό διανυσματικών αναπαραστάσεων λέξεων με χρήση των Κατανεμημένων Σημασιολογικών Μοντέλων λαμβάνεται υπόψη η συνεμφάνιση των λέξεων σε πηγές κειμένου, γεγονός το οποίο καθιστά τα μοντέλα αυτά κατάλληλα για την αξιολόγηση της σημασιολογικής σχετικότητας [132].

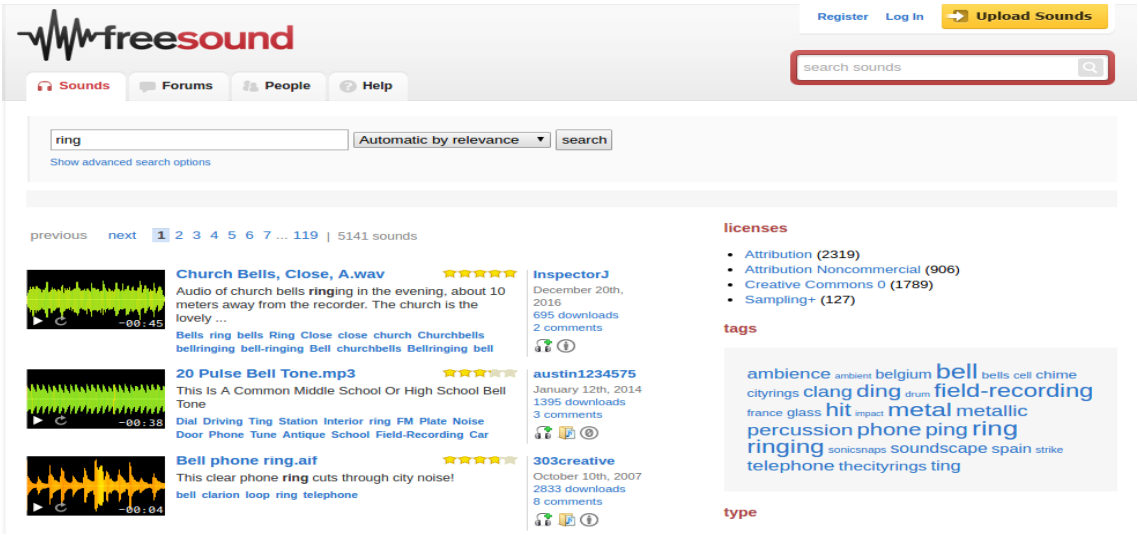
Πώς θα αξιολογούσε κάποιος τη σημασιολογική σχετικότητα των λέξεων (κιθάρα - πιάνο) και πώς συγκρίνεται με τη σημασιολογική σχετικότητα των λέξεων (κιθάρα - τύμπανο); Αντίστοιχα, σχετίζονται σημασιολογικά οι λέξεις (σειρήνα - ασθενοφόρο); Είναι προφανές ότι στις παραπάνω περιπτώσεις, η αξιολόγηση της σημασιολογικής ομοιότητας/σχετικότητας καθορίζεται σε μεγάλο βαθμό από τις ακουστικές ιδιότητες των λέξεων. Παρόλο που και οι τρεις λέξεις (κιθάρα, πιάνο, τύμπανο) αναφέρονται σε μουσικά όργανα, το άκουσμα της κιθάρας είναι πιο σχετικό με το άκουσμα του πιάνου σε σχέση με το άκουσμα του τυμπάνου, επομένως η σχετικότητα του ζεύγους (κιθάρα - πιάνο) αναμένεται να είναι μεγαλύτερη από τη σχετικότητα του ζεύγους (κιθάρα - τύμπανο). Για την αξιολόγηση, λοιπόν, της σημασιολογικής ομοιότητας/σχετικότητας λέξεων με βάση τις ακουστικές τους ιδιότητες θα γίνει χρήση του ακουστικού-σημασιολογικού (ADSM) μοντέλου.

4.1.1 Περιγραφή Αλγορίθμου/Δεδομένων

Ο αλγόριθμος για τον υπολογισμό των αναπαραστάσεων λέξεων με βάση τα ακουστικά τους χαρακτηριστικά περιγράφεται αναλυτικά στην Ενότητα 3.1. Εδώ θα γίνει αναφορά στα δεδομένα που χρησιμοποιήθηκαν για τη δημιουργία των σημασιολογικών αναπαραστάσεων και θα γίνει αξιολόγηση των μοντέλων (με διαφορετικούς συνδυασμούς παραμέτρων) στο ζήτημα υπολογισμού της σημασιολογικής ομοιότητας λέξεων.

Δεδομένα και Εξαγωγή Χαρακτηριστικών

Για τη δημιουργία του ακουστικού-σημασιολογικού (ADSM) μοντέλου είναι απαραίτητη η εύρεση ενός συνόλου από ηχητικά αποσπάσματα, καθένα από τα οποία συνοδεύονται από περιγραφικές λέξεις (tags). Στα πλαίσια αυτής της διπλωματικής έγινε χρήση 4474 ηχητικών αποσπασμάτων τα οποία εντοπίστηκαν στην online μηχανή αναζήτησης Freesound [133] με χρήση του Freesound API.



Σχήμα 4.1: Η μηχανή αναζήτησης ηχητικών αποσπασμάτων Freesound.

Τα ηχητικά αποσπάσματα ήταν κωδικοποιημένα στην καθιερωμένη μορφή ανοιχτού κώδικα OGG και δεν περιορίζονται μόνο σε αποσπάσματα μουσικής ή φωνής αλλά περιλαμβάνουν ήχους γενικού τύπου όπως ειδοποιήσεις ξυπνητηριών, βήματα ανθρώπων, θόρυβο πόλης κλπ. Σε γενικές γραμμές τα περισσότερα αποσπάσματα είναι μικρής διάρκειας και καθένα από αυτά συνοδεύονται από ένα ή περισσότερα tags. Περισσότερα στατιστικά για τα αποσπάσματα περιλαμβάνονται στον Πίνακα 4.1. Σημειώνεται ότι διατηρήθηκαν τα tags που δεν περιλαμβάνουν μόνο ψηφία (μερικά tags για παράδειγμα αναφερόταν σε ημερομηνίες π.χ. 1990) και εμφανίζονται περισσότερες από 5 φορές στο σύνολο των αποσπασμάτων.

Για την εξαγωγή χαρακτηριστικών, όλα τα αποσπάσματα μετατράπηκαν στη μορφή WAV και ο ρυθμός δειγματοληψίας κάθε αποσπάσματος άλλαξε (εφόσον ήταν απαραίτητο) στα 44.1kHz. Το μέγεθος του χρονικού παραθύρου για την εξαγωγή χαρακτηριστικών ορίζεται ως η μεταβλητή L και θα διερευνηθεί κατά πόσο επηρεάζει την απόδοση του ADSM

Συνολικός αριθμός αποσπασμάτων	4474	Συνολικός αριθμός από tags	37203
Ελάχιστη διάρκεια	0.1s	Μέσος αριθμός tags ανά απόσπασμα	8
Μέγιστη διάρκεια	120s	Μέσος αριθμός αποσπασμάτων ανά tag	40
Μέση διάρκεια	16.6s	Αριθμός από μοναδικά tags	940

Πίνακας 4.1: Στατιστικά χαρακτηριστικά των ηχητικών αποσπασμάτων που εντοπίστηκαν μέσω της μηχανής αυτόματης αναζήτησης Freesound. Κάθε απόσπασμα συνοδεύεται από μία ή περισσότερες περιγραφικές λέξεις tags.

μοντέλου στο συγκεκριμένο πρόβλημα. Για τη μελέτη του παραδοσιακού ADSM μοντέλου, τα διανύσματα χαρακτηριστικών που κατασκευάστηκαν αποτελούνται από 39 χαρακτηριστικά, εκ των οποίων 13 συντελεστές MFCC (όπου ο πρώτος συντελεστής έχει αντικατασταθεί από τη φασματική ενέργεια του σήματος), 13 πρώτες παράγωγοι και 13 δεύτερες παράγωγοι των συντελεστών.

Υπολογισμός των bag-of-audio-words αναπαραστάσεων

Ο υπολογισμός των ακουστικών λέξεων γίνεται ξεχωριστά για καθέναν από τους χώρους χαρακτηριστικών S_1,S_2,S_3 όπως περιγράφεται στην Ενότητα 3.1.5. Αντί του κλασσικού αλγορίθμου k-means έγινε χρήση μίας βελτιστοποιημένης μεθόδου (mini-batch k-means) με χρήση της οποίας μειώνεται σημαντικά το υπολογιστικό κόστος. Πιο συγκεκριμένα, αντί να τρέξει ο αλγόριθμος στο σύνολο των δεδομένων, τρέχει διαδοχικά σε τυχαία υποσύνολά τους (batches) και σε κάθε επανάληψη, τα κεντροειδή του αλγορίθμου ανανεώνονται με βάση την ελάττωση της παραγώγου (gradient descent). Το πλήθος των ακουστικών λέξεων (κεντροειδή του mini-batch k-means) σε κάθε χώρο χαρακτηριστικών ορίζεται ως k_1,k_2,k_3 αντίστοιχα και θα διερευνηθεί πώς οι μεταβλητές αυτές επηρεάζουν την απόδοση του ADSM μοντέλου στο πρόβλημα αξιολόγησης της σημασιολογικής ομοιότητας.

Στη συνέχεια, γίνεται υπολογισμός των bag-of-audio-words αναπαραστάσεων για τα ηχητικά αποσπάσματα με τις μεθόδους soft encoding και hard encoding που περιγράφονται στην Ενότητα 3.1.6. Τέλος, οι αναπαραστάσεις των tags προκύπτουν ως ο μέσος όρος (ανά σημείο) των bag-of-audio-words αναπαραστάσεων των αποσπασμάτων που σχετίζονται με αυτά. Οι τελικές αναπαραστάσεις για τα tags προκύπτουν εφαρμόζοντας (προαιρετικά) μείωση διαστασιμότητας με χρήση της PCA στον πίνακα που προκύπτει αν τοποθετηθούν (σε γραμμές) οι αναπαραστάσεις του συνόλου T των διαθέσιμων tags.

Εκτίμηση της απόδοσης του ADSM μοντέλου

Όπως αναφέρθηκε νωρίτερα, η ποιότητα των bag-of-audio-words αναπαραστάσεων που προκύπτουν από το ADSM μοντέλο θα εκτιμηθεί μέσω της αξιολόγησης της σημασιολογικής ομοιότητας λέξεων. Η ομοιότητα δύο λέξεων (tags) υπολογίζεται ως η ομοιότητα συνημιτόνου (cosine similarity) των αντίστοιχων διανυσματικών (bag-of-audio-words) αναπαραστάσεων.

Για την εκτίμηση της απόδοσης του ADSM μοντέλου γίνεται χρήση τεσσάρων διαφορετικών συνόλων δεδομένων που αποτελούνται από ζεύγη λέξεων και μία αριθμητική τιμή για κάθε ζεύγος, η οποία αντιπροσωπεύει το δείκτη ομοιότητας/σχετικότητας των αντίστοιχων

λέξεων. Το πρώτο σύνολο δεδομένων ονομάζεται MEN [1] και επικεντρώνεται στη σημασιολογική σχετικότητα (semantic relatedness) των λέξεων. Το δεύτερο ονομάζεται SimLex-999 (SIM) [132] και επικεντρώνεται στη σημασιολογική ομοιότητα (semantic similarity) των λέξεων. Οι αριθμητικές τιμές που περιλαμβάνονται στα MEN και Simlex-999 έχουν δοθεί από ανθρώπους. Το τρίτο σύνολο δεδομένων ονομάζεται CDSM και έχει προέλθει μέσω των προβλέψεων ενός state-of-the-art κατανεμημένου σημασιολογικού μοντέλου [21]. Το τέταρτο σύνολο δεδομένων ονομάζεται word2vec καθώς οι προβλέψεις έχουν προκύψει από το γνωστό σημασιολογικό μοντέλο word2vec [54, 55].

Για να χρησιμοποιηθεί ένα ζεύγος λέξεων των συνόλων δεδομένων για την αξιολόγηση του ADSM μοντέλου, είναι απαραίτητο κάθε μία από τις δύο λέξεις να περιλαμβάνεται στο σύνολο T των διαθέσιμων tags. Αγνοώντας όλα τα υπόλοιπα ζεύγη, το πλήθος των διαθέσιμων ζευγών ανά σύνολο δεδομένων αποτυπώνεται στον Πίνακα 4.2.

Σύνολο Δεδομένων	MEN	SimLex-999	CDSM	word2vec
Ζεύγη Λέξεων	157	44	1084	785

Πίνακας 4.2: Πλήθος Διαθέσιμων ζευγών λέξεων για κάθε σύνολο δεδομένων.

Παρόλο που στα σύνολα δεδομένων CDSM και word2vec οι αριθμητικοί δείκτες ομοιότητας έχουν προκύψει αυτόματα, χρησιμοποιούνται ως πραγματικές τιμές (groundtruth) διότι έχουν μεγάλη συσχέτιση με τις αξιολογήσεις ανθρώπων. Επίσης, σε αντίθεση με τα σύνολα δεδομένων MEN και SimLex-999, τα CDSM και word2vec περιλαμβάνουν αξιολογήσεις για σημαντικά περισσότερα ζεύγη λέξεων.

Για την εκτίμηση της απόδοσης του ADSM μοντέλου ως προς ένα σύνολο δεδομένων, αρχικά γίνεται προσδιορισμός της σημασιολογικής ομοιότητας για κάθε ζεύγος λέξεων του συνόλου. Έπειτα, υπολογίζεται ο βαθμός συσχέτισης των προβλέψεων με τις πραγματικές groundtruth τιμές που παρέχονται από το σύνολο δεδομένων. Εδώ, ο βαθμός συσχέτισης υπολογίζεται ως η απόλυτη τιμή του Spearman συντελεστή συσχέτισης. Η τιμή που προκύπτει κυμαίνεται από 0 έως 1. Η τιμή 0 υποδηλώνει ότι υπάρχει μηδενική συσχέτιση ανάμεσα στις προβλέψεις του ADSM μοντέλου και στις πραγματικές τιμές ενώ η τιμή 1 υποδηλώνει ότι υπάρχει πλήρης συσχέτιση. Επομένως όσο μεγαλύτερη είναι η τιμή του συντελεστή συσχέτισης, τόσο καλύτερη είναι η απόδοση του ADSM μοντέλου.

4.1.2 Πειραματικά Αποτελέσματα

Αρχικά, θα γίνει αναφορά στα πειραματικά αποτελέσματα που προκύπτουν από τη χρήση μόνο του χώρου χαρακτηριστικών S_1 (MFCC συντελεστές, 1οι και 2οι παράγωγοι) ενώ στη συνέχεια θα αναφερθούν αποτελέσματα και για τους άλλους χώρους χαρακτηριστικών καθώς και για το συνδυασμό τους μέσω της επέκτασης του ADSM μοντέλου (βλ. Ενότητα 3.1.8). Τα πειράματα θα γίνουν ως προς το μήκος του χρονικού παραθύρου για την εξαγωγή χαρακτηριστικών, το πλήθος των ακουστικών λέξεων, και τον αριθμό διαστάσεων κατά τη μείωση διαστασιμότητας. Τα αποτελέσματα θα σχολιάζονται για τα σύνολα δεδομένων CDSM και word2vec καθώς περιλαμβάνουν σημαντικά περισσότερα ζεύγη λέξεων από τα MEN και

SimLex-999 και ως αποτέλεσμα, η στατιστική σημαντικότητα (statistical significance) αυτών των αποτελεσμάτων είναι μεγαλύτερη από αυτά που αφορούν τα τελευταία δύο σύνολα δεδομένων.

Μήκος Χρονικού Παραθύρου

Έγινε πειραματισμός με διαφορετικές τιμές για το μήκος L και το βήμα H του χρονικού παραθύρου κατά την εξαγωγή χαρακτηριστικών ενώ οι άλλες παράμετροι διατηρούνται σταθερές. Οι τιμές για το μήκος κυμαίνονται από 25ms μέχρι 1000ms ενώ για το χρονικό βήμα κυμαίνονται από 10ms έως 400ms αναλογικά με το μήκος παραθύρου. Τα αποτελέσματα που αναφέρονται στον Πίνακα 4.3 έχουν προκύψει χρησιμοποιώντας $k_1 = 100$ ακουστικές λέξεις, ενώ δεν πραγματοποιήθηκε μείωση της διαστασιμότητας με χρήση της μεθόδου SVD καθώς το πλήθος των μειωμένων διαστάσεων θα διερευνηθεί ως ξεχωριστή παράμετρος του μοντέλου.

k_1	SVD	L (ms)	H (ms)	MEN	SimLex-999	CDSM	word2vec
100	-	25	10	0.397	0.327	0.321	0.264
		50	20	0.320	0.179	0.299	0.281
		100	40	0.373	**0.348**	0.319	0.279
		250	100	0.378	0.278	0.320	**0.291**
		500	200	**0.401**	0.286	0.307	0.280
		1000	400	0.385	0.282	**0.323**	0.254

Πίνακας 4.3: Απόδοση του ADSM μοντέλου (Spearman συντελεστής συσχέτισης) για διαφορετικές τιμές μήκους και βήματος του χρονικού παραθύρου κατά την εξαγωγή χαρακτηριστικών.

Παρατηρείται ότι η απόδοση του μοντέλου είναι σχετικά εύρωστη στο εύρος των διαφορετικών τιμών μήκους και βήματος χρονικού παραθύρου. Η μεγαλύτερη τιμή του συντελεστή συσχέτισης για το CDSM παρατρηρείται για $L = 1000ms$ χωρίς όμως να είναι σημαντικά μεγαλύτερη από τους συντελεστές συσχέτισης για διαφορετικές τιμές του μήκους παραθύρου. Για το σύνολο δεδομένων word2vec ο συντελεστής συσχέτισης λαμβάνει μέγιστη τιμή για $L = 250ms$. Αξίζει να σημειωθεί ότι όσο μεγαλώνει το μήκος και το βήμα του χρονικού παραθύρου, τόσο μειώνεται το πλήθος των διανυσμάτων χαρακτηριστικών που χρησιμοποιούνται για τη δημιουργία των ακουστικών λέξεων μέσω της ομαδοποίησής τους (clustering). Επομένως, μία καλή τακτική για τον υπολογισμό των ακουστικών λέξεων είναι να διαμορφώνεται το πλήθος των κεντροειδών του αλγορίθμου k-means (εδώ k_1) αντιστρόφως ανάλογα με το μήκος και το βήμα του παραθύρου.

Πλήθος Ακουστικών Λέξεων

Στη συνέχεια έγινε υπολογισμός της απόδοσης του ADSM μοντέλου για διαφορετικές τιμές πλήθους k_1 των ακουστικών λέξεων. Οι τιμές κυμαίνονται από 100 μέχρι 550. Το μήκος και το βήμα του χρονικού παραθύρου για την εξαγωγή χαρακτηριστικών ορίζονται ως $L = 25ms$ και $H = 10ms$ αντίστοιχα. Παρατηρείται ότι σε γενικές γραμμές, μεγαλύτερο πλήθος ακουστικών λέξεων είναι προτιμητέο για το συγκεκριμένο μήκος και βήμα χρονικού παραθύρου

(τα οποία λαμβάνουν σχετικά μικρές τιμές), γεγονός που επιβεβαιώνει την παρατήρηση στην προηγούμενη υποενότητα. Η καλύτερη απόδοση επιτυγχάνεται για πλήθος ακουστικών λέξεων $k = 500$ τόσο για το σύνολο CDSM όσο και για το word2vec.

k_1	SVD	L (ms)	H (ms)	MEN	SimLex-999	CDSM	word2vec
100				0.397	0.327	0.321	0.264
200				0.376	**0.367**	0.356	0.320
250				0.399	0.269	0.334	0.293
300				**0.432**	0.355	0.365	0.311
350	-	25	10	0.409	0.329	0.342	0.306
400				0.403	0.334	0.360	0.320
450				0.417	0.31	0.361	0.323
500				0.398	0.285	**0.373**	**0.333**
550				0.214	0.197	0.365	0.331

Πίνακας 4.4: Απόδοση του ADSM μοντέλου (Spearman συντελεστής συσχέτισης) για διαφορετικές τιμές πλήθους ακουστικών λέξεων (κεντροειδή του αλγορίθμου k-means).

Πλήθος SVD Διαστάσεων

Μία επιπλέον παράμετρος του μοντέλου είναι το πλήθος των διαστάσεων μετά από μείωση της διαστασιμότητας των bag-of-audio-words αναπαραστάσεων με τη μέθοδο PCA (σε συνδυασμό με τη μέθοδο SVD - βλ. Ενότητα 2.3). Η απόδοση του ADSM μοντέλου υπολογίστηκε για πλήθος διαστάσεων που κυμαίνεται από 10 έως 260 διαστάσεις. Το μήκος και το βήμα του χρονικού παραθύρου ορίστηκαν ως $L = 25ms$, $H = 10ms$ αντίστοιχα και το πλήθος των ακουστικών λέξεων ορίστκηκε ως $k_1 = 300$.

k_1	SVD	L (ms)	H (ms)	MEN	SimLex-999	CDSM	word2vec
300	10	25	10	0.300	0.329	0.364	**0.330**
	50			0.409	0.338	0.372	0.326
	90			0.435	0.332	**0.375**	0.313
	130			0.432	0.350	0.374	0.318
	170			**0.437**	**0.369**	0.371	0.315
	210			0.434	0.351	0.370	0.316
	260			0.432	0.353	0.368	0.316
	-			0.432	0.355	0.365	0.311

Πίνακας 4.5: Απόδοση του ADSM μοντέλου (Spearman συντελεστής συσχέτισης) για διαφορετικό πλήθος διαστάσεων μετά από μείωση της διαστασιμότητας.

Σύμφωνα με τον Πίνακα 4.5, η απόδοση του ADSM μοντέλου βελτιώνεται σημαντικά μετά από τη μείωση διαστασιμότητας. Πιο συγκεκριμένα, για το σύνολο δεδομένων CDSM επιτυγχάνεται μέγιστος συντελεστής συσχέτισης 0.375 μετά από μείωση στις 90 διαστάσεις

ενώ για το word2vec επιτυγχάνεται μέγιστος συντελεστής συσχέτισης 0.330 μετά από μείωση στις 10 διαστάσεις. Η τελευταία γραμμή του Πίνακα 4.5 αντιστοιχεί στην απόδοση του ADSM μοντέλου χωρίς μείωση της διαστασιμότητας.

Αποτελέσματα Προγενέστερων Δημοσιεύσεων

Αναφορά στην απόδοση του παραδοσιακού ADSM μοντέλου έχει γίνει σε δύο προγενέστερες δημοσιεύσεις [3, 4] τα αποτελέσματα των οποίων συνοψίζονται στον Πίνακα 4.6.

k_1	SVD	L (ms)	H (ms)	MEN	SimLex-999	CDSM	word2vec
100	60	250	100	0.402	0.233	n/a	n/a
300	-	250	100	0.325	0.161	n/a	n/a

Πίνακας 4.6: Απόδοση του παραδοσιακού ADSM μοντέλου (Spearman συντελεστής συσχέτισης) όπως αυτή αναφέρεται στις δημοσιεύσεις [3] (πρώτη γραμμή) και [4] (δεύτερη γραμμή).

Επέκταση του ADSM μοντέλου

Στην Ενότητα 3.1.8 έγινε πρόταση μίας μεθόδου για την επέκταση του ADSM μοντέλου με χρήση πολλαπλών χώρων χαρακτηριστικών και τη σύμπτυξή τους με βάρη ανάλογα με τη 'φύση' του ήχου. Εδώ, έγινε δημιουργία των ακόλουθων χώρων χαρακτηριστικών:

- S_1: 13 συντελεστές MFCC (όπου ο πρώτος συντελεστής έχει αντικατασταθεί από τη φασματική ενέργεια του σήματος), οι πρώτες και δεύτεροι παράγωγοι (39 χαρακτηριστικά συνολικά).
- S_2: Η θεμελιώδης συχνότητα (F0) του σήματος.
- S_3: χρωμόγραμμα (chromagram), φασματική ροή (spectral flux), ποσοστό σημείων μηδενισμού (zero-crossing-rate), φασματικό κεντροειδές (spectral centroid), ακουστική ευκρίνεια (brightness), φασματική διάδοση (spectral spread), φασματική ασυμμετρία (spectral skewness), φασματική κύρτωση (spectral kurtosis), συχνότητα κάτω από την οποία βρίσκεται το 85% των συχνοτήτων (roll-off with 85% threshold), συχνότητα κάτω από την οποία βρισκεται το 95% των συχνοτήτων (roll-off with 95% threshold), φασματική εντροπία (spectral entropy), φασματική ομαλότητα (spectral flatness), τραχύτητα (roughness), ασυμμετρία (irregulatiry) και μη-αρμονικότητα (inharmonicity). Συνολικά παράγονται 27 χαρακτηριστικά.

Περισσότερες λεπτομέρειες για κάθε ένα από τα παραπάνω ακουστικά χαρακτηριστικά δίνονται στην Ενότητα 3.1.2. Τα χαρακτηριστικά S_1,S_2 εξήχθησαν με χρήση της βιβλιοθήκης Librosa [134] ενώ τα S_3 εξήχθησαν με χρήση του εργαλείου MIR Toolbox [135].

Υπενθυμίζεται ότι τα βάρη για τη σύμπτυξη των χώρων χαρακτηριστικών υπολογίζονται ανάλογα με την ταξινόμηση των ηχητικών αποσπασμάτων στις κλάσεις 'μουσική', 'φωνή', 'απόσπασμα γενικού τύπου'. Στα πλαίσια αυτής της διπλωματικής έγινε χρήση των Μηχανών Διανυσματικής Υποστήριξης (Support Vector Machines - SVM) με γραμμικό πυρήνα,

Κλάση	u_1	u_2	u_3
Μουσική	0.3	0.2	0.5
Φωνή	0.8	0.2	0.0
Γενικού Τύπου Απόσπασμα	0.3	0.0	0.7

Πίνακας 4.7: Βάρη για τη σύμπτυξη των τριών χώρων χαρακτηριστικών S_1, S_2 και S_3. Διαφορετικός συνδυασμός βαρών χρησιμοποιείται ανάλογα με την ταξινόμηση ενός αποσπάσματος στις κλάσεις 'μουσική', 'φωνή', 'γενικού τύπου απόσπασμα'.

χρησιμοποιώντας τη βιβλιοθηκη pyAudioAnalysis [136]. Περισσότερες λεπτομέρειες για τον ταξινομητή και τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευσή του μπορούν να βρεθούν στην αντίστοιχη δημοσίευση [137]. Σημειώνεται ότι τα δεδομένα αυτά δεν περιλαμβάνουν κάποιο από τα 4477 ηχητικά αποσπάσματα που περιγράφονται στην Ενότητα 4.1.1. Αφού τα δεδομένα διαχωρίστηκαν σε 10 ανεξάρτητες ομάδες, ο υπολογισμός των βαρών για τη σύμπτυξη των χώρων χαρακτηριστικών πραγματοποιήθηκε σε κάθε ομάδα ξεχωριστά μέσω της εξαντλητικής αναζήτησης στο σύνολο όλων των πιθανών συνδυασμών βαρών (με ακρίβεια ενός δεκαδικού ψηφίου) που αθροίζονται στην μονάδα. Έτσι, τα βάρη u_1,u_2,u_3 που χρησιμοποιήθηκαν για τη σύμπτυξη των χώρων χαρακτηριστικών S_1,S_2,S_3 αντίστοιχα περιλαμβάνονται στον Πίνακα 4.7.

Χώρος Χαρακτηριστικών	k	SVD	MEN	SimLex-999	CDSM	word2vec
S_1	300	-	0.416	0.235	0.333	0.332
S_2			0.308	0.313	0.269	0.248
S_3			0.418	0.205	0.278	0.315
S_{123}			**0.468**	**0.387**	**0.388**	**0.382**
S_1		90	0.436	0.209	0.283	0.320
S_2			0.302	0.34	0.275	0.26
S_3			0.422	0.252	0.343	0.337
S_{123}			**0.480**	**0.374**	**0.402**	**0.401**
S_1	400	-	0.457	0.24	0.298	0.309
S_2			0.304	0.334	0.283	0.259
S_3			0.423	0.300	0.384	0.343
S_{123}			**0.462**	**0.437**	**0.404**	**0.379**
S_1		90	0.427	0.317	0.375	0.331
S_2			0.314	0.351	0.278	0.254
S_3			0.46	0.225	0.293	0.302
S_{123}			**0.477**	**0.407**	**0.416**	**0.407**

Πίνακας 4.8: Απόδοση του ADSM μοντέλου (Spearman συντελεστής συσχέτισης) για κάθε χώρο χαρακτηριστικών ξεχωριστά (S_1,S_2,S_3) καθώς και της επέκτασης του ADSM μοντέλου μέσω της σύμπτυξης των τριών αναπαραστάσεων (S_{123}).

Στον Πίνακα 4.8, γίνεται αναφορά στην απόδοση του ADSM μοντέλου με ανεξάρτητη χρήση των τριών χώρων χαρακτηριστικών (S_1,S_2,S_3) αλλά και με τη σύμπτυξη των τριών χώρων (S_{123}). Το πλήθος των ακουστικών λέξεων για κάθε χώρο ορίστηκε ως $k_1 = k_2 = k_3 = k = 300$. Το μήκος και βήμα του χρονικού παραθύρου για την εξαγωγή χαρακτηριστικών ορίστηκαν $L = 250ms$ και $H = 100ms$ αντίστοιχα.

Παρατηρείται ότι με χρήση της μεθόδου για τη σύμπτυξη των αναπαραστάσεων, ο συντελεστής συσχέτισης λαμβάνει σημαντικά μεγαλύτερη τιμή σε σχέση με το παραδοσιακό ADSM μοντέλο. Για παράδειγμα, στην περίπτωση όπου το πλήθος ακουστικών λέξεων είναι $k = 300$ και δεν πραγματοποιηθεί μείωση της διαστασιμότητας (πρώτες τέσσερις γραμμές του Πίνακα 4.8), με τη σύμπτυξη των αναπαραστάσεων επιτυγχάνεται σχετική βελτίωση της απόδοσης κατά 12% για το MEN, 23.6% για το SimLex-999, 16.5% για το CDSM και 15.1% για το word2vec. Όσον αφορά τη χρήση των τριών χώρων ξεχωριστά, δεν υπάρχει ξεκάθαρος 'νικητής', καθώς φαίνεται ότι επιτυγχάνεται παρόμοια κατά μέσο όρο απόδοση τόσο με χρήση του χώρου S_1 όσο και με χρήση του S_2. Η κατά μέσο όρο απόδοση του επεκτεταμένου μοντέλου για τα σύνολο δεδομένων CDSM και word2vec είναι 0.412 (για $k = 900$ και με μείωση διαστασιμότητας στις 90 διαστάσεις).

Πολυτροπική Σύμπτυξη

Έως εδώ, έχει γίνει αξιολόγηση της σημασιολογικής ομοιότητας λέξεων με χρήση του ADSM μοντέλου. Με τη μέθοδο της πολυτροπικής σύμπτυξης (multimodal fusion, βλ. Ενότητα 3.3), επιτυγχάνεται η σύμπτυξη των αναπαραστάσεων από διαφορετικά μοντέλα σε μία κοινή αναπαράσταση. Στη συνέχεια, θα γίνει σύμπτυξη των αναπαραστάσεων λέξεων που προκύπτουν από το σημασιολογικό (DSM), ακουστικό-σημασιολογικό (ADSM) και οπτικό-σημασιολογικό (VDSM) μοντέλο και οι προκύπτουσες αναπαραστάσεις θα χρησιμοποιηθούν για την αξιολόγηση της σημασιολογικής ομοιότητας λέξεων.

Αναμένεται ότι η πληροφορία που παρέχεται από το ακουστικό-σημασιολογικό (ADSM) μοντέλο θα είναι χρήσιμη κυρίως στις περιπτώσεις λέξεων με ακουστικές ιδιότητες, όπως για παράδειγμα η λέξη 'βιολί'. Αντίθετα, σε περιπτώσεις λέξεων χωρίς ακουστικές ιδιότητες όπως για παράδειγμα η λέξη 'δημοκρατία', δεν είναι τόσο χρήσιμη η πληροφορία από το ADSM μοντέλο όσο η πληροφορία από το παραδοσιακό DSM μοντέλο. Για τη σύγκριση της σχετικής επίδρασης κάθε μοντέλου στις αντίστοιχες κατηγορίες λέξεων προστίθενται τρία νέα σύνολα δεδομένων. Το πρώτο σύνολο δεδομένων ονομάζεται AMEN και είναι το υποσύνολο του συνόλου δεδομένων MEN, όπου έχουν διατηρηθεί όλες οι λέξεις που έχουν ηχητικές ιδιότητες. Το δεύτερο σύνολο δεδομένων ονομάζεται TMEN και είναι το συμπληρωματικό του συνόλου AMEN ως προς το MEN, δηλαδή τις λέξεις που δεν έχουν ηχητικές ιδιότητες. Το τρίτο σύνολο δεδομένων ονομάζεται ASLex και είναι υποσύνολο του SimLex-999, όπου έχουν διατηρηθεί όλες οι λέξεις που έχουν ηχητικές ιδιότητες. Τα σύνολα AMEN και ASLex παρέχονται σε προηγούμενη δημοσίευση [4] ενώ το TMEN κατασκευάζεται εύκολα γνωρίζοντας τα MEN και AMEN. Βέβαια, το TMEN περιλαμβάνει σημαντικά περισσότερο πλήθος λέξεων σε σύγκριση με το AMEN. Για να είναι συγκρίσιμα τα αποτελέσματα μεταξύ των AMEN και

Μοντέλο	**MEN**	**AMEN**	**TMEN**	**SimLex-999**	**ASLex**
ADSM	0.433	0.554	0.532	0.352	0.292
DSM	0.774	0.762	**0.812**	0.427	0.398
VDSM	0.233	0.435	0.181	0.248	0.269
ADSM&DSM	**0.783**	0.815	0.759	0.475	0.424
ADSM&VDSM	0.470	0.632	0.438	0.401	0.348
DSM&VDSM	0.762	0.814	0.772	0.481	**0.497**
ADSM&DSM&VDSM	0.776	**0.827**	0.798	**0.502**	0.476
Ζεύγη Λέξεων	1533	141	135	207	100

Πίνακας 4.9: Απόδοση (Spearman συντελεστής συσχέτισης) της μεθόδου Early Fusion για την πολυτροπική σύμπτυξη των αναπαραστάσεων.

TMEN πραγματοποιείται τυχαία επιλογή λέξεων από το TMEN ώστε να έχει το ίδιο μέγεθος με το TMEN. Καθώς οι λέξεις επιλέγονται τυχαία, η επιλογή και η αξιολόγηση των μοντέλων ως προς το TMEN πραγματοποιείται 10 φορές και η τελική τιμή απόδοσης προκύπτει ως ο μέσος όρος των 10 επιμέρους τιμών.

Για το ADSM μοντέλο έγινε επιλογή των παραμέτρων $k = 300$, $L = 250$, $H = 100$. Ως ακουστικά χαρακτηριστικά χρησιμοποιήθηκαν τα χαρακτηριστικά του συνόλου S_1 και για τη δημιουργία των bag-of-audio-words αναπαραστάσεων έγινε χρήση της μεθόδου hard encoding. Ως DSM μοντέλο επιλέχθηκε το μοντέλο CDSM [21] με αναπαραστάσεις 300 διαστάσεων. Για την εκπαίδευση του VDSM μοντέλου έγινε χρήση του συνόλου εικόνων ESP-Game dataset[1], το οποίο περιλαμβάνει 100.000 εικόνες κάθε μία από τις οποίες συνοδεύεται από ένα ή περισσότερα tags. Τα οπτικά χαρακτηριστικά δημιουργήθηκαν με εφαρμογή του αλγορίθμου SIFT στον χώρο HSV οδηγώντας σε bag-of-visual-words αναπαραστάσεις 300 διαστάσεων.

Καθώς τα DSM, ADSM και VDSM μοντέλα κατασκευάζονται με χρήση διαφορετικών συνόλων δεδομένων προκύπτουν αναπαραστάσεις διαφορετικών συνόλων λέξεων για κάθε μοντέλο. Επομένως, για να γίνει η πολυτροπική σύμπτυξη των αναπαραστάσεων, διατηρείται η τομή των συνόλων λέξεων, δηλαδή 1613 λέξεις για τις οποίες υπάρχουν διαθέσιμες σημασιολογικές αναπαραστάσεις και από τα τρία μοντέλα. Έπειτα, πραγματοποιείται σύμπτυξη των αναπαραστάσεων λέξεων με τις μεθόδους Early Fusion και Late Fusion (βλ. Ενότητα 3.3). Μετά την εφαρμογή της μεθόδου Early Fusion πραγματοποιείται μείωση της διαστασιμότητας των τελικών αναπαραστάσεων στις 300 διαστάσεις για δύο λόγους. Πρώτον, για να είναι δίκαιη η αξιολόγηση όλων των μοντέλων ως προς το πλήθος των διαστάσεων και δεύτερον για να γίνει 'ανάμειξη' της πληροφορίας από διαφορετικά μοντέλα και να βρεθούν οι 'κρυφές' κοινές διαστάσεις μεταξύ των DSM, ADSM και VDSM μοντέλων.

Στον Πίνακα 4.9 παρουσιάζεται η απόδοση της πολυτροπικής σύμπτυξης με τη μέθοδο Early Fusion. Παρατηρείται ότι για όλα τα σύνολα αξιολόγησης, με εξαίρεση το TMEN,

[1] http://www.espgame.org

Μοντέλο	MEN	AMEN	TMEN	SimLex-999	ASLex
ADSM	0.433	0.554	0.532	0.352	0.292
DSM	**0.774**	0.762	**0.812**	0.427	0.398
VDSM	0.233	0.435	0.181	0.248	0.269
ADSM&DSM	0.741	0.719	0.718	0.406	0.317
ADSM&VDSM	0.474	0.635	0.428	0.405	0.340
DSM&VDSM	0.762	**0.814**	0.737	**0.478**	**0.492**
ADSM&DSM&VDSM	0.459	0.639	0.308	0.403	0.345
Ζεύγη Λέξεων	1533	141	135	207	100

Πίνακας 4.10: Απόδοση (Spearman συντελεστής συσχέτισης) της μεθόδου Late Fusion για την πολυτροπική σύμπτυξη των αναπαραστάσεων.

μέσω της πολυτροπικής σύμπτυξης επιτυγχάνεται καλύτερη απόδοση σε σύγκριση με την απόδοση κάθε μοντέλου ξεχωριστά. Για το TMEN, είναι λογικό να μη βελτιώνεται η απόδοση με χρήση της πολυτροπικής σύμπτυξης, καθώς περιλαμβάνει λέξεις με λεξιλογικές και όχι ηχητικές ή οπτικές ιδιότητες. Επίσης, η σχετική βελτίωση της απόδοσης είναι μεγαλύτερη στην περίπτωση του συνόλου αξιολόγησης AMEN σε σύγκριση με το MEN, το οποίο ήταν αναμενόμενο καθώς οι αναπαραστάσεις των μοντέλων ADSM και VDSM παρέχουν χρήσιμη πληροφορία κυρίως για τις λέξεις που έχουν ακουστικές και οπτικές ιδιότητες αντίστοιχα. Μεγαλύτερο μέρος τέτοιων λέξεων παρατηρείται στο σύνολο AMEN σε σύγκριση με το MEN. Στον Πίνακα 4.10 παρουσιάζεται η απόδοση της πολυτροπικής σύμπτυξης με τη μέθοδο Late Fusion. Συγκρίνοντας τους δύο πίνακες, παρατηρείται ότι σε όλα τα σύνολα αξιολόγησης, η απόδοση της μεθόδου Early Fusion είναι σημαντικά καλύτερη από την απόδοση της μεθόδου Late Fusion για τη σύμπτυξη των αναπαραστάσεων.

4.1.3 Συμπεράσματα

Σε αυτή την ενότητα έγινε εφαρμογή του ADSM μοντέλου στην αξιολόγηση της σημασιολογικής ομοιότητας των λέξεων με βάση τις ακουστικές τους ιδιότητες. Επιπλέον, έγινε διερεύνηση της επίδρασης διαφορετικών παραμέτρων στην απόδοση του ADSM μοντέλου καθώς και πειραματισμός με την επέκταση του μοντέλου με σκοπό τη σύμπτυξη διαφορετικών αναπαραστάσεων. Τα βάρη για τη σύμπτυξη των αναπαραστάσεων ορίστηκαν μέσω της ταξινόμησης καθενός από τα αποσπάσματα σε μία από τις κλάσεις 'μουσική', 'φωνή', 'γενικού τύπου απόσπασμα'. Διαπιστώθηκε ότι η μέθοδος αυτή οδήγησε σε σχετική βελτίωση της απόδοσης του μοντέλου (συντελεστής συσχέτισης ως προς αξιολογήσεις ανθρώπων) μέχρι και 23.6%. Επίσης, βρέθηκε ότι η μείωση της διαστασιμότητας των bag-of-audio-words αναπαραστάσεων επιφέρει βελτίωση της απόδοσης, τόσο του παραδοσιακού ADSM μοντέλου όσο και της επέκτασής του. Ο αλγόριθμος και ορισμένα από τα αποτελέσματα αυτής της ενότητας δημοσιεύτηκαν στο συνέδριο Interspeech το Σεπτέμβριο του 2016 [138].

4.2 Αυτόματος Χαρακτηρισμός Ηχητικών Αποσπασμάτων

Τα μεταδεδομένα που συχνά συνοδεύουν ηχητικά αποσπάσματα έχουν αξιοποιηθεί με επιτυχία για τη βελτίωση της απόδοσης σε πολλές εφαρμογές, ειδικά στον τομέα Εξαγωγής Πληροφορίας από τη Μουσική (Music Information Retrieval) [90, 91, 139]. Συνήθως τα μεταδεδομένα ηχητικών αποσπασμάτων συναντώνται σε δύο μορφές: ελεύθερο κείμενο, για παράδειγμα κείμενο μίας ιστοσελίδας όπου σχολιάζεται το απόσπασμα, ή λέξεις (tags) που περιγράφουν το περιεχόμενο του αποσπάσματος και έχουν προστεθεί είτε από το δημιουργό του αποσπάσματος είτε από άλλους χρήστες. Τα tags προτιμώνται έναντι του ελεύθερου κείμενου καθώς παρέχουν μία άμεση περιγραφή του αποσπάσματος ενώ το κείμενο συνήθως χαρακτηρίζεται ως 'θορυβώδες' διότι μόνο ένα μέρος του κειμένου είναι άμεσα συνδεδεμένο με το περιεχόμενο του αποσπάσματος. Επιπλέον, τα tags έχουν αποδειχτεί ιδιαίτερα χρήσιμα καθώς χρησιμοποιούνται σε μεγάλο βαθμό για την ηλεκτρονική ταξινόμηση των μουσικών κομματιών ανά στυλ, καλλιτέχνη, μουσικό όργανο κλπ. Βέβαια, δε συνοδεύονται πάντα τα ηχητικά αποσπάσματα από tags. Επομένως, έχει προκύψει η ανάγκη για τον αυτόματο χαρακτηρισμό των ηχητικών αποσπασμάτων (auto-tagging).

Στη βιβλιογραφία έχουν προταθεί πολλές διαφορετικές μέθοδοι για τον αυτόματο χαρακτηρισμό των ηχητικών αποσπασμάτων. Για παράδειγμα, στην περίπτωση μουσικών κομματιών, εφόσον είναι γνωστοί οι καλλιτέχνες τους, έχει γίνει αξιοποίηση της ομοιότητας μεταξύ των καλλιτεχνών για την πρόβλεψη των πιο αντιπροσωπευτικών tags [140]. Επίσης, η πρόβλεψη των tags έχει γίνει με χρήση γλωσσικών μοντέλων ώστε να ληφθεί υπόψη η σημασιολογική σχέση μεταξύ των tags [141]. Εφόσον είναι πιθανό να μην είναι διαθέσιμα τα μεταδεδομένα των ηχητικών αποσπασμάτων, συνήθως ο αυτόματος χαρακτηρισμός γίνεται μέσω της ανάλυσης του ακουστικού τους περιεχομένου. Για παράδειγμα, σε προγενέστερη εφαρμογή [142, 143] κάθε μουσικό απόσπασμα αναπαρίσταται ως πολυτροπική σημασιολογική κατανομή ως προς ένα λεξιλόγιο από tags.

Το ADSM μοντέλο (βλ. Ενότητα 3.1) επιτρέπει τον υπολογισμό αναπαραστάσεων ήχων και λέξεων ως διανύσματα στον ακουστικό-σημασιολογικό χώρο. Το γεγονός ότι τόσο ηχητικά αποσπάσματα όσο και λέξεις λαμβάνουν bag-of-audio-words αναπαραστάσεις, επιτρέπει την εφαρμογή του ADSM μοντέλου στο πρόβλημα του αυτόματου χαρακτηρισμού ηχητικών αποσπασμάτων (audio autotagging). Στη συνέχεια θα γίνει περιγραφή του αλγορίθμου καθώς και των δεδομένων για τον αυτόματο χαρακτηρισμό ηχητικών αποσπασμάτων με χρήση του ADSM μοντέλου.

4.2.1 Περιγραφή Αλγορίθμου/Δεδομένων

Δεδομένα και Εξαγωγή Χαρακτηριστικών

Για το σκοπό αυτής της εφαρμογής έγινε χρήση του συνόλου δεδομένων MagnaTagATune. Το MagnaTagATune αποτελείται από 25,863 ηχητικά αποσπάσματα των 30 δευτερολέπτων τα περισσότερα πό τα οποία συνοδεύονται από tags. Συνολικά, υπάρχουν 188 μοναδικά tags. Τα μεγαλύτερο μέρος των αποσπασμάτων είναι μουσικά κομμάτια, για τα οποία υπάρχουν

δικαιώματα διαμοιρασμού καθώς είναι μέρος της συλλογής Magnatune[2].

Ως μέρος της προετοιμασίας, όλα τα αποσπάσματα μετατράπηκαν στη μορφή WAV με ρυθμό δειγματοληψίας 22.05kHz. Το μέγεθος και το βήμα του χρονικού παραθύρου για την εξαγωγή χαρακτηριστικών ορίστηκαν ως $L = 250ms$ και $H = 100ms$ αντίστοιχα. Για τη δημιουργία του ADSM μοντέλου χρησιμοποιήθηκαν δύο διαφορετικοί τύποι διανυσμάτων χαρακτηριστικών.

Ο πρώτος τύπος (χαρακτηριστικά EchoNest) περιλαμβάνει 24 χαρακτηριστικά, 12 εκ των οποίων είναι τα χαρακτηριστικά Χρωμογράμματος (chromagram features) και όπως έχει αναφερθεί στην Ενότητα 3.1.2 συμβολίζουν την σχετική επικράτεια καθενός από τους 12 χρωματικούς μουσικούς τόνους στο αντίστοιχο μέρος του αποσπάσματος. Τα 12 χαρακτηριστικά που απομένουν από τα 24 EchoNest χαρακτηριστικά αντιστοιχούν στους συντελεστές 12 συναρτήσεων βάσης οι οποίες αναπαριστούν την υφή του ήχου και είναι παρόμοια με τους συντελεστές MFCC. Η εξαγωγή των EchoNest χαρακτηριστικών έγινε με χρήση του EchoNest API 1.0[3]. Ο δεύτερος τύπος χαρακτηριστικών (χαρακτηριστικά MFCCdd) περιλαμβάνει 39 συντελεστές, εκ των οποίων οι πρώτοι 13 είναι οι συντελεστές MFCC (βλ. Ενότητα 3.1.2) οι επόμενοι 13 είναι οι πρώτες παράγωγοι και οι τελευταίοι 13 είναι οι δεύτερες παράγωγοι των MFCCs. Και στις δύο περιπτώσεις (Echonest και MFCCdd), τα διανύσματα χαρακτηριστικών όλων των αποσπασμάτων έχουν κανονικοποιηθεί ώστε να λαμβάνουν μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση.

Αλγόριθμος για Αυτόματο Χαρακτηρισμό των Αποσπασμάτων

Για τον αυτόματο χαρακτηρισμό των αποσπασμάτων του συνόλου δεδομένων Magnatagatune θα γίνει χρήση του ADSM μοντέλου. Ακολουθώντας τη διαδικασία που περιγράφεται στην Ενότητα 3.1.6 τα ηχητικά αποσπάσματα αναπαριστώνται ως ένα ιστόγραμμα ακουστικών λέξεων:

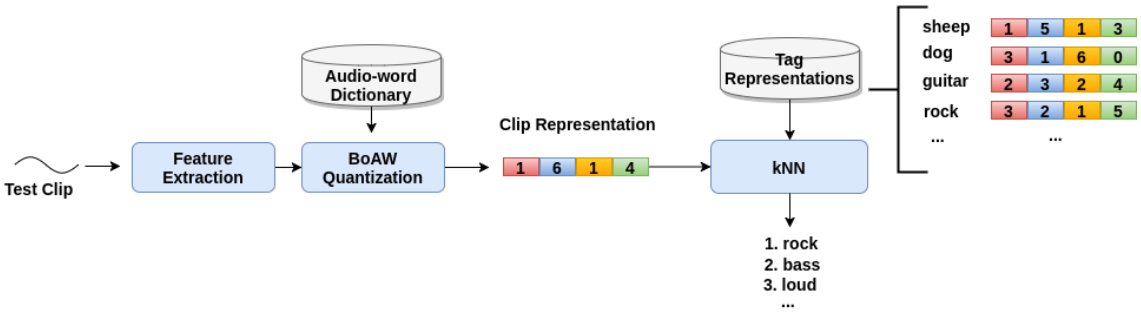
$$r_c = (w_1, w_2, ..., w_k), \tag{4.1}$$

όπου $w_i \in [0, 1]$ και $\sum_{i=1}^{k} w_i = 1$. Το βάρος w_i υποδηλώνει τη συνεισφορά της i-οστής ακουστικής λέξης στην αναπαράσταση του αποσπάσματος. Υπενθυμίζεται ότι τέτοιου τύπου αναπαραστάσεις αναφέρονται ως bag-of-audio-words αναπαραστάσεις και μπορούν να προκύψουν με μεθόδους όπως η hard encoding (βλ. Ενότητα 3.1.6), η soft encoding (βλ. Ενότητα 3.1.6) ή μέσω της σύμπτυξης πολλαπλών χώρων χαρακτηριστικών (βλ. Ενότητα 3.1.8). Εδώ, έγινε χρήση της μεθόδου hard encoding με πλήθος ακουστικών λέξεων $k = 300$.

Η bag-of-audio-words αναπαράσταση r_j ενός tag j υπολογίζεται όπως περιγράφεται στην Ενότητα 3.1.7 ως ο μέσος όρος (ανά σημείο) των bag-of-audio-words αναπαραστάσεων των ηχητικών αποσπασμάτων που σχετίζονται με το tag j.

Έχοντας υπολογίσει τις bag-of-audio-words αναπαραστάσεις για τα 188 tags, ο αυτόματος χαρακτηρισμός ενός ηχητικού αποσπάσματος c μπορεί να πραγματοποιηθεί με άμεσο τρόπο. Αρχικά, υπολογίζεται η bag-of-audio-words αναπαράσταση r_c του αποσπάσματος c. Στη

[2]http://magnatune.com/
[3]http://developer.echonest.com/



Σχήμα 4.2: Απεικόνιση του αλγορίθμου auto-tagging με χρήση του μοντέλου ADSM.

συνέχεια, υπολογίζεται η ομοιότητα συνημιτόνου μεταξύ της αναπαράστασης r_c και της αναπαράστασης r_j καθενός από τα 188 tags:

$$s_{cj} = \frac{r_c \cdot r_j}{|r_c| \cdot |r_j|} \tag{4.2}$$

Τα N tags που περιγράφουν το ηχητικό απόσπασμα είναι αυτά που αντιστοιχούν στις N μεγαλύτερες τιμές ομοιότητας s_{cj} με το απόσπασμα.

4.2.2 Πειραματικά Αποτελέσματα

Παραδείγματα και οπτικοποίηση

Όπως αναφέρθηκε προηγουμένως, τα περισσότερα από τα 25,863 ηχητικά αποσπάσματα του συνόλου MagnaTagATune συνοδεύονται από tags που έχουν οριστεί από ανθρώπους. Έστω το ηχητικό απόσπασμα c. Για την αυτόματη πρόβλεψη των N πιθανότερων tags γίνεται δημιουργία του ADSM μοντέλου αγνοώντας το απόσπασμα c. Στη συνέχεια υπολογίζεται η bag-of-audio-words αναπαράσταση r_c του αποσπάσματος c και τα N πιθανότερα tags προκύπτουν ως τα N κοντινότερα tags με χρήση της ομοιότητας συνημιτόνου (Εξίσωση 4.2). Η ίδια διαδικασία επαναλαμβάνεται για κάθε ένα από τα 25,863 ηχητικά αποσπάσματα.

Μερικά παραδείγματα εφαρμογής του παραπάνω αλγορίθμου σε αποσπάσματα του MagnaTagATune περιλαμβάνονται στον Πίνακα 4.11. Παρατηρείται ότι πολλά από τα tags περιλαμβάνονται και στα πραγματικά αλλά και στα προβλεπόμενα από τον αλγόριθμο tags. Επίσης πολλά προβλεπόμενα tags τα οποία δεν εμφανίζονται ως πραγματικά tags έχουν παρόμοια έννοια με τα πραγματικά tags. Για παράδειγμα, στο απόσπασμα με ID '48010' το προβλεπόμενο tag "quiet" έχει παρόμοια έννοια με το πραγματικό tag "silence". Ακόμη, αντιπροσωπευτικά tags προβλέπονται και για αποσπάσματα τα οποία δεν συνοδεύονται από πραγματικά tags (π.χ. απόσπασμα με ID '19920').

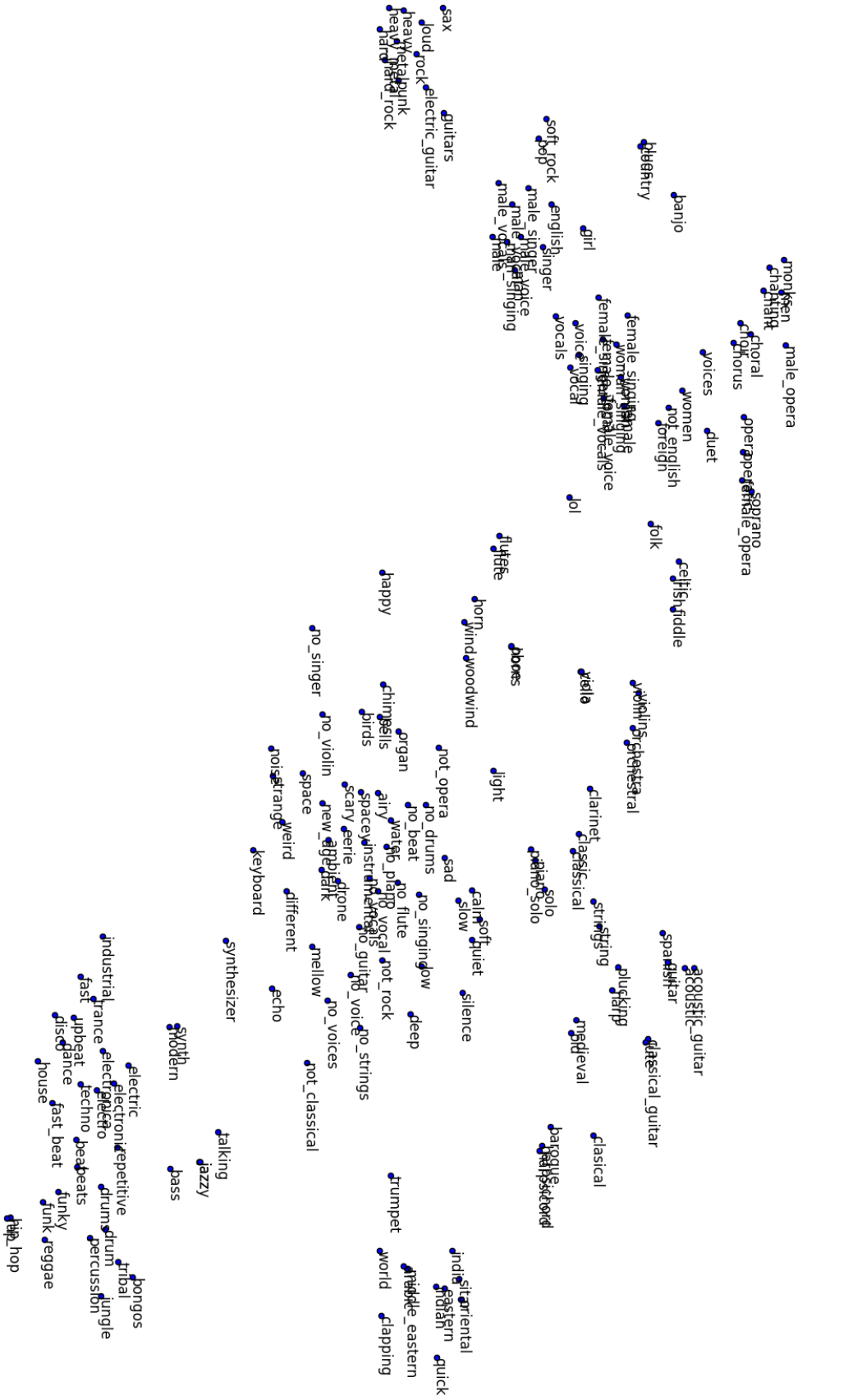
Ενδιαφέρον παρουσιάζει επίσης η οπτικοποίηση των bag-of-audio-words αναπαραστάσεων των tags. Για να πραγματοποιηθεί οπτικοποίηση των αναπαραστάσεων είναι απαραίτητο να γίνει μείωση της διαστασιμότητας. Εδώ, αντί της PCA έγινε εφαρμογή της μεθόδου t-Distributed Stochastic Neighbor Embedding (t-SNE) [144]. Στόχος της μεθόδου είναι η μείωση της διαστασιμότητας με τέτοιο τρόπο ώστε να διατηρηθεί όσο το δυνατόν περισσότερο η έννοια της ομοιότητας μεταξύ των δεδομένων. Γι' αυτό το σκοπό, οι τιμές ομοιότητας

ID	Πραγματικά Tags	Προβλεπόμενα Tags (N=5)
3843	**indian**, **sitar**	**sitar**, **indian**, eastern, india, oriental
9531	rock, **heavy**, **heavy metal**, **loud**, fast, hard rock, **metal**	hard, **loud**, **heavy**, **heavy metal**, **metal**
13526	bass, **drums**, drum, **funky**, **reggae**	**funky**, beat, **drums**, **reggae**, funk
15380	**classical**, **solo**, **cello**, **violin**, strings	**cello**, viola, **violin**, **solo**, **classical**
19920	-	orchestra, violins, flutes, fiddle, violin
21725	choir, **choral**, **men**, man	monks, chant, chanting, **men**, **choral**
29231	**acoustic**, **guitar**	classical guitar, **guitar**, **acoustic**, lute, spanish
43390	**rock**, loud, **pop**, vocals, **male vocals**	**male vocals**, **pop**, male vocal, male singer, **rock**
48010	silence	low, soft, no singing, quiet, wind
57081	**piano**	piano solo, **piano**, classic, solo, classical

Πίνακας 4.11: Παραδείγματα εφαρμογής του προτεινόμενου αλγορίθμου για τον αυτόματο χαρακτηριστμό ηχητικών αποσπασμάτων. Κάθε παράδειγμα αντιστοιχεί σε ένα ηχητικό απόσπασμα του συνόλου δεδομένων Magnatagatune, όπου ID είναι το αναγνωριστικό του ηχητικού αποσπάσματος.

μετατρέπονται σε από κοινού πιθανότητες και επιδιώκεται η ελαχιστοποίηση της απόκλισης Kullback-Leibler μεταξύ των πιθανοτήτων του αρχικού χώρου και των πιθανοτήτων του μειωμένης διαστασιμότητας χώρου. Σημαντικό χαρακτηριστικό της μεθόδου t-SNE είναι η χρήση μη-κυρτής (non-convex) συνάρτησης κόστους. Για την ελαχιστοποίηση της συνάρτησης κόστους εφαρμόζεται η βελτιστοποίηση με κάθοδο της παραγώγου (gradient descent), η οποία αρχικοποιείται με τυχαίο τρόπο. Ως εκ τούτου, διαφορετικές εκτελέσεις της μεθόδου t-SNE δίνουν διαφορετικά αποτελέσματα. Η οπτικοποίηση των αναπαραστάσεων των tags απεικονίζεται στο Σχήμα 4.3. Παρατηρώντας το σχήμα, κάποιες από τις συστάδες στις οποίες έχουν ομαδοποιηθεί τα tags είναι οι ακόλουθες:

- loud, rock, electric guitar, punk, metal, hard
- make singer, male, male voice, male vocals, man singing
- female singing, woman, female, female voice, girl
- india, sitar, oriental, eastern, indian
- not opera, no drums, no singing, no beat, no flute, no piano, no rock, no voice

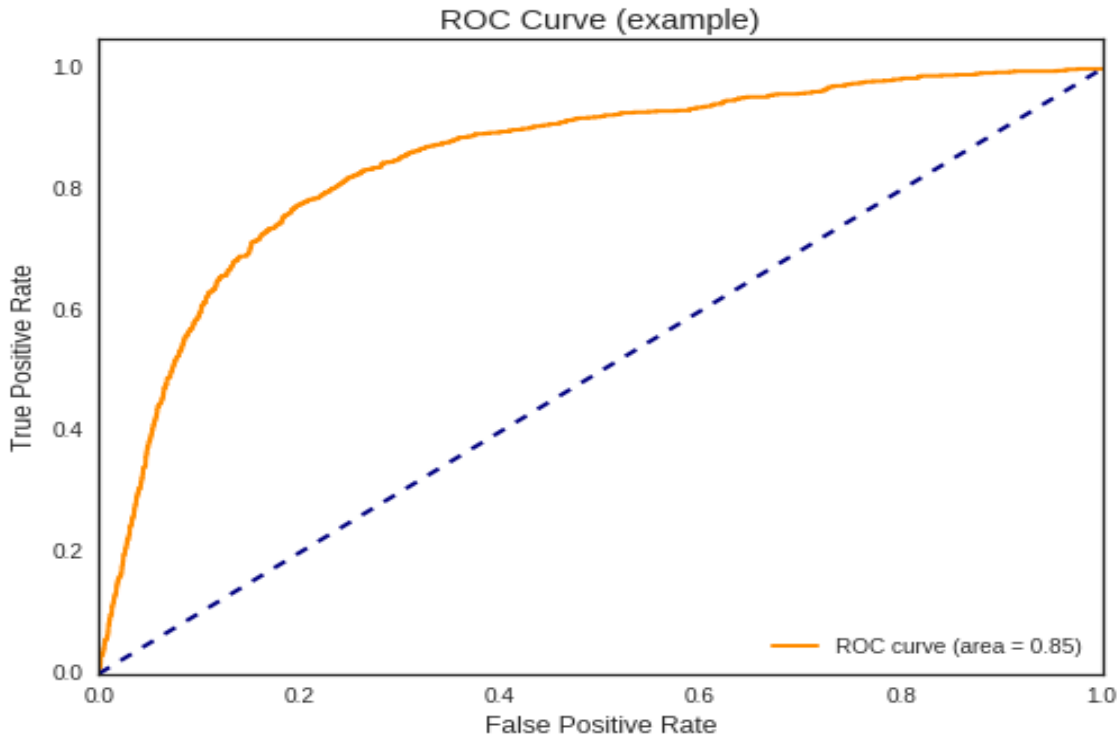


Σχήμα 4.3: Οπτικοποίηση των tags με τη μέθοδο t-SNE.

Αξιολόγηση του αλγορίθμου

Το πρόβλημα του αυτόματου χαρακτηρισμού ηχητικών αποσπασμάτων κατατάσσεται στα προβλήματα ταξινόμησης σε πολλαπλές κατηγορίες (multi-label classification), γιατί κάθε απόσπασμα περιγράφεται ταυτόχρονα από πολλαπλές λέξεις (tags). Για την αξιολόγηση του αλγορίθμου, τα πραγματικά δεδομένα (groundtruth) κωδικοποιούνται με τη μορφή ενός διανύσματος διάστασης T, όπου κάθε διάσταση αντιστοιχεί σε ένα διακεκριμένο tag και λαμβάνει δυαδικές τιμές. Πιο συγκεκριμένα, αν ένα tag υπάρχει στο σύνολο των πραγματικών tags, τότε η αντίστοιχη διάσταση λαμβάνει μοναδιαία τιμή, διαφορετικά λαμβάνει μηδενική τιμή (one-hot vector). Η αναπαράσταση των προβλέψεων του αλγορίθμου auto-tagging γίνεται με αντίστοιχο τρόπο, δηλαδή ως διανύσματα διάστασης T με τη διαφορά ότι οι επιμέρους διαστάσεις των διανυσμάτων δε λαμβάνουν δυαδικές τιμές αλλά οποιαδήποτε τιμή στο διάστημα $[0, 1]$. Όσο μεγαλύτερη είναι η τιμή που προβλέπεται από τον αλγόριθμο (αναφέρεται ως τιμή αξιοπιστίας - confidence score), τόσο πιο αντιπροσωπευτικό θεωρείται το αντίστοιχο tag για το ηχητικό απόσπασμα. Επομένως, αν ζητείται η πρόβλεψη των N πιο αντιπροσωπευτικών tags, διατηρούνται τα N tags με τις μεγαλύτερες τιμές αξιοπιστίας. Εναλλακτικά, η πρόβλεψη των tags μπορεί να γίνει με χρήση μίας τιμής αποκοπής (cutoff value). Έτσι, ένα tag θεωρείται αντιπροσωπευτικό του κομματιού μόνο αν η τιμή αξιοπιστίας είναι μεγαλύτερη από την τιμή αποκοπής.

Για την αξιολόγηση του αλγορίθμου στο σύνολο δεδομένων MagnaTagATune, από το σύνολο των 188 tags διατηρούνται τα 50 πιο συχνά εμφανιζόμενα tags, όπως αναφέρεται και σε παλαιότερες δημοσιεύσεις [145, 146, 147, 148]. Επομένως το πραγματικό διάνυσμα αλλά και το προβλεπόμενο (από τον αλγόριθμο auto-tagging) διάνυσμα θα έχουν διάσταση $T = 50$. Για κάθε ένα από τα 25,863 μουσικά κομμάτια γίνεται εφαρμογή του αλγορίθμου auto-tagging και υπολογισμός του προβλεπόμενου διανύσματος. Η μετρική που χρησιμοποιείται στην παρούσα διπλωματική ονομάζεται Έμβαδόν Περιοχής Κάτω από την Καμπύλη Λειτουργικού Χαρακτηριστικού Δέκτη' (Area Under Receiver Operating Characteristic Curve - AUC). Η Καμπύλη Λειτουργικού Χαρακτηριστικού Δέκτη (ROC curve) είναι ιδιαίτερα χρήσιμη γραφική απεικόνιση της απόδοσης ενός ταξινομητή. Ορίζεται για προβλήματα ταξινόμησης σε δύο κλάσεις, ωστόσο γενικεύεται και για μεγαλύτερο πλήθος κλάσεων. Ο οριζόντιος άξονας του γραφήματος συμβολίζει το ποσοστό λανθασμένων προβλέψεων (false positive rate) ενώ ο κατακόρυφος άξονας συμβολίζει το ποσοστό ορθών προβλέψεων (true positive rate) του ταξινομητή. Τα ποσοστά αυτά εξαρτώνται από την τιμή αποκοπής (cutoff value), επομένως κάθε σημείο της καμπύλης αντιστοιχεί σε διαφορετικό σημείο αποκοπής. Ένα παράδειγμα της καμπύλης ROC παρατίθεται στο Σχήμα 4.4. Εφόσον είναι επιθυμητό να μεγιστοποιηθεί το ποσοστό ορθών προβλέψεων και να ελαχιστοποιηθεί το ποσοστό λανθασμένων προβλέψεων, το ιδανικότερο σημείο στην καμπύλη ROC θα ήταν το σημείο $(0, 1)$ (πάνω-αριστερά). Η μετρική που χρησιμοποιείται λοιπόν για την εκτίμηση της απόδοσης ενός ταξινομητή είναι το εμβαδό κάτω από την καμπύλη ROC. Όσο μεγαλύτερο είναι το εμβαδό τόσο καλύτερη θεωρείται η απόδοση του ταξινομητή με μέγιστη τιμή $AUC = 1.0$. Σημειώνεται ότι η τυχαία πρόβλεψη τιμών αντιστοιχεί στη μπλε διαγώνιο του σχήματος, οδηγώντας σε εμβαδό $AUC = 0.5$.



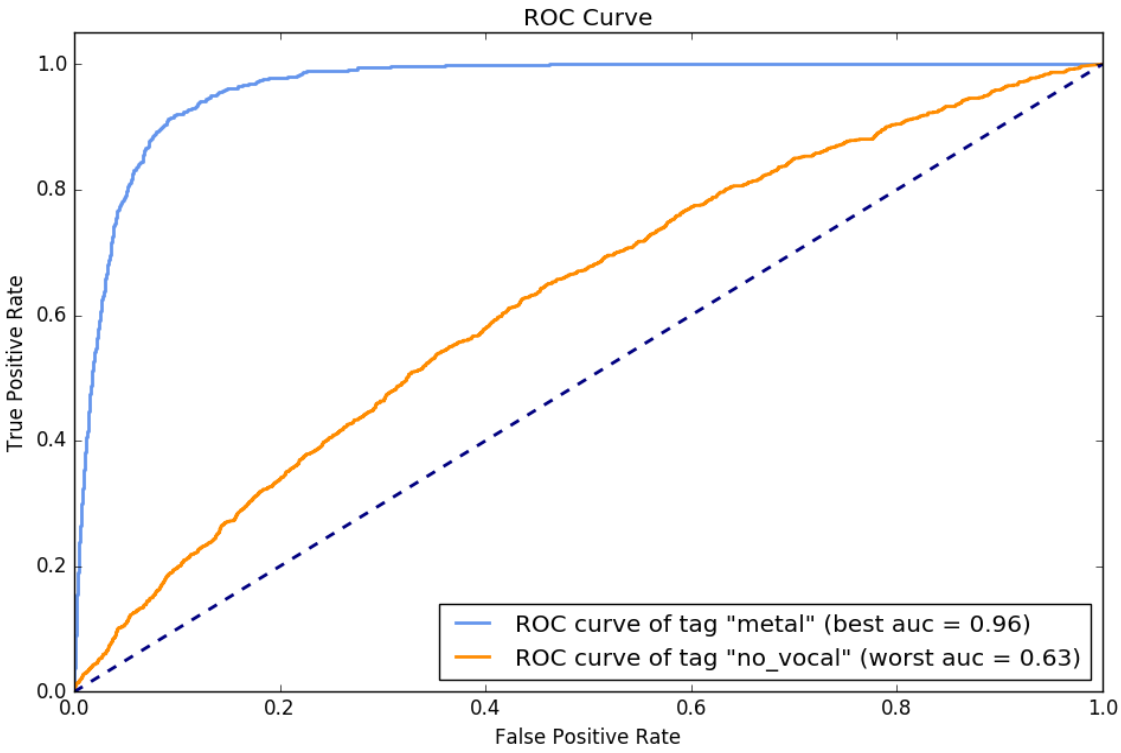
Σχήμα 4.4: Παράδειγμα της καμπύλης Receiver Operating Characteristic (ROC).

k	**EchoNest**	**MFCCdd**
300	0.809	0.806
800	0.826	0.814

Πίνακας 4.12: Απόδοση του αλγορίθμου auto-tagging (τιμή AUC) για διαφορετικές παραμέτρους και χαρακτηριστικά του ADSM μοντέλου.

Στην περίπτωση του παρόντος προβλήματος, μία καμπύλη ROC σχεδιάζεται για κάθε tag ξεχωριστά οδηγώντας σε 50 διακριτές τιμές AUC. Η τελική τιμή απόδοσης ορίζεται ως ο μέσος όρος των 50 τιμών AUC. Επιπλέον, καθώς οι αναπαραστάσεις των tags υπολογίζονται σύμφωνα με το ADSM μοντέλο με χρήση του ίδιου συνόλου δεδομένων (MagnaTagATune), η αξιολόγηση του αλγορίθμου πραγματοποιείται με χρήση της τεχνικής 10-fold cross-validation, ώστε να είναι σίγουρο ότι τα μουσικά κομμάτια προς αξιολόγηση δεν χρησιμοποιούνται για την εκπαίδευση του ADSM μοντέλου. Στον Πίνακα 4.12 παρουσιάζεται η απόδοση του αλγορίθμου auto-tagging για μοντέλα με διαφορετικές παραμέτρους.

Για ορισμένα tags επιτυγχάνεται ιδιαίτερα ψηλή τιμή AUC ενώ για άλλα χαμηλότερη. Για παράδειγμα, στην περίπτωση όπου χρησιμοποιούνται EchoNest χαρακτηριστικά για την εκπαίδευση του ADSM μοντέλου και το πλήθος των ακουστικών λέξεων είναι $k = 300$, η τιμή AUC που αντιστοιχεί στο tag “metal” είναι 0.96 ενώ η τιμή AUC για το tag “no vocal” είναι 0.63. Οι αντίστοιχες ROC καμπύλες απεικονίζονται στο Σχήμα 4.5. Στον Πίνακα 4.2.2 αναφέρονται οι τιμές AUC για κάθε ένα από τα 50 tags όπως υπολογίζονται από το ίδιο



Σχήμα 4.5: Η καμπύλη Receiver Operating Characteristic (ROC) για δύο tags του συνόλου δεδομένων MagnaTagATune.

μοντέλο.

tag	AUC	tag	AUC	tag	AUC	tag	AUC
metal	0.965	choral	0.946	choir	0.942	opera	0.941
rock	0.927	harp	0.906	harpsichord	0.900	cello	0.897
dance	0.891	beats	0.881	beat	0.877	flute	0.875
violin	0.870	loud	0.868	techno	0.862	country	0.847
piano	0.838	classical	0.828	pop	0.827	solo	0.826
classic	0.823	quiet	0.823	sitar	0.821	drums	0.818
man	0.808	strings	0.799	male	0.795	male vocal	0.786
guitar	0.786	electronic	0.786	male voice	0.783	ambient	0.775
soft	0.773	fast	0.768	indian	0.767	singing	0.766
woman	0.757	female	0.757	new age	0.755	female vocal	0.755
vocal	0.747	female voice	0.747	synth	0.738	weird	0.737
vocals	0.733	voice	0.728	slow	0.727	no voice	0.642
no vocals	0.629	no vocal	0.626				

Πίνακας 4.13: Απόδοση (AUC) του αλγορίθμου auto-tagging ($k = 300$, features=EchoNest) για κάθε ένα από τα 50 πιο συχνά εμφανιζόμενα tags του συνόλου MagnaTagATune.

Παρατηρείται ότι τα τρία tags με τη μικρότερη τιμή AUC (no vocal, no vocals, no voice)

tag	**AUC**	**tag**	**AUC**	**tag**	**AUC**	**tag**	**AUC**
metal	0.969	choral	0.953	choir	0.949	opera	0.943
rock	0.93	harpsichord	0.916	cello	0.916	harp	0.911
dance	0.907	flute	0.904	beats	0.891	violin	0.89
beat	0.885	country	0.88	techno	0.873	loud	0.873
piano	0.849	classical	0.848	classic	0.845	solo	0.844
sitar	0.841	quiet	0.839	pop	0.839	man	0.828
drums	0.827	strings	0.815	male	0.813	male voice	0.808
guitar	0.807	male vocal	0.805	electronic	0.8	female voice	0.796
female vocal	0.795	singing	0.788	indian	0.787	ambient	0.784
woman	0.784	fast	0.783	soft	0.783	female	0.781
new age	0.778	vocal	0.776	vocals	0.767	weird	0.764
synth	0.764	voice	0.75	slow	0.733	no voice	0.674
no vocal	0.664	no vocals	0.66				

Πίνακας 4.14: Απόδοση (AUC) του αλγορίθμου auto-tagging ($k = 300$, features=MFCCdd) για κάθε ένα από τα 50 πιο συχνά εμφανιζόμενα tags του συνόλου MagnaTagATune.

έχουν την ίδια έννοια και είναι οι αρνήσεις άλλων tags. Επομένως, είναι λογικό να 'μπερδεύουν' τον αλγόριθμο auto-tagging και δεδομένου ότι έχουν σημαντικά μικρότερη τιμή AUC από τα άλλα 47 tags, χωρίς την ύπαρξή τους η απόδοση θα ήταν σημαντικά καλύτερη. Στην επόμενη ενότητα θα γίνει περιγραφή της χρήσης του αλγορίθμου auto-tagging για την αξιολόγηση της ομοιότητας μουσικών κομματιών.

4.3 Αξιολόγηση της Μουσικής Ομοιότητας

Μία διαδεδομένη εφαρμογή του τομέα Ανάκτησης Μουσικής Πληροφορίας είναι η 'Ερώτηση με παραδείγματα' (query-by-example), όπου ο χρήστης θέτει ως ερώτημα ένα μουσικό κομμάτι και η μηχανή αναζήτησης του επιστρέφει μία ταξινομημένη λίστα από διαφορετικά μουσικά κομμάτια, παρόμοια με το κομμάτι που τέθηκε ως ερώτημα. Μία ακόμη εφαρμογή είναι η δημιουργία λιστών αναπαραγωγής, στις οποίες συνήθως περιλαμβάνονται μουσικά κομμάτια από ίδιο ή παρόμοιο είδος μουσικής. Η πληροφορία ενός κομματιού που παρέχεται στη μορφή μεταδεδομένων π.χ. για το είδος της μουσικής ή για το όνομα του καλλιτέχνη είναι ιδιαίτερα σημαντική στις παραπάνω εφαρμογές, καθώς κομμάτια είδου είδους ή καλλιτέχνη συνήθως θεωρούνται παρόμοια. Συχνά μάλιστα λαμβάνονται υπόψη οι προτιμήσεις άλλων χρηστών για την πρόταση νέων κομματιών ή καλλιτεχνών σε έναν χρήστη, τεχνική που υπάγεται στον τομέα του Συνεργατικού Φιλτραρίσματος (Collaborative Filtering).

Ωστόσο, η χρήση των μεταδεδομένων για την αυτόματη πρόταση μουσικών κομματιών παρουσιάζει ορισμένα μειονεκτήματα. Έστω για παράδειγμα ένας νέος καλλιτέχνης, του οποίου τα έργα δεν έχουν λάβει κριτικές από χρήστες. Με χρήση των κλασσικών αλγορίθμων συνεργατικού φιλτραρίσματος, αυτός ο καλλιτέχνης δε θα προβληθεί ποτέ σε άλλους χρήστες καθώς θα προτιμώνται πάντα άλλοι καλλιτέχνες με περισσότερες κριτικές. Το πρόβλημα αυτό, το οποίο αναφέρεται ως πρόβλημα ψυχρής εκκίνησης (cold start problem) είναι δυνατό να επιλυθεί με μεθόδους ανάλυσης του περιεχομένου των μουσικών κομματιών. Μία μέθοδος είναι η πρόταση μουσικών κομματιών αξιολογώντας τη μεταξύ τους μουσική ομοιότητα [149].

Αρχικό βήμα για την αξιολόγηση της μουσικής ομοιότητας είναι η αναπαράσταση μουσικών αποσπασμάτων ως προς συγκεκριμένα ακουστικά χαρακτηριστικά, όπως για παράδειγμα τα χαρακτηριστικά Χρωμογράμματος (Chromagram features). Στη συνέχεια, η ομοιότητα δύο κομματιών υπολογίζεται με τη σύγκριση των αντίστοιχων αναπαραστάσεων με χρήση κάποιας μετρικής απόστασης π.χ. Ευκλείδια απόσταση. Βέβαια, η μετρική απόστασης δεν επιλέγεται πάντα αυθαίρετα, αλλά μέσω τεχνικών μηχανικής μάθησης ώστε οι αξιολογήσεις της μουσικής ομοιότητας να προσεγγίζουν αποδοτικά την πραγματική (groundtruth) τιμή μουσικής ομοιότητας [5, 6, 150]. Για το σκοπό αυτό έχει γίνει πρόταση μιας σειράς μεθόδων για τη συλλογή πραγματικών τιμών μουσικής ομοιότητας [151] αλλά και μεθόδων για την αξιολόγησης των αλγορίθμων [152].

Σε αυτή την ενότητα θα γίνει εφαρμογή του ακουστικού-σημασιολογικού (ADSM) μοντέλου (βλ. Ενότητα 3.1.7) για την αξιολόγηση της ομοιότητας μεταξύ μουσικών κομματιών του συνόλου MagnaTagATune. Επιπλέον θα γίνει περιγραφή ενός αλγορίθμου για τη σύμπτυξη ακουστικών και σημασιολογικών αναπαραστάσεων των μουσικών κομματιών με σκοπό τη βελτίωση των προβλέψεων.

4.3.1 Περιγραφή Αλγορίθμου/Δεδομένων

Δεδομένα Μουσικής Ομοιότητας

Το σύνολο μουσικών κομματιών που χρησιμοποιήθηκε είναι το Magnatagatune, για το οποίο έχει ήδη γίνει η περιγραφή στην Ενότητα 4.2.1. Εκτός από τα μουσικά κομμάτια και τα σχετικά tags, στο σύνολο δεδομένων περιλαμβάνονται και αξιολογήσεις ανθρώπων ως προς τη σχετική ομοιότητα των μουσικών κομματιών οι οποίες έχουν ανακτηθεί από ένα πρόσθετο μέρος της δοκιμασίας TagATune [153]. Σε αυτό το μέρος της δοκιμασίας, ο χρήστης ακούει τρία διαφορετικά μουσικά κομμάτια και καλείται να επιλέξει το κομμάτι που ακούγεται πιο αταίριαστο ως προς τα άλλα δύο κομμάτια (leave one out game). Εφόσον για κάθε τριάδα κομματιών ψηφίζουν πολλοί χρήστες, τα δεδομένα σώζονται με τη μορφή ιστογράμματος ψήφων για κάθε τριάδα και ως το πιο αταίριαστο κομμάτι επιλέγεται αυτό με τις περισσότερες ψήφους. Οι ψήφοι σώζονται στη μορφή (a, b, c), όπου c είναι το αναγνωριστικό του αταίριαστου κομματιού (outlier) ενώ a και b τα αναγνωριστικά των άλλων δύο κομματιών. Αν ως $d(x, y)$ συμβολιστεί η 'απόσταση' μεταξύ των μουσικών κομματιών x και y όπως την αντιλαμβάνεται ο άνθρωπος, τότε η τριπλέτα (a, b, c) υποδηλώνει τις σχέσεις:

$$d(a, b) < d(a, c) \tag{4.3}$$

και

$$d(b, a) < d(b, c) \tag{4.4}$$

Εδώ θεωρείται ότι η μετρική απόστασης είναι συμμετρική, δηλαδή $d(a, b) = d(b, a)$. Λόγω της υποκειμενικής άποψης κάθε χρήστη, είναι πιθανό να υπάρχουν τριπλέτες που έρχονται σε αντίθεση. Για παράδειγμα αν εκτός από την τριπλέτα (a, b, c) υπάρχει και η τριπλέτα (b, c, a) στο σύνολο δεδομένων τότε θα ισχύουν και οι σχέσεις:

$$d(b, c) < d(b, a) \tag{4.5}$$

και

$$d(c, b) < d(c, a), \tag{4.6}$$

το οποίο είναι αδύνατο καθώς οι σχέσεις 4.4 και 4.5 έρχονται σε αντίθεση. Μία πρώτη σκέψη για την αποφυγή τέτοιου είδους ασυνεπειών είναι η διαγραφή τριπλετών οι οποίες επαναλαμβάνονται ως αντιμεταθέσεις των αναγνωριστικών κομματιών. Ωστόσο, τέτοιου είδους διαγραφές οδηγούν σε απάλειψη μόνο των άμεσων ασυνεπειών ενώ είναι πιθανό να υπάρχουν και άλλες ασυνέπειες, π.χ. $d(a, b) < d(b, c) < d(b, e) < d(a, b)$. Για να γίνει, λοιπόν, άπαλειψη όλων των συνεπειών κατασκευάζεται ένας κατευθυνόμενος γράφος ως εξής:

- κόμβοι: ένας κόμβος του γράφου αντιστοιχεί σε ένα ζεύγος μουσικών αποσπασμάτων, π.χ. (a, b)
- ακμές: δύο κόμβοι (a, b) και (a, c) ενώνονται με ακμή μόνο αν ισχύει $d(a, b) < d(a, c)$.

Έτσι, η τριπλέτα (a, b, c) μετατρέπεται σε τρεις κόμβους (a, b),(a, c),(b, c), μία ακμή από τον κόμβο (a, b) στον (a, c) και μία ακμή από τον (a, b) στον (b, c). Με βάση τα παραπάνω

η ύπαρξη ασυνέπειας ισοδυναμεί με την ύπαρξη κύκλου στον παραγόμενο γράφο. Επομένως, για την απάλειψη όλων των ασυνεπειών απαιτείται η μετατροπή του γράφου σε ακυκλικό [154]. Το πρόβλημα υπολογισμού του μέγιστου ακυκλικού υπογράφου (maximum acyclic subgraph) χαρακτηρίζεται ως NP-hard. Ωστόσο είναι δυνατό να βρεθεί μία προσεγγιστική λύση μέσω του παρακάτω αλγορίθμου:

Αλγόριθμος: Προσεγγιστικός υπολογισμός μέγιστου ακυκλικού γράφου

Είσοδος: Κατευθυνόμενος γράφος $G = (V, E)$
Έξοδος: Ακυκλικός γράφος G'
$E' \leftarrow \emptyset$
Για κάθε ακμή $(u, v) \in E$ (με τυχαία σειρά) **κάνε**:
  Αν ο γράφος $E' \cup (u, v)$ είναι ακυκλικός, **τότε**:
    $E' \leftarrow E' \cup (u, v)$
  Τέλος αν
Τέλος για κάθε
$G' \leftarrow (V, E')$

Ο αρχικός γράφος αποτελείται από 15300 ακμές, από τις οποίες οι 1598 είναι μοναδικές. Εφόσον η σειρά προσπέλασης των ακμών επιλέγεται με τυχαίο τρόπο, ο αλγόριθμος επαναλαμβάνεται για 10 φορές και διατηρούνται οι ακμές που έχουν επιλεχθεί και στις 10 προσεγγιστικές λύσεις. Τελικά, από τις 1598 ακμές διατηρήθηκαν οι 860 κάθε μία από τις οποίες σώζεται πάλι με μορφή τριπλέτας και ορίζει έναν μοναδικό περιορισμό (constraint) αποστάσεων. Η τριπλέτα (x, y, z) ορίζει τον περιορισμό $d(x, y) < d(x, z)$. Σε προγενέστερη δημοσίευση [5] οι 860 περιορισμοί διαχωρίστηκαν σε 10 μη επικαλυπτόμενα σύνολο αξιολόγησης (86), ενώ για κάθε σύνολο αξιολόγησης τα εναπομείναντα 774 στοιχεία χρησιμοποιούνται ως σύνολο εκπαίδευσης. Σημειώνεται ότι δεν υπάρχει επικάλυψη μεταξύ των συνόλων εκπαίδευσης και αξιολόγησης. Τα επιλεγμένα σύνολα εκπαίδευσης και αξιολόγησης διατίθενται ηλεκτρονικά[4] ώστε να χρησιμοποιηθούν ως κοινό σημείο αναφοράς.

Εξαγωγή Χαρακτηριστικών

Η προεπεξεργασία των δεδομένων και η εξαγωγή χαρακτηριστικών από το σύνολο δεδομένων MagnaTagATune έχουν περιγραφεί αναλυτικά στην προηγούμενη εφαρμογή (Ενότητα 4.2.1). Υπενθυμίζεται ότι εξάγονται δύο τύποι διανυσμάτων χαρακτηριστικών: τα χαρακτηριστικά EchoNest και τα χαρακτηριστικά MFCCdd και η απόδοση του προτεινόμενου αλγορίθμου θα αξιολογηθεί ξεχωριστά για κάθε σύνολο χαρακτηριστικών.

Η μέθοδος AUDIO

Για την αξιολόγηση της μουσικής ομοιότητας, πρώτο βήμα είναι η δημιουργία των bag-of-audio-words αναπαραστάσεων για τα μουσικά κομμάτια του MagnatagATune (βλ.Ενότητα 3.1.6).

[4] http://mirg.city.ac.uk/datasets/ismir2012/index1.html

Έτσι, η ομοιότητα μεταξύ δύο μουσικών κομματιών a και b υπολογίζεται ως η ομοιότητα συνημιτόνου μεταξύ των bag-of-audio-words αναπαραστάσεων r_a και r_b αντίστοιχα:

$$sim(a, b) = \frac{r_a \cdot r_b}{|r_a| \cdot |r_b|} \tag{4.7}$$

Η απόσταση των δύο κομματιών υπολογίζεται ως $d(a, b) = 1 - sim(a, b)$. Με αυτή τη μέθοδο αξιολόγησης της μουσικής ομοιότητας δε λαμβάνονται υπόψη τα tags των μουσικών κομματιών αλλά μόνο τα ακουστικά τους χαρακτηριστικά. Στις επόμενες ενότητες θα αναφέρεται ως η μέθοδος AUDIO.

Οι μέθοδοι ADSM-REALTAG και ADSM-AUTOTAG

Με χρήση του ADSM μοντέλου λαμβάνονται υπόψη και τα tags των μουσικών κομματιών για την αξιολόγηση της μουσικής ομοιότητας. Πρώτα, γίνεται δημιουργία των bag-of-audio-words αναπαραστάσεων για τα tags (βλ. Ενότητα 3.1.7). Η αναπαράσταση ενός μουσικού κομματιού με τη μέθοδο ADSM-REALTAG προκύπτει ως ο μέσος όρος (ανά σημείο) των bag-of-audio-words αναπαραστάσεων των tags που χαρακτηρίζουν το κομμάτι. Στη συνέχεια, η αξιολόγηση της ομοιότητας γίνεται μέσω της ομοιότητας συνημιτόνου, όπως ακριβώς έγινε και στη μέθοδο AUDIO. Βέβαια, η μέθοδος ADSM-REALTAG δεν μπορεί να εφαρμοστεί σε μουσικά κομμάτια για τα οποία δεν υπάρχουν διαθέσιμα tags. Το πρόβλημα αυτό επιλύεται αν αντί των πραγματικών tags γίνει αυτόματη πρόβλεψη των πιο αντιπροσωπευτικών tags (βλ. Ενότητα 4.2) και στη συνέχεια υπολογισμός της αναπαράστασης ενός κομματιού με βάση τα προβλεπόμενα tags. Η μέθοδος αυτή αναφέρεται ως ADSM-AUTOTAG και ο αριθμός N των tags που θα επιστρέφονται για κάθε κομμάτι θα μελετηθεί ως παράμετρος του πειράματος.

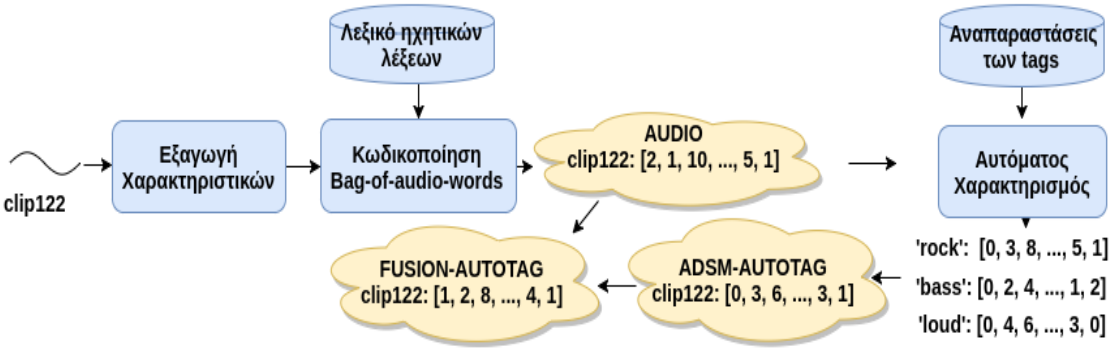
Οι μέθοδοι FUSION-REALTAG και FUSION-AUTOTAG

Υπάρχουν περιπτώσεις για τις οποίες η δημιουργία αναπαραστάσεων για μουσικά κομμάτια με τις μεθόδους ADSM-REALTAG και ADSM-AUTOTAG οδηγεί σε ανακριβείς προβλέψεις. Για παράδειγμα, αν δύο μουσικά κομμάτια περιγράφονται από το ίδιο σύνολο από tags τότε οι δύο αναπαραστάσεις είναι ακριβώς ίδιες και έτσι προκύπτει μοναδιαία τιμή ομοιότητας μεταξύ των δύο κομματιών παρόλο που τα κομμάτια δεν έχουν ίδιο άκουσμα. Για να αποφευχθούν τέτοιου είδους περιπτώσεις προτείνεται η σύμπτυξη των αναπαραστάσεων που προκύπτουν από τις μεθόδους AUDIO και ADSM-REALTAG. Έτσι, σύμφωνα με τη μέθοδο FUSION-AUTOTAG, οι αναπαραστάσεις r_c και r'_c ενός μουσικού κομματιού που προκύπτουν από τις μεθόδους AUDIO και ADSM-AUTOTAG αντίστοιχα συνδυάζονται μέσω του σχήματος σταθμισμένου μέσος (ανά σημείο):

$$r''_c = w \cdot r'_c + (1 - w) \cdot r_c \tag{4.8}$$

ή ως το σταθμισμένο επαυξημένο διάνυσμα:

$$r''_c = (w \cdot r'_c, (1 - w) \cdot r_c), \tag{4.9}$$



Σχήμα 4.6: Σχηματική αναπαράσταση των μεθόδων AUDIO, ADSM-AUTOTAG και FUSION-AUTOTAG για το παράδειγμα ενός μουσικού αποσπάσματος (clip 122).

όπου w είναι η τιμή βάρους και θα μελετηθεί ως πειραματική παράμετρος. Στο Σχήμα 4.6 περιλαμβάνεται η σχηματική αναπαράσταση της μεθόδου FUSION-AUTOTAG. Ψπονοείται ότι το Λεξικό Ηχητικών λέξεων έχει ήδη εκπαιδευτεί και οι αναπαραστάσεις των tags έχουν υπολογιστεί ως bag-of-audio-words αναπαραστάσεις με χρήση του ADSM μοντέλου. Με αντίστοιχο τρόπο προκύπτει η μέθοδος FUSION-REALTAG μέσω της σύμπτυξης των αναπαραστάσεων που υπολογίζονται από τις μεθόδους AUDIO και ADSM-REALTAG.

Αλγόριθμος Αξιολόγησης

Όπως αναφέρθηκε νωρίτερα, τα δεδομένα σημασιολογικής ομοιότητας είναι διαχωρισμένα σε 10 σύνολα εκπαίδευσης και αντίστοιχα 10 σύνολα αξιολόγησης. Με χρήση αυτών των συνόλων εφαρμόζεται διασταυρωμένη επικύρωση 10-πτυχών (10-fold cross-validation). Πιο συγκεκριμένα, για κάθε πτυχή (fold) τα 774 δεδομένα εκπαίδευσης χρησιμοποιούνται για τη δημιουργία του Λεξικού Ακουστικών Λέξεων[5] και τον υπολογισμό των αναπαραστάσεων των tags με χρήση του ADSM μοντέλου. Στη συνέχεια γίνεται υπολογισμός των bag-of-audio-words αναπαραστάσεων για τα μουσικά κομμάτια που βρίσκονται στο σύνολο αξιολόγησης και οι τιμές μουσικής ομοιότητας ($sim(x, y)$) υπολογίζονται μέσω της ομοιότητας συνημιτόνου των αντίστοιχων αναπαραστάσεων. Ένα δεδομένο αξιολόγησης (περιορισμός αποστάσεων) (a, b, c) ικανοποιείται αν για τις αποστάσεις των μουσικών κομματιών ισχύει: $d(a, b) < d(a, c)$. Διαφορετικά, ο περιορισμός δεν ικανοποιείται. Η απόδοση του μοντέλου για την συγκεκριμένη πτυχή ορίζεται ως το ποσοστό των περιορισμών που ικανοποιούνται. Ακολουθώντας την ίδια διαδικασία για κάθε μία από τις 10 πτυχές, η απόδοση του μοντέλου υπολογίζεται ως ο μέσος όρος των 10 επιμέρους ποσοστών. Για να εξασφαλιστεί η αξιοπιστία των αποτελεσμάτων, η τελική απόδοση του μοντέλου υπολογίζεται επαναλαμβάνοντας τη διαδικασία (10-fold cross-validation) 10 φορές και παίρνοντας το μέσο όρο των αποδόσεων.

[5]Κάθε δεδομένο εκπαίδευσης (τριπλέτα) περιλαμβάνει τρία μουσικά κομμάτια. Ωστόσο, για υπολογιστική αποδοτικότητα έγινε επιλογή μόνο 1000 κομματιών από το σύνολο εκπαίδευσης για τη δημιουργία του Λεξικού Ακουστικών Λέξεων.

4.3.2 Πειραματικά Αποτελέσματα

Στον Πίνακα 4.15 περιλαμβάνονται οι τιμές απόδοσης των μεθόδων που έχουν αναφερθεί στη βιβλιογραφία (βλ. [6] για σύντομη περίληψη των μεθόδων και αποτελεσμάτων) ακολουθώντας την ίδια μέθοδο αξιολόγησης. Οι μέθοδοι SVM και MLR περιγράφονται στην [5] ενώ οι Euclidean και RITML περιγράφονται στην [6] και συγκρίνονται με τις πρώτες δύο.

Μέθοδος Βιβλιογραφίας	**Απόδοση Χαρακτηριστικά EchoNest**
Euclidean	0.598
RITML	0.711
SVM	**0.712**
MLR	0.689

Πίνακας 4.15: Απόδοση των μεθόδων που περιγράφονται στη βιβλιογραφία [5, 6] ως το ποσοστό των περιορισμών αποστάσεων που ικανοποιούνται.

Στον Πίνακα 4.16 γίνεται αναφορά στην απόδοση των προτεινόμενων μεθόδων στο ζήτημα αξιολόγησης της μουσικής ομοιότητας. Οι πρώτες δύο στήλες αναφέρονται στη χρήση των χαρακτηριστικών EchoNest για τη δημιουργία των bag-of-audio-words αναπαραστάσεων ενώ οι τελευταίες δύο στη χρήση των χαρακτηριστικών MFCCdd. Το πλήθος των ακουστικών λέξεων ορίστηκε ως $k = 300$ ενώ έγινε και δοκιμή μείωσης της διαστασιμότητας στις 10 διαστάσεις με χρήση της τεχνικής SVD. Οι μέθοδοι FUSION-REALTAG και FUSION-AUTOTAG εφαρμόστηκαν με τη μορφή σταθμισμένου μέσου [6], όπου $w = 0.9$. Το πλήθος των προβλεπόμενων tags από τις μεθόδους ADSM-AUTOTAG και FUSION-AUTOTAG ορίστηκε ως $N = 20$. Εφόσον η αξιολόγηση των αλγορίθμων γίνεται με ίδιο τρόπο στην παρούσα εργασία και στις προγενέστερες δημοσιεύσεις, τα αποτελέσματα που αναφέρονται στους Πίνακες 4.15 και 4.16 είναι άμεσα συγκρίσιμα. Το μεγαλύτερο ποσοστό (0.731) επιτυγχάνεται με χρήση της μεθόδου FUSION-REALTAG και είναι σημαντικά μεγαλύτερη από τα ποσοστά που αναφέρονται στον Πίνακα 4.15. Επίσης, συγκρίνοντας τις μεθόδους AUDIO και ADSM-REALTAG παρατηρείται ότι λαμβάνοντας υπόψη την σημασιολογική πληροφορία των tags επιτυγχάνεται σημαντική βελτίωση της απόδοσης (έως και 12.7% σχετική αύξηση του ποσοστού περιορισμών που ικανοποιούνται).

Συγκρίνοντας τις μεθόδους ADSM-REALTAG και ADSM-AUTOTAG διαπιστώνεται ότι είτε χρησιμοποιώντας τα πραγματικά tags είτε προβλέποντας τα $N = 20$ πιο αντιπροσωπευτικά tags επιτυγχάνεται εξίσου καλή απόδοση. Το γεγονός αυτό υποδεικνύει ότι ο αλγόριθμος auto-tagging (βλ. Ενότητα 4.2) προβλέπει ιδιαίτερα αντιπροσωπευτικά tags για το σύνολο δεδομένων MagnaTagATune. Εξάλλου, η ποιότητα των προβλέψεων επιβεβαιώνεται και μέσω των παραδειγμάτων που αποτυπώθηκαν στον Πίνακα 4.11. Άμεσα προκύπτει ότι οι προβλέψεις του αλγορίθμου auto-tagging είναι δυνατό να χρησιμοποιηθούν για τον χαρακτηρισμό μουσι-

[6] Η απόδοση των μεθόδων με χρήση του επαυξημένου διανύσματος έδωσε παρόμοια αποτελέσματα σε σχέση με αυτά που αναφέρονται στον Πίνακα 4.16

Προτεινόμενη Μέθοδος	**EchoNest**		**MFCCdd**	
	k=300	svd=10	k=300	svd=10
AUDIO	0.613	0.644	0.636	0.646
ADSM-REALTAG	0.705	0.719	**0.717**	**0.720**
FUSION-REALTAG	**0.720**	**0.731**	0.681	0.684
ADSM-AUTOTAG	0.705	0.705	0.693	0.696
FUSION-AUTOTAG	0.705	0.709	0.662	0.672

Πίνακας 4.16: Απόδοση των προτεινόμενων μεθόδων ως το ποσοστό των περιορισμών απόστασης (distance constraints) που ικανοποιούνται.

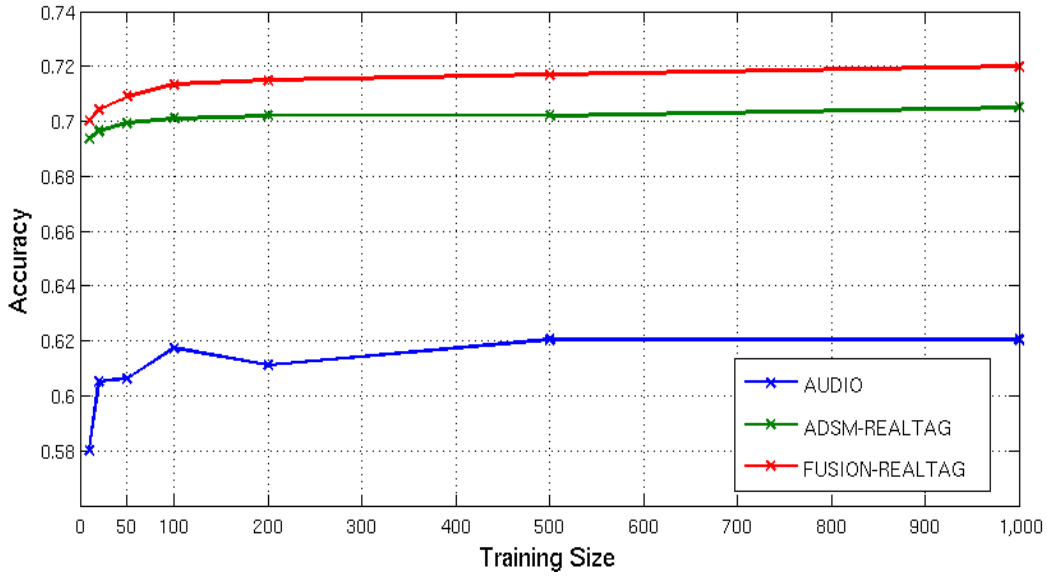
κών αποσπασμάτων χωρίς μεταδεδομένα ή για τη δημιουργία καλύτερων χαρακτηρισμών για αποσπάσματα με φτωχά μεταδεδομένα.

Αξίζει επίσης να σημειωθεί ότι παρόλο που δε γίνεται αξιοποίηση των δεδομένων μουσικής ομοιότητας που παρέχονται στα σύνολα εκπαίδευσης (με τη μορφή περιορισμών απόστασης), επιτυγχάνεται καλύτερη απόδοση σε σχέση με αλγορίθμους της βιβλιογραφίας που τα χρησιμοποιούν. Επομένως, ο προτεινόμενος αλγόριθμος μη επιβλεπόμενης μάθησης είναι δυνατό να εφαρμοστεί σε οποιοδήποτε σύνολο δεδομένων, χωρίς να είναι απαραίτητη η ύπαρξη πραγματικών δεδομένων μουσικής ομοιότητας.

Στο Σχήμα 4.7 γίνεται απεικόνιση της απόδοσης των μεθόδων AUDIO, ADSM-REALTAG και FUSION-REALTAG σε συνάρτηση με το πλήθος των μουσικών αποσπασμάτων που χρησιμοποιούνται για την κατασκευή του Λεξικού Ακουστικών Λέξεων. Σημειώνεται ότι τα μουσικά αποσπάσματα επιλέγονται με τυχαίο τρόπο από το σύνολο εκπαίδευσης. Είναι ενδιαφέρον το γεγονός ότι ακόμα και μικρός αριθμός από μουσικά αποσπάσματα (50-200) είναι αρκετός για τη δημιουργία του Λεξικού Ακουστικών Λέξεων, οδηγώντας σε ικανοποιητική απόδοση. Για παράδειγμα, χρησιμοποιώντας 200 μουσικά κομμάτια και τη μέθοδο FUSION-REALTAG, επιτυγχάνεται ποσοστό 0.717. Επιπλέον, σύμφωνα με το σχήμα οι μέθοδοι ADSM-REALTAG και FUSION-REALTAG είναι πιο εύρωστες σε σύγκριση με τη μέθοδο AUDIO στην περίπτωση ελλιπών δεδομένων.

4.3.3 Συμπεράσματα

Σε αυτή την ενότητα έγινε αξιολόγηση της μουσικής ομοιότητας με χρήση του ακουστικού-σημασιολογικού μοντέλου (ADSM). Χρησιμοποιήθηκαν και συγκρίθηκαν μουσικά κομμάτια από το σύνολο δεδομένων MagnaTagATune σε συνδυασμό από πραγματικές τιμές μουσικής ομοιότητας. Κατά την αξιολόγηση της μουσικής ομοιότητας με χρήση του ADSM μοντέλου λαμβάνεται υπόψη και η σημασιολογική πληροφορία των μουσικών κομματιών που παρέχεται ως περιγραφικές λέξεις (tags), οδηγώντας σε σημαντικά καλύτερη απόδοση σε σύγκριση με τη χρήση μόνο των ακουστικών χαρακτηριστικών. Επιπλέον, η καλύτερη απόδοση παρατηρείται μέσω της προτεινόμενης μεθόδου για τη σύμπτυξη σημασιολογικών και ακουστικών αναπαραστάσεων. Παρόλο που τα πραγματικά δεδομένα μουσικής ομοιότητας προς εκπαίδευση δε



Σχήμα 4.7: Απόδοση των μεθόδων AUDIO, ADSM-REALTAG και FUSION-REALTAG σε συνάρτηση με το πλήθος των μουσικών αποσπασμάτων που χρησιμοποιούνται για την κατασκευή του Λεξικού Ακουστικών Λέξεων. Χαρακτηριστικά: EchoNest, $k = 300$.

λαμβάνονται υπόψη από τις προτεινόμενες μεθόδους (μη επιβλεπόμενη μάθηση), επιτυγχάνεται σημαντικά καλύτερη απόδοση από μεθόδους που προτείνονται στη βιβλιογραφία. Τέλος, οι μέθοδοι που λειτουργούν με αυτόματη πρόβλεψη των tags έχουν παρόμοια απόδοση με τις μεθόδους που χρησιμοποιούν τα διαθέσιμα πραγματικά tags, γεγονός που υποδεικνύει την ποιότητα του προτεινόμενου αλγορίθμου autotagging και επιτρέπει την εφαρμογή των μεθόδων σε σύνολα δεδομένων χωρίς την ύπαρξη μεταδεδομένων. Τα πειραματικά αποτελέσματα που παρουσιάστηκαν σε αυτή την ενότητα έχουν αντληθεί από το [138].

Κεφάλαιο 5

Επίλογος

5.1 Συμπεράσματα

Σε αυτή τη διπλωματική εργασία έγινε περιγραφή των πολυτροπικών σημασιολογικών μοντέλων για τη δημιουργία διανυσματικών αναπαραστάσεων λέξεων. Κίνητρο για τη δημιουργία πολυτροπικών αναπαραστάσεων είναι η αντιμετώπιση του προβλήματος εδραίωσης (grounding) των παραδοσιακών κειμενικών αναπαραστάσεων στην ανθρώπινη αντίληψη, το οποίο συχνά αναφέρεται ως Symbol Grounding Problem. Αντίθετα με τα παραδοσιακά σημασιολογικά μοντέλα, τα πολυτροπικά σημασιολογικά μοντέλα εμπνέονται από την ανθρώπινη αντίληψη της σημασιολογίας λέξεων, κατά την οποία ο άνθρωπος αντιλαμβάνεται τη σημασία μίας λέξης συνδυάζοντας τα ερεθίσματα όλων των αισθήσεών του. Εδώ έγινε περιγραφή του ακουστικού-σημασιολογικού (ADSM) μοντέλου και του οπτικού-σημασιολογικού (VDSM) μοντέλου για τη δημιουργία αναπαραστάσεων λέξεων με βάση τα ακουστικά και οπτικά τους χαρακτηριστικά αντίστοιχα. Επιπλέον, παρουσιάστηκαν μέθοδοι για τη σύμπτυξη των παραπάνω μοντέλων με τα παραδοσιακά σημασιολογικά μοντέλα με σκοπό τη δημιουργία μίας κοινής πολυτροπικής αναπαράστασης στην οποία κωδικοποιούνται τα κειμενικά, ακουστικά και οπτικά χαρακτηριστικά των λέξεων. Τα πολυτροπικά σημασιολογικά μοντέλα χρησιμοποιήθηκαν για τρία διαφορετικά ζητήματα.

Το πρώτο ζήτημα που μελετήθηκε είναι η αξιολόγηση της σημασιολογικής ομοιότητας λέξεων. Αρχικά, έγινε δημιουργία του ακουστικού-σημασιολογικού (ADSM) μοντέλου και διερεύνηση της επίδρασης διαφορετικών πειραματικών παραμέτρων στην απόδοση του μοντέλου. Κύρια συμπεράσματα είναι ότι το πλήθος των ακουστικών λέξεων πρέπει να διαμορφώνεται λαμβάνοντας αντιστρόφως ανάλογη τιμή από το μέγεθος του χρονικού παραθύρου κατά την εξαγωγή χαρακτηριστικών και ότι είναι ιδιαίτερα αποδοτική η μείωση της διαστασιμότητας με τη μέθοδο PCA (σε συνδυασμό με την SVD). Επίσης, η μέθοδος soft encoding για τη δημιουργία bag-of-audio-words αναπαραστάσεων δεν έδωσε καλύτερα αποτελέσματα από τη μέθοδο hard encoding, γεγονός το οποίο αποτελεί κίνητρο για περαιτέρω διερεύνηση του προβλήματος, όπως θα αναφερθεί στην επόμενη ενότητα. Έπειτα, διαπιστώθηκε ότι η προτεινόμενη μέθοδος για την παραμετρική σύμπτυξη πολλαπλών ακουστικών χώρων με κριτήριο τη φύση του ήχου επιφέρει βελτίωση των state of the art αποτελεσμάτων που αναφέρονται στη βιβλιογραφία

(σχετική αύξηση του συντελεστή συσχέτισης Spearman κατά 23.6%). Τέλος, παρατηρήθηκε ότι η σύμπτυξη των σημασιολογικών (DSM), ακουστικών-σημασιολογικών (ADSM) και οπτικών-σημασιολογικών (VDSM) μοντέλων με τη μέθοδο Early Fusion οδηγεί σε σημαντικά καλύτερη απόδοση σε σχέση με αυτή κάθε μοντέλου ξεχωριστά ξεπερνώντας σε απόδοση ακόμα και το state of the art distributional μοντέλο (CDSM).

Το δεύτερο ζήτημα που μελετήθηκε είναι ο αυτόματος χαρακτηρισμός μουσικών αποσπασμάτων (auto-tagging). Το ADSM μοντέλο παρέχει έναν άμεσο τρόπο υπολογισμού tags για μουσικά κομμάτια. Με οπτικοποίηση των αναπαραστάσεων των tags αλλά και με τον υπολογισμό της τιμής AUC (Area Under the Receiver Operating Characteristic Curve) διαπιστώθηκε ότι ο προτεινόμενος αλγόριθμος auto-tagging προβλέπει tags, τα οποία είναι ιδιαίτερα αντιπροσωπευτικά στην περίπτωση του συνόλου δεδομένων MagnaTagATune.

Τέλος, έγινε πειραματισμός ως προς το ζήτημα της αξιολόγησης μουσικής ομοιότητας. Αρχικά, η αξιολόγηση πραγματοποιήθηκε αποκλειστικά με χρήση των ακουστικών χαρακτηριστικών των κομματιών και αποκλειστικά με χρήση των tags που τα περιγράφουν. Στη συνέχεια, με τη χρήση του ADSM μοντέλου έγινε ταυτόχρονη αξιοποίηση των tags αλλά και των μουσικών κομματιών για την αξιολόγηση της μουσικής ομοιότητας, οδηγώντας σε σημαντικά καλύτερη απόδοση σε σχέση με την απόδοση στην περίπτωση που χρησιμοποιούνται μόνο ακουστικά χαρακτηριστικά ή μόνο τα tags. Ο προτεινόμενος αλγόριθμος (ο οποίος λειτουργεί χωρίς επίβλεψη) οδήγησε σε σχετική βελτίωση της απόδοσης που αναφέρεται στη βιβλιογραφία ως state of the art κατά 2.7%, ξεπερνώντας την απόδοση αλγορίθμων οι οποίοι λειτουργούν με επίβλεψη. Επιπλέον, διαπιστώθηκε ότι η αντικατάσταση των πραγματικών tags με tags που προβλέπονται αυτόματα με χρήση του προτεινόμενου αλγορίθμου auto-tagging οδηγεί σε εξίσου καλή απόδοση. Ως αποτέλεσμα, ο προτεινόμενος αλγόριθμος για την αξιολόγηση της μουσικής ομοιότητας σε συνδυασμό με τον αλγόριθμο auto-tagging είναι δυνατό να εφαρμοστούν σε σύνολα δεδομένων χωρίς την απαίτηση μεταδεδομένων.

5.2 Μελλοντική Έρευνα

Η ικανοποιητική απόδοση των πολυτροπικών μοντέλων στα τρία παραπάνω ζητήματα αποτελεί κίνητρο για την περαιτέρω μελέτη στον συγκεκριμένο τομέα. Έτσι, έχουν προκύψει πολλές ιδέες σχετικά με την τροποποίηση των πολυτροπικών μοντέλων αλλά και με την επέκταση των υπαρχόντων μεθόδων για την πολυτροπική σύμπτυξη.

Μία ιδέα αφορά την τροποποίηση της μεθόδου υπολογισμού ακουστικών λέξεων και οπτικών λέξεων για τα ADSM και VDSM μοντέλα αντίστοιχα. Υπενθυμίζεται ότι στα προτεινόμενα μοντέλα, οι ακουστικές και οπτικές λέξεις υπολογίζονται ως τα κεντροειδή του αλγορίθμου ομαδοποίησης k-means. Μία πιθανή τροποποίηση είναι η αντικατάσταση του αλγορίθμου ομαδοποίησης k-means από έναν αλγόριθμο dictionary learning. Για παράδειγμα, ο αλγόριθμος k-SVD [155], ο οποίος αποτελεί γενίκευση του αλγορίθμου k-means, τρέχει επαναληπτικά στο σύνολο των δεδομένων εκπαίδευσης και ανανεώνει τα στοιχεία του λεξικού (atoms, αντίστοιχα clusters του k-means) ώστε να επιτευκτεί βέλτιστη αναπαράσταση των δεδομένων, ικανοποιώντας ταυτόχρονα κάποιους περιορισμούς αραιότητας των αναπαραστάσεων (sparsity

constraints). Η εύρεση του βέλτιστου λεξικού για την κωδικοποίηση των αναπαραστάσεων είναι μη-κυρτό (non-convex) πρόβλημα και ως αποτέλεσμα ο k-SVD δεν εγγυάται την εύρεση της ολικά βέλτιστης λύσης. Ωστόσο, στην πράξη, ο αλγόριθμος k-SVD έχει λειτουργήσει αποδοτικά [156], επομένως θεωρείται εύλογη η χρήση του για τη δημιουργία πολυτροπικών αναπαραστάσεων.

Επίσης, θα διερευνηθεί η σύγκριση των μεθόδων soft encoding και hard encoding ως προς το πλήθος των δειγμάτων προς ομαδοποίηση με τον αλγόριθμο k-means. Όπως αναφέρθηκε νωρίτερα, η προτεινόμενη μέθοδος soft encoding δεν έδωσε σημαντικά καλύτερα αποτελέσματα από τη μέθοδο hard encoding, κατά την οποία κάθε χρονικό τμήμα ενός αποσπάσματος αντιστοιχίζεται σε ένα μόνο κεντροειδές του k-means. Σκοπός, λοιπόν του πειράματος είναι η επιβεβαίωση ή απόρριψη της υπόθεσης ότι όσο αυξάνεται το πλήθος δειγμάτων προς ομαδοποίηση, τόσο μειώνεται η βελτίωση της μεθόδου soft encoding έναντι της μεθόδου hard encoding.

Τα ADSM και VDSM μοντέλα ορίστηκαν ως επεκτάσεις των Κατανεμημένων Σημασιολογικών Μοντέλων, τα οποία υπενθυμίζεται ότι μοντελοποιούν τις λέξεις μετρώντας το πλήθος συνεμφανίσεων σε πηγές κειμένων. Τα μοντέλα Word2vec και Glove έχουν εφαρμοστεί με επιτυχία για τη δημιουργία αναπαραστάσεων λέξεων. Επίσης, μοντέλα βασισμένα σε αυτά έχουν εφαρμοστεί με επιτυχία σε πολλά άλλα προβλήματα, όπως για παράδειγμα τα μοντέλα dna2vec [157] και node2vec [158]. Επομένως, κρίνεται σκόπιμη η επέκταση αυτών των δύο μοντέλων για τη δημιουργία πολυτροπικών μοντέλων ήχου και εικόνας.

Επίσης, κατά τη δημιουργία του ADSM μοντέλου θεωρήσαμε ότι συγκεκριμένα χαρακτηριστικά π.χ. MFCCs είναι κατάλληλα για την αναπαράσταση του ηχητικού σήματος και την εφαρμογή των αναπαραστάσεων σε συγκεκριμένα προβλήματα. Μία πιο σύγχρονη μέθοδος έχει να κάνει με την αυτόματη εκμάθηση χαρακτηριστικών ως τον βέλτιστο μετασχηματισμό του ηχητικού σήματος δοθέντος ενός προβλήματος. Για το σκοπό αυτό χρησιμοποιούνται βαθιά νευρωνικά δίκτυα (deep neural networks), τα οποία αναφέρονται ως end-to-end μοντέλα και έχουν χρησιμοποιηθεί σε πολλές εφαρμογές, όπως ο αυτόματος χαρακτηρισμός μουσικών αποσπασμάτων [146] και η αναγνώρισης φωνής [159].

Ένας σημαντικός περιορισμός των αναπαραστάσεων ηχητικών αποσπασμάτων σύμφωνα με την υπόθεση bag-of-audio-words είναι ότι αγνοούν τη χρονική συσχέτιση των δειγμάτων, μοντελοποιώντας μόνο τη στατιστική κατανομή του συχνοτικού περιεχομένου των δειγμάτων [62]. Μία ιδέα για την συμπερίληψη της χρονικής συσχέτισης των δειγμάτων είναι η προσθήκη ενός επιπλέον σταδίου μετά την εξαγωγή των αναπαραστάσεων των χρονικών τμημάτων κάθε αποσπάσματος, όπου οι αναπαραστάσεις θα εισάγονται σε ένα αναδρομικό νευρωνικό δίκτυο (Recurrent Neural Network - RNN), όπως για παράδειγμα το δίκτυο Long-Short Term Memory (LSTM). Τα LSTMs αλλά και η επέκτασή τους που ονομάζεται Bi-directional LSTMs (BLSTMs) έχουν ήδη εφαρμοστεί με επιτυχία στην αναπαράσταση ηχητικών αποσπασμάτων με ποικίλες εφαρμογές όπως η αναγνώριση συναισθήματος από φωνή [160, 161, 162, 163], η αναγνώριση ηχητικών γεγονότων [164, 165] κλπ.

Η σύμπτυξη των πολυτροπικών αναπαραστάσεων με τις μεθόδους Early Fusion και Late Fusion έχει δώσει ιδιαίτερα ενδιαφέροντα αποτελέσματα στο πρόβλημα αξιολόγησης της ση-

μασιολογικής ομοιότητας λέξεων. Κατά την περιγραφή αυτών των δύο μεθόδων (βλ. Ενότητα 3.3), θεωρήθηκε ότι κατά τη σύμπτυξη με βάρη γίνεται χρήση του ίδιου συνδυασμού βαρών για κάθε αναπαράσταση. Μία πιθανή επέκταση των δύο μεθόδων είναι η χρήση τιμών βάρους οι οποίες προσδιορίζονται από τη φύση της λέξης προς αναπαράσταση. Πιο συγκεκριμένα, υπάρχουν λέξεις οι οποίες είναι περισσότερο σχετικές με τον ήχο παρά με την εικόνα (π.χ. 'jazz'), άλλες λέξεις που είναι πιο σχετικές με την εικόνα (π.χ. 'black') και άλλες που δεν είναι σχετικές ούτε με ήχο ούτε με εικόνα (π.χ. 'democracy'). Έτσι, στις περιπτώσεις λέξεων σχετικών με ήχο έχει νόημα η αντιστοίχηση μεγαλύτερου βάρους στο ακουστικό-σημασιολογικό μοντέλο κ.ο.κ. Προτείνεται λοιπόν η χρήση συνδυασμών βαρών για τη σύμπτυξη, οι οποίοι διαφέρουν από λέξη σε λέξη και καθορίζονται με βάση τη σύνδεση κάθε λέξης με τα οπτικά και ακουστικά ερεθίσματα που προκαλεί στον άνθρωπο. Ένας τρόπος για τον καθορισμό της 'ακουστικότητας' και 'ηχητικότητας' των λέξεων είναι η χρήση διαθέσιμων συνόλων δεδομένων όπως το Sensicon [166]. Βέβαια, αντί του χειροκίνητου καθορισμού των βαρών για τη σύμπτυξη είναι δυνατό να γίνει χρήση βαθιών νευρωνικών δικτύων (DNNs) ή αλγορίθμων μάθησης πολλαπλοτήτων (Manifold Learning) ώστε κατά τη σύμπτυξη των αναπαραστάσεων να ληφθούν υπόψη και μη γραμμικές σχέσεις μεταξύ των επιμέρους αναπαραστάσεων. Το κίνητρο για τη χρήση νευρωνικών δικτύων δεν είναι μονάχα η πολύ καλή απόδοσή τους σε πληθώρα προβλημάτων, αλλά και η υποστήριξη της θέσης ότι το grounding problem θα μπορούσε να επιλυθεί με τη φιλοσοφία του συνδετισμού [167, 168]. Στην περίπτωση που γίνει χρήση βαθιών νευρωνικών δικτύων έχει ενδιαφέρον η μοντελοποίηση της 'ακουστικότητας' και 'ηχητικότητας' των λέξεων ως επιπλέον όρων κανονικοποίησης (regularization terms) για την εκπαίδευση των μοντέλων.

Τέλος, ιδιαίτερο ενδιαφέρον για την πολυτροπική σύμπτυξη παρουσιάζουν οι περιπτώσεις όπου υπάρχει έλλειψη πληροφορίας ως προς κάποιο πολυτροπικό μοντέλο. Πώς μπορεί να γίνει, για παράδειγμα, εκμάθηση μίας κοινής πολυτροπικής αναπαράστασης για μία λέξη όταν είναι διαθέσιμες οι αναπαραστάσεις των DSM και ADSM μοντέλων αλλά όχι η αναπαράσταση του VDSM μοντέλου; Μία λύση του προβλήματος αυτού είναι η χρήση βαθιών νευρωνικών δικτύων για τον υπολογισμό της άγνωστης αναπαράστασης (έξοδος του δικτύου) δοθέντων των γνωστών αναπαραστάσεων (είσοδοι του δικτύου). Για την εκπαίδευση του δικτύου θα γίνει χρήση των δεδομένων όπου όλες οι αναπαραστάσεις είναι διαθέσιμες. Η διαδικασία αυτή θα μπορούσε να εφαρμοστεί και στο πρόβλημα ανάκτησης πολυμεσικής πληροφορίας, όπου για παράδειγμα δίνεται ως ερώτημα ένα ηχητικό απόσπασμα και αναζητάται κάποια εικόνα σχετική με το απόσπασμα.

Φτάνοντας στο τέλος αυτής της διπλωματικής ελπίζω οι παραπάνω ιδέες να υλοποιηθούν ή να αποτελέσουν έμπνευση για νέες ιδέες στον τομέα των πολυτροπικών σημασιολογικών αναπαραστάσεων.

Bibliography

[1] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(2014):1–47, 2014.

[2] Stephen Clark. Vector space models of lexical meaning. *Handbook of Contemporary Semantics–second edition. Wiley-Blackwell*, 2012.

[3] Alessandro Lopopolo and Emiel van Miltenburg. Sound-based distributional models. *IWCS 2015*, page 70, 2015.

[4] Douwe Kiela and Stephen Clark. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *EMNLP*, pages 2461–2470, 2015.

[5] Daniel Wolff, Sebastian Stober, Andreas Nürnberger, and Tillman Weyde. A systematic comparison of music similarity adaptation approaches. In *ISMIR*, pages 103–108. FEUP Edições, 2012.

[6] Daniel Wolff, Andrew MacFarlane, and Tillman Weyde. Comparative music similarity modelling using transfer learning across user groups. In *ISMIR*, pages 24–30, 2015.

[7] Arthur M Glenberg and Michael P Kaschak. Grounding language in action. *Psychonomic bulletin & review*, 9(3):558–565, 2002.

[8] Lawrence W Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645, 2008.

[9] Friedemann Pulvermüller. Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6(7):576–582, 2005.

[10] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[11] George A Miller and Walter G Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991.

[12] Hinrich Schütze. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123, 1998.

[13] Marco Pennacchiotti, Diego De Cao, Roberto Basili, Danilo Croce, and Michael Roth. Automatic induction of framenet lexical units. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 457–465. Association for Computational Linguistics, 2008.

[14] Ekaterina Shutova, Tim Van de Cruys, and Anna Korhonen. Unsupervised metaphor paraphrasing using a vector space model. In *COLING (Posters)*, pages 1121–1130, 2012.

[15] Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *ACL*, pages 236–244, 2008.

[16] Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429, 2010.

[17] Alessandro Lenci. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31, 2008.

[18] Timothy T Rogers and James L McClelland. *Semantic cognition: A parallel distributed processing approach*. MIT press, 2004.

[19] Brian Murphy, Partha Talukdar, and Tom Mitchell. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 114–123. Association for Computational Linguistics, 2012.

[20] Alexandros Potamianos. Cognitive multimodal processing: from signal to behavior. In *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*, pages 27–34. ACM, 2014.

[21] Elias Iosif, Spiros Georgiladakis, and Alexandros Potamianos. Cognitively motivated distributional representations of meaning. In *10th Language Resources and Evaluation Conference (LREC)*, 2016.

[22] Arthur M Glenberg and David A Robertson. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language*, 43(3):379–401, 2000.

[23] Lawrence W Barsalou, Ava Santos, W Kyle Simmons, and Christine D Wilson. Language and simulation in conceptual processing. *Symbols, embodiment, and meaning*, pages 245–283, 2008.

[24] Max M Louwerse. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2):273–302, 2011.

[25] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[26] Reinhard Rapp. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the ninth machine translation summit*, pages 315–322, 2003.

[27] Stephen Clark. Vector space models of lexical meaning. *Handbook of Contemporary Semantic Theory, The*, pages 493–522, 2015.

[28] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.

[29] Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

[30] Will Lowe. Towards a theory of semantic space. In *Proceedings of the twenty-third annual conference of the cognitive science society*, pages 576–581, 2001.

[31] Elias Iosif and Alexandros Potamianos. Unsupervised semantic similarity computation between terms using web documents. *IEEE Transactions on Knowledge and Data Engineering*, 22(11):1637–1647, 2010.

[32] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208, 1996.

[33] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[34] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[35] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[36] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247, 2014.

[37] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[38] R Harald Baayen and Rochelle Lieber. Word frequency distributions and lexical semantics. *Computers and the Humanities*, 30(4):281–291, 1996.

[39] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[40] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[41] John A Bullinaria and Joseph P Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526, 2007.

[42] Peter D Turney. The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, 33:615–655, 2008.

[43] Douwe Kiela and Stephen Clark. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL*, pages 21–30, 2014.

[44] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29. ACM, 1996.

[45] Peter D Turney and Michael L Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[46] Patrick Pantel and Dekang Lin. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619. ACM, 2002.

[47] Karl 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*, 2(11):559–572, 1901.

[48] Christos H Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 159–168. ACM, 1998.

[49] Burghard B Rieger. *On distributed representation in word semantics*. International Computer Science Institute Berkeley, CA, 1991.

[50] Sebastian Padó and Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

[51] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[52] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[53] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[54] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[55] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[56] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[57] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[58] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344, 2014.

[59] Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems*, pages 2265–2273, 2013.

[60] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014.

[61] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[62] Alastair Porter, Dmitry Bogdanov, Robert Kaye, Roman Tsukanov, and Xavier Serra. Acousticbrainz: a community platform for gathering music information obtained from audio. In *International Society for Music Information Retrieval Conference (ISMIR'15)*, 2015.

[63] Robert M French and Christophe Labiouse. Four problems with extracting human semantics from large text corpora. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 316–322. Erlbaum Mahwah, NJ, 2002.

[64] Stephen Clark and Stephen Pulman. Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium: Quantum Interaction*, pages 52–55, 2007.

[65] Peter D Turney. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416, 2006.

[66] Johansen Alexander, McCann Bryan, Bradbury James, and Socher Richard. Learning when to skim and when to read arxiv paper coming soon. *arXiv*, 2017.

[67] Harvey Fletcher. Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure. *Journal of the Acoustical Society of America*, 1934.

[68] David J Benson. Music: a mathematical offering. *The Mathematical Intelligencer*, 30(1):76–77, 2008.

[69] Francis Rumsey and Tim McCormick. *Sound and recording: an introduction*. CRC Press, 2012.

[70] James H Martin and Daniel Jurafsky. Speech and language processing. *International Edition*, 710, 2000.

[71] Alan V Oppenheim, Alan S Willsky, and Syed Hamid Nawab. *Signals and systems*, volume 2. Prentice-Hall Englewood Cliffs, NJ, 1983.

[72] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[73] Todor Ganchev, Nikos Fakotakis, and George Kokkinakis. Comparative evaluation of various mfcc implementations on the speaker verification task. In *Proceedings of the SPECOM*, volume 1, pages 191–194, 2005.

[74] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, 2000.

[75] Brian CJ Moore. *An introduction to the psychology of hearing*. Brill, 2012.

[76] Emilia Gómez. Tonal description of music audio signals. *Department of Information and Communication Technologies*, 2006.

[77] Daniel PW Ellis and Graham E Poliner. Identifyingcover songs' with chroma features and dynamic programming beat tracking. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1429. IEEE, 2007.

[78] Sebastian Ewert, Meinard Muller, and Peter Grosche. High resolution audio synchronization using chroma onset features. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1869–1872. IEEE, 2009.

[79] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences*, volume 17, pages 97–110. Amsterdam, 1993.

[80] Reinier Plomp and Willem Johannes Maria Levelt. Tonal consonance and critical bandwidth. *The journal of the Acoustical Society of America*, 38(4):548–560, 1965.

[81] William A Sethares. *Tuning, timbre, spectrum, scale*. Springer Science & Business Media, 2005.

[82] Dimitrios Giannoulis, Michael Massberg, and Joshua D Reiss. Parameter automation in a dynamic range compressor. *Journal of the Audio Engineering Society*, 61(10):716–726, 2013.

[83] Gal Chechik, Eugene Ie, Martin Rehn, Samy Bengio, and Dick Lyon. Large-scale content-based audio retrieval from text queries. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 105–112. ACM, 2008.

[84] Yusuke Uchida, Shigeyuki Sakazawa, Motilal Agrawal, and Murat Akbacak. Kddi labs and sri international at trecvid 2010: Content-based copy detection. In *TRECVID*, 2010.

[85] Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Dan Ellis, Shih-Fu Chang, Subhabrata Bhattacharya, and Mubarak Shah. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *TRECVID*, volume 2, pages 3–2, 2010.

[86] Stephanie Pancoast and Murat Akbacak. Bag-of-audio-words approach for multimedia event classification. In *Interspeech*, pages 2105–2108, 2012.

[87] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[88] Stephanie Pancoast and Murat Akbacak. Softening quantization in bag-of-audio-words. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1370–1374. IEEE, 2014.

[89] Mark Levy and Mark Sandler. A semantic space for music derived from social tags. *Austrian Compuer Society*, 1:12, 2007.

[90] Yu-Ching Lin, Yi-Hsuan Yang, and Homer H Chen. Exploiting online music tags for music emotion classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 7(1):26, 2011.

[91] Paul Lamere. Social tagging and music information retrieval. *Journal of new music research*, 37(2):101–114, 2008.

[92] Douglas R Turnbull, Luke Barrington, Gert Lanckriet, and Mehrdad Yazdani. Combining audio content and social context for semantic music discovery. In *Proceedings*

of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pages 387–394. ACM, 2009.

[93] Malcolm Slaney. Semantic-audio retrieval. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 4, pages IV–4108. IEEE, 2002.

[94] Thomas L Blum, Douglas F Keislar, James A Wheaton, and Erling H Wold. Method and article of manufacture for content-based analysis, storage, retrieval, and segmentation of audio information, June 29 1999. US Patent 5,918,223.

[95] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42, 2012.

[96] Jonathan T Foote. Content-based retrieval of music and audio. In *Voice, Video, and Data Communications*, pages 138–147. International Society for Optics and Photonics, 1997.

[97] George Tzanetakis and Perry Cook. *Manipulation, analysis and retrieval systems for audio signals*. Princeton University Princeton, NJ, USA, 2002.

[98] Adam Berenzweig, Daniel PW Ellis, and Steve Lawrence. Anchor space for classification and similarity measurement of music. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 1, pages I–29. IEEE, 2003.

[99] Thomas Lidy and Andreas Rauber. Evaluation of feature extractors and psychoacoustic transformations for music genre classification. In *ISMIR*, pages 34–41, 2005.

[100] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. 2004.

[101] Eric Scheirer and Malcolm Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1331–1334. IEEE, 1997.

[102] Dimitrios Dimitriadis, Petros Maragos, and Alexandros Potamianos. Robust am-fm features for speech recognition. *IEEE signal processing letters*, 12(9):621–624, 2005.

[103] Björn Schuller, Brüning JB Schmitt, Dejan Arsic, Stephan Reiter, Manfred Lang, and Gerhard Rigoll. Feature selection and stacking for robust discrimination of speech, monophonic singing, and polyphonic music. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 840–843. IEEE, 2005.

[104] Nima Mesgarani, Malcolm Slaney, and Shihab A Shamma. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):920–930, 2006.

[105] Thierry Bertin-Mahieux and Daniel PW Ellis. Large-scale cover song recognition using hashed chroma landmarks. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 117–120. IEEE, 2011.

[106] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[107] Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206. ACM, 2007.

[108] Josef Sivic, Andrew Zisserman, et al. Video google: A text retrieval approach to object matching in videos. In *iccv*, volume 2, pages 1470–1477, 2003.

[109] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 2161–2168. Ieee, 2006.

[110] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Image classification using random forests and ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[111] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[112] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics, 2012.

[113] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[114] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[115] Mark D Fairchild. Status of cie color appearance models. In *9th Congress of the International Color Association*, pages 550–553. International Society for Optics and Photonics, 2002.

[116] Bela Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981.

[117] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[118] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[119] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1458–1465. IEEE, 2005.

[120] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[121] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. *Journal of machine learning research*, 3(Feb):1107–1135, 2003.

[122] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[123] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013.

[124] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010.

[125] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010.

[126] Lewis D Griffin, M Husni Wahab, and Andrew J Newell. Distributional learning of appearance. *PloS one*, 8(2):e58074, 2013.

[127] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[128] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[129] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[130] Yoshua Bengio, Ian J Goodfellow, and Aaron Courville. Deep learning. *Nature*, 521:436–444, 2015.

[131] Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[132] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 2016.

[133] Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *Proceedings of the 21st acm international conference on multimedia*, pages 411–412. ACM, 2013.

[134] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, 2015.

[135] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[136] Theodoros Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12):e0144610, 2015.

[137] Theodoros Giannakopoulos. Study and application of acoustic information for the detection of harmful content, and fusion with visual information. *Department of Informatics and Telecommunications, vol. PhD. University of Athens, Greece*, 2009.

[138] Giannis Karamanolakis, Elias Iosif, Athanasia Zlatintsi, Aggelos Pikrakis, and Alexandros Potamianos. Audio-based distributional representations of meaning using a fusion of feature encodings. *Interspeech 2016*, pages 3658–3662, 2016.

[139] Mark Levy and Mark Sandler. Music information retrieval using social tags and audio. *IEEE Transactions on Multimedia*, 11(3):383–395, 2009.

[140] Joon Hee Kim, Brian Tomasik, and Douglas Turnbull. Using artist similarity to propagate semantic information. In *ISMIR*, volume 9, pages 375–380, 2009.

[141] Michael I Mandel, Razvan Pascanu, Douglas Eck, Yoshua Bengio, Luca M Aiello, Rossano Schifanella, and Filippo Menczer. Contextual tag inference. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 7(1):32, 2011.

[142] Klaus Seyerlehner, Markus Schedl, Peter Knees, and Reinhard Sonnleitner. A refined block-level feature set for classification, similarity and tag prediction. *Extended Abstract to MIREX*, 2011.

[143] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008.

[144] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[145] Philippe Hamel, Simon Lemieux, Yoshua Bengio, and Douglas Eck. Temporal pooling and multiscale learning for automatic annotation and ranking of music audio. In *ISMIR*, pages 729–734, 2011.

[146] Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6964–6968. IEEE, 2014.

[147] Jordi Pons, Olga Slizovskaia, Rong Gong, Emilia Gómez, and Xavier Serra. Timbre analysis of music audio signals with convolutional neural networks. *arXiv preprint arXiv:1703.06697*, 2017.

[148] Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*, 2016.

[149] Brian McFee, Luke Barrington, and Gert Lanckriet. Learning content similarity for music recommendation. *IEEE transactions on audio, speech, and language processing*, 20(8):2207–2218, 2012.

[150] Sebastian Stober and Andreas Nürnberger. An experimental comparison of similarity adaptation approaches. In *International Workshop on Adaptive Multimedia Retrieval*, pages 96–113. Springer, 2011.

[151] Daniel PW Ellis, Brian Whitman, Adam Berenzweig, and Steve Lawrence. The quest for ground truth in musical artist similarity. In *ISMIR*. Paris, France, 2002.

[152] Beth Logan, Daniel PW Ellis, and Adam Berenzweig. Toward evaluation techniques for music similarity. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003): Workshop on the Evaluation of Music Information Retrieval Systems*, 2003.

[153] Edith Law and Luis Von Ahn. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1197–1206. ACM, 2009.

[154] Brian McFee and Gert RG Lanckriet. Heterogeneous embedding for subjective artist similarity. In *ISMIR*, pages 513–518, 2009.

[155] Michal Aharon, Michael Elad, and Alfred Bruckstein. *rmk*-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[156] Ron Rubinstein, Alfred M Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[157] Patrick Ng. dna2vec: Consistent vector representations of variable-length k-mers. *arXiv preprint arXiv:1701.06279*, 2017.

[158] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864. ACM, 2016.

[159] Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals. Learning the speech front-end with raw waveform cldnns. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[160] Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Bjorn Schuller, and Shrikanth Narayanan. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing*, 3(2):184–198, 2012.

[161] Jinkyu Lee and Ivan Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In *INTERSPEECH*, pages 1537–1540, 2015.

[162] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. Representation learning for speech emotion recognition. *Interspeech 2016*, pages 3603–3607, 2016.

[163] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention.

[164] Yun Wang, Leonardo Neves, and Florian Metze. Audio-based multimedia event detection using deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 2742–2746. IEEE, 2016.

[165] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. Recurrent neural networks for polyphonic sound event detection in real life recordings. In

Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, pages 6440–6444. IEEE, 2016.

[166] Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. Sensicon: An automatically constructed sensorial lexicon. 2014.

[167] Stevan Harnad. Grounding symbols in the analog world with neural nets. *Think*, 2(1):12–78, 1993.

[168] Morten H Christiansen and Nick Chater. Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences*, 5(2):82–88, 2001.