Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
τομεας σηματων, ελεγχου και ρομποτικης

Μη Παράλληλη, από Πολλά σε Πολλά, Μετατροπή Συναισθηματικής Ομιλίας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κωνσταντίνου Κλάψα Πουλογιάννη

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

ΕΡΓΑΣΤΗΡΙΟ ΕΠΕΞΕΡΓΑΣΙΑΣ ΦΩΝΗΣ ΚΑΙ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ
Αθήνα, Νοέμβριος 2023

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Επεξεργασίας Φωνής και Φυσικής Γλώσσας

Μη Παράλληλη, από Πολλά σε Πολλά, Μετατροπή Συναισθηματικής Ομιλίας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κωνσταντίνου Κλάψα Πουλογιάννη

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 1η Νοεμβρίου, 2023.

.....	.....	.....
Αλέξανδρος Ποταμιάνος	Γιώργος Στάμου	Κωνσταντίνος Τζαφέστας
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2023

.....

Κωνσταντινος Κλαψας Πουλογιαννης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Στην μητέρα μου

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με το πρόβλημα της μετατροπής συναισθηματικής φωνής, όπου το ζητούμενο είναι να μετατραπεί μία εκφώνηση που ειπώθηκε με ένα δεδομένο συναίσθημα σε μία εκφώνηση που ακούγεται σαν να ειπώθηκε με ένα άλλο δεδομένο συναίσθημα, χωρίς να παραμορφωθεί το περιεχόμενο της πρότασης. Επιπλέον, το μοντέλο που επιλύει αυτό το πρόβλημα εκπαιδεύεται χωρίς τη βοήθεια ενός παράλληλου συνόλου δεδομένων, όπου η ίδια έκφραση έχει ειπωθεί με διαφορετικά συναισθήματα, και χωρίς καμία πληροφορία κειμένου. Επομένως, η μόνη απαίτηση για την εργασία αυτή είναι ένα σύνολο δεδομένων συναισθηματικής ομιλίας, όχι κατ' ανάγκη μεταγραμμένο, αλλά με επισημειωμένα τα συναισθήματα.

Η αρχιτεκτονική που χρησιμοποιήθηκε ως βάση για την παρούσα εργασία βασίζεται στο StarGAN-VC, ένα μοντέλο βαθύ νευρωνικού δικτύου που μαθαίνει από πολλά σε πολλά αντιστοιχίσεις μεταξύ των φασματικών χαρακτηριστικών των πεδίων του συνόλου δεδομένων. Η εκπαίδευση γίνεται χρησιμοποιώντας το πλαίσιο των GAN, όπου το μοντέλο μετατροπής προσπαθεί να ξεγελάσει ένα μοντέλο διάκρισης ώστε να αντιληφθεί την έξοδό του ως διαφορετικό πεδίο από αυτό της εισόδου. Η αρχική χρήση αυτού του μοντέλου ήταν στην μετατροπή χροιάς φωνής του ομιλητή αλλά εμείς το εφαρμόζουμε στη μετατροπή συναισθήματος.

Στην συνέχεια, προτείνεται μια τροποποίηση αυτής της αρχιτεκτονικής, στην οποία η ομιλία εισόδου μετασχηματίζεται πρώτα σε έναν ανεξάρτητο από το συναίσθημα χώρο, διατηρώντας όμως όλο το περιεχόμενο της ομιλίας, πριν από την αποκωδικοποίηση στο συναίσθημα-στόχο. Ο μετασχηματισμός σε αυτόν τον ουδέτερο χώρο γίνεται με την βοήθεια ανταγωνιστικής εκπαίδευσης.

Δεδομένου ότι η θεμελιώδης συχνότητα είναι σημαντικό χαρακτηριστικό της συναισθηματικής ομιλίας, και επειδή και στα δύο προηγούμενα μοντέλα ο μετασχηματισμός της γίνεται από απλή κανονικοποίηση στο ζητούμενο συναίσθημα, δοκιμάζεται και μια περαιτέρω προσαρμογή, στην οποία η θεμελιώδης συχνότητα του σήματος μετασχηματίζεται με νευρωνικά δίκτυα.

Διεξάγουμε αντικειμενική αξιολόγηση στα μοντέλα, σε δύο βάσεις δεδομένων, μια ελληνική και μια αγγλική, με πέντε και επτά συναισθήματα αντίστοιχα. Η αξιολόγηση αποτελείται από μετρικές ανακατασκευής καθώς και την αξιολόγηση της ποιότητας και την ταξινόμηση συναισθήματος από προεκπαιδευμένα νευρωνικά μοντέλα. Επίσης, διεξάγουμε υποκειμενική αξιολόγηση στην ελληνική βάση δεδομένων, για την οποία χρησιμοποιούμε 25 ακροατές οι οποίοι βαθμολογούν την ποιότητα των συνθετικών προτάσεων καθώς και το συναίσθημα με το οποίο πιστεύουν ότι ειπώθηκε.

Με βάση τις αντικειμενικές αξιολογήσεις και στις δύο βάσεις δεδομένων, η ικανότητα μετατροπής συναισθημάτων του προτεινόμενου μοντέλου φαίνεται να υπερέχει του βασικού μοντέλου, όμως με ταυτόχρονη μικρή μείωση της ποιότητας. Αντίστοιχα, το μοντέλο που αξιοποιεί την θεμελιώδη συχνότητα έχει ακόμα καλύτερη ικανότητα μετατροπής, με αντίστοιχα μεγαλύτερη πτώση ποιότητας. Οι υποκειμενικές αξιολογήσεις φαίνεται να υποστηρίζουν αυτά τα συμπεράσματα, με την διαφορά ότι δεν δείχνουν σημαντική διαφορά μεταξύ του βασικού μοντέλου και του προτεινόμενου σε ό,τι αφορά την ποιότητα.

Λέξεις Κλειδιά —Συναισθηματική μετατροπή φωνής, Συναισθηματική ομιλία, Μοντελοποίηση ομιλίας, Generative Adversarial Networks, Μηχανική Μάθηση, Βαθιά Νευρωνικά Δίκτυα, Autoencoders, Μετατροπή φωνής, Μη παράλληλη μετατροπή

Abstract

This thesis tackles the problem of emotional voice conversion, where the task is to convert an utterance spoken with a given emotion to an utterance which sounds like it was uttered with another given emotion, without distorting the content of the sentence. Moreover, the model that solves this task is to be trained without the aid of a parallel dataset where the same utterance has been spoken with different emotions, and without any textual information. Therefore, the only requirement for this work is an emotional speech corpus, not necessarily transcribed, but with labeled emotions.

The architecture used as a baseline for this work is based on StarGAN-VC, a deep neural network model that learns many to many mappings between the spectral features of the domains of the dataset. The training is done using the GAN framework, where the conversion model tries to fool a discriminator model to perceive it's output as a different domain than the one on its input. The original use of this model was in speaker voice conversion but we apply it to emotion conversion.

Furthermore, a modification of this architecture is proposed, in which the input speech is first transformed into an emotion independent space, while retaining the content of the utterance, before being decoded to the target emotion. The transformation to this emotion independent space is done with the use of adversarial training.

Since fundamental frequency plays a very prominent role in emotional speech, and since in both of these models its transformation is done by simple normalization to the target emotion, a further adaptation is tested, in which the fundamental frequency of the signal is transformed by neural networks.

We conduct objective evaluation on the models in two databases, one Greek and one English, with five and seven emotions respectively. The evaluation consists of reconstruction metrics as well as quality assessment and emotion classification from pre-trained neural models. We also conduct subjective evaluation on the Greek database, for which we use 25 listeners who rate the quality of the synthesized sentences as well as the emotion with which they believe it was said.

Based on the objective evaluations on both databases, the emotion conversion ability of the proposed model seems to outperform the baseline model, but with a slight decrease in quality. Similarly, the model utilizing the fundamental frequency has an even better conversion ability, with a correspondingly larger drop in quality. The subjective evaluations seem to support these conclusions, except that they do not show a significant difference between the baseline model and the proposed model in terms of quality.

Keywords — Emotional Voice Conversion, Emotional Speech, Speech Modeling, Generative Adversarial Networks, Machine Learning, Deep Neural Networks, Autoencoders, Voice Conversion, Non parallel Conversion

Ευχαριστίες

Θα ήθελα καταρχάς να ευχαριστήσω τον καθηγητή κ. Α. Ποταμιάνο για την επίβλεψη αυτής της διπλωματικής εργασίας καθώς και για την ευκαιρία που μου έδωσε να την εκπονήσω στο εργαστήριο Επεξεργασίας Φωνής και Φυσικής Γλώσσας. Επίσης ευχαριστώ τους Γ. Παρασκευόπουλο και Ν.Έλληνα για την καθοδήγησή τους στην εργασία και για τις πολλές ενδιαφέρουσες συζητήσεις. Τέλος θα ήθελα να ευχαριστήσω την οικογένεια μου για την υποστήριξη της καθ όλη την μεγάλη διάρκεια των σπουδών μου καθώς και τους φίλους μου και ιδιαιτέρως αυτούς που απείλησαν ότι δεν θα μου ξαναμιλήσουν αν δεν τελειώσω επιτέλους την διπλωματική εργασία, χωρίς τους οποίους μάλλον δεν θα έπαιρνα ποτέ πτυχίο.

Κωνσταντίνος Κλάψας Πουλογιάννης
Οκτώβριος 2023

Περιεχόμενα

Λίστα Σχημάτων

Κατάλογος Πινάκων

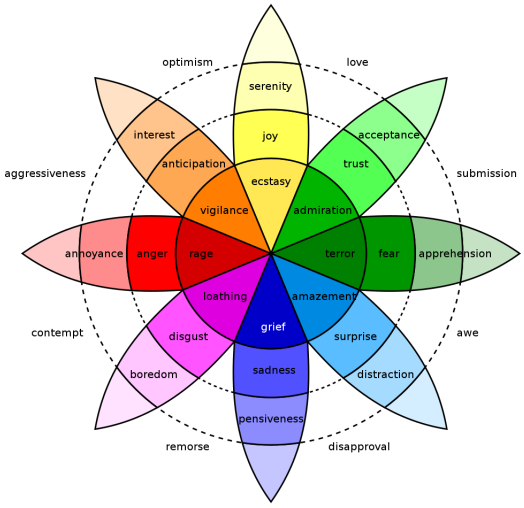
Κεφάλαιο 1

Εισαγωγή

1.1 Συναίσθημα στην Ομιλία

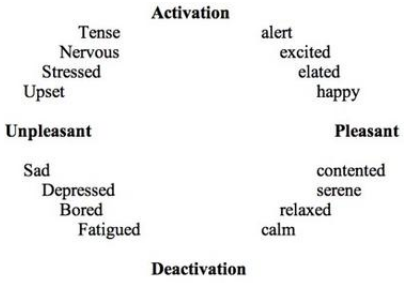
Σύμφωνα με τον Ekman, τα συναισθήματα είναι διακριτά, διαχωρίσιμα όσο αφορά την φυσιολογία και καθολικά καθώς είναι αναγνωρίσιμα και από κουλτούρες που δεν έχουν έρθει σε επαφή με τα μέσα μαζικής ενημέρωσης ώστε να μάθουν τις διασυνδέσεις συναισθημάτων-εκφράσεων που υπάρχουν στις υπόλοιπες κοινωνίες. Με μελέτες εκφράσεων προσώπου εντόπισε τα εξής 6 βασικά συναισθήματα: θυμός (anger), απέχθεια (disgust), φόβος (fear), χαρά (happiness), λύπη (sadness) και έκπληξη (surprise) [1].

Ο Robert Plutchik επέκτεινε το μοντέλο του Ekman προτείνοντας 8 πρωτεύοντα συναισθήματα, τα οποία όμως αποτελούν 4 ζευγάρια αντιθέτων: χαρά (joy) με λύπη (sadness), θυμός (anger) με φόβο (fear), εμπιστοσύνη (trust) με απέχθεια (disgust) και έκπληξη (surprise) με προσμονή (anticipation) [2]. Τα βασικά συναισθήματα σύμφωνα με τον Plutchik μπορούν να απεικονιστούν σε ένα τροχό συναισθημάτων που φαίνεται στο Σχήμα 1.1.1 και δείχνει όλα τα βασικά συναισθήματα σε διαφορετικές εντάσεις. Περίπλοκα συναισθήματα μπορεί να προκύ-ψουν από τον συνδυασμό βασικών συναισθημάτων, και 8 συνδυασμοί από δύο εκ των βασικών συναισθημάτων απεικονίζονται και στο σχήμα.



Σχήμα 1.1.1: Συναισθήματα κατά τον Plutchik

Μια διαφορετική προσέγγιση μοντελοποίησης των συναισθημάτων προκύπτει υιοθετώντας ένα συνεχές μοντέλο. Το πιο συνηθισμένο τέτοιο μοντέλο είναι το δισδιάστατο μοντέλο το οποίο απεικονίζεται στο Σχήμα 1.1.2 [3]. Ο κατακόρυφος άξονας ονομάζεται arousal και αντιστοιχεί στο πόση "ενέργεια" περιλαμβάνει το συναίσθημα, και ο οριζόντιος άξονας ονομάζεται valence και αντιστοιχεί στο πόσο θετικό ή αρνητικό είναι το συναίσθημα. Οι ενδιάμεσες περιοχές του σχήματος περιλαμβάνουν τα πιο ουδέτερα συναισθήματα.



Σχήμα 1.1.2: Δισδιάστατη αναπαράσταση των συναισθημάτων

Στην περίπτωση της ομιλίας, θεωρούμε πως τα συναισθήματα αποτελούν κομμάτι της προσωδίας, των

χαρακτηριστικών δηλαδή της ομιλίας που δεν συσχετίζονται με το γλωσσολογικό περιεχόμενο. Φυσικά στην πραγματική ζωή η προσωδία χρησιμοποιείται σε συνδυασμό με την επιλογή των λέξεων για την έκφραση των συναισθημάτων, και δεν μπορεί εύκολα να θεωρηθεί ανεξάρτητη της εννοιολογικής σημασίας της πρότασης. Ταυτόχρονα όμως, σε σχετικά πειράματα [4] έχει δειχθεί ότι οι άνθρωποι μπορούν να ξεχωρίσουν συναισθήματα στην ομιλία ακόμα και αν έχει ειπωθεί χωρίς γλωσσολογικό περιεχόμενο, κάτι που σε μεγάλο βαθμό δικαιολογεί αυτή την υπόθεση. Χαρακτηριστικά της προσωδίας τα οποία περιέχουν πληροφορία για το συναίσθημα είναι η θεμελιώδης συχνότητα (ύψος της φωνής), η ενέργεια, ο ρυθμός αλλά και χαρακτηριστικά όπως τα formants.

Στα πλαίσια αυτής της εργασίας, τα συναισθήματα θεωρούνται διακριτές κατηγορίες και χωρίς να υπάρχει επικάλυψη μεταξύ τους. Παρόλο που στην γενική περίπτωση ομιλίας αυτή η υπόθεση δεν είναι σωστή, στην δική μας περίπτωση δικαιολογείται από τις βάσεις δεδομένων που χρησιμοποιούνται, οι οποίες περιλαμβάνουν ομιλίες οι οποίες έχουν ηχογραφηθεί με σκοπό να εκφράζουν ένα συγκεκριμένο συναίσθημα. Για την επικάλυψη όλου του εύρους των μη διακριτών συναισθημάτων χρειάζεται ένα μεγάλο dataset, το οποίο όμως γενικά δεν υπάρχει διαθέσιμο. Μια εξαίρεση είναι το dataset IEMOCAP [5] που όμως έχει σημαντικό πρόβλημα επισήμανσης των συναισθημάτων, χαμηλή σχετικά ποιότητα της ηχογράφησης αλλά και πολλούς ηθοποιούς (μερικές φορές και στην ίδια πρόταση), όλα χαρακτηριστικά που περιπλέκουν το πρόβλημα. Η χρησιμοποίηση διακριτών και αμοιβαία αποκλειομένων συναισθημάτων έγινε επομένως σαν μια υπόθεση απλοποίησης του προβλήματος, αντίστοιχα με την υπόθεση ανεξαρτησίας γλωσσολογικού περιεχομένου και συναισθήματος.

Η επιλογή του συγκεκριμένου συνόλου συναισθημάτων που θα χρησιμοποιηθούν, καθορίζεται από την βάση των δεδομένων στην οποία γίνεται η εκπαίδευση και η αξιολόγηση. Οι δύο βάσεις που χρησιμοποιούμε είναι οι CVSP_EAV [6] και TESS [7] οι οποίες θα αναλυθούν με λεπτομέρεια στην συνέχεια. Η πιο πλούσια σε αριθμό συναισθημάτων είναι το TESS η οποία περιλαμβάνει τα συναισθήματα θυμός, χαρά, λύπη, ουδέτερο, απέχθεια, ευχάριστη έκπληξη και φόβος, τα οποία αντιστοιχούν αρκετά καλά με τα συναισθήματα του Ekman. Δυστυχώς όμως, λόγω άλλων προβλημάτων αυτής της βάσης, για τα περισσότερα πειράματα μας χρησιμοποιούμε κυρίως την βάση CVSP_EAV η οποία περιέχει μόνο τα τέσσερα πρώτα από αυτά τα συναισθήματα, το οποίο περιορίζει σε κάποιο βαθμό τα συμπεράσματα που μπορούμε να βγάλουμε.

1.2 Μετατροπή Φωνής

Με την ανάπτυξη των νέων μεθόδων μηχανικής μάθησης, έχει καταστεί εφικτή η επεξεργασία δεδομένων με βάση ιεραρχικά υψηλών χαρακτηριστικών τους. Ένα τέτοιο πεδίο εφαρμογών είναι η μετατροπή δεδομένων από ένα πεδίο σε κάποιο άλλο, ειδικά σε περιπτώσεις όπου τα πεδία δεν είναι καλώς ορισμένα, και έχουν οριστεί μόνο έμμεσα, με βάση ασαφείς κατηγορίες που έχουν προκύψει από την φυσική γλώσσα των ανθρώπων. Παραδείγματα τέτοιων εφαρμογών είναι η μετατροπή μουσικής σε διαφορετικό είδος ή η μετατροπή ενός πίνακα σε διαφορετικό στυλ ζωγραφικής, δυο παραδείγματα τα οποία ενδεικνύουν την ασάφεια που μπορεί να υπάρχει στον ορισμό των κατηγοριών των δεδομένων. Ένα παράδειγμα που η μετατροπή είναι πολύ πιο εύκολα ορισμένη είναι η υπερδειγματοληψία εικόνων ή ήχου σε υψηλότερη ανάλυση.

Πιο συγκεκριμένα, η μετατροπή φωνής ασχολείται με την μετατροπή των χαμηλών χαρακτηριστικών μιας ομιλίας, με τέτοιο τρόπο ώστε να αλλάξουν συγκεκριμένα υψηλότερα χαρακτηριστικά της ομιλίας αλλά όλα τα υπόλοιπα να παραμείνουν ίδια. Παραδείγματα αποτελούν την μετατροπή χροιάς ενός ομιλητή (voice conversion) ή την αντιγραφή χροιάς του ομιλητή (voice cloning), την μετατροπή της προφοράς του ομιλητή (accent conversion), ή την μετατροπή του συναισθήματος με το οποίο ειπώθηκε η πρόταση (emotional voice conversion), το θέμα που θα μας απασχολήσει σε αυτή την εργασία.

Σε γενικές γραμμές η μετατροπή φωνής (και γενικά οποιοδήποτε δεδομένων) χωρίζεται σε δύο κατηγορίες, στην παράλληλη και στην μη παράλληλη. Στην πρώτη περίπτωση, υπάρχει μια διαθέσιμη βάση δεδομένων η οποία περιλαμβάνει ζευγάρια από την ζητούμενη μετατροπή. Ο τρόπος εκπαίδευσης ενός τέτοιου μοντέλου είναι γενικά πιο απλός, καθώς κατά την διάρκεια της εκπαίδευσης, υπάρχει διαθέσιμος ο στόχος της μετατροπής. Αντίθετα, στην περίπτωση της μη παράλληλης βάσης δεδομένων, έχουμε δεδομένα από κάθε ξεχωριστό πεδίο, αλλά χωρίς καμμία συσχέτιση μεταξύ τους. Κατά συνέπεια, είναι αρκετά πιο δύσκολο να εκπαιδευτεί ένα μοντέλο που να εκτελεί την μετατροπή καθώς δεν είναι προφανές ποια χαρακτηριστικά του δεδομένου πρέπει να αλλοιωθούν κατά την μετατροπή και ποια όχι.

Μια διαφορετική ταξινόμηση της μετατροπής μπορεί να γίνει ανάλογα με το πόσα πεδία είναι δυνατόν να καλύψει το μοντέλο που εκτελεί την μετατροπή. Ένα προς ένα μετατροπή (one to one) είναι η περίπτωση που έχουμε

ένα μόνο πεδίο εισόδου και ένα εξόδου (π.χ. η μετατροπή της ομιλίας ενός συγκεκριμένου ομιλητή σε έναν συγκεκριμένο άλλο, ή η μετατροπή ομιλίας από άντρα σε γυναίκα). Αντίστοιχα, από πολλά σε πολλά (many to many) ονομάζεται η μετατροπή στην οποία η είσοδος είναι μια από έναν συγκεκριμένο αριθμό από πεδία, και η έξοδος είναι επίσης ένα από τα ίδια πεδία (η έξοδος ονομάζεται και στόχος). Με τον ίδιο τρόπο, ορίζονται και οι μετατροπές από ένα σε πολλά ή από πολλά σε ένα. Μία τελευταία περίπτωση είναι η μετατροπή από οποιοδήποτε σε οποιοδήποτε (any to any), φυσικά και σε συνδυασμό με τις προηγούμενες περιπτώσεις. Αυτό αφορά τις περιπτώσεις που τα πεδία είναι μεν διακριτά, αλλά δεν είναι απαραίτητα μόνο αυτά που χρησιμοποιούνται κατά την εκπαίδευση των μοντέλων. Αυτή είναι και η πιο δύσκολη περίπτωση, καθώς απαιτεί το μοντέλο να μάθει τα χαρακτηριστικά με τέτοιο τρόπο ώστε να γενικεύεται σε πεδία που δεν έχει δει ποτέ. Ένα τέτοιο παράδειγμα είναι η μετατροπή του ομιλητή σε ομιλητή που δεν έχει δει ποτέ κατά την εκπαίδευση. Στην περίπτωση που η μετατροπή αυτή γίνεται σε εφαρμογή σύνθεσης ομιλίας από κείμενο (text to speech) αυτό αναφέρεται σαν κλωνοποίηση ομιλητή (voice cloning).

Η περίπτωση που θα μας απασχολήσει σε αυτή την εργασία είναι η περίπτωση της μη παράλληλης, από πολλά σε πολλά μετατροπή συναισθηματικής ομιλίας. Αυτό επιλέχθηκε αφενός γιατί τα συναισθήματα στην εργασία αυτή μοντελοποιήθηκαν σαν διακριτά και σχετικά περιορισμένα, αφετέρου γιατί η παράλληλη βάση δεδομένων είναι αρκετά περιοριστική.

Επιπλέον, τα μόνα χαρακτηριστικά που μπορεί να χρησιμοποιήσει το μοντέλο είναι η κυματομορφή και όποια ακουστικά χαρακτηριστικά μπορούν να εξαχθούν άμεσα από αυτή, καθώς και το συναίσθημα της ομιλίας εισόδου. Το τελευταίο χαρακτηριστικό είναι επίσης ένα μη ρεαλιστικό χαρακτηριστικό για οποιαδήποτε πιθανή εφαρμογή της εργασίας, αλλά αν δεχθούμε ότι η ταξινόμηση συναισθημάτων ομιλίας είναι ένα ευκολότερο πρόβλημα από την μετατροπή των συναισθημάτων, μπορούμε να υποθέσουμε ότι έχουμε στην διάθεσή μας ένα τέτοιο μοντέλο ταξινόμησης το οποίο θα δημιουργήσει μικρή σχετικά διαφορά στην ποιότητα των μοντέλων.

1.3 Συνεισφορά Εργασίας

Η συνεισφορά αυτής της εργασίας, είναι στην αξιοποίηση μοντέλων που έχουν ήδη αναπτυχθεί για την μετατροπή ομιλητή στην μετατροπή συναισθήματος, και στην μετέπειτα βελτίωση αυτών των μοντέλων. Η αξιολόγηση των μοντέλων γίνεται με την χρήση αντικειμενικών μετρικών αξιολόγησης, προεκπαιδευμένων μοντέλων μηχανικής μάθησης και υποκειμενικής αξιολόγησης των δειγμάτων από ακροατές.

Προτείνουμε δυο διαφορετικούς τρόπους αντιμετώπισης του προβλήματος, ο πρώτος εκ των οποίων προσφέρει καλύτερη ικανότητα μετατροπής, χωρίς σημαντική επιδείνωση της ποιότητας της ομιλίας και ο δεύτερος σημαντικά καλύτερη ικανότητα μετατροπής που όμως συνοδεύεται από ανάλογη πτώση ποιότητας.

Παρόλο που δεν υπάρχει κάποια προφανής αξιοποίηση των μοντέλων αυτής της εργασίας με οικονομικό ή πρακτικό όφελος, αποτελεί ένα αρκετά ενδιαφέρον ερευνητικό θέμα, σε μεγάλο βαθμό εξαιτίας της υποκειμενικής φύσης των συναισθημάτων. Επιπλέον, οι μέθοδοι που αναπτύσσονται και συγκρίνονται για την αντιμετώπιση του προβλήματος, είναι πιθανόν να έχουν εφαρμογή και σε άλλα προβλήματα μετατροπής φωνής.

1.4 Περίγραμμα Εργασίας

Η οργάνωση αυτής της εργασίας είναι ως εξής:

- Στο Κεφάλαιο 2 αναλύονται τα χαρακτηριστικά της ανθρώπινης φωνής τα οποία χρησιμοποιούνται σε αυτήν την εργασία καθώς και ο τρόπος εξαγωγής τους από τα ηχητικά σήματα, για την μετέπειτα αξιοποίησή τους από τον υπολογιστή.
- Στο Κεφάλαιο 3 περιγράφονται οι βασικές τεχνικές της μηχανικής μάθησης καθώς και βασικές δομές νευρωνικών δικτύων, οι οποίες είναι απαραίτητες για την κατανόηση της εργασίας.
- Στο Κεφάλαιο 4 γίνεται μια σύντομη επισκόπηση της χρήσης νευρωνικών δικτύων για την επεξεργασία φωνής.
- Το Κεφάλαιο 5 κάνει μια εισαγωγή στην δομή των Generative Adversarial Networks καθώς και σε σημαντικές παραλλαγές οι οποίες χρησιμοποιούνται εκτενώς στην εργασία.

- Το Κεφάλαιο 6 περιλαμβάνει την συνεισφορά μας στο πεδίο της μετατροπής συναισθηματικής ομιλίας με επεξήγηση των μοντέλων μας και του τρόπου αξιολόγησης τους, και σχολίαση των αποτελεμάτων της αξιολόγησης.
- Το Κεφάλαιο 7 συνοψίζει τα αποτελέσματα της εργασίας καθώς και πιθανές μελλοντικές προεκτάσεις.

Κεφάλαιο 2

Χαρακτηριστικά Φωνής

2.1 Εισαγωγή

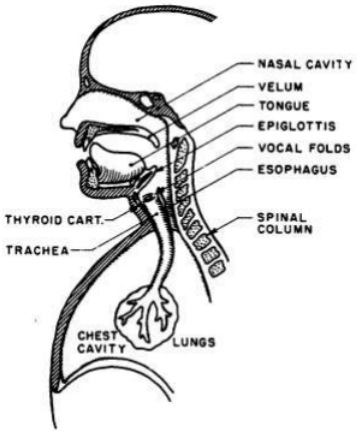
Η φωνή, όπως και κάθε ήχος, αποτελείται από ταλαντώσεις πίεσης που διαδίδονται σε ένα μέσο μετάδοσης όπως ο αέρας. Αυτές οι ταλαντώσεις μεταδίδονται στο τύμπανο του αυτιού, κάτι που μετατρέπει την ακουστική ενέργεια του σήματος σε μηχανική ενέργεια της μεμβράνης του τυμπάνου. Η μηχανική αυτή ενέργεια μεταβιβάζεται με την βοήθεια των ακουστικών οστών στο έσω αυτί, όπου η δόνηση του κοχλία προκαλεί δόνηση στο όργανο του Corti με αποτέλεσμα τη μετατροπή της ενέργειας σε ηλεκτρική. Τέλος, πραγματοποιείται η μεταβίβαση της ηλεκτρικής ταλάντωσης μέσω του κοχλιακού νεύρου στον ακουστικό φλοιό του εγκεφάλου που βρίσκεται στον κροταφικό λοβό [8]. Εκεί θα γίνει η αντίληψη του ήχου ως το αίσθημα της ακοής.

Τα μέλη του σώματος τα οποία είναι υπεύθυνα για την παραγωγή της ανθρώπινης φωνής φαίνονται στο Σχήμα 2.1.1. Η βάση για την ομιλία είναι η ελεγχόμενη κίνηση του αέρα από τους πνεύμονες. Ο αέρας αφού βγει από τους πνεύμονες περνάει από το λάρυγγα, ο οποίος περιέχει τις φωνητικές χορδές. Στην συνέχεια, κινείται προς τα έξω μέσα από τη στοματική κοιλότητα και εξέρχεται από τα χείλη ή από τη ρινική κοιλότητα όπου εξέρχεται από τα ρουθούνια. Η φωνητική οδός είναι λοιπόν η περιοχή που ακολουθεί το ρεύμα αέρος από το λάρυγγα έως ότου βγει έξω από το ανθρώπινο σώμα.

Όταν οι φωνητικές χορδές είναι ανοικτές, όπως είναι όταν αναπνέουμε, τότε ο αέρας περνάει ανεμπόδιστος στη φωνητική οδό και ο παραγόμενος φθόγγος είναι άηχος. Αντίθετα, κατά την άρθρωση των ηχηρών φθόγγων οι φωνητικές χορδές είναι έτσι τοποθετημένες ώστε να εφάπτονται μεταξύ τους με αποτέλεσμα ο αέρας από τους πνεύμονες να βρίσκει εμπόδιο. Η πίεση πίσω από τις φωνητικές χορδές αυξάνεται μέχρι αυτές να ανοίξουν. Με το άνοιγμά τους όμως η πίεση μειώνεται και το γεγονός αυτό σε συνδυασμό με την ελαστική φύση των φωνητικών χορδών προκαλεί το κλείσιμό τους μέχρι να ανοίξουν και πάλι και να επαναληφθεί ο κύκλος. Το αποτέλεσμα αυτής της διαδικασίας είναι η δημιουργία φώνησης, ενός περιοδικού ήχου πλούσιου σε αρμονικές συχνότητες.

Η άρθρωση των συμφώνων και των φωνηέντων έχει σημαντικές διαφορές. Κατά την άρθρωση των συμφώνων η ροή του αέρα από τους πνεύμονες συναντά κάποιο εμπόδιο στη φωνητική οδό από τα όργανα που συμμετέχουν στην άρθρωση των φθόγγων, τους αρθρωτές. Τα σύμφωνα διαφοροποιούνται μεταξύ τους ανάλογα με το σημείο στη φωνητική οδό όπου εμφανίζεται το εμπόδιο στη ροή του αέρα, τον λεγόμενο τόπο άρθρωσης και το είδος του εμποδίου, τον λεγόμενο τρόπο άρθρωσης. Παράδειγμα τόπων άρθρωσης είναι τα χείλη, τα δόντια και ο ουρανίσκος. Παράδειγμα τρόπων άρθρωσης είναι όταν η δίοδος του αέρα κλείνει τελείως, όταν ο αέρας διαφεύγει από την μύτη και όταν η ροή του αέρα κάνει τους αρθρωτές να πάλλονται τρεις-τέσσερις φορές, να ενώνονται δηλαδή και απομακρύνονται.

Κατά την άρθρωση των φωνηέντων, η ροή του αέρα από τους πνεύμονες θέτει σε κίνηση τις φωνητικές χορδές αλλά, σε αντίθεση με ό,τι συμβαίνει κατά την άρθρωση των συμφώνων, δε συναντά κανένα εμπόδιο στην πορεία της στη στοματική κοιλότητα. Οι διαφορές ποιότητας ανάμεσα στα φωνήεντα είναι αποτέλεσμα του σχήματος της στοματικής κοιλότητας και πιο συγκεκριμένα της θέσης της γλώσσας και του σχήματος των χειλιών κατά την παραγωγή τους.



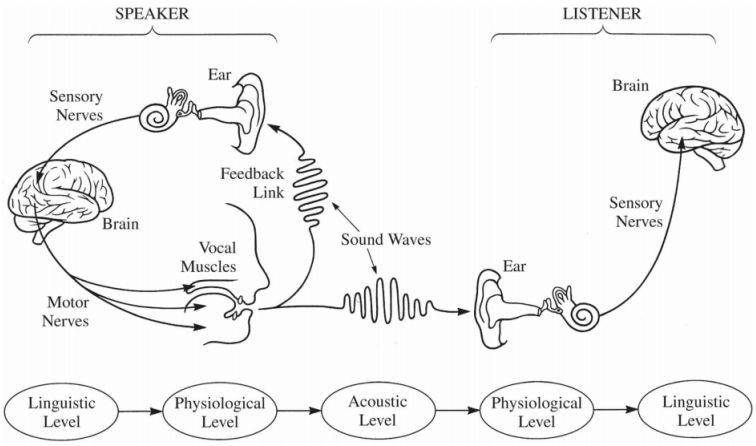
Σχήμα 2.1.1: Σχηματική αναπαράσταση της ανθρώπινης φωνητικής οδού [9]

Μία χρήσιμη απλουστευμένη μοντελοποίηση της διαδικασίας παραγωγής φωνής είναι το μοντέλο πηγής-φίλτρου (source filter model) [10]. Σύμφωνα με αυτό το μοντέλο η παραγωγή φωνής χωρίζεται στην πηγή που δημιουργεί ένα περιοδικό ή μη απεριοδικό σήμα, και στο φίλτρο το οποίο είναι ένα γραμμικό φίλτρο που ενισχύει και εξασθενεί τις απαραίτητες συχνότητες ώστε το τελικό σήμα να έχει τα σωστά χαρακτηριστικά. Συγκεκριμένα, η πηγή είναι οι φωνητικές χορδές που μπορούν να δημιουργήσουν ένα περιοδικό σήμα όταν είναι σφιγμένες με τον τρόπο που περιγράφηκε προηγουμένως, ή ένα μη περιοδικό σήμα (λευκό θόρυβο) όταν είναι χαλαρές. Το φίλτρο είναι το υπόλοιπο της φωνητικής οδού που αλλάζει σχήμα με κατάλληλο χειρισμό του φάρυγγα, του στόματος και της ρινικής κοιλότητας, αλλάζοντας τα αντίστοιχα χαρακτηριστικά του φίλτρου.

Η διαδικασία παραγωγής και αντίληψης φωνής που αναφέρθηκαν παραπάνω αποτελούν κομμάτια της λεγόμενης αλυσίδας φωνής [11] η οποία φαίνεται στο Σχήμα 2.1.2. Ο σκοπός της ομιλίας είναι η μετάδοση μηνυμάτων από τον ομιλητή στον ακροατή και σύμφωνα με τον φορμαλισμό της αλυσίδας φωνής γίνεται σε 3 επίπεδα. Τα επίπεδα είναι τα εξής: γλωσσολογικό επίπεδο (linguistic level) το οποίο περιλαμβάνει την αναπαράσταση του ζητούμενου μηνύματος στην ομιλούμενη γλώσσα, επίπεδο φυσιολογίας (physiological level) στο οποίο οι συνιστώσες της φωνητικής οδού παράγουν τους αντίστοιχους ήχους συμπεριλαμβάνοντας και την πληροφορία συναισθηματικής φόρτισης και διάρκειας, ακουστικό επίπεδο (acoustic level) στο οποίο το κωδικοποιημένο μήνυμα εξέρχεται μέσω της φωνής και λαμβάνεται τόσο από τον ομιλητή όσο και από τον ακροατή.

Στην συνέχεια, η αντίληψη της ομιλίας γίνεται με την αποκωδικοποίηση του σήματος φωνής, πρώτα σε επίπεδο φυσιολογίας μέσω των μηχανισμών που βρίσκονται στο αυτί, και στην συνέχεια σε γλωσσολογικό επίπεδο αφού γίνει η ερμηνεία των νευρικών σημάτων από τον εγκέφαλο.

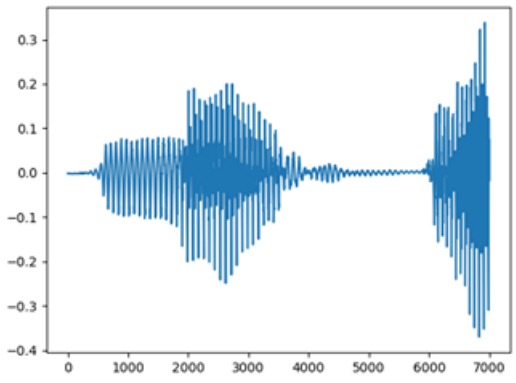
Η αναπαράσταση της πληροφορίας στο γλωσσολογικό επίπεδο, τόσο στον ομιλητή όσο και στον ακροατή είναι διακριτή. Αντίθετα στα άλλα δύο στάδια είναι συνεχής και περιλαμβάνει και επιπλέον χαρακτηριστικά από την γλωσσολογική πληροφορία, όπως τα χαρακτηριστικά του ομιλητή, την συναισθηματική του κατάσταση, την προφορά του κλπ. Η επιτυχής μοντελοποίηση αυτών των χαρακτηριστικών αποτελεί μεγάλο μέρος της σύγχρονης έρευνας πάνω στην παραγωγή συνθετικής ομιλίας. Παρ' όλ' αυτά , ένα μεγάλο μέρος της πληροφορίας που υπάρχει στο ακουστικό επίπεδο είναι πλεονάζον και μπορούμε να πετύχουμε καλή συμπίεση του σήματος επιλέγοντας τα κατάλληλα ακουστικά χαρακτηριστικά.



Σχήμα 2.1.2: Αναπαράσταση της αλυσίδας φωνής [11]

2.2 Ακουστικά Χαρακτηριστικά

Σύμφωνα με την θεωρία πληροφορίας όπως αναπτύχθηκε από τον Shannon [12], ένα οποιοδήποτε μήνυμα πληροφορίας που αναπαρίσταται σαν μία ακολουθία διακριτών συμβόλων, μπορεί να ποσοτικοποιηθεί με βάση το πληροφοριακό περιεχόμενό του σε bit, όπου ο ρυθμός μετάδοσης μετριέται σε bits/δευτερόλεπτο. Στην παραγωγή ομιλίας, η πληροφορία που μεταδίδεται κωδικοποιείται σε μορφή αναλογικής κυματομορφής, η οποία μπορεί να μεταδοθεί, εγγραφεί, επεξεργαστεί και τελικά αποκωδικοποιηθεί από έναν ακροατή. Η στοιχειώδης αναλογική μορφή του μηνύματος είναι μία ακουστική κυματομορφή η οποία αποκαλείται σήμα φωνής.



Σχήμα 2.2.1: Σήμα φωνής

Στην περίπτωση των σημάτων φωνής, όπως και σε κάθε μορφή αναλογικής πληροφορίας, επιβάλλεται να γίνει μια μετατροπή από την συνεχή κυματομορφή σε μια ψηφιακή αναπαράσταση. Αυτό επιτυγχάνεται με την διαδικασία της δειγματοληψίας (sampling) κατά την οποία εξάγονται δείγματα από την αναλογική κυματομορφή κατά περιοδικά διαστήματα πολύ μικρής διάρκειας. Τα δείγματα που εξάγονται διακριτοποιούνται (quantization) σε ένα προκαθορισμένο σύνολο τιμών οι οποίες συνήθως είναι είτε γραμμικές στο εύρος τιμών των σημάτων είτε λογαριθμικές. Η αναπαράσταση αυτή των χαρακτηριστικών λέγεται PCM από τα αρχικά των λέξεων Pulse Code Modulation (παλμοκωδική διαμόρφωση).

Ένα σημαντικό θεώρημα της δειγματοληψίας είναι ότι ένα αναλογικό σήμα μπορεί να ανακατασκευαστεί με απόλυτη ακρίβεια αν ο ρυθμός δειγματοληψίας είναι τουλάχιστον διπλάσιος από την μεγαλύτερη συχνότητα που υπάρχει στο αναλογικό σήμα [13]. Ο ρυθμός αυτός ονομάζεται ρυθμός Nyquist. Το θεώρημα συνεπάγεται ότι τα μόνα σφάλματα που μπορεί να προκύψουν κατά την ανακατασκευή αναλογικού σήματος από την ψηφιακή αναπαράσταση είναι τα σφάλματα κβάντισης.

Επειδή οι συχνότητες της ανθρώπινης ακοής κυμαίνονται μεταξύ 20 και 20 kHz, η συχνότητα δειγματοληψίας που επιβεβαιώνει ότι όλο το ακουστικό περιεχόμενο οποιουδήποτε ήχου αντιληπτού από τον άνθρωπο θα παραμείνει αναλλοίωτο κατά την ψηφιοποίηση, είναι περίπου 40kHz (ή περισσότερο). Επειδή όμως τα σήματα φωνής σπάνια περιέχουν συχνότητες μεγαλύτερες από 8kHz, συνήθεις ρυθμοί δειγματοληψίας για την επεξεργασία φωνής είναι 16kHz ή 24kHz. Στην παρούσα εργασία όλοι οι ρυθμοί δειγματοληψίας είναι 16kHz.

Ένα παράδειγμα σήματος φωνής φαίνεται στο Σχήμα 2.2.1, όπου στον άξονα του χρόνου φαίνεται ο αριθμός των δειγμάτων. Όπως φαίνεται, τα σήματα φωνής είναι αρκετά περίπλοκα στην αρχική τους μορφή, και για αυτό τον λόγο συνήθως χρειάζεται να εξαχθούν χαρακτηριστικά τα οποία μοντελοποιούν καλύτερα τους παράγοντες της φωνής που μας ενδιαφέρουν για κάθε επικείμενο πρόβλημα. Συγκεκριμένα, αφού τα σήματα φωνής έχουν πολύ υψηλό αρμονικό περιεχόμενο και έντονο περιοδικό χαρακτήρα, χρησιμοποιείται συχνά μια αναπαράσταση σε πεδίο συχνοτήτων αντί για πεδίο χρόνου. Αυτή η αναπαράσταση έχει το επιπλέον πλεονέκτημα ότι αφαιρεί σημαντικό ποσοστό από την πλεονάζουσα πληροφορία που υπάρχει λόγο της συχνής επανάληψης των δειγμάτων στον χρόνο. Για παράδειγμα η αναπαράσταση ενός ημιτόνου, που έχει άπειρη διάρκεια στο πεδίο του χρόνου, μπορεί να γίνει δίνοντας δύο μόνο αριθμούς στο πεδίο της συχνότητας. Πολλές από τις τεχνικές που χρησιμοποιούνται στην εξαγωγή χαρακτηριστικών, προέκυψαν από τα πεδία ψηφιακής επεξεργασίας φωνής, με εφαρμογές όπως η αναγνώριση φωνής, ή από ψυχοακουστικές μελέτες.

2.2.1 Mel Spectrograms

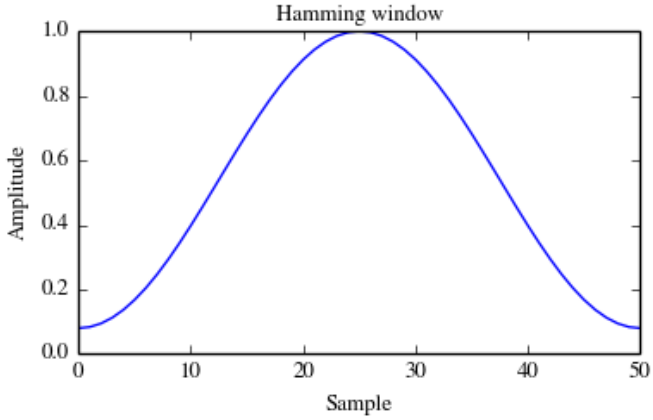
Το Mel Spectrogram είναι μια αναπαράσταση των συχνοτήτων του ηχητικού σήματος σε επίπεδο πλαισίου (frame). Το κάθε πλαίσιο αποτελεί ένα μικρό απόσπασμα του σήματος και έχει διάσταση συνήθως 20-50 ms. Η αιτιολόγηση της διάσπασης του σήματος σε μικρά πλαίσια είναι το γεγονός ότι μια οποιαδήποτε ειπωμένη πρόταση αποτελείται από μία ακολουθία φωνημάτων. Αν σπάσουμε το σήμα σε αρκετά μικρά κομμάτια, τότε κάθε ένα κομμάτι θα αντιστοιχεί σε ένα μόνο φώνημα. Επιπλέον, επιλέγοντας αρκετά μικρή διάρκεια για το πλαίσιο,

μπορούμε να εξασφαλίσουμε ότι τα χαρακτηριστικά του σήματος κατά την διάρκειά του δεν θα αλλάξουν σε σημαντικό βαθμό.

Σε κάθε πλαίσιο, υπολογίζεται αρχικά ο διακριτός μετασχηματισμός Fourier (Discrete Fourier Transform- DFT) με σκοπό να μετασχηματιστεί το σήμα από το πεδίο του χρόνου στο πεδίο της συχνότητας. Ο DTFT ορίζεται ως εξής για ένα τυχόν σήμα x στο παράθυρο m:

$$X_m[k] = \sum_{n=0}^{N-1} x_m(n) e^{\frac{-j2\pi k}{N} n} \tag{2.2.1}$$

Ο μετασχηματισμός Fourier στην γενική του περίπτωση υποθέτει σήμα άπειρου μεγέθους, αλλά εμείς έχουμε περιορίσει το σήμα σε διάρκεια N χρονικών στιγμών. Αν απλά αποκόψουμε το πλαίσιο από την θέση του στο σήμα, ο μετασχηματισμός Fourier ουσιαστικά βλέπει ένα σήμα που απότομα παίρνει τιμή μηδέν για τιμές χρόνου μικρότερες από 0 και μεγαλύτερες από N. Αυτό δημιουργεί ασυνέχειες και κατά συνέπεια στην έξοδο θα υπάρχουν συχνοτικά χαρακτηριστικά που δεν υπήρχαν στο αρχικό σήμα. Αυτό το φαινόμενο λέγεται Spectral Leakage. Για να μην υπάρχουν ασυνέχειες στο σήμα, κάθε πλαίσιο πολλαπλασιάζεται με μία συνάρτηση που ονομάζεται συνάρτηση παραθύρου η οποία ξεκινάει ομαλά από το μηδέν (ή από πολύ μικρή τιμή) και καταλήγει συμμετρικά πάλι στην τιμή που ξεκίνησε στο τέλος του πλαισίου.



Σχήμα 2.2.2: Ένα παράθυρο Hamming

Επειδή η παραθύρωση του σήματος σημαίνει ότι αναγκαστικά κάποιες τιμές του θα αλλάξουν, για να μην χαθεί πληροφορία, χρησιμοποιείται η τεχνική overlap-add. Τα πλαίσια επιλέγονται με τέτοιο τρόπο ώστε δύο διαδοχικά παράθυρα να έχουν επικάλυψη μεταξύ τους. Η επικάλυψη αυτή μπορεί να μέγεθος όσο το μισό ενός παραθύρου ή μπορεί και να είναι μικρότερη, π.χ. της τάξης των 10 ms για παράθυρα με μέγεθος 50 ms. Για να γίνει η σωστή ανακατασκευή του σήματος πρέπει τα παράθυρα να αθροίζουν στην μονάδα στις επικαλυπτόμενες περιοχές του σήματος.

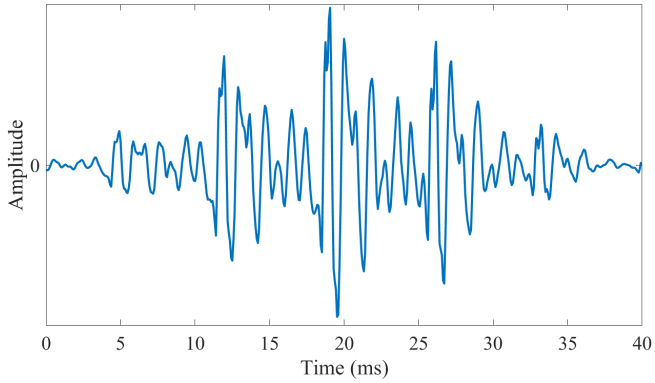
Υπάρχουν πολλές επιλογές για συνάρτηση παραθύρου που ικανοποιούν την πάνω ιδιότητα, αλλά μία πολύ συνηθισμένη κατηγορία παραθύρων είναι τα παράθυρα που προκύπτουν από το άθροισμα συνημιτόνων και είναι της μορφής:

$$w(n) = \sum_{k=0}^{K} (-1)^k a_k \cos(\frac{2\pi k n}{N}), 0 \leq n \leq N \tag{2.2.2}$$

Στην πράξη, χρησιμοποιείται συνήθως μόνο ένα συνημίτονο, (K=1) με μορφή:

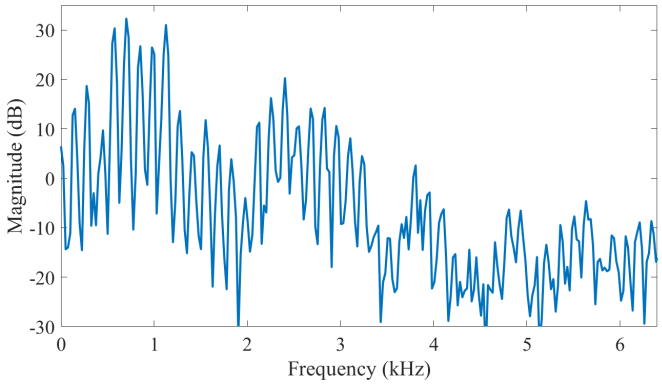
$$w(n) = a_0 - (1 - a_0) \ \cos(\frac{2\pi n}{N}), 0 \leq n \leq N \tag{2.2.3}$$

Για $a_0 = 0.54$ έχουμε το λεγόμενο παράθυρο Hamming το οποία φαίνεται στο Σχήμα 2.2.2, και για $a_0 = 0.5$ έχουμε το παράθυρο Hann, το οποίο μερικές φορές αποκαλείται και Hanning λόγω της ομοιότητας του ονόματος με το Hamming. Επίσης στο Σχήμα 2.2.3 φαίνεται η επίδραση του πολλαπλασιασμού του παραθύρου στο σήμα.



Σχήμα 2.2.3: Παραθυροποιημένο σήμα ήχου μικρής διάρκειας

Επειδή το πλάτος ενός μετασχηματισμού Fourier ενός σήματος ήχου (ο μετασχηματισμός Fourier δίνει στην γενική περίπτωση μιγαδικό αριθμό) παίρνει τιμές που καλύπτουν ένα μεγάλο εύρος από τάξεις μεγέθους στις διαφορετικές συχνότητες, συνήθως αναπαρίσταται στην λογαριθμική κλίμακα των decibels (dB) (Σχήμα 2.2.4).



Σχήμα 2.2.4: Λογαριθμικό φάσμα σήματος

Στην συνέχεια, για κάθε πλαίσιο και κάθε συχνότητα βρίσκεται η πυκνότητα φάσματος ισχύος του σήματος (Power Spectrum). Η πυκνότητα φάσματος ισχύος, επίσης φασματική ισχύς, ορίζεται ως ο μετασχηματισμός Fourier της συνάρτησης αυτοσυσχέτισης του σήματος:

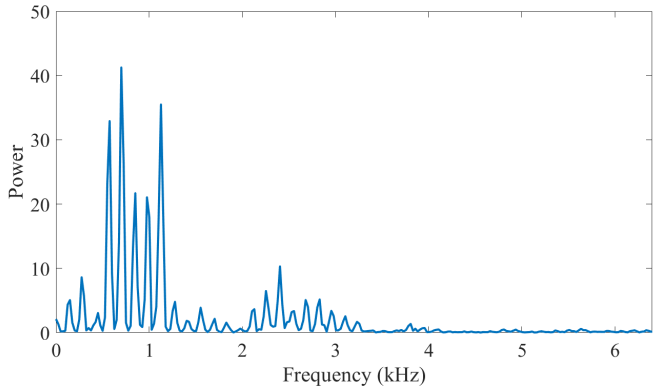
$$P(f) = \mathcal{F}\{x(t) * x(-t)\} \tag{2.2.4}$$

όπου $\mathcal{F}$ είναι ο μετασχηματισμός Fourier και $*$ είναι η πράξη της συνέλιξης.

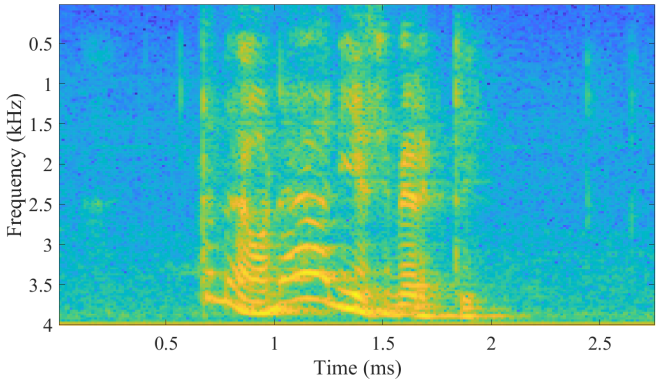
Από τις ιδιότητες του μετασχηματισμού Fourier μπορούμε να δούμε εύκολα ότι η πυκνότητα φάσματος ισχύος (σε ντετερμινιστικά σήματα) μπορεί να υπολογιστεί απλά παίρνοντας το τετράγωνο του μέτρου του φάσματος:

$$\mathcal{F}\{x(t) * x(-t)\} = \mathcal{F}(f) \cdot \mathcal{F}^*(f) = |\mathcal{F}(f)|^2 \tag{2.2.5}$$

Αυτό δίνει ένα σήμα που φαίνεται στο Σχήμα 2.2.5. Η αναπαράσταση που προκύπτει όταν βάλουμε όλα τα πλαίσια στην σειρά είναι το σπεκτρόγραμμα ή αλλιώς φασματογράφημα, το οποίο φαίνεται στο Σχήμα 2.2.6. Στον έναν άξονα φαίνεται ο χρόνος στον οποίο βρίσκεται το κάθε πλαίσιο και στον άλλον όλες οι συχνότητες του σήματος, ενώ το χρώμα του γραφήματος αντιστοιχεί στην ισχύ, με το κίτρινο να αντιστοιχεί σε μεγαλύτερη ισχύ απ' ότι το μπλε.



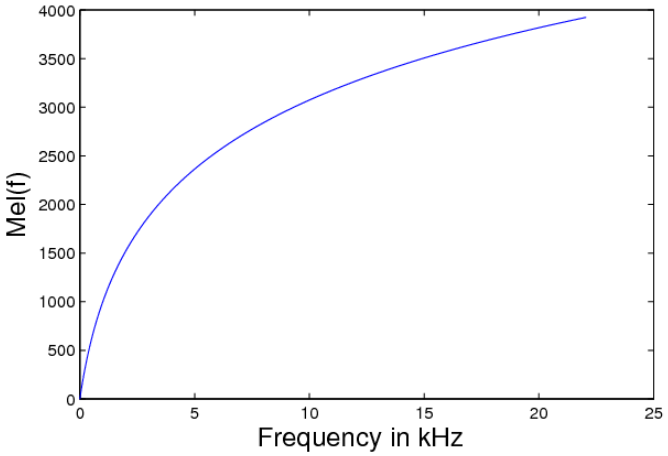
Σχήμα 2.2.5: Πυκνότητα φάσματος ισχύος ενός σήματος ήχου [14]



Σχήμα 2.2.6: Ένα Spectrogram ενός σήματος φωνής

Επειδή η ανθρώπινη αντίληψη των συχνοτήτων δεν συνάδει με την γραμμική κλίμακα, συνήθως γίνεται μετατροπή της συχνότητας στην κλίμακα Mel (Σχήμα 2.2.7), σκοπός της οποίας είναι μια σταθερή διαφορά σε αυτή, να αντιστοιχεί σε σταθερή αντιληπτική διαφορά. Οι άνθρωποι μπορούν να διακρίνουν καλύτερα μεταβολές σε μικρές συχνότητες απ' ότι σε μεγάλες, και για αυτό το λόγο, η κλίμακα Mel είναι λογαριθμική. Η μετατροπή δεν μπορεί να βρεθεί ακριβώς αφού η αντίληψη των συχνοτήτων είναι υποκειμενική, αλλά ένας συνηθισμένος τύπος ο οποίος έχει βρεθεί εμπειρικά με βάση ψυχοακουστικά πειράματα είναι:

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \tag{2.2.6}$$

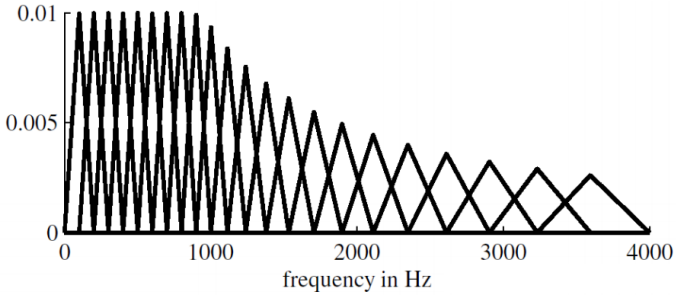


Σχήμα 2.2.7: Η κλίμακα Mel

Η μετατροπή του σήματος σε κλίμακα Mel γίνεται διαλέγοντας έναν αριθμό από συχνότητες που έχουν ίσες αποστάσεις στην κλίμακα Mel και κατά συνέπεια αντιστοιχούν σε ίσες αντιληπτικές αποστάσεις, και χρησιμοποιώντας ένα φίλτρο για την κάθε μια. Το φίλτρο βρίσκει έναν σταθμισμένο όρο από όλες τις συχνότητες γύρω από την κεντρική. Συγκεκριμένα, ο τύπος είναι:

$$P_m[r] = \frac{\sum_{k=L_r}^{U_r} |V_r(k) X_m(k)|^2}{\sum_{k=L_r}^{U_r} |V_r(k))|^2} \tag{2.2.7}$$

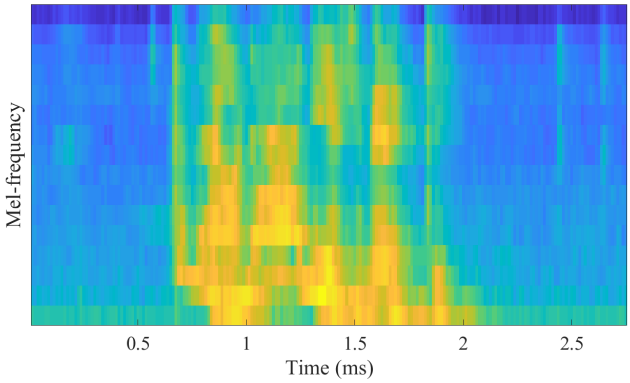
όπου $V_r[k]$ είναι η συνάρτηση στάθμισης για το r-οστό φίλτρο, το οποίο εκτείνεται από τον δείκτη L_r μέχρι U_r στις συχνότητες σε κλίμακα Hertz. Τα φίλτρα που χρησιμοποιούνται είναι συνήθως τριγωνικά για κάθε κρίσιμη ζώνη συχνότητας που επιθυμούμε όπως φαίνονται στο Σχήμα 2.2.8. Μπορούμε να δούμε ότι τα κέντρα των φίλτρων έχουν μεγαλύτερη απόσταση μεταξύ τους όσο μεγαλώνει η συχνότητα, κάτι που οφείλεται στο γεγονός ότι η κλίμακα Mel είναι λογαριθμική.



Σχήμα 2.2.8: Τα τριγωνικά φίλτρα που χρησιμοποιούνται στην εξαγωγή χαρακτηριστικών [15]

Το μετασχηματισμένο σπεκτρόγραμμα στην κλίμακα Mel ονομάζεται Mel Spectrogram και είναι ένα βασικό χαρακτηριστικό πολλών μοντέλων επεξεργασίας και παραγωγής φωνής. Φαίνεται στο Σχήμα 2.2.9.

Ένα μεγάλο μειονέκτημα των Mel χαρακτηριστικών, καθώς και όλων των χαρακτηριστικών που προκύπτουν από ενέργειες του φάσματος που προκύπτει από τον μετασχηματισμό Fourier, είναι ότι η φάση του αρχικού σήματος χάνεται όταν υπολογίζεται το μέτρο των μιγαδικών αριθμών του φάσματος στον τύπο 2.2.7 και κατά συνέπεια, το σήμα φωνής δεν είναι ανακατασκευάσιμο. Αυτό μπορεί να αντιμετωπιστεί είτε με διάφορους κλασικούς αλγορίθμους όπως ο Griffin Lim [16], ο οποίος προσπαθεί να κάνει ανακατασκευή της φάσης με μια επαναληπτική διαδικασία, είτε με τη βοήθεια από Vocoders, νευρωνικούς ή μη, όπως θα αναλυθούν στη συνέχεια.

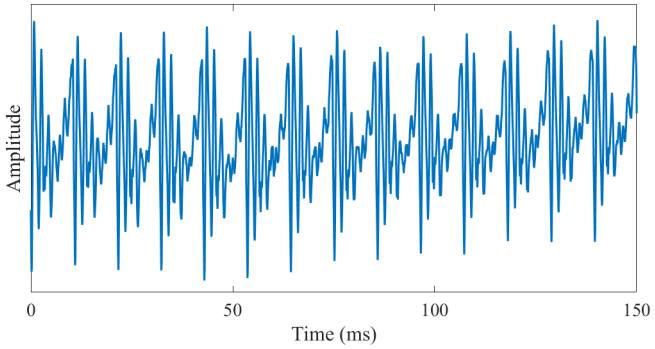


Σχήμα 2.2.9: Ένα Mel Spectrogram ενός σήματος ήχου

2.2.2 Θεμελιώδης Συχνότητα (F0)

Με βάση την ανάλυση Fourier κάθε περιοδικό σήμα μπορεί να γραφτεί σαν άθροισμα από ημιτονοειδή σήματα διακριτών συχνοτήτων. Εν γένει, το άθροισμα αυτό μπορεί να είναι και απείρων όρων, αλλά όλοι οι όροι είναι ακέραια πολλαπλάσια μίας συχνότητας. Αυτοί οι όροι ονομάζονται αρμονικές συχνότητες και ο πρώτος (χαμηλότερος) όρος ονομάζεται θεμελιώδης συχνότητα (F0).

Τα σήματα φωνής, προφανώς δεν είναι απολύτως περιοδικά αλλά σε μικρές χρονικές κλίμακες, οι έμφωνες περιοχές ομιλίας είναι σχεδόν περιοδικές, όπως μπορεί να φανεί στο Σχήμα 2.2.10. Κατά συνέπεια, αφού η φωνή διαμεριστεί σε πλαίσια, η θεμελιώδης συχνότητα μπορεί να προσεγγιστεί για κάθε πλαίσιο, και να αποτελέσει σημαντικό χαρακτηριστικό της φωνής.



Σχήμα 2.2.10: Μία έμφωνη περιοχή ενός σήματος φωνής. Σχήμα από [14].

Η θεμελιώδης συχνότητα μετριέται σε Hz αν και μερικές φορές μετατρέπεται σε ημιτόνια ή τόνους για καλύτερη κατανόηση. Τυπικές θεμελιώδεις συχνότητες για ομιλία κυμαίνονται μεταξύ 80-450 Hz, με τους άντρες να έχουν γενικά χαμηλότερες συχνότητες από τις γυναίκες. Για παράδειγμα, στο Σχήμα 2.2.10, η θεμελιώδης συχνότητα είναι 93 Hz. Η θεμελιώδης συχνότητα μοντελοποιεί το ύψος της φωνής και κατά συνέπεια, στις τραγουδιστές φωνές, αντιστοιχεί στην πρώτη αρμονική της νότας που τραγουδιέται ανά πάσα στιγμή. Ένα ενδιαφέρον ψυχοακουστικό φαινόμενο είναι ότι στην πραγματικότητα η αίσθηση του ύψους της φωνής παραμένει η ίδια ακόμα και αν αφαιρεθεί η θεμελιώδης συχνότητα [14], για παράδειγμα περνώντας το σήμα από ένα υψιπερατό φίλτρο. Φαίνεται ότι ο εγκέφαλος καταφέρνει να ανακατασκευάσει την θεμελιώδη συχνότητα από τις αρμονικές, ένα φαινόμενο που δεν έχει εξηγηθεί πλήρως.

Η θεμελιώδης συχνότητα έχει μεγάλη σημασία για την μοντελοποίηση του συναισθήματος, καθώς το ύψος μιας ομιλίας συχνά αλλάζει με εκφραστικούς σκοπούς, όπως για έμφαση ή για απορία. Η συσχέτιση των διαφορετικών συναισθημάτων με το F0, μπορεί να φανεί και από τα αποτελέσματα της έρευνας [17], στην οποία μελετήθηκαν

τα διαφορετικά εύρη συχνοτήτων για διαφορετικά συναισθήματα στην Γερμανική γλώσσα. Οι διαφορές αυτές, μετρημένες σε ημιτόνια, φαίνονται στο Σχήμα 2.2.11.

Η εκτίμηση του F0 μπορεί να γίνει με διαφορετικούς τρόπους είτε στο πεδίο της συχνότητας, είτε στο πεδίο του χρόνου. Για παράδειγμα, μπορεί να γίνει εκτίμηση χρησιμοποιώντας τα Mel Spectrograms (ή τα Cepstrum που αναλύονται στην συνέχεια), καθώς το F0 θα φαίνεται σαν το μέγιστο στην συχνότητα F0=Fs/T, και σε όλα τα ακέραια πολλαπλάσια της, όπου Fs είναι η συχνότητα δειγματοληψίας. Παρ' όλ' αυτά, στην πράξη χρειάζονται πιο σύνθετες προσεγγίσεις, καθώς υπάρχουν πολλά πιθανά σφάλματα, όπως το σφάλμα οκτάβας που προκύπτει όταν μία αρμονική συνιστώσα (συνήθως η δεύτερη) έχει μεγαλύτερη τιμή από το F0 και θεωρείται λανθασμένα ως η θεμελιώδης. Επίσης, η εκτίμηση του F0 έχει νόημα μόνο στους έμφωνους ήχους οπότε ο εκτιμητής θεμελιώδους περιόδου πρέπει να λαμβάνει μια απόφαση ταξινόμησης μεταξύ έμφωνων και μη πλαισίων (σε αυτήν τη περίπτωση συνήθως θέτεται F0 = 0 στα μη έμφωνα πλαίσια). Ένα συχνά χρησιμοποιούμενο εργαλείο είναι το Praat [18] αλλά στα πλαίσια αυτής της εργασίας, χρησιμοποιείται ο WORLD Vocoder που αναλύεται στην συνέχεια.

	S	F	N	B	A	H
1-4	4,78	8,59	9,50	11,92	16,04	16,13
1-3	3,68	5,97	5,99	8,48	9,10	9,55
2-4	3,33	6,43	7,15	7,58	12,30	12,34
2-3	2,23	3,80	3,64	4,14	5,38	5,75

Σχήμα 2.2.11: Διαφορές στα εύρη του F0 ανά συναίσθημα. Τα γράμματα αντιστοιχούν σε S – θλίψη (sadness), F – φόβος (fear), N – ουδέτερη (neutral), B – ανία (boredom), A – θυμός (anger), H – χαρά (happiness). Όπου δεν υπάρχει κατακόρυφη γραμμή μεταξύ διπλανών κελιών, η διαφορά μεταξύ τους δεν είναι στατιστικά σημαντική ($p < 0.05$).

2.2.3 Cepstrum Χαρακτηριστικά

Το λογαριθμικό φάσμα είναι ένα συνεχές σήμα, κάτι που οφείλεται στην εξομάλυνση των παραθύρων. Επίσης, έχει και περιοδικό χαρακτήρα, όπως μπορούμε να δούμε και στο Σχήμα 2.2.4, καθώς οι βασικές του κορυφές βρίσκονται στα ακέραια πολλαπλάσια της θεμελιώδης συχνότητας. Τα πιο σημαντικά χαρακτηριστικά του όμως, τα οποία είναι και υπεύθυνα π.χ. για την διαφοροποίηση των φωνημάτων, είναι στην μακροσκοπική δομή της συνάρτησης. Για αυτό το λόγο, πολλές φορές προσπαθούμε να προσεγγίσουμε την φασματική περιβάλλουσα αντί να δουλεύουμε με τα Mel Spectrograms.

Γενικά, η περιβάλλουσα ενός σήματος ταλάντωσης ορίζεται ως η ομαλή καμπύλη που αποτελεί το περίγραμμα των ακραίων τιμών της. Αποτελεί μια γενίκευση της έννοιας του πλάτους, καθώς σε κάθε χρονική στιγμή η περιβάλλουσα δείχνει το στιγμιαίο πλάτος. Η φασματική περιβάλλουσα είναι η περιβάλλουσα του λογαριθμικού φάσματος του σήματος και περιέχει την πληροφορία των μακροσκοπικών χαρακτηριστικών της.

Ο τρόπος που μπορούμε να πάρουμε πληροφορία για την φασματική περιβάλλουσα είναι χρησιμοποιώντας τον αντίστροφο μετασχηματισμό Fourier παίρνοντας τα Cepstrum χαρακτηριστικά:

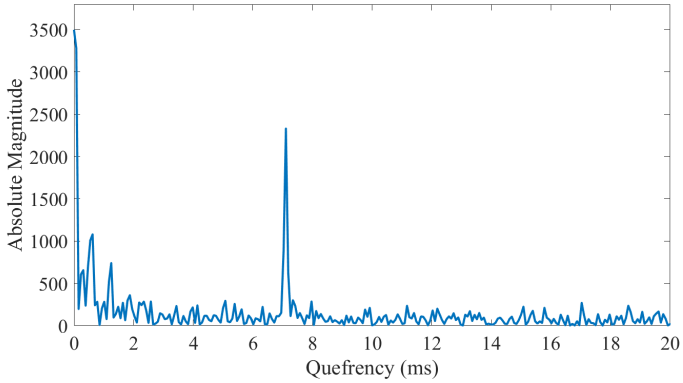
$$\mathcal{C}\{x\} = \left|\mathcal{F}^{-1}\left\{\log\left(|\mathcal{F}\{x(t)\}|^2\right)\right\}\right|^2, \tag{2.2.8}$$

όπου $\mathcal{C}\{x\}$ δηλώνει το Cepstrum του σήματος, $\mathcal{F}\{x\}$ είναι ο μετασχηματισμός Fourier και $\mathcal{F}^{-1}\{x\}$ είναι ο αντίστροφος μετασχηματισμός Fourier.

Το όνομα Cepstrum προκύπτει από την λέξη spectrum για να αντικατοπτρίζεται το γεγονός ότι αποτελεί μια περίπλοκη αναδιάταξη μετασχηματισμών. Επειδή προκύπτει από αντίστροφο μετασχηματισμό από το πεδίο της συχνότητας, μετριέται στον άξονα του χρόνου, όπως φαίνεται και στο Σχήμα 2.2.12. Κατά αναλογία με την λέξη Cepstrum, και για να φαίνεται ότι ο άξονας του χρόνου σε αυτά τα χαρακτηριστικά δεν αντιστοιχεί σε πραγματικό χρόνο, αυτός ο άξονας πολλές φορές ονομάζεται quefrency (αντί για frequency που σημαίνει συχνότητα).

Οι τιμές του Cepstrum που βρίσκονται στους χαμηλούς χρόνους δίνουν πληροφορία για τα χαρακτηριστικά του λογαριθμικού φάσματος που αλλάζουν με αργό ρυθμό, δηλαδή τα χαρακτηριστικά της φασματικής περιβάλλουσας που θέλουμε να μοντελοποιήσουμε. Επιπλέον, σε πιο μεγάλους χρόνους μπορούμε να δούμε χαρακτηριστικά του

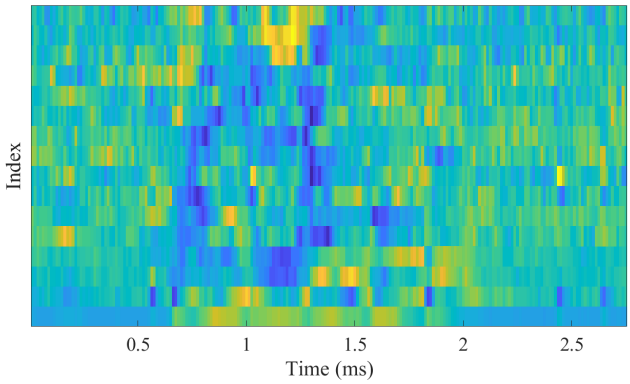
συχνοτικού περιεχομένου του φάσματος, με πιο ξεκάθαρη την θεμελιώδη συχνότητα η οποία στο Σχήμα 2.2.12 αντιστοιχεί στην μεγάλη τιμή στα περίπου 7 ms. Τα 7 ms αντιστοιχούν σε συχνότητα περίπου 143 Hz οπότε έχουμε μία εκτίμηση του F0 του σήματος 2.2.4.



Σχήμα 2.2.12: Cepstrum σήματος φωνής

Μια επιπλέον αναπαράσταση της φασματικής περιβάλλουσας είναι τα mfcc χαρακτηριστικά (Mel-Frequency Cepstral Coefficients). Αντί να χρησιμοποιηθεί αντίστροφος μετασχηματισμός Fourier στο λογαριθμικό φάσμα, πρώτα γίνεται η μετατροπή του φάσματος σε συχνότητα Mel με την συστοιχία φίλτρων όπως αναλύθηκε προηγουμένως. Στην συνέχεια γίνεται διακριτός μετασχηματισμός συνημιτόνου (DCT) στα χαρακτηριστικά των Mel συχνοτήτων και το αποτέλεσμα είναι τα mfcc χαρακτηριστικά (Σχήμα 2.2.13).

Ο διακριτός μετασχηματισμός ημιτόνου έχει αντίστοιχο ρόλο με τον αντίστροφο μετασχηματισμό Fourier στα Cepstrum χαρακτηριστικά με το επιπλέον πλεονέκτημα ότι γίνεται αποσυσχέτιση του σήματος καθώς οι mfcc συντελεστές δεν έχουν μεγάλη συσχέτιση μεταξύ τους. Ο αριθμός των συντελεστών που μπορούν να εξαχθούν από αυτή την διαδικασία εξαρτάται από το πρόβλημα αλλά συνήθως χρησιμοποιούνται οι 12-13 πρώτοι, οι οποίοι περιέχουν την πληροφορία κυρίως για της χαμηλές συχνότητες.

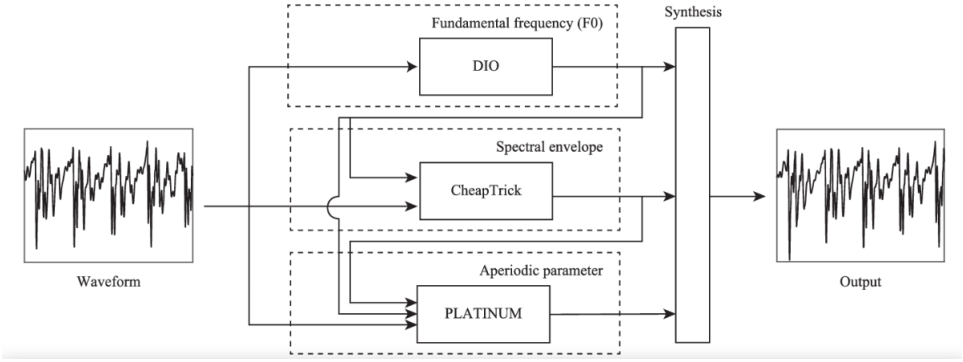


Σχήμα 2.2.13: Mfcc χαρακτηριστικά

Τα mfcc χαρακτηριστικά έχουν το πλεονέκτημα ότι περιλαμβάνουν την πληροφορία για τα περισσότερα χαρακτηριστικά της ομιλίας που μας ενδιαφέρουν όπως π.χ. την ταυτότητα των φωνημάτων. Για αυτό τον λόγο χρησιμοποιούνται συνέχεια σε εφαρμογές όπως η αναγνώριση φωνής. Παρ' όλ' αυτά, επειδή μεγάλο μέρος της πληροφορίας του σήματος έχει χαθεί (η φάση των μετασχηματισμών και οι υψηλές συνιστώσες του DCT), δεν είναι εύκολο να γίνει ανακατασκευή από τα mfcc χαρακτηριστικά και σαν συνέπεια δεν προσφέρονται για σύνθεση φωνής. Στα πλαίσια αυτής της εργασίας χρησιμοποιούνται μόνο για αξιολόγηση αποτελεσμάτων.

2.3 WORLD Vocoder

Η επιλογή των χαρακτηριστικών που θα χρησιμοποιηθούν για την επίλυση κάθε προβλήματος, είναι πολύ σημαντικό κομμάτι της διαδικασίας της μηχανικής μάθησης. Στην ιδανική περίπτωση, τα διάφορα χαρακτηριστικά που χρησιμοποιούνται αντιστοιχούν σε διαφορετικές ιδιότητες της ομιλίας, με τέτοιο τρόπο ώστε το κάθε χαρακτηριστικό να μην περιέχει πληροφορία για κάθε άλλο. Με αυτόν το τρόπο, τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους και μπορούμε να επεξεργαστούμε το κάθε ένα από αυτά με διαφορετικό τρόπο. Ένας τρόπος εξαγωγής τέτοιων χαρακτηριστικών είναι ο WORLD Vocoder [19].



Σχήμα 2.3.1: Σχηματική απεικόνιση του WORLD Vocoder

Όπως διατυπώνει και το όνομά του, ο WORLD Vocoder αποτελεί έναν κωδικοποιητή φωνής (vocoder) ο οποίος είναι ένα σύστημα ανάλυσης και σύνθεσης της φωνής. Οι Vocoders, αποτελούνται από το κομμάτι της κωδικοποίησης του σήματος σε μια αναπαράσταση η οποία είναι πιο εύκολα διαχειρίσιμη, και το κομμάτι της αποκωδικοποίησης η οποία ανακατασκευάζει το σήμα με όση μικρότερη απόκλιση από το αρχικό γίνεται. Σε ορισμένες περιπτώσεις, οι Vocoders δεν κάνουν πιστή αναπαραγωγή της αρχικής κυματομορφής αλλά έχουν διατηρήσει μόνο την αντιληπτή ποιότητα κατά την αναπαραγωγή της φωνής από την ψηφιακή αναπαράσταση.

Σε όλες τις περιπτώσεις χαρακτηριστικών που αναφέραμε, χρειάζεται ένας vocoder για να επιστρέψουμε στην κανονική αναπαράσταση του σήματος φωνής. Πολλοί Vocoders βασίζονται σε κλασικούς αλγορίθμους επεξεργασίας φωνής αν και τα τελευταία χρόνια οι Vocoders που χρησιμοποιούνται βασίζονται κυρίως σε νευρωνικά δίκτυα.

Ιστορικά, οι Vocoders χρησιμοποιούνταν για συμπίεση των σημάτων με σκοπό την καλύτερη μετάδοσή τους, με μικρότερο ρυθμό μετάδοσης δεδομένων (bandwidth). Επιπλέον, οι Vocoders βρήκαν και μεγάλο πεδίο χρήσης στον τομέα της μουσικής, με συγκροτήματα όπως οι Krafwerk ή οι Daft Punk να αποτελούν χαρακτηριστικά παραδείγματα. Καθώς η μονάδα αποκωδικοποίησης του Vocoder (που ονομάζεται Voder) μπορεί να χρησιμοποιηθεί με κάποιο σήμα F0 διαφορετικό από το αρχικό, αυτό το σήμα μπορεί να δοθεί από κάποιο εξωτερικό σήμα, π.χ. ένα synthesizer. Στην πράξη, αυτό δημιουργεί μια ομιλία η οποία δεν ακούγεται τελείως φυσική και έχει έναν χαρακτηριστικό "ρομποτικό" χαρακτήρα, ο οποίος μετά μπορεί να χρησιμοποιηθεί στην μουσική σύνθεση. Βλέπουμε ουσιαστικά ότι αυτή η χρήση των Vocoders βασίζεται στην διαμέριση των φωνητικών χαρακτηριστικών και την ανεξάρτητη επεξεργασία ή τροποποίησή τους, κάτι που έχει μεγάλη σημασία και στο πρόβλημα της μετατροπής συναισθηματικής ομιλίας. Επίσης, στην προκειμένη περίπτωση, η δυσκολία πιστής ανακατασκευής του σήματος από τον Vocoder είναι θεμιτή καθώς χρησιμοποιείται σαν εκφραστικό μέσο.

Ο WORLD Vocoder, είναι βασισμένος στο σύστημα Tandem-STRAIGHT [20] και όπως φαίνεται και από το Σχήμα 2.3.1, αποτελείται από τρία ξεχωριστά υποσυστήματα, τα οποία υπολογίζουν την θεμελιώδη συχνότητα, την φασματική περιβάλλουσα (spectral envelope) μέσω του cepstrum και επιπλέον απεριοδικές παραμέτρους.

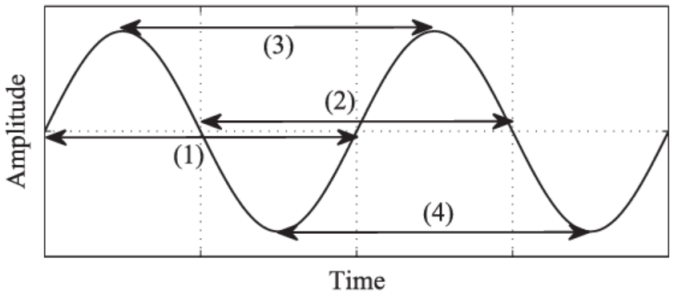
Συγκεκριμένα, τα συστήματα είναι:

DIO

Ο αλγόριθμος DIO [21] είναι αλγόριθμος εκτίμησης της θεμελιώδης συχνότητας και δουλεύει σε τρία βήματα. Αρχικά το σήμα περνάει από μια σειρά από βαθυπερατά φίλτρα με διαφορετικές συχνότητες αποκοπής, με την

λογική ότι αν ένα σήμα απομείνει μόνο με την θεμελιώδη συνιστώσα, τότε θα είναι ένα ημιτονικό σήμα με περίοδο $T_0 = 1/F_0$. Στη συνέχεια, υπολογίζονται όλα τα πιθανά F0 που έχουν προκύψει από τα διάφορα βαθυπερατά φίλτρα, καθώς και μετρικές αξιοπιστίας για κάθε μία απο αυτές. Συγκεκριμένα, υπολογίζεται η συχνότητα ως το αντίστροφο των τεσσάρων αποστάσεων που φαίνονται στο Σχήμα 2.3.2, (μέγιστο, ελάχιστο, θετική και αρνητική αλλαγή προσήμου). Σε ένα ημιτονικό σήμα, οι τέσσερες αυτές τιμές θα έπρεπε να δίνουν την ίδια ακριβώς συχνότητα, την συχνότητα του ημιτόνου, οπότε η μέση τιμή της τυπικής απόκλισης αυτών των τιμών χρησιμοποιείται σαν αξιοπιστία. Στο τελευταίο βήμα, η τιμή με την μικρότερη τυπική απόκλιση επιλέγεται. Η όλη διαδικασία επαναλαμβάνεται σε κάθε πλαίσιο, και σαν αποτέλεσμα έχουμε μια τιμή F0 για κάθε πλαίσιο.

την ίδια ακριβώς συχνότητα, την συχνότητα του ημιτόνου, οπότε η μέση τιμή της τυπικής απόκλισης αυτών των τιμών χρησιμοποιείται σαν αξιοπιστία. Στο τελευταίο βήμα, η τιμή με την μικρότερη τυπική απόκλιση επιλέγεται. Η όλη διαδικασία επαναλαμβάνεται σε κάθε πλαίσιο, και σαν αποτέλεσμα έχουμε μια τιμή F0 για κάθε πλαίσιο.



Σχήμα 2.3.2: Οι τέσσερις μετρικές της F0 σύμφωνα με τον αλγόριθμο DIO

CheapTrick

Ο αλγόριθμος Cheaptrick [22] χρησιμοποιείται για να υπολογιστεί η φασματική περιβάλλουσα κάθε πλαισίου με την βοήθεια των cepstrum χαρακτηριστικών και βασίζεται πάνω στην ιδέα του Pitch Synchronous analysis [23]. Ακολουθεί η διαδικασία που γίνεται σε κάθε πλαίσιο.

Αρχικά, χρησιμοποιείται ένα παράθυρο Hanning με μήκος $3T_0$ όπου το T_0 είναι το αντίστροφο του F0 που έχει υπολογιστεί στο προηγούμενο βήμα και υπολογίζεται η ισχύς του σήματος:

$$\begin{aligned}\int_0^{3T_0} (y(t)w(t))^2 dt &= \int_0^{T_0} y^2(t)w^2(t)dt + \int_0^{T_0} y^2(t)w^2(t+T_0)dt + \int_0^{T_0} y^2(t)w^2(t+2T_0)dt \\ &= \int_0^{T_0} y^2(t)\big(w^2(t) + w^2(t+T_0) + w^2(t+2T_0)\big)dt \\ &= 1.125\int_0^{T_0} y^2(t)dt,\end{aligned} \tag{2.3.1}$$

όπου η τελευταία ισότητα προκύπτει από την αντικατάσταση του τύπου του παράθυρου Hanning και χρησιμοποιώντας τριγωνομετρικές ιδιότητες. Από αυτή την εξίσωση βλέπουμε ότι η παραθύρωση του σήματος δεν επηρεάζει την ολική ισχύ κάποιου πλαισίου πέρα από τον πολλαπλασιασμό από μια σταθερά.

Στην συνέχεια, αφού γίνει εξαγωγή της φασματικής ισχύς όπως αναλύθηκε προηγουμένως, γίνεται εξομάλυνση (smoothing) με διαδικασία ανάλογη με αυτή που γίνεται για την εξαγωγή του Mel Spectrogram. Η διαφορά είναι ότι δεν χρησιμοποιείται Mel κλίμακα για τις συχνότητες και τα φίλτρα που χρησιμοποιούνται είναι τετραγωνικά αντί για τριγωνικά όπως συνηθίζεται:

$$P_s(\omega) = \frac{3}{2\omega}\int_{-\frac{\omega}{3}}^{\frac{\omega}{3}} P(\omega+\lambda)d\lambda, \tag{2.3.2}$$

Ο βασικός λόγος που γίνεται αυτό το φιλτράρισμα είναι για να αποφευχθούν τα μηδενικά στην φασματική ισχύ που θα οδηγήσουν σε άπειρες τιμές όταν θα παρθεί ο λογάριθμος του φάσματος.

Το τρίτο βήμα του αλγορίθμου είναι η εξάλειψη του περιοδικού χαρακτήρα του λογαριθμικού φάσματος, που προκύπτει από την θεμελιώδη συχνότητα. Η διαδικασία είναι αντίστοιχη με την εξαγωγή της φασματικής περιβάλλουσας παίρνοντας τις μικρότερες τιμές του cepstrum και ουσιαστικά αποτελείται από το φιλτράρισμα από δύο εξειδικευμένα φίλτρα στα cepstrum χαρακτηριστικά και σε έναν ακόμα μετασχηματισμό Fourier ώστε να ξαναγυρίσουμε στο πεδίο της συχνότητας:

$$\begin{aligned} p_s(\tau) &= \mathcal{F}^{-1}[\log\left(P_s(\omega)\right)], \\ l_q(\tau) &= \tilde{q}_0 + 2\tilde{q}_1 \cos\left(\frac{2\pi\tau}{T_0}\right), \\ l_s(\tau) &= \frac{sin(\pi f_0 \tau)}{\pi f_0 \tau}, \\ \log P_l(\omega) &= \mathcal{F}[l_s(\tau) l_q(\tau) p_s(\tau)] \end{aligned} \tag{2.3.3}$$

όπου $P_l(\omega)$ είναι το τελικό φάσμα ισχύος και l_q, l_s είναι τα δύο φίλτρα που χρησιμοποιούνται.

Το φίλτρο l_s ουσιαστικά αντιστοιχεί σε συνέλιξη με τετραγωνικό παλμό στο πεδίο της συχνότητας, κάτι που αντιστοιχεί στην επιπλέον εξομάλυνση του λογαριθμικού σήματος. Το φίλτρο l_s εξαφανίζει τις συνιστώσες της συχνότητας που αντιστοιχούν στα πολλαπλάσια της θεμελιώδης συχνότητας, αφαιρώντας όλη την περιοδική πληροφορία που έχει μείνει στο σήμα. Οι τιμές των σταθερών q_0, q_1 έχουν υπολογιστεί εμπειρικά και είναι $q_0 = 1.18$, $q_1 = -0.09$.

PLATINUM

Ο αλγόριθμος Platinum [24] υπολογίζει σε κάθε παράθυρο όλα τα μη περιοδικά στοιχεία που έχουν μείνει μετά την μοντελοποίηση της θεμελιώδους συχνότητας και των φασματικών χαρακτηριστικών. Υπολογίζει το λεγόμενο σήμα διέγερσης (excitation signal) το οποίο είναι ανάλογο με την πηγή στο μοντέλο πηγής φίλτρου που αναφέρθηκε στην εισαγωγή του κεφαλαίου. Τα χαρακτηριστικά της περιβάλλουσας που υπολογίστηκαν στο προηγούμενο βήμα θεωρούνται ως η κρουστική απόκριση του φίλτρου και ο αλγόριθμος προσπαθεί να βρει το σήμα διέγερσης ($x_p(t)$) που οδηγεί στο αρχικό σήμα με βάση αυτήν την κρουστική απόκριση. Αυτό προκύπτει σχετικά εύκολα από τον αντίστροφο μετασχηματισμό Fourier του αρχικού σήματος φωνής διαιρεμένου με τον μετασχηματισμό της κρουστικής απόκρισης:

$$\begin{aligned} x_p(t) &= \mathcal{F}^{-1}[X_p(\omega)], \\ X_p(\omega) &= \frac{X(\omega)}{S_m(\omega)}, \end{aligned} \tag{2.3.4}$$

όπου $X(\omega)$ είναι το αρχικό σήμα παραθυροποιημένο με παράθυρο διάρκειας $2T_0$ και $S_m(\omega)$ είναι ο μετασχηματισμός της κρουστικής απόκρισης. Το $S_m(\omega)$ αποτελείται από τον μετασχηματισμό των cepstral χαρακτηριστικών (φάσμα ισχύος που υπολογίστηκε πριν) αφού πρώτα αφαιρεθούν όλες οι αρνητικές συνιστώσες ώστε να εξασφαλιστεί ότι το σύστημα είναι ελάχιστης φάσης:

$$\begin{aligned} c(\tau) &= \mathcal{F}^{-1}[\log(P_l(\omega))], \\ c_m(\tau) &= \begin{cases} 2c(\tau) & (\tau > 0) \\ c(\tau) & (\tau = 0) \\ 0 & (\tau < 0) \end{cases}, \\ S_m(\omega) &= exp(\mathcal{F}[c_m(\tau)]) \end{aligned} \tag{2.3.5}$$

Τα απεριοδικά χαρακτηριστικά έχουν ουσιαστικά σκοπό να καλύψουν ό,τι δεν έχει πιαστεί από τα προηγούμενα βήματα, κάτι που σημαίνει ότι στην πράξη περιέχουν κυρίως πληροφορία για άφωνους ήχους της ομιλίας, όπως τα σύμφωνα. Επίσης, τα απεριοδικά χαρακτηριστικά μπορεί να περιέχουν και όποιο θόρυβο υπάρχει στην ομιλία ή και στην ηχογράφηση.

Ο WORLD Vocoder δεν είναι αντιστρέψιμος αλγόριθμος αλλά η διαδικασία σύνθεσης επιτρέπει να ανακατασκευαστεί το σήμα με μικρό σφάλμα. Παρατηρήσαμε βέβαια, ότι σε περιπτώσεις όπου στο σήμα ομιλίας υπάρχουν χαρακτηριστικά που δεν είναι αμιγώς ομιλία (π.χ. γέλιο, αναφιλητό ή αναφωνήσεις) ο αλγόριθμος εισάγει διάφορα σφάλματα κατά την ανακατασκευή. Οι περιπτώσεις που αυτό δημιουργεί σημαντικά προβλήματα στις βάσεις δεδομένων που χρησιμοποιούμε είναι βέβαια πολύ λίγες.

Κεφάλαιο 3

Θεωρητικό Υπόβαθρο Μηχανικής Μάθησης

3.1 Εισαγωγή

Η Μηχανική Μάθηση (Machine Learning) είναι ένα πεδίο της τεχνητής νοημοσύνης που στοχεύει στη δημιουργία συστημάτων που μπορούν να μαθαίνουν και να αξιοποιούν χρήσιμες πληροφορίες για την επίλυση διαφόρων προβλημάτων, χωρίς να απαιτείται ο ρητός προγραμματισμός τους. Οι αλγόριθμοι εκπαίδευσης μηχανικής μάθησης χρησιμοποιούν ένα σύνολο από δεδομένα ώστε να βελτιώσουν την απόδοση τους στο ζητούμενο πρόβλημα, με την βοήθεια μεθόδων στατιστικής και μαθηματικής μοντελοποίησης. Η χρησιμότητα τους έγκειται στο ότι μπορούν να μάθουν να λύνουν προβλήματα τα οποία οι άνθρωποι δεν μπορούν (ή μπορούν με μεγάλη δυσκολία) να λύσουν με τις μηχανιστικές μεθόδους των κλασικών αλγορίθμων, για λόγους περιπλοκότητας ή μεγάλης απαιτούμενης λεπτομέρειας.

Στην μηχανική μάθηση, οι μέθοδοι γενικά ταξινομούνται σε διάφορες κατηγορίες ανάλογα με το είδος προβλήματος που επιλύουν, το είδος δεδομένων που χρειάζονται και τις διαφορετικές εισόδους και εξόδους. Δύο πολύ σημαντικές κατηγορίες είναι η επιβλεπόμενη μάθηση (supervised learning) και η μη επιβλεπόμενη μάθηση (supervised learning).

Επιβλεπόμενη Μάθηση

Στην επιβλεπόμενη μάθηση ο σκοπός είναι η μάθηση μιας συνάρτησης απεικόνισης (mapping) από ένα σύνολο εισόδου σε ένα σύνολο εξόδου. Δίνεται ένα σύνολο από δεδομένα εκπαίδευσης x (που στη γενική περίπτωση είναι διανύσματα) σε κάθε ένα από τα οποία αντιστοιχεί μία τιμή y (επίσης διάνυσμα στη γενική περίπτωση) οι οποίες ονομάζονται ετικέτες (labels). Σκοπός της εκπαίδευσης του μοντέλου είναι η προσέγγιση της συνάρτηση απεικόνισης $Y = f(X)$. Η κατηγορία ονομάζεται επιβλεπόμενη μάθηση, επειδή για κάθε δεδομένο υπάρχει σωστή απάντηση την οποία έχουμε πάντα στη διάθεσή μας κατά την διάρκεια της εκπαίδευσης. Έτσι μπορούμε να εκτιμήσουμε την απόδοση του μοντέλου σε κάθε βήμα της εκπαίδευσης και να χρησιμοποιήσουμε αυτή την γνώση για να βελτιώσουμε το μοντέλο.

Φυσικά, σκοπός της εκπαίδευσης είναι η συνάρτηση απεικόνισης να λειτουργεί αποδοτικά και σε δεδομένα που δεν έχει δει κατά την διάρκεια της εκπαίδευσης. Οι υποθέσεις που κάνουμε για το εκάστοτε πρόβλημα που μας δικαιολογούν ότι το μοντέλο που έχουμε επιλέξει θα γενικεύει και σε καινούργια δεδομένα ονομάζεται inductive bias. Μερικές φορές, αυτό δεν συμβαίνει και το μοντέλο έχει πολύ καλή απόδοση στα δεδομένα εκπαίδευσης, αλλά κακή απόδοση στα καινούργια δεδομένα. Αυτό το φαινόμενο λέγεται υπερπροσαρμογή (over-fitting).

Δύο σημαντικές υποκατηγορίες επιβλεπόμενης μάθησης είναι η παλινδρόμηση (regression) και η ταξινόμηση (classification).

- Παλινδρόμηση: Το πρόβλημα εκτίμησης μιας συνεχούς ποσότητας.
- Ταξινόμηση: Το πρόβλημα ανάθεσης των δεδομένων σε διαφορετικές διακριτές κατηγορίες που ονομάζονται κλάσεις.

Μη Επιβλεπόμενη Μάθηση

Στην περίπτωση της μη επιβλεπόμενης μάθησης, δίνεται ένα σύνολο από δεδομένα, αλλά χωρίς να δίνεται τίποτε παραπάνω π.χ. κάποια περαιτέρω γνώση για το κάθε δεδομένο. Ο στόχος της μη επιβλεπόμενης μάθηση ποικίλει αλλά σε γενικές γραμμές βασίζεται στην ικανότητα των μοντέλων να βρουν μοτίβα τα οποία εμφανίζονται συχνά πάνω στα δεδομένα χωρίς καμία επιπλέον επισήμανση των δεδομένων από ανθρώπους.

Ένα παράδειγμα είναι οι αλγόριθμοι ομαδοποίησης (clustering) στόχος των οποίων είναι να χωρίσουν τα δεδομένα σε διαφορετικές ομάδες με τέτοιο τρόπο ώστε τα δεδομένα κάθε ομάδας να είναι πιο όμοια μεταξύ τους απ'ό,τι δεδομένα από διαφορετικές ομάδες. Αυτό γίνεται με μεθόδους όπως το k-means ή τα Self organizing Maps.

Ένα άλλο παράδειγμα είναι η κωδικοποίηση της των δεδομένων σε κάποιο χώρο διαφορετικής διάστασης, συνήθως μικρότερης από την αρχική (dimensionality reduction). Αυτό μπορεί να συμβαίνει για λόγους οικονομίας χώρου, για λόγους οπτικοποίησης των δεδομένων, η σαν ενδιάμεσο βήμα σε κάποιο άλλο πρόβλημα. Ιδανικά, θέλουμε κάθε διάσταση της μικρότερης αναπαράστασης, κάθε δηλαδή feature, να αντιστοιχεί σε κάποια ερμηνεύσιμη περιγραφή του δεδομένου. Σχετικοί μέθοδοι είναι οι Autoencoders ή πιο παραδοσιακές μέθοδοι όπως η Ανάλυση σε Κύριες Συνιστώσες (Principal Component Analysis - PCA).

Τέλος, μια κατηγορία πολύ σχετική με την παρούσα εργασία είναι τα γεννητικά μοντέλα (Generative Models). Στόχος αυτών των μοντέλων είναι η μίμηση της διαδικασία δημιουργίας των δεδομένων εκπαίδευσης. Στην ιδανική περίπτωση, ένα τέτοιο μοντέλο θα πρέπει να μπορεί να δημιουργήσει νέα δεδομένα τα οποία, αν και τεχνητά, δεν μπορούν να διαχωριστούν από τα πραγματικά. Σαν είσοδος αυτών των μοντέλων πολλές φορές χρησιμοποιείται ένα τυχαίο σήμα από μία δεδομένη κατανομή πιθανότητας, έτσι ώστε τα μοντέλα να μην παράγουν το ίδιο δεδομένο κάθε φορά αλλά να έχουν την απαραίτητη ποικιλία. Έτσι τα μοντέλα μαθαίνουν μία συνάρτηση απεικόνισης από την δοσμένη κατανομή (που ονομάζεται latent κατανομή), στην κατανομή των δεδομένων.

Όταν η είσοδος είναι και αυτή ένα δεδομένο που όμως ανήκει σε διαφορετική ομάδα από την επιθυμητή έξοδο, τότε το πρόβλημα λέγεται πρόβλημα μετατροπής (conversion) και μπορεί να είναι supervised ή unsupervised ανάλογα με τα αν υπάρχει κάποια δοσμένη αντιστοιχία μεταξύ των δεδομένων που ανήκουν σε διαφορετικές ομάδες.

3.2 Συνάρτηση Σφάλματος

Όπως είδαμε στην προηγούμενη ενότητα, ο στόχος των αλγορίθμων επιβλεπόμενης μάθησης είναι η προσέγγιση μίας συνάρτησης από τα δεδομένα σε κάποια δοσμένη τιμή. Ο τρόπος που επιτυγχάνεται αυτός ο σκοπός είναι πολύ συχνά (όπως στην περίπτωση των νευρωνικών δικτύων) χρησιμοποιώντας μια επιπλέον συνάρτηση που αντιπροσωπεύει την απόδοση του μοντέλου. Η συνάρτηση αυτή είναι συνάρτηση των δεδομένων του προβλήματος, αλλά και των παραμέτρων του μοντέλου που χρησιμοποιείται. Στη συνέχεια, με μια διαδικασία που θα εξηγηθεί αργότερα, βρίσκουμε ένα σύνολο από παραμέτρους στις οποίες η συνάρτηση παίρνει μια επιθυμητή τιμή, συνήθως ελάχιστο ή μέγιστο. Όταν αυτή η συνάρτηση ποσοτικοποιεί το σφάλμα της συνάρτησης που προσεγγίζει το μοντέλο στα δεδομένα, λέγεται συνάρτηση σφάλματος ή απώλεια και ο σκοπός της εκπαίδευσης είναι η ελαχιστοποίηση της.

Δεδομένου ενός συνόλου εκπαίδευσης με δεδομένα $x_i,\ i = 1, \ldots, n$, labels $y_i,\ i = 1, \ldots, n$, ενός μοντέλου που υλοποιεί την συνάρτηση f η οποία με παραμέτρους θ προσπαθεί να προβλέψει τα labels από τα δεδομένα, αρχικά ορίζουμε μια συνάρτηση κόστους ανά δεδομένο $L(y, \hat{y})$ όπου $\hat{y}_i = f(x_i, \theta)$. Στη συνέχεια, ορίζουμε τη συνολική απώλεια του μοντέλου με παραμέτρους θ πάνω σε όλο το σύνολο των δεδομένων ως τη μέση απώλεια πάνω σε όλα τα δεδομένα εκπαίδευσης.

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i, \theta)) \tag{3.2.1}$$

Ο στόχος τώρα της εκπαίδευσης είναι να υπολογίσουμε την τιμή του θ για την οποία έχουμε το ελάχιστο σφάλμα, δηλαδή:

$$\hat{\theta} = \arg\min_{\theta} L(\theta) = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i, \theta)) \tag{3.2.2}$$

Παρατηρούμε ότι η συνάρτηση σφάλματος είναι συνάρτηση μόνο των παραμέτρων του μοντέλου αφού η εξάρτηση από κάθε δεδομένο εξαφανίστηκε αθροίζοντας. Φυσικά η εξάρτηση από τα δεδομένα παραμένει έμμεση από τον ορισμό της συνάρτησης. Ελπίζουμε ότι με κατάλληλη επιλογή των δεδομένων, η συνάρτηση είναι αρκετά γενική ώστε να αποτελεί καλή μετρική για το πρόβλημα, και ότι ένα μοντέλο που πετυχαίνει ικανοποιητική τιμή σε αυτή τη συνάρτηση θα λύνει ικανοποιητικά το πρόβλημα και για εισόδους που δεν έχει ξαναδεί.

Η επιλογή της συνάρτησης σφάλματος είναι ένα πολύ σημαντικό κομμάτι της μηχανικής μάθησης. Συνήθως επιλέγεται ώστε να αντιπροσωπεύει ένα μέτρο απόστασης μεταξύ των δύο τιμών, που στην γενική περίπτωση είναι διανύσματα, με τέτοιο τρόπο ώστε όταν το διάνυσμα που βγάζει το μοντέλο είναι ίδιο με το ζητούμενο, το σφάλμα είναι 0. Κοινές επιλογές για συνάρτηση σφάλματος είναι οι p-νόρμες οι οποίες ορίζονται:

$$||x - y||_p = \left(\sum_{i=1}^{\dim(x)} |x_i - y_i|^p\right)^{1/p} \tag{3.2.3}$$

Ιδιαίτερα συνηθισμένη είναι η 1-νόρμα που λέγεται και απόλυτο σφάλμα, καθώς και το τετράγωνο της 2-νόρμας που αντιστοιχεί στην Ευκλείδεια απόσταση και ονομάζεται τετραγωνικό σφάλμα. Άλλες συναρτήσεις σφάλματος θα αναλυθούν αργότερα.

Οι συναρτήσεις σφάλματος, χρησιμοποιούνται και στη μη επιβλεπόμενη μάθηση, απλά σε αυτή τη περίπτωση δεν αντιστοιχούν απλά σε μια μετρική απόσταση από τις εξόδους του μοντέλου με τις επιθυμητές εξόδους, και είναι πολύ διαφορετικές μεταξύ τους ανάλογα με το πρόβλημα. Η βασική ιδέα πάλι παραμένει ότι η συνάρτηση αντιπροσωπεύει το πόσο καλά το μοντέλο λύνει το ζητούμενο πρόβλημα.

Στα περισσότερα προβλήματα, ο τρόπος που βρίσκονται οι ζητούμενες τιμές των παραμέτρων θ είναι με την μέθοδο Gradient Descent. Ο τρόπος που επιλέγουμε την μοντελοποίηση της συνάρτησης είναι έτσι ώστε η έξοδος του μοντέλου να είναι παραγωγίσιμη ως προς τις παραμέτρους. Επίσης, η συνάρτηση σφάλματος επιλέγεται να είναι και αυτή παραγωγίσιμη και σαν αποτέλεσμα, μπορούμε να βρούμε την παράγωγο της συνάρτησης σφάλματος ως προς την κάθε παράμετρο του μοντέλου. Επειδή οι παράμετροι είναι περισσότεροι από μία, ουσιαστικά βρίσκουμε το διάνυσμα που αποτελεί την κλίση της συνάρτησης σφάλματος, το οποίο μας δείχνει την κατεύθυνση στην οποία η συνάρτηση σφάλματος μεγαλώνει με τον μεγαλύτερο ρυθμό. Στη συνέχεια αλλάζουμε τις τιμές των παραμέτρων ώστε να κινηθούν προς στην αντίθετη κατεύθυνση από την κλίση, κάτι που θα μειώσει την τιμή της συνάρτησης σφάλματος (υποθέτοντας πως το βήμα που κάναμε είναι αρκετά μικρό). Στο τέλος της εκπαίδευσης, ελπίζουμε ότι θα έχουμε καταλήξει σε ένα τοπικό ελάχιστο της συνάρτησης και ιδανικά σε ένα τοπικό ελάχιστο με όσο το δυνατόν μικρότερη τιμή σφάλματος. Αυτό βέβαια δεν εξασφαλίζεται μαθηματικά εκτός από σε πολύ περιορισμένες περιπτώσεις.

Συγκεκριμένα, οι τιμές των παραμέτρων στην επανάληψη n ανανεώνονται ως εξής:

$$\theta_{n+1} = \theta_n - \gamma \nabla L(\theta_n) \tag{3.2.4}$$

όπου το γ ονομάζεται ρυθμός μάθησης.

Επειδή η συνάρτηση κόστους εξαρτάται από όλα τα δεδομένα και το κόστος υπολογισμού της σε χρόνο και μνήμη είναι μεγάλο, συνήθως η βελτιστοποίηση της συνάρτησης γίνεται μεσω του stochastic gradient descent που προσεγγίζει την κλίση της συνάρτησης χρησιμοποιώντας μόνο μερικά δείγματα από τα δεδομένα:

$$\theta_{n+1} = \theta_n - \gamma \nabla \frac{1}{b} \sum_{i=1}^{b} L(y_i, f(x_i, \theta_n)) \tag{3.2.5}$$

όπου το b ονομάζεται batch size.

Παράμετροι όπως ο ρυθμός μάθησης και το batch size, που καθορίζονται πριν την εκπαίδευση και δεν αλλάζουν ανάλογα με την διαδικασία του Gradient Descent ονομάζονται υπερπαράμετροι.

Στην πράξη, χρησιμοποιούνται πολλοί διαφορετικοί αλγόριθμοι βελτιστοποίησης, οι οποίοι αποτελούν παραλλαγές του stochastic gradient descent και εξασφαλίζουν καλύτερη ταχύτητα σύγκλισης, μεγαλύτερη πιθανότητα σύγκλισης σε επιθυμητό σημείο ή άλλα επιθυμητά χαρακτηριστικά. Αυτό συχνά γίνεται αλλάζοντας το ρυθμό μάθησης κατά την διάρκεια της εκπαίδευσης με κάποιο προκαθορισμένο τρόπο. Για παράδειγμα, καθώς το μοντέλο γίνεται καλύτερο, ο ρυθμός μάθησης πολύ συχνά μειώνεται έτσι ώστε το μοντέλο να μάθει πιο πολλές λεπτομέρειες καθώς δεν χρειάζονται μεγάλες αλλαγές στα βάρη. Ένας πολύ συνηθισμένος αλγόριθμος βελτιστοποίησης είναι ο Adam [25].

3.3 Έννοιες από Θεωρία Πληροφορίας

Η θεωρία πληροφορίας έχει πολλές έννοιες που έχουν άμεση εφαρμογή στο πεδίο της μηχανικής μάθησης. Ίσως η πιο σημαντική έννοια είναι αυτή της εντροπίας η οποία περιγράφει το ελάχιστο αναμενόμενο αριθμό των bits που χρειάζονται για να επικοινωνήσεις ένα δείγμα από μια κατανομή πιθανότητας, με κατάλληλη δυαδική κωδικοποίηση των τιμών της συνάρτησης. Ορίζεται ως:

$$H(x) = -\sum_{x} P(x) \log(P(x)) \tag{3.3.1}$$

όταν η p είναι μια διακριτή συνάρτηση κατανομής πιθανότητας και ως:

$$H(x) = -\int_x p(x)\log(p(x))dx \tag{3.3.2}$$

όταν η p είναι μια συνάρτηση μάζας πιθανότητας. Αν η εντροπία είναι μεγάλη τότε αυτό σημαίνει πως η κατανομή παίρνει πολλές τιμές με μικρή πιθανότητα την κάθε μία, ενώ αν είναι μικρή σημαίνει πως παίρνει λιγότερες τιμές με μεγάλη πιθανότητα. Κατά συνέπεια μπορεί να χρησιμοποιηθεί σαν μέτρο αβεβαιότητας της κατανομής.

Η συνάρτηση cross entropy ορίζεται ως:

$$H(p,q) = -\sum_x p(x)\log(q(x)) \tag{3.3.3}$$

και αντιστοιχεί στον αναμενόμενο αριθμό bits που χρειάζονται για να επικοινωνήσεις ένα δείγμα από την κατανομή p όταν ο κώδικας έχει βελτιστοποιηθεί για την κατανομή q. Μια σημαντική υποπερίπτωση της cross entropy συνάρτησης είναι η binary cross entropy, όταν η κατανομή p παίρνει δύο τιμές με πιθανότητες $y, (1-y)$. Αυτή η συνάρτηση χρησιμοποιείται συχνά σαν συνάρτηση σφάλματος όταν έχουμε πρόβλημα δυαδικής ταξινόμησης.

Έστω πως έχουμε ένα dataset το οποίο αποτελείται από τα δείγματα $x_i,\ i = 1,\dots,n$ στα οποία αντιστοιχούν τα labels $y_i,\ i = 1,\dots,n$ τα οποία αντιστοιχούν στην πιθανότητα το δείγμα i να ανήκει στην κλάση 1. Φυσικά, αφού για τα δεδομένα εκπαίδευσης σε supervised learning ξέρουμε με βεβαιότητα σε ποια κλάση ανήκουν, έχουμε ότι $y_i \in \{0,\ 1\}\ , \forall i \in 1,\dots,n$. Επίσης θεωρούμε πως για να προβλέψουμε την κάθε τιμή y_i έχουμε έναν ταξινομητή που μοντελοποιούμε σαν συνάρτηση $f(x)$ που εκτιμά την πιθανότητα $f(x) = P(y=1|x)$, και ορίζουμε σαν $\hat{y}_i \equiv f(x_i)$, τις εξόδους αυτής της συνάρτησης στα δείγματα του προβλήματος. Συνήθως η έξοδος του ταξινομητή αποτελείται από την σύνθεση μιας άλλης συνάρτησης (π.χ. γραμμικής συνάρτησης στην περίπτωση της λογιστικής παλινδρόμησης) με την λογιστική συνάρτηση η οποία ορίζεται ως:

$$\sigma(x) = \frac{e^x}{1+e^x} \tag{3.3.4}$$

Έτσι εξασφαλίζουμε ότι θα ισχύει $0 < f(x) < 1$, το οποίο είναι ζητούμενο αφού οι έξοδοι της συνάρτησης αντιπροσωπεύουν συναρτήσεις πιθανότητας.

Τότε, η συνάρτηση σφάλματος που χρησιμοποιείται είναι:

$$L(y,\hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}[\ y_i\log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\ ] \tag{3.3.5}$$

Αντίστοιχα, στην περίπτωση όπου έχουμε περισσότερες από δύο κλάσεις, έστω m, ο ταξινομητής βγάζει σαν έξοδο m τιμές που εκτιμούν τις πιθανότητες $f(x)_j = P(y=j|x), \forall j \in 1,\dots,m$. Σε αυτή την περίπτωση, για να εξασφαλίσουμε ότι $\sum_{j=1}^{m} f(x)_j = 1$ ώστε ο ταξινομητής εκτιμά έγκυρη συνάρτηση κατανομής πιθανότητας, χρησιμοποιούμε την συνάρτηση softmax $\sigma : \mathbb{R}^m \to \mathbb{R}^m$ που ορίζεται ως:

$$\sigma(\boldsymbol{z})_j = \frac{e^{z_j}}{\sum_i^m e^{z_i}}, \text{για } i = 1,\dots,m \text{ και } \boldsymbol{z} = (z_1,\dots,z_m) \in \mathbb{R}^m \tag{3.3.6}$$

Η συνάρτηση softmax είναι συνεχής (και μάλιστα παραγωγίσιμη) προσέγγιση της διακριτής συνάρτησης argmax η οποία, αν υποθέσουμε ότι το διάνυσμα $\boldsymbol{z}$ έχει τη μέγιστη τιμή του στη θέση κ, θα επιστρέψει ένα διάνυσμα ίδιου μεγέθους με το $\boldsymbol{z}$, που θα έχει παντού μηδενικά εκτός από την θέση κ. Η συνάρτηση softmax έχει επίσης την θεμιτή ιδιότητα ότι είναι αμετάβλητη από την πρόσθεση μιας σταθεράς σε κάθε κλάση, όπως προκύπτει άμεσα από τον τύπο.

Η συνάρτηση κόστους που χρησιμοποιείται στην ταξινόμηση με πολλές κλάσεις, είναι:

$$L(y,\hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\ y_{ij}\log(\hat{y}_{ij}) \tag{3.3.7}$$

όπου $y_{ij} = f(x_i)_j$ και y_{ij} είναι 1 ανν το x_i ανήκει στην κλάση j.

Η εξίσωση 3.3.7 μπορεί να γραφτεί και σαν:

$$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \log(\hat{\boldsymbol{y}}_i) \tag{3.3.8}$$

όπου $\hat{\boldsymbol{y}}_i = f(x_i) \in \mathbb{R}^m$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im}) \in \mathbb{R}^m$ και το γινόμενο μεταξύ των διανυσμάτων είναι το εσωτερικό γινόμενο. Η μορφή του $\boldsymbol{y}$ που αποτελείται από μηδενικά παντού εκτός από την κλάση στην οποία ανήκει το δείγμα στην οποία έχει ένα ονομάζεται one hot vector.

Η πιθανοφάνεια (likelihood) που αποδίδει το μοντέλο στις σωστές κλάσεις στο σύνολο των δεδομένων είναι (υποθέτοντας ανεξάρτητα δεδομένα) :

$$\mathcal{L}(D) = \prod_{i}^{n}\prod_{j=1}^{m} P(y_i = j|x_i)^{y_{ij}} \tag{3.3.9}$$

και ο λογάριθμος της (log-likelihood) είναι:

$$\begin{aligned}
\log \mathcal{L}(D) &= \sum_{i=1}^{n}\sum_{j=1}^{m} \log\left(P(y_i = j|x_i)^{y_{ij}}\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} \log\left(P(y_i = j|x_i)\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} \log(\hat{y}_{ij}) \\
&= -H(y, \hat{y})
\end{aligned} \tag{3.3.10}$$

Άρα, μπορούμε να συμπεράνουμε ότι η μεγιστοποίηση της πιθανοφάνειας είναι ισοδύναμη με την ελαχιστοποίηση του σφάλματος cross entropy αφού ο λογάριθμος είναι μονότονη συνάρτηση.

Τέλος, ορίζουμε την έννοια της απόκλισης Kullback-Leibler η οποία δίνεται από τον τύπο:

$$KL(p||q) = -\sum_{x} p(x)\frac{\log(q(x))}{\log(p(x))}$$

Η απόκλιση Kullback-Leibler αντιστοιχεί στον αναμενόμενο αριθμό από επιπλέον bits που χρειάζονται για να επικοινωνήσεις ένα δείγμα από την κατανομή p όταν ο κώδικας έχει βελτιστοποιηθεί για την κατανομή q και μπορεί να χρησιμοποιηθεί σαν μέτρο απόστασης μεταξύ δύο κατανομών. Από τον ορισμό, μπορούμε να δούμε ότι:

$$KL(p||q) = H(p, q) - H(p)$$

Κατα συνέπεια, όταν η συνάρτηση p είναι δεδομένη όπως στην περίπτωση του binary cross entropy loss, μπορούμε να δούμε ότι το να βελτιστοποιούμε την $H(p, q)$ και την $KL(p||q)$ είναι ισοδύναμα.

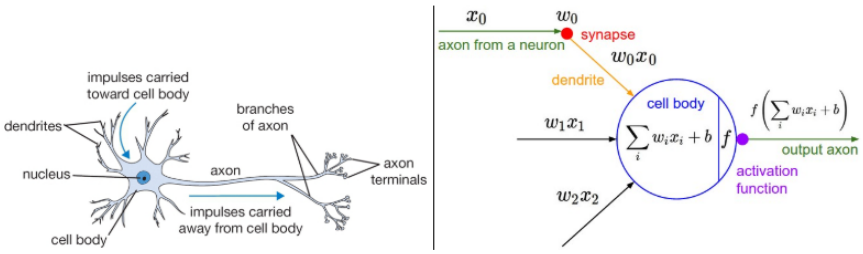
3.4 Τεχνητά Νευρωνικά Δίκτυα

3.4.1 Εισαγωγή

Τα τεχνητά νευρωνικά δίκτυα (στο εξής νευρωνικά δίκτυα) είναι μία μέθοδος μηχανικής μάθησης που έχει αποκτήσει μεγάλη δημοτικότητα τα τελευταία χρόνια. Έχουν πολύ ευρύ πεδίο εφαρμογής και έχουν χρησιμοποιηθεί με επιτυχία σε εφαρμογές επεξεργασίας και παραγωγής ήχου, όρασης υπολογιστών, επεξεργασίας και αυτόματης μετάφρασης κειμένου κλπ. Χρησιμοποιούν ως έμπνευση την δομή που έχουν οι βιολογικοί νευρώνες του εγκεφάλου και αρχικά είχαν ως σκοπό την μοντελοποίηση τους, αν και πλέον χρησιμοποιούνται σχεδόν αποκλειστικά σε προβλήματα μηχανικής μάθησης.

Η βασική υπολογιστική μονάδα του εγκεφάλου, είναι ένα ειδικό κύτταρο που ονομάζεται νευρώνας. Κάθε νευρώνας συνδέεται με άλλους νευρώνες με την βοήθεια ηλεκτρο-χημικών σήματων. Η είσοδος των σημάτων

γίνεται με την βοήθεια ηλεκτρικού δυναμικού στους δενδρίτες, και η έξοδος γίνεται μέσω του άξονα, τα τερματικά του οποίου συνδέονται με τους δενδρίτες ενός άλλου νευρώνα με τη βοήθεια συνδέσεων που ονομάζονται συνάψεις. Κάθε σύναψη μπορεί να διεγείρει ή να αναστείλει το δυναμικό που έχει στην είσοδο της, και μάλιστα κατά διαφορετικό ποσοστό από σύναψη σε σύναψη. Αυτό μοντελοποιείται σαν μια γραμμική συνάρτηση των σημάτων εισόδου, όπως φαίνεται στο Σχήμα 3.4.1. Όταν η τιμή της εισόδου ξεπεράσει κάποια δεδομένη τιμή, ο άξονας του νευρώνα θα εμφανίσει μια ηλεκτρική ώση που ονομάζεται δυναμικό ενέργειας (action potential). Αυτό θα διατηρηθεί για ένα μικρό χρονικό διάστημα (της τάξης των millisecond). Το δυναμικό ενέργειας είναι μια απάντηση του τύπου "όλα ή ουδέν" (δηλαδή, ή θα παραχθεί στην πλήρη του μορφή, ή καθόλου) και η μοντελοποίηση του γίνεται μέσω μιας συνάρτησης που ονομάζεται συνάρτηση ενεργοποίησης (activation function). Στην απλούστερη μορφή της είναι απλά μια δυαδική συνάρτηση που βγάζει 1 όταν ο άξονας βρίσκεται στο δυναμικό ενέργειας και 0 στο δυναμικό ηρεμίας. Πολλές φορές όμως χρησιμοποιούνται συναρτήσεις που αντιστοιχούν στο ρυθμό με τον οποίο ο νευρώνας στέλνει σήματα εξόδου και είναι συνεχείς. Ο εγκέφαλος αποτελείται από περίπου 80 δισεκατομμύρια νευρώνες ο κάθε ένας από τους οποίους συνδέεται με περίπου 7 χιλιάδες νευρώνες.



Σχήμα 3.4.1: Ένας βιολογικός νευρώνας (αριστερά) και η αντιστοιχία του με τον τεχνητό νευρώνα (δεξιά). Σχήμα από [26].

3.4.2 Perceptron

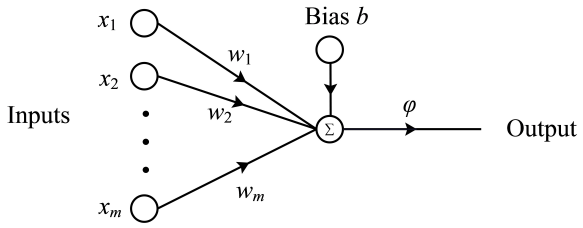
Ο αλγόριθμος Perceptron είναι η απλούστερη μορφή ενός νευρωνικού δικτύου. Αποτελεί έναν αλγόριθμο που κάνει δυαδική ταξινόμηση των δεδομένων εισόδου, τα οποία θεωρούμε ότι αποτελούνται από δύο κλάσεις και ότι αναπαριστώνται από διανύσματα $\mathbf{x} \in \mathbb{R}^m$. Οι παράμετροι του Perceptron αποτελούνται από ένα διάνυσμα βαρών w_i, $i = 1 \dots m$, το οποίο πολλαπλασιάζεται εσωτερικά με το διάνυσμα εισόδου, και ενός αριθμού b το οποίο ονομάζεται bias και προστίθεται στο εσωτερικό γινόμενο του διανύσματος εισόδου με τα βάρη. Ο μετασχηματισμός αυτός μαζί με το bias ονομάζεται αφινικός. Στην συνέχεια, το αποτέλεσμα περνάει από μία συνάρτηση ενεργοποίησης ϕ (οι διάφορες συναρτήσεις ενεργοποίησης αναλύονται στη συνέχεια) και βγάζει την έξοδο y. Στην απλούστερη περίπτωση, η έξοδος της ϕ είναι 1 όταν η είσοδος είναι θετική και 0 όταν είναι αρνητική, κάτι που αντιστοιχεί στις δύο κλάσεις του προβλήματος. Η αρχιτεκτονική του φαίνεται στο Σχήμα 3.4.2 και η εξίσωση του είναι:

$$y = \phi(\mathbf{w}^T\mathbf{x} + b) = \phi(\sum_i^m w_i x_i + b) \tag{3.4.1}$$

Η εξίσωση 3.4.1 περιγράφει την εξίσωση ενός υπερεπιπέδου στον χώρο $\mathbb{R}^m$.Τα δεδομένα τοποθετούνται στην μια κλάση ή την άλλη ανάλογα με τη πλευρά του υπερεπιπέδου στην οποία βρίσκονται. Κατά συνέπεια, ο αλγόριθμος Perceptron μπορεί να ταξινομήσει μόνο πρότυπα που είναι γραμμικώς διαχωρίσιμα. Τα περισσότερα σύνθετα προβλήματα όμως, δεν είναι γραμμικώς διαχωρίσιμα και απαιτούν ισχυρότερες αρχιτεκτονικές από το απλό Perceptron. Ένα πολύ απλό παράδειγμα προβλήματος που δεν μπορεί να λυθεί με γραμμικό επίπεδο είναι το πρόβλημα XOR το οποίο περιλαμβάνει δύο κλάσεις με τα δεδομένα $(0,0)$ $(1,1)$ και $(1,0)$ $(0,1)$ αντίστοιχα.

3.4.3 Βαθιά Νευρωνικά Δίκτυα

Τα βαθιά νευρωνικά δίκτυα (Deep Neural Networks) αποτελούν μια γενίκευση του Perceptron που μπορεί να προσεγγίσει πιο περίπλοκες μη γραμμικές συναρτήσεις και κατά συνέπεια να λύσει περισσότερα προβλήματα.

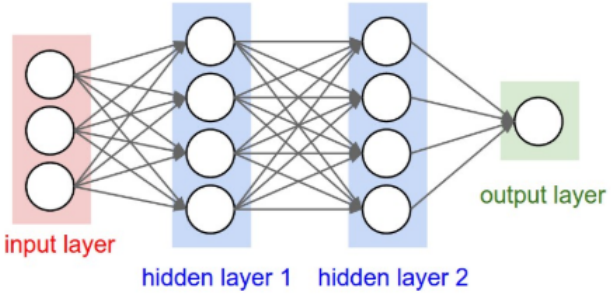


Σχήμα 3.4.2: Αρχιτεκτονική ενός μοντέλου Perceptron

Μάλιστα, έχει αποδειχθεί ότι για κάθε δυνατή συνάρτηση πολλών μεταβλητών μπορεί να κατασκευαστεί ένα νευρωνικό δίκτυο κατάλληλου μεγέθους το οποίο την προσεγγίζει με οποιαδήποτε ζητούμενη ακρίβεια.

Η αρχιτεκτονική των βαθιών νευρωνικών δικτύων αποτελείται από πολλά επίπεδα, κάθε ένα εκ των οποίων αποτελείται από έναν αριθμό νευρώνων παρόμοιων με αυτόν του Perceptron. Πιο συγκεκριμένα, η αρχιτεκτονική φαίνεται στο Σχήμα 3.4.3 και αποτελείται από τα εξής:

- Επίπεδο εισόδου (*input layer)* το οποίο απλά περνάει την πληροφορία της εισόδου στα επόμενα επίπεδα
- Κρυφά επίπεδα (*hidden layer*) τα οποία μετασχηματίζουν την έξοδο του προηγούμενου επιπέδου
- Επίπεδο εξόδου (*output layer*)



Σχήμα 3.4.3: Σχηματική αναπαράσταση ενός νευρωνικού δικτύου με 2 κρυφά επίπεδα. Σχήμα από [26].

Κάθε κρυφό επίπεδο μετασχηματίζει την έξοδο του προηγούμενου επιπέδου με έναν αφινικό μετασχηματισμό ακολουθούμενο από μια συνάρτηση ενεργοποίησης όπως περιγράφεται στην εξίσωση:

$$y_i^l = \phi^l(\sum_j^m w_{ij}^l y_j^{l-1} + b_i^l) \tag{3.4.2}$$

όπου ϕ είναι η συνάρτηση ενεργοποίησης, w_{ij} είναι το βάρος της σύνδεσης του νευρώνα i με τον νευρώνα j, b_i είναι το bias που προστίθεται στον νευρώνα i, y_j είναι η έξοδος j και ο εκθέτης l αναπαριστά το κρυφό επίπεδο.

Με χρήση διανυσμάτων και πινάκων μπορούμε να γράψουμε την εξίσωση 3.4.2 ως:

$$\mathbf{y}^l = \phi^l(\mathbf{W}^l\mathbf{y}^{l-1} + \mathbf{b}^l) \tag{3.4.3}$$

Back Propagation

Η εκπαίδευση των νευρωνικών δικτύων γίνεται με τον αλγόριθμο Gradient Descent όπως αναλύθηκε σε προηγούμενη ενότητα. Για να δουλέψει ο αλγόριθμος, χρειάζεται σε κάθε επανάληψη να υπολογίσουμε την μερική παράγωγο της συνάρτησης σφάλματος ως προς όλες τις παραμέτρους όλων των επιπέδων του δικτύου.

Αυτό μπορεί να επιτευχθεί αποτελεσματικά με τον αλγόριθμο Back Propagation. Αρχικά βρίσκεται η παράγωγος της συνάρτησης σφάλματος ως προς τις παραμέτρους του τελευταίου επιπέδου με άμεσο τρόπο. Στη συνέχεια, χρησιμοποιώντας τον κανόνα της αλυσίδας $\frac{df}{dx} = \frac{df}{dy}\frac{dy}{dx}$ μπορούμε να βρούμε τις παραγώγους των παραμέτρων του αμέσως προηγούμενου επιπέδου επαναχρησιμοποιώντας τις μερικές παραγώγους που έχουμε ήδη υπολογίσει και τις γνωστές παραγώγους των συναρτήσεων ενεργοποίησης και των αφινικών μετασχηματισμών. Συνεχίζουμε αυτή την διαδικασία μέχρι το πρώτο επίπεδο και στο τέλος ανανεώνουμε τα βάρη με κατάλληλο τρόπο ώστε να μειωθεί το σφάλμα. Η όλη διαδικασία επαναλαμβάνεται μέχρι να βρεθούμε σε ένα ζητούμενο τοπικό ελάχιστο, ολοκληρώνοντας την εκπαίδευση.

Κανονικοποίηση (Regularization)

Όπως αναφέρθηκε και προηγουμένως, η αφελής ελαχιστοποίηση της συνάρτησης σφάλματος, μπορεί να οδηγήσει στο φαινόμενο του overfitting δηλαδή ένα δίκτυο που απλά απομνημονεύει τα δεδομένα εισόδου και έχει κακή απόδοση στο test set. Το ίδιο πρόβλημα υπάρχει όταν το μοντέλο μαθαίνει χαρακτηριστικά που υπάρχουν στο dataset αλλά είναι προϊόν τυχαίου θορύβου των συγκεκριμένων δεδομένων και δεν γενικεύει σε δεδομένα που δεν έχουν συναντηθεί στην εκπαίδευση. Το Regularization είναι η διαδικασία η οποία έχει στόχο την βελτίωση της ικανότητας γενίκευσης των μοντέλων και γίνεται με διάφορους τρόπους.

Ένας συχνός τρόπος Regularization είναι να προσθέσουμε στην συνάρτηση σφάλματος έναν επιπλέον όρο ως εξής:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i, \theta_n)) + \lambda R(\theta) \tag{3.4.4}$$

όπου το λ είναι υπερπαράμετρος.

Οι όροι κανονικοποίησης αποσκοπούν στο να μειωθεί η περιπλοκότητα του μοντέλου και πολύ συχνά κωδικοποιούν μια εκ των προτέρων γνώση (prior) που έχουμε για το σύστημα. Συχνές επιλογές είναι οι:

- L2 norm:
$$R(\theta) = \frac{1}{2}\lambda||\theta||_2^2 = \frac{1}{2}\lambda w^T w \tag{3.4.5}$$
όπου W το διάνυσμα βαρών του νευρωνικού δικτύου.

  Η L2 νόρμα, περιορίζει τα βάρη του δικτύου ώστε να μην γίνουν πολύ μεγάλα με τέτοιο τρόπο ώστε τα μεγαλύτερα βάρη να τιμωρούνται περισσότερο από τα μικρά και είναι ισοδύναμη με το να μειώνουμε τα βάρη κατά ένα ποσοστό σε κάθε επανάληψη (weight decay).

- L1 norm:
$$R(\theta) = \lambda||\theta||_1 \tag{3.4.6}$$

  Η L1 νόρμα, περιορίζει όλα τα βάρη σε μικρότερες τιμές ανεξαρτήτως από το μέγεθος τους και σαν αποτέλεσμα οδηγεί σε μικρότερη πυκνότητα στο δίκτυο, κάτι που αναφέρεται ως μεγαλύτερο sparsity. Αυτός ο τρόπος (Regularization) λέγεται και lasso [27].

Μια διαφορετική προσέγγιση που είναι πολύ δημοφιλής είναι το dropout [28], το οποίο σε κάθε επανάληψη μηδενίζει κάθε βάρος ενός δικτύου με κάποια δεδομένη πιθανότητα. Όταν το μοντέλο αξιολογείται, χρησιμοποιούνται όλα τα βάρη κανονικά. Η βασική ιδέα είναι ότι αυτή η προσέγγιση είναι σαν να εκπαιδεύουμε πολλά (συγκεκριμένα εκθετικά πολλά) μικρότερα νευρωνικά δίκτυα με κοινούς παραμέτρους (τα οποία προκύπτουν με την αφαίρεση συνδέσεων μεταξύ επιπέδων) και στο τέλος βρίσκουμε τον μέσο όρο όλων των προβλέψεων.

Για να αξιολογήσουμε την απόδοση ενός μοντέλου, με τρόπο που δεν είναι ευαίσθητος στο overfitting, διαλέγουμε ένα υποσύνολο των δεδομένων (σύνολο επαλήθευσης ή Validation set) τα οποία δεν χρησιμοποιούμε καθόλου κατά την διάρκεια της εκπαίδευσης. Το σύνολο επαλήθευσης χρησιμοποιείται επίσης για την επιλογή των υπερπαραμέτρων της εκπαίδευσης. Επίσης χρησιμοποιείται και ένα επιπλέον υποσύνολο δεδομένων το οποίο ονομάζεται test set, το οποίο δεν συμμετέχει ούτε στην διαδικασία εύρεσης των υπερπαραμέτρων, και χρησιμοποιείται μόνο για την τελική αξιολόγηση.

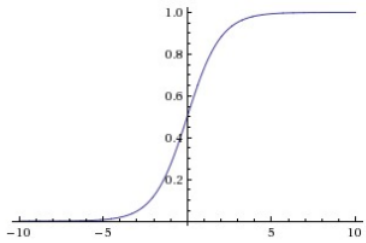
Συναρτήσεις Ενεργοποίησης

Επειδή η σύνθεση αφινικών μετασχηματισμών είναι αφινικός μετασχηματισμός, αν δεν υπήρχαν οι συναρτήσεις ενεργοποίησης, όλο το δίκτυο θα μπορούσε να αντικατασταθεί με ένα ισοδύναμο με μόνο ένα επίπεδο. Έτσι ο ρόλος των συναρτήσεων ενεργοποίησης είναι να προσθέτουν μια μη-γραμμικότητα στο δίκτυο και είναι απαραίτητος για να μπορεί το νευρωνικό δίκτυο να προσεγγίσει περίπλοκες συναρτήσεις.

Πολύ συχνές επιλογές για την συνάρτηση ενεργοποίησης είναι:

- *Σιγμοειδής (sigmoid)*

  Η σιγμοειδής συνάρτηση στα μαθηματικά, είναι οποιαδήποτε συνάρτηση η οποία έχει σχήμα όπως το 3.4.4 αλλά στο πλαίσιο της μηχανικής μάθησης αναφέρεται στην λογιστική συνάρτηση η οποία ορίστηκε στην εξίσωση 3.3.4 και είναι ιστορικά από τις πρώτες συναρτήσεις που χρησιμοποιήθηκαν σαν συνάρτηση ενεργοποίησης.



Σχήμα 3.4.4: Η σιγμοειδής συνάρτηση

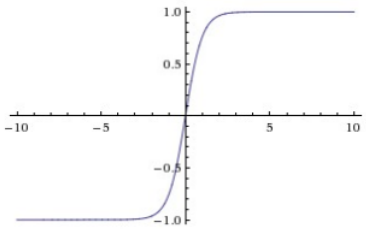
  Το πλεονέκτημα της συνάρτησης αυτής είναι ότι η παράγωγός της είναι $\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$ και σαν συνέπεια μπορεί να υπολογιστεί σχετικά γρήγορα κατά το back propagation χρησιμοποιώντας τις τιμές της συνάρτησης. Επειδή όμως η παράγωγός της είναι πάντα μικρότερη από το 1, όταν χρησιμοποιείται σε νευρωνικά δίκτυα με πολλά επίπεδα, οι παράγωγοι της συνάρτησης σφάλματος ως προς τις παραμέτρους των πρώτων επιπέδων έχει γίνει σχεδόν μηδέν με αποτέλεσμα να μην μπορούν να ανανεωθούν με τέτοιο τρόπο ώστε να μειωθεί το σφάλμα. Αυτό συμβαίνει επειδή σύμφωνα με τον κανόνα της αλυσίδας, όταν παραγωγίζουμε τις παραμέτρους ενός επιπέδου ως προς το επόμενο, κάθε παράγωγος πολλαπλασιάζεται με έναν αριθμό μικρότερο του ενός και η παράγωγος των πρώτων επιπέδων μειώνεται εκθετικά ως προς τον αριθμό των επιπέδων. Το πρόβλημα αυτό ονομάζεται vanishing gradient problem. Για αυτό το λόγο, στα βαθιά νευρωνικά δίκτυα, η σιγμοειδής συνάρτηση χρησιμοποιείται σχεδόν αποκλειστικά όταν θέλουμε να περιορίσουμε έναν αριθμό στο διάστημα $[0, 1]$, όπως για παράδειγμα στην δυαδική ταξινόμηση.

- *Υπερβολική εφαπτομένη (tanh)*

  Αυτή η συνάρτηση φαίνεται στο Σχήμα 3.4.5 και ορίζεται ως:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3.4.7}$$

  Σε αυτή την περίπτωση, η παράγωγος της συνάρτησης είναι $\frac{d}{dx}\tanh(x) = 1 - \tanh^2(x)$ οπότε έχουμε το ίδιο πλεονέκτημα και μειονέκτημα που είχαμε με την λογιστική συνάρτηση. Η διαφορά είναι ότι η έξοδος της υπερβολικής εφαπτομένης είναι στο $[-1, 1]$ αντί στο $[0, 1]$ και χρησιμοποιείται αναλόγως.

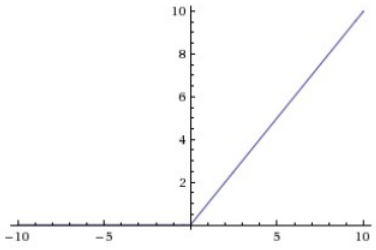


Σχήμα 3.4.5: Η συνάρτηση υπερβολικής εφαπτομένης tanh

- *ReLU (Rectified Linear Unit)*

  Αυτή η συνάρτηση φαίνεται στο Σχήμα 3.4.6 και ορίζεται ως:

$$f(x) = \max(x, 0) \tag{3.4.8}$$



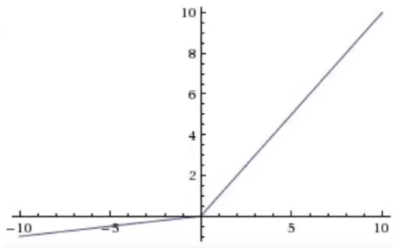
Σχήμα 3.4.6: Η συνάρτηση ReLU

Τα πλεονεκτήματα της συνάρτησης ReLU είναι ότι είναι πάρα πολύ γρήγορη να υλοποιηθεί, αφού ουσιαστικά αποτελείται μόνο από μια σύγκριση, και οτι αποφεύγει το vanishing gradient πρόβλημα. Αυτό εξηγείται από το ότι η παράγωγος της ReLU είναι ή 1, ή 0, ανάλογα με το πρόσημο της εισόδου και δεν μειώνεται η πληροφορία της παραγώγου κατά την διάρκεια της παραγώγισης ως προς τα πρώτα επίπεδα. Το μειονέκτημα που έχει η ReLU είναι ότι αν κάποιος νευρώνας δέχεται μόνο αρνητικές εισόδους, η έξοδος είναι πάντα μηδέν, και επειδή η παράγωγος είναι 0 στις αρνητικές τιμές, δεν υπάρχει τρόπος να αξιοποιηθεί η πληροφορία της εισόδου σε κανένα στάδιο της εκπαίδευσης. Σε αυτή την περίπτωση, η ReLU έχει κολλήσει και λέμε ότι είναι νεκρή.

- Leaky ReLU

  Αυτή η συνάρτηση φαίνεται στο Σχήμα 3.4.7 και ορίζεται ως:

$$f(x) = \begin{cases} \alpha x, & x < 0 \\ x, & x \geq 0 \end{cases} \tag{3.4.9}$$

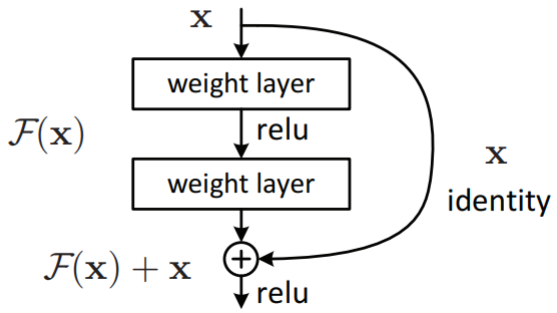


Σχήμα 3.4.7: Η συνάρτηση leaky ReLU

όπου το α είναι μια μικρή παράμετρος (π.χ. $\alpha = 0.01$). Η συμπεριφορά αυτής της συνάρτησης είναι πολύ παρόμοια με αυτήν της ReLU, με την διαφορά ότι αν η εκπαίδευση κολλήσει στην αρνητική περιοχή, υπάρχει πιθανότητα να ξεκολλήσει γιατί τώρα η παράγωγος στις αρνητικές τιμές δεν είναι 0.

Residual Connections

Παρά την χρήση συναρτήσεων όπως ReLU το vanishing gradient πρόβλημα εξακολουθεί να υπάρχει όταν εκπαιδεύονται αρχιτεκτονικές με πολύ μεγάλο βάθος. Ένας τρόπος επίλυσης αυτού του προβλήματος είναι με residual connections (ή αλλιώς skip connections) [29], οι οποίες φαίνονται στο Σχήμα 3.4.8. Αντί να περάσουμε το σήμα από πολλά διαφορετικά επίπεδα, χρησιμοποιούμε μια παράκαμψη και αφήνουμε το σήμα να περάσει στην έξοδο αλλά προσθέτοντας του την έξοδο των ενδιαμέσων επιπέδων. Η ιδέα είναι ότι ο άμεσος δρόμος είναι ισοδύναμος με το νευρωνικό δίκτυο που προκύπτει αν αφαιρέσουμε τα ενδιάμεσα επίπεδα και η εκπαίδευση είναι λιγότερο πιθανό να έχει το πρόβλημα. Το Back Propagation μπορεί να εκπαιδεύσει τα προηγούμενα επίπεδα χωρίς οι παράγωγοι να γίνονται πολύ μικρές, αλλά παράλληλα τα ενδιάμεσα επίπεδα μπορούν να αλλάξουν το σήμα που περνάει χωρίς να επηρεάζουν την εκπαίδευση του υπόλοιπου δικτύου σε δραματικό βαθμό.



Σχήμα 3.4.8: Απεικόνιση ενός residual connection

3.4.4 Autoencoders

Ένας Autoencoder είναι ένα νευρωνικό δίκτυο που μαθαίνει να απεικονίζει την είσοδό του στον εαυτό της, αφού την περάσει από ένα ενδιάμεσο διάνυσμα [30]. Αποτελείται από δύο διαφορετικά μικρότερα δίκτυα, τον Encoder και τον Decoder. Ο στόχος του πρώτου είναι να κωδικοποιήσει την αναπαράσταση της εισόδου στον χώρο του ενδιάμεσου διανύσματος. Αποτελεί δηλαδή μια απεικόνιση (mapping) $f : \mathbb{R}^N \to \mathbb{R}^M$ όπου N είναι η διάσταση των δεδομένων εισόδου και M η ενδιάμεση διάσταση. Ο ρόλος του Decoder είναι να υλοποιεί την αντίστροφη απεικόνιση $g : \mathbb{R}^M \to \mathbb{R}^N$. Στην ιδανική περίπτωση ισχύει $g(f(x)) = x$ δηλαδή η διαδικασία δεν χάνει καμμία πληροφορία από την αρχική είσοδο. Η εκπαίδευση των Autoencoders γίνεται όπως και στα υπόλοιπα νευρωνικά δίκτυα χρησιμοποιώντας σαν συνάρτηση σφάλματος μία μετρική απόστασης της εξόδους με την είσοδο, όπως την L1 ή L2 νόρμα.

Στην γενική περίπτωση ισχύει $M < N$. Αυτό συμβαίνει επειδή τα δεδομένα με τα οποία ασχολούμαστε, όπως εικόνες ή σήματα ήχου, έχουν συγκεκριμένη δομή και μεγάλο βαθμό από επαναλαμβανόμενα μοτίβα και η εντροπία τους είναι μικρότερη από τον αριθμό bits που χρησιμοποιούνται στην κωδικοποίηση τους χωρίς καμμία επεξεργασία. Έτσι, με την απεικόνιση σε έναν χώρο μικρότερων διαστάσεων, το νευρωνικό δίκτυο αναγκάζεται να βρει καλύτερη κωδικοποίηση από αυτή που υπάρχει εξ αρχής στα δεδομένα, αφαιρώντας πλεονασμούς και άχρηστες πληροφορίες που υπάρχουν στην αρχική αναπαράσταση. Αν αποτύχει να το κάνει αυτό, θα χαθεί χρήσιμη πληροφορία και αναγκαστικά η έξοδος του δικτύου δεν θα είναι ίδια με την είσοδο.

Οι Autoencoders χρησιμοποιούνται σε προβλήματα dimensionality reduction σαν εναλλακτική επιλογή από μεθόδους όπως η PCA που είναι γραμμική και δεν έχει τόση δυνατότητα γενίκευσης [31, 32] . Αυτό μπορεί να χρησιμοποιηθεί σαν ενδιάμεσο στάδιο ενός πιο περίπλοκου συστήματος καθώς για να καταφέρει ο Autoencoder να μειώσει την διάσταση των δεδομένων, θα πρέπει η κάθε διάσταση της κωδικοποίησης να αντιστοιχεί σε κάποια σημαντική πληροφορία που υπάρχει στο αρχικό σήμα (feature learning). Αυτή η πληροφορία μπορεί να χρησιμοποιηθεί πιο εύκολα από τα επόμενα στάδια του συστήματος. Ιδανικά, κάθε διάσταση θα πρέπει να

απεικονίζει κάποια ξεχωριστή πληροφορία, να είναι σημασιολογικά ερμηνεύσιμη, και να μην υπάρχει μεγάλη συσχέτιση μεταξύ των διαστάσεων. Αυτή η ιδιότητα λέγεται disentanglement.

Επίσης, οι Autoencoders χρησιμοποιούνται σαν de-noising συστήματα όπου ο σκοπός είναι να πάρουμε σαν είσοδο δεδομένα με μη θεμιτό θόρυβο και να ανακτήσουμε το αρχικό δεδομένο χωρίς τον θόρυβο. Η διαδικασία εκπαίδευσης τους είναι παρόμοια μόνο που η συνάρτηση σφάλματος είναι μεταξύ της εξόδου και του δεδομένου χωρίς θόρυβο αντί της εισόδου.

Τέλος, μια σημαντική παραλλαγή των Autoencoders είναι οι Variational Autoencoders [33]. Η διαφορά με τους κλασικούς, είναι ότι η ενδιάμεση αναπαράσταση της εισόδου αποτελεί την παραμετροποίηση μιας κατανομής πιθανότητας, την οποία υποθέτουμε σαν την κρυφή (latent) κατανομή που δημιουργεί τα δεδομένα. Συνήθως η κατανομή πιθανότητας είναι μια πολυδιάστατη Gaussian και οι παράμετροι που χρησιμοποιούνται είναι η μέση τιμή και η διασπορά. Μετά την εκπαίδευση των δικτύων, μπορούμε να κάνουμε δειγματοληψία από την κρυφή κατανομή και αφού το δείγμα περάσει από τον Decoder να έχουμε ένα καινούργιο δείγμα από τα δεδομένα, το οποίο δεν υπήρχε στο dataset. Οι Variational Autoencoders δηλαδή, αποτελούν γεννητικό μοντέλο.

3.4.5 Συνελικτικά Νευρωνικά Δίκτυα

Τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks ή CNNs) [34, 35] είναι μια ιδιαίτερη μορφή αρχιτεκτονικής βαθιών νευρωνικών δικτύων που χρησιμοποιούνται ιδιαίτερα στην επεξεργασία εικόνων και φωνής [36]. Βασίζονται στην πράξη της (διακριτής) συνέλιξης η οποία οποία είναι μια πράξη μεταξύ δυο συναρτήσεων που συμβολίζεται με το σύμβολο $*$ και ορίζεται μαθηματικά ως:

$$(f * g)(n) = \sum_{m=-\infty}^{\infty} f(m)g(n-m) \tag{3.4.10}$$

Συνήθως, η μία από τις δύο συναρτήσεις που χρησιμοποιούνται στην συνέλιξη ονομάζεται φίλτρο. Η πράξη της συνέλιξης είναι ιδιαίτερα χρήσιμη στην επεξεργασία σημάτων καθώς η έξοδος ενός γραμμικού χρονικώς αμετάβλητου συστήματος μπορεί να γραφεί σαν την συνέλιξη της εισόδου με κάποιο φίλτρο. Επίσης, οι συνελίξεις έχουν την χρήσιμη ιδιότητα ότι ο μετασχηματισμός Fourier της συνέλιξης δύο σημάτων είναι το γινόμενο των μετασχηματισμών Fourier των επιμέρους σημάτων.

Στο πλαίσιο της μηχανικής μάθησης, ένα συνελικτικό επίπεδο υπολογίζει την συνέλιξη της εισόδου με έναν δεδομένο αριθμό από φίλτρα, οι τιμές των οποίων αποτελούν τις παραμέτρους του επιπέδου. Στη περίπτωση που η είσοδος αποτελείται από περισσότερα από ένα σήματα (κανάλια), όπως προκύπτει για παράδειγμα αν είναι η έξοδος ενός προηγούμενου συνελικτικού επιπέδου, κάθε ένα από τα φίλτρα συνελίσσεται με κάθε είσοδο και η έξοδος είναι το σταθμισμένο άθροισμα όλων των εξόδων. Οι γραμμικοί όροι στάθμισης είναι και αυτοί παράμετροι. Η εξίσωση που δίνει την έξοδο ενός επιπέδου είναι:

$$y(C_{out_i}) = bias(C_{out_i}) + \sum_{k=1}^{C_{in}} weight(C_{out_i}, k) * x(k) \tag{3.4.11}$$

οπου όπου C_{in} C_{out} τα κανάλια εισόδου και εξόδου αντίστοιχα (τα κανάλια εξόδου είναι όσα και τα φίλτρα), το $weight(C_{out_i}, k)$ αντιστοιχεί στο φίλτρο με είσοδο κανάλι k και έξοδο i, και x είναι το σήμα εισόδου.

Η συνέλιξη μπορεί να γίνει σε μία διάσταση ή σε περισσότερες, ανάλογα με το πρόβλημα. Στην επεξεργασία εικόνας, το πιο σύνηθες είναι οι δισδιάστατες συνελίξεις, ενώ στην επεξεργασία φωνής χρησιμοποιούνται και δισδιάτατες αλλά και μονοδιάστατες συνελίξεις.

Η διαδικασία της συνέλιξης μπορεί να περιγραφεί ως εξής:

Έχουμε ένα φίλτρο το οποίο ξεκινάει από την αρχή του σήματος εισόδου, και στη συνέχεια το μετακινούμε μία-μία θέση μέχρι να φτάσει στο τέλος. Στην περίπτωση δισδιάστατου σήματος, η μετακίνηση γίνεται αρχικά στην πρώτη διάσταση και μετά κατά ένα βήμα στην δεύτερη διάσταση και ούτω καθεξής. Έτσι το φίλτρο θα περάσει από όλες τις δυνατές θέσεις στις οποίες υπάρχει επικάλυψη με το σήμα. Όταν υπάρχει μόνο μερική επικάλυψη, συνήθως θεωρείται πως το σήμα έχει μηδενικά στα σημεία που δεν ορίζεται. Σε κάθε θέση που βρίσκεται το φίλτρο πολλαπλασιάζουμε τις τιμές του φίλτρου με τις τιμές του σήματος που βρίσκονται σε αυτή την θέση και προσθέτουμε τα αποτελέσματα. Έχουμε έτσι μια τιμή για κάθε δυνατή θέση του φίλτρου με τέτοιο

τρόπο ώστε να επικαλύπτεται μερικώς με το σήμα και η έξοδος θα έχει παραπλήσιες διαστάσεις με το αρχικό σήμα. Για να επιλέξουμε τις ακριβείς διαστάσεις του σήματος εξόδου, τοποθετούμε τον απαιτούμενο αριθμό από μηδενικά είτε στην αρχή είτε στο τέλος του σήματος (padding).

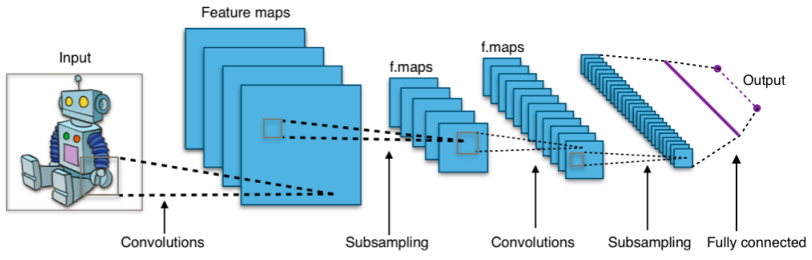
Αξίζει να σημειωθεί πως όταν το φίλτρο μετακινείται, δεν είναι απαραίτητο να μετακινηθεί κατά ένα βήμα. Ο αριθμός των βημάτων που γίνονται σε κάθε μετακίνηση λέγεται stride και έχει την ιδιότητα ότι όταν κάνουμε συνέλιξη με stride κ, το σήμα θα έχει μέγεθος περίπου όσο η είσοδος δια κ (το ακριβές σχήμα εξαρτάται απο το μήκος του φίλτρου). Αυτό σημαίνει ότι γίνεται μείωση μεγέθους στο σήμα (down-sampling), κάτι που χρησιμοποιείται σε δομές παρόμοιες με αυτές των Autoencoders. Σε πολλές αρχιτεκτονικές αντί να χρησιμοποιηθεί stride μεγαλύτερο του ένα, χρησιμοποιείται το pooling [37] το οποίο περιλαμβάνει να πάρουμε το μέγιστο ή την μέση τιμή από μια γειτονιά σημείων.

Η αντίστροφη διαδικασία της συνέλιξης ονομάζεται deconvolution ή transposed convolution. Το deconvolution με stride κ αυξάνει το μέγεθος του σήματος κατά κ αντί να το μειώνει (up sampling) και χρησιμοποιείται για να επαναφέρει το σήμα στην αρχική του μορφή σε δομές όπως των Autoencoders.

Κάθε συνέλιξη μπορεί να θεωρηθεί σαν ένα πλήρως συνδεδεμένο νευρωνικό επίπεδο με την επιπλέον υπόθεση ότι κάθε νευρώνας εξόδου ενώνεται με συνδέσεις που έχουν τα ίδια βάρη αλλά με διαφορετικούς νευρώνες εισόδου και ότι ο αριθμός των συνδέσεων που δεν είναι μηδέν αντιστοιχεί στο μήκος του φίλτρου. Αυτή η υπόθεση δικαιολογείται στην περίπτωση των εικόνων και του ήχου καθώς αυτά τα δεδομένα έχουν υψηλή περιοδικότητα και τα συνελικτικά δίκτυα μπορούν να αναγνωρίσουν τα επαναλαμβανόμενα μοτίβα που υπάρχουν στα δεδομένα. Επίσης, τα δεδομένα μπορούν να υπάρξουν σε οποιοδήποτε σημείο του αρχικού σήματος, και η αναγνώρισή τους διευκολύνεται από το γεγονός ότι τα βάρη είναι ίδια σε κάθε σημείο που τοποθετούμε το φίλτρο. Επειδή στην πράξη κάθε φίλτρο έχει μικρό μέγεθος (π.χ. 5) και ο αριθμός των παραμέτρων για μία συνέλιξη είναι πολύ μικρός, τα συνελικτικά νευρωνικά δίκτυα μπορούν να βρουν πολλά περισσότερα μοτίβα χρησιμοποιώντας μεγάλο αριθμό συνελίξεων απ'ό,τι θα ήταν εφικτό με πλήρως συνεκτικά επίπεδα.

Η δομή των βαθιών συνεκτικών δικτύων αποτελείται από μια σειρά από συνελίξεις και φαίνεται στο Σχήμα 3.4.9. Η ερμηνεία των βαθιών CNNs βασίζεται στην ιεραρχική δομή τους. Τα πρώτα επίπεδα αντιστοιχούν σε αναγνώριση απλών δομικών χαρακτηριστικών όπως ακμές και γωνίες στην περίπτωση των εικόνων, ή κάποιας δεδομένης συχνότητας στην περίπτωση της φωνής. Σε κάθε επόμενο επίπεδο, χρησιμοποιούμε την έξοδο των προηγουμένων για να ανακαλύψουμε όλο και πιο σύνθετα χαρακτηριστικά όπως σχημάτων ή φωνημάτων. Μια αξιόλογη προσπάθεια ερμηνείας ενός ολοκλήρου συνελικτικού νευρωνικού δικτύου βρίσκεται στο [38].

Σε ένα βαθύ συνελικτικό δίκτυο, το κάθε σημείο της εξόδου μπορεί να επηρεαστεί από έναν περιορισμένο μόνο αριθμό από δείγματα της εισόδου. Για παράδειγμα, για ένα δίκτυο ενός επίπεδο (με dilation 1) κάθε σημείο της εξόδου έχει προκύψει από την επεξεργασία τόσων δειγμάτων της εισόδου όσο και το μέγεθος του πυρήνα. Ο αριθμός των δειγμάτων εισόδου που επηρεάζουν ένα δείγμα εξόδου λέγεται receptive field. Το receptive field παρουσιάζει μια θεμελιώδη αδυναμία των συνελικτικών δικτύων, καθώς είναι αδύνατον να μοντελοποιήσουν χρονικές εξαρτήσεις μεγαλύτερες από το receptive field. Για δίκτυα μεγαλύτερου βάθους βέβαια, το receptive field αυξάνεται, όπως και όταν χρησιμοποιούμε μεγαλύτερο dilation, και κατά συνέπεια ένα πιο βαθύ δίκτυο μπορεί να μοντελοποιήσει μεγαλύτερες εξαρτήσεις.



Σχήμα 3.4.9: Απεικόνιση ενός βαθιού CNN. Εικόνα από wikipedia.

Κεφάλαιο 4

Επεξεργασία Φωνής με Νευρωνικά Δίκτυα

4.1 Εισαγωγή

Με την ανάπτυξη νέων μεθόδων μηχανικής μάθησης, πολλές από τις εφαρμογές κλασικής επεξεργασίας φωνής έχουν μετατεθεί σε εφαρμογές νευρωνικών δικτύων. Σημαντικά τέτοια παραδείγματα είναι οι εφαρμογές σύνθεσης φωνής από κείμενο (text to speech), η αναγνώριση φωνής που σημαίνει την αναγνώριση των λέξεων από ένα σήμα ομιλίας (ASR, Automatic Speech Recognition) και οι Vocoders. Επιπλέον εφαρμογές είναι η αναγνώριση συναισθήματος, η μετατροπή ομιλίας (σε όλες τις δυνατές παραλλαγές), η αναγνώριση ομιλητή, η σύνθεση τραγουδιστής φωνής κλπ.

Ένα πολύ συνηθισμένο πρότυπο που χρησιμοποιείται για την επεξεργασία φωνής με βάση τα νευρωνικά δίκτυα είναι τα Auto-regressive μοντέλα. Auto-regressive μοντέλα, ονομάζονται τα μοντέλα που παράγουν την ζητούμενη έξοδο ένα δείγμα την φορά και σε κάθε στιγμή χρησιμοποιούν σαν είσοδο όλη την προηγούμενη έξοδο τους. Έχουν χρησιμοποιηθεί και σε επεξεργασία εικόνας [39] αλλά η κύρια εφαρμογή τους είναι σε δεδομένα που δεν έχουν σταθερό μέγεθος όπως το κείμενο [40] ή ο ήχος.

Μία πολύ σημαντική δομή Auto-regressive νευρωνικών δικτύων είναι τα αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Networks) [41]. Τα αναδρομικά νευρωνικά δίκτυα έχουν την ιδιότητα ότι παίρνουν σαν είσοδο, εκτός από οποιοδήποτε χαρακτηριστικό χρειάζονται, την έξοδο που είχαν βγάλει στο αμέσως προηγούμενο βήμα. Κατά συνέπεια, είναι πολύ χρήσιμα για την μοντελοποίηση διαδικασιών που εξαρτώνται από τον χρόνο όπως ο ήχος. Παρόλο που θεωρητικά τα RNN μπορούν να μοντελοποιήσουν εξαρτήσεις μεταξύ δύο οποιοδήποτε χρονικών στιγμών, στην πράξη αποτυγχάνουν να κρατήσουν την πληροφορία για μεγάλο αριθμό βημάτων, κάτι που περιορίζει τις εφαρμογές τους. Μια αρχιτεκτονική που αντιμετωπίζει σε μεγάλο βαθμό αυτό το πρόβλημα είναι τα Long Short Memory Cells ή αλλιώς LSTMs [42] που περνάνε σε κάθε χρονική στιγμή μια πληροφορία από κάποια προηγούμενη χρονική στιγμή που αποκαλείται cell state. Σε κάθε βήμα, το LSTM αποφασίζει αν θα γράψει κάποια καινούργια πληροφορία στο cell state ανάλογα με την είσοδό του, το προηγούμενο cell state και την προηγούμενη είσοδο. Έτσι μπορεί να διατηρήσει την πληροφορία για όσο χρονικό διάστημα χρειάζεται.

4.2 Text to Speech

Μια σημαντική εφαρμογή στην επεξεργασία φωνής στην οποία χρησιμοποιούνται σχεδόν αποκλειστικά Auto-regressive μοντέλα είναι η παραγωγή φωνής από δοσμένο κείμενο. Αυτού του είδους τα προβλήματα που απαιτούν τον μετασχηματισμό από μία ακολουθία σε άλλη ακολουθία ονομάζονται sequence to sequence [43]. Η πιο συνηθισμένη δομή ενός τέτοιου μοντέλου είναι ένας Auto-regressive Encoder όπως π.χ. ένα LSTM, που παίρνει σαν είσοδο το μεταβλητού μεγέθους κείμενο, και βγάζει σαν έξοδο ένα σήμα επίσης μεταβλητού μεγέθους, και ένας Auto-regressive Decoder που με είσοδο την έξοδο του Encoder και όλες τις προηγούμενες εξόδους βγάζει τη ζητούμενη φωνή.

Στην πράξη, χρησιμοποιείται επιπλέον ο μηχανισμός attention (προσοχής) [41, 44] για να μπορέσει ο Decoder να εστιάσει στο σωστό σημείο του κειμένου εισόδου για κάθε έξοδο που βγάζει. Ο μηχανισμός αυτός βρίσκει ένα σταθμισμένο άθροισμα από όλες τις τιμές ενός σήματος μεταβλητής διάρκειας, όπου κάθε βάρος αντιστοιχεί στο πόσο "προσοχή" δίνεται στην αντίστοιχη χρονική στιγμή. Τα βάρη είναι κανονικοποιημένα και υπολογίζονται χρησιμοποιώντας την συνάρτηση softmax στις ενδιάμεσες τιμές του Decoder αφού έχει επεξεργαστεί την προηγούμενη έξοδο. Φυσικά έχουν προταθεί πάρα πολλές αλλαγές στον ακριβή τρόπο που λειτουργεί ο μηχανισμός attention για την σύνθεση φωνή [45, 46].

Ένας επιπλέον τρόπος να αντιστοιχηθούν οι ακολουθίες κειμένου και ήχου είναι με την μοντελοποίηση της διάρκειας της ακολουθίας των φωνημάτων [47, 48]. Συνήθως πρώτα εκπαιδεύεται ένας κλασικός αλγόριθμος αντιστοίχησης κειμένου-ομιλίας, όπως το Montreal Force Aligner (MFA) [49] ο οποίος βασίζεται σε κρυφές μαρκοβιανές αλυσίδες, και στην συνέχεια το νευρωνικό δίκτυο μαθαίνει να προβλέπει αυτές τις διάρκειες των φωνημάτων όταν πρέπει να εκτελέσει την σύνθεση. Σε άλλες περιπτώσεις η αντιστοιχία γίνεται με κάποιο αλγόριθμο δυναμικού προγραμματισμού [50].

Ένα πολύ σημαντικό μοντέλο για Text to Speech το οποίο είναι βασισμένο στον μηχανισμό attention, είναι το Tacotron [51]. Πάνω στην αρχιτεκτονική του Tacotron βασίζονται και πάρα πολλά μοντέλα που μοντελοποιούν το συναίσθημα. Συνήθως η αρχιτεκτονική που χρησιμοποιείται είναι παρόμοια με αυτή των Autoencoderκαι χρησιμοποιείται ένας δεύτερος Encoder (Reference Encoder) που έχει σκοπό να κωδικοποιήσει όλη την πληροφορία που αφορά το σήμα ήχου και δεν υπάρχει στο κείμενο, δηλαδή την προσωδία. Η προσωδία

περιλαμβάνει και το συνάισθημα αλλά και διάφορα άλλα χαρακτηριστικά όπως η θεμελιώδης συχνότητα F0, ή η ένταση. Συχνά, η διαδικασία είναι όσο πιο unsupervised γίνεται και τα σχετικά χαρακτηριστικά που μοντελοποιούν την προσωδία μαθαίνονται μόνο από το μοντέλο.

Σημαντικά παραδείγματα είναι τα:

- *End to End Prosody Transfer*

  Σε αυτή την δημοσίευση [52], ένα LSTM μοντέλο μαθαίνει να συνοψίζει όλη την πληροφορία ενός reference σήματος βγάζοντας ένα διάνυσμα σταθερού μεγέθους.

  Επειδή όλη η γλωσσολογική πληροφορία βρίσκεται στον Text Encoder η μόνη πληροφορία που απομένει να αποκωδικοποιήσει ο Reference Encoder είναι αυτή που αφορά την προσωδία. Μετά την εκπαίδευση, μπορεί να χρησιμοποιηθεί οποιαδήποτε εκφώνηση με οποιοδήποτε κείμενο, μεταφέροντας την προσωδία της εκφώνησης στο κείμενο.

- *Global Style Tokens (GSTs)*

  Τα GSTs [53] χρησιμοποιούν έναν αριθμό από σταθερά embeddings (tokens) το κάθε ένα εκ των οποίων, αν πετύχει η εκπαίδευση, αντιστοιχεί σε κάποιο στυλ ομιλίας π.χ. υψηλή φωνή, μπάσα φωνή, ψιθύρισμα κλπ. Σε κάθε εκφώνηση χρησιμοποιείται ένας Encoder στο σήμα, που έχει σαν έξοδο ένα διάνυσμα σταθερού μεγέθους, όπως και στο [52], και με τον μηχανισμό attention βρίσκεται ένας συνδυασμός από τα tokens ο οποίος αποτελεί την είσοδο του Decoder. Τα tokens ξεκινάνε αρχικοποιημένα τυχαία και μαθαίνονται κατά την διάρκεια της εκπαίδευσης με το Back Propagation. Στην συνέχεια, μπορεί να χρησιμοποιηθεί είτε μια οποιαδήποτε εκφώνηση στον Reference Encoder έτσι ώστε το κείμενο να ειπωθεί με το στυλ αυτής της εκφώνησης, είτε ένας συγκεκριμένος συνδυασμός από τα tokens.

- *Gaussian Mixture Variational Autoencoder*

  Αυτό το μοντέλο [54] βασίζεται πάνω στην λογική των Variational Autoencoders. Ο Reference Encoder μαθαίνει να απεικονίζει το δοσμένο σήμα σε μία πιθανοτική κατανομή. Η κατανομή που χρησιμοποιείται σαν latent κατανομή για τους Autoencoders, είναι βασισμένη σε μείγμα Γκαουσιανών κατανομών, που σημαίνει ότι έχει κωδικοποιημένη μια ιεραρχική δομή.

- *VQ-VAE*

  Μια διαφορετική δομή πάνω στην οποία μπορεί να λειτουργήσει ένα μοντέλο Autoencoder, αυτή η εργασία [55] βασίζεται πάνω στους Quantized VAE στους οποίους οι latent κατανομές είναι διακριτές αντί για συνεχείς. Επιπλέον, αυτό το μοντέλο χρησιμοποιεί και ένα έξτρα Auto-regressive μοντέλο με σκοπό να προβλέπει την latent κατανομή (που αντιστοιχεί στην προσωδία) κατά την σύνθεση.

- *Fastspeech-2*

  Το Fastspeech-2 [47] είναι ένα μοντέλο που βασίζεται στην πρόβλεψη της διάρκειας των φωνημάτων για την αντιστοίχιση του κειμένου με την ομιλίας, αλλά έχει και επιπλέον δίκτυα τα οποία κάνουν πρόβλεψη διαφόρων προσωδιακών χαρακτηριστικών, όπως το F0 ή η ενέργεια του σήματος. Η πρόβλεψη αυτή γίνεται με βάση Auto-regressive δίκτυα.

Τέλος, μια πιο πρόσφατη αλλά σημαντική κατηγορία νευρωνικών δικτύων είναι αυτά που είναι βασισμένα σε normalizing flows [56]. Τα δίκτυα αυτού του είδους, μαθαίνουν έναν μετασχηματισμό ο οποίος είναι εξ ορισμού αντιστρέψιμος, ο οποίος απεικονίζει την κατανομή των δεδομένων σε μία κατανομή με γνωστή συνάρτηση πιθανοφάνειας. Επειδή ο μετασχηματισμός είναι αντιστρέψιμος, μπορούμε να υπολογίσουμε με ακρίβεια την πιθανότητα κάθε δεδομένου (με βάση την κατανομή βάσης που υποθέσαμε) και κατά συνέπεια να εκπαιδεύσουμε ένα νευρωνικό δίκτυο με σκοπό να μεγιστοποιήσει αυτή την πιθανότητα.

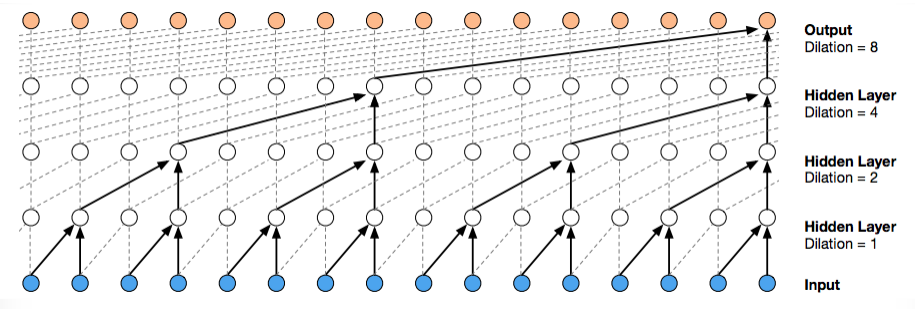
Αν και αυτού του είδους τα δίκτυα χρησιμοποιούνται κυρίως σε vocoders, έχουν χρησιμοποιηθεί και σε text to speech εφαρμογές. Παραδείγματα είναι το GlowTTS [57] το οποίο μοντελοποιεί τον μετασχηματισμό από το κείμενο στην φωνή με flow δίκτυο και το VITS [58] το οποίο αποτελεί ουσιαστικά έναν Autoencoder ο οποίος χρησιμοποιεί τα flows, με σκοπό να χρησιμοποιεί πιο περίπλοκη prior κατανομή η οποία έχει σκοπό να πιάσει όλη την διακύμανση στην προσωδία της ομιλίας.

4.3 Vocoders

Επειδή τα περισσότερα μοντέλα δεν λειτουργούν στο πεδίο του χρόνου αλλά σε κάποιο άλλο πεδίο χαρακτηριστικών όπως τα Mel Spectrograms, μετά από οποιαδήποτε σύνθεση φωνής χρειάζεται μία τελική μετατροπή σε κυματομορφή. Αυτό μπορεί να γίνει με κάποιον κλασικό Vocoder όπως τον WORLD Vocoder που αναλύθηκε προηγουμένως, ή μπορεί να γίνει χρησιμοποιώντας ένα νευρωνικό δίκτυο. Επειδή γενικά κατά την κωδικοποίηση σε χαρακτηριστικά χάνεται πληροφορία, όπως για παράδειγμα η πληροφορία της φάσης στον μετασχηματισμό Fourier, τα νευρωνικά δίκτυα μπορούν να φανούν χρήσιμα καθώς μπορούν να "αποθηκεύσουν" όποια επιπλέον πληροφορία χρειάζεται για τον αντίστροφο μετασχηματισμό στα βάρη τους κατά την διάρκεια της εκπαίδευσης.

Ένα πολύ δημοφιλές μοντέλο που χρησιμοποιείται κυρίως για Vocoder είναι το Wavenet [59]. Το Wavenet είναι ένα Auto-regressive μοντέλο που βασίζεται πλήρως σε συνελικτικά νευρωνικά δίκτυα. Δεδομένου ενός σήματος, προβλέπει την τιμή του σήματος την επόμενη χρονική στιγμή κάνοντας μια σειρά από συνελίξεις. Επειδή όπως είδαμε στην Ενότητα 3.4.5, για κάθε δείγμα που χρειάζεται να παραχθεί, το μοντέλο χρησιμοποιεί μόνο έναν συγκεκριμένο αριθμό από από προηγούμενα δείγματα, (receptive field) χρησιμοποιείται συνέλιξη με dilation, με σκοπό να αυξηθεί αυτός ο αριθμός όσο το δυνατόν γίνεται χωρίς να αυξηθούν υπερβολικά οι παράμετροι. Όταν ένα φίλτρο έχει dilation κ, σημαίνει ότι το φίλτρο της συνάρτησης έχει μόνο μία μηδενική παράμετρο για κάθε κ μηδενικά. Για κάθε επίπεδο συνέλιξης, το dilation αυξάνεται εκθετικά. Αυτό, όπως φαίνεται και στο Σχήμα 4.3.1 μεγαλώνει το παράθυρο που βλέπει κάθε δείγμα εκθετικά ως προς το βάθος του δικτύου, αλλά οι παράμετροι δεν είναι περισσότεροι απ'ό,τι θα ήταν χωρίς καθόλου dilations. Στην πράξη, για να παράξει ήχο από το μηδέν, ξεκινάει παίρνοντας στην είσοδο ένα σήμα μόνο με μηδενικά.

Στην αρχική του μορφή το Wavenet παράγει ήχο χωρίς καμμία είσοδο. Ο ήχος αυτός έχει τα σωστά στατιστικά της φωνής αλλά δεν έχει καθόλου συνοχή εκτός από το πολύ τοπικό επίπεδο και ακούγεται σαν ασυναρτησίες. Μια πολύ απλή αλλαγή όμως είναι να χρησιμοποιηθεί κάποιο σήμα σαν conditioning. Όταν χρησιμοποιηθεί για παράδειγμα Mel Spectrograms, το μοντέλο στην ουσία τα μετατρέπει σε κυματομορφή, αλλά με φάση η οποία έχει τα σωστά στατιστικά στοιχεία της φωνή και κατά συνέπεια ακούγεται πολύ ρεαλιστική.



Σχήμα 4.3.1: Η δομή των συνελίξεων στο Wavenet

Το Wavenet είναι αρκετά γρήγορο στην εκπαίδευση καθώς όλα τα δείγματα παράγονται ταυτόχρονα, με την διαδικασία της συνέλιξης. Στην διαδικασία της σύνθεσης όμως, πρέπει να δημιουργηθεί ένα δείγμα την φορά, κάτη που την κάνει πολύ πιο αργή διαδικασία. Μία λύση σε αυτό το πρόβλημα δίνεται από το Parallel Wavenet [60] το οποίο εκπαιδεύει δύο δίκτυα, ένα Teacher και ένα Student και στην συνέχεια εκπαιδεύει τον Student έτσι ώστε να παράξει την ίδια κατανομή με τον Teacher, μεταφέροντας δηλαδή την γνώση του Teacher στον Student. Ο Teacher είναι ένα κανονικό Wavenet και εκπαιδεύεται με τον κανονικό τρόπο εκπαίδευσης ενώ ο Student είναι ένα μοντέλο βασισμένο πάνω στο Inverse Auto-regressive Flow [61], το οποίο είναι ένα είδους normalizing flow δίκτυο. Σκοπός της εκπαίδευσης είναι να ελαχιστοποιηθεί η Kullback-Leibler απόσταση μεταξύ των κατανομών του Teacher και του Student. Όλες οι πράξεις τόσο στην σύνθεση όσο και στην εκπαίδευση γίνονται παράλληλα με αποτέλεσμα να είναι πολύ πιο γρήγορο από το αρχικό μοντέλο.

Μία διαφορετική προσέγγιση είναι αυτή του Wave-RNN [62] το οποίο χρησιμοποιεί ένα RNN για να παράξει ήχο. Αρχικά παράγει τα πρώτα 8 bits, για να έχει μία πρώτη προσέγγιση του ζητούμενου δείγματος ήχου, και στην συνέχεια με βάση αυτά παράγει τα επόμενα 8 ώστε να προσθέσει περισσότερες λεπτομέρειες στον ήχο. Επιτυγχάνει γρήγορη ταχύτητα στην παραγωγή παράγοντας πολλά δείγματα ταυτόχρονα και χρησιμοποιώντας αραιά (sparce) νευρωνικά δίκτυα, δηλαδή δίκτυα που τα περισσότερα στοιχεία στους πίνακες είναι μηδέν.

Τα normalizing flows έχουν επίσης χρησιμοποιηθεί σε μεγάλο βαθμό σαν Vocoders, με μοντέλα όπως το WaveGlow [63] ή το WaveFlow [64] τα οποία αφού μετατρέψουν το σπεκτρόγραμμα σε αρχείο με το σωστός μήκος μιας ομιλίας (μια μετατροπή σταθερής σχέσης μήκους) μοντελοποιούν την μετατροπή σε κανονικό σήμα ήχου.

Τέλος, ένα πολύ μεγάλο μέρος των μοντέρνων Vocoders είναι βασισμένο πάνω στα GANs τα οποία θα αναλυθούν στο επόμενο κεφάλαιο.

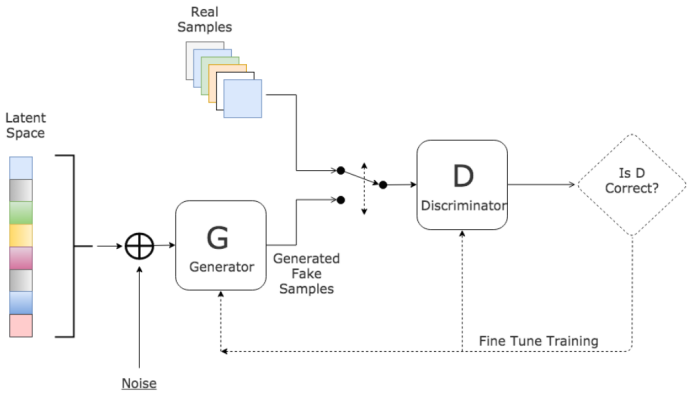
Κεφάλαιο 5

Generative Adversarial Networks

5.1 Εισαγωγή

Ένα Generative Adversarial Network (GAN) [65] είναι ένα generative μοντέλο μηχανικής μάθησης το οποίο έχει ως σκοπό την δειγματοληψία (sampling) από μία δεδομένη κατανομή, στην οποία όμως έχει πρόσβαση μόνο μέσω ενός σύνολου δειγμάτων (δεδομένων) και όχι μέσω αναλυτικής μορφής συνάρτησης πυκνότητας πιθανότητας. Αποτελείται από δύο νευρωνικά δίκτυα, τον Generator και τον Discriminator. Ο Generator παίρνει σαν είσοδο ένα τυχαίο δείγμα από μία κατανομή ανεξάρτητη από την κατανομή των δεδομένων (π.χ. μία πολυδιάστατη γκαουσιανή) και βγάζει σαν έξοδο ένα δείγμα από τη ζητούμενη κατανομή με την ζητούμενη πιθανότητα. Ουσιαστικά μαθαίνει μία απεικόνιση από την δοσμένη κατανομή στην κατανομή των δεδομένων. Για παράδειγμα, αν η ζητούμενη κατανομή είναι μια κατανομή πάνω στις εικόνες που απεικονίζουν ένα ανθρώπινο πρόσωπο, όταν πάρουμε ένα δείγμα από τον Generator, θα πάρουμε ένα καινούργιο πρόσωπο το οποίο δεν αντιστοιχεί σε κανένα από τα δοσμένα δεδομένα. Ο Discriminator παίρνει σαν είσοδο ένα δεδομένο και προσπαθεί να εκτιμήσει την πιθανότητα αυτό το δεδομένο να προήλθε από την πραγματική κατανομή ή αν είναι έξοδος του Generator. Βγάζει δηλαδή σαν έξοδο έναν αριθμό από 0 εώς 1, ο οποίος αντιστοιχεί στην πιθανότητα το δείγμα να είναι αληθινό. Η βασική αρχιτεκτονική φαίνεται σχηματικά στο Σχήμα 5.1.1.

Η εκπαίδευση των δικτύων γίνεται ως εξής: Ο Discriminator προσπαθεί να μειώσει το σφάλμα της εξόδου με τα labels, τα οποία είναι 1 για πραγματικά δεδομένα και 0 για ψεύτικα, και ο Generator προσπαθεί να αυξήσει το σφάλμα του Discriminator δημιουργώντας στην πορεία όλο και πιο ρεαλιστικά δείγματα έτσι ώστε να καταφέρει να τον μπερδέψει. Στην αρχή της εκπαίδευσης κανένα από τα δύο δίκτυα δεν κάνει την δουλειά του ικανοποιητικά και ο Generator παράγει δεδομένα που αποτελούνται ουσιαστικά από τυχαίο θόρυβο ενώ ο Discriminator δεν μπορεί να τα ξεχωρίσει από τα πραγματικά δεδομένα. Στη συνέχεια όμως ο Discriminator βελτιώνεται και αναγνωρίζει τις εξόδους του Generator ως ψεύτικες, κάτι που αναγκάζει τον Generator να αλλάξει την έξοδο του έτσι ώστε να γίνει πιο ρεαλιστική.



Σχήμα 5.1.1: Η αρχιτεκτονική ενός GAN. Σχήμα από [66]

Στο τέλος μιας επιτυχημένης εκπαίδευσης, ο Generator έχει προσεγγίσει την ζητούμενη κατανομή τόσο καλά που τα δεδομένα που βγάζει σαν έξοδο δεν μπορούνε να διαχωριστούν από τα πραγματικά και μπορούμε να τον χρησιμοποιήσουμε μόνο του για να πάρουμε καινούργια δείγματα. Πιο συγκεκριμένα οι συναρτήσεις σφάλματος για την εκπαίδευση του μοντέλου είναι:

$$L(D) = -\mathbb{E}_{x\sim p_{data}}[\log(D(x))] - \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \tag{5.1.1}$$

$$L(G) = \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \tag{5.1.2}$$

όπου D, G είναι οι Discriminator, Generator αντίστοιχα, p_{data} είναι η κατανομή των δεδομένων που προσπαθούμε να προσεγγίσουμε και p_z είναι η κατανομή του θορύβου που χρησιμοποιείται σαν είσοδος του Generator. Μπορούμε να δούμε ότι ο Discriminator ουσιαστικά κάνει δυαδική ταξινόμηση χρησιμοποιώντας binary cross entropy σαν συνάρτηση σφάλματος.

Κατά την διάρκεια της εκπαίδευσης, χρησιμοποιείται ο αλγόριθμος Back Propagation για την ανανέωση των βαρών των νευρωνικών δικτύων. Η ανανέωση γίνεται πρώτα για κ βήματα για τις παραμέτρους του Discriminator

ώστε να έχει προλάβει να μάθει τα λάθη που κάνει σε αυτό το σημείο ο Generator και να παρέχει χρήσιμη πληροφορία, και μετά ένα βήμα για τις παραμέτρους του Generator. Στην πράξη, πολλές φορές χρησιμοποιείται $\kappa = 1$, δηλαδή τα δίκτυα εκπαιδεύονται εναλλασσόμενα.

Από την συνάρτηση σφάλματος, μπορούμε να δούμε ότι για έναν δεδομένο Generator η βέλτιστη τιμή του Discriminator είναι:

$$D^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \tag{5.1.3}$$

αφού:

$$\begin{aligned} L(D) &= -\mathbb{E}_{x\sim p_{data}}[\log(D(x))] - \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \\ &= -\int_x p_{data}(x)\log(D(x))dx \; - \int_z p_z(z)\log(1 - D(G(z)))dz \\ &= -\int_x p_{data}(x)\log(D(x)) \; + p_g(x)\log(1 - D(x)dx \end{aligned} \tag{5.1.4}$$

και για κάθε $(a, b) \in \mathbb{R}^2 \setminus \{0, 0\}$, η συνάρτηση $y \to a\ \log(y) + b\ \log(1 - y)$ πετυχαίνει το μέγιστο της στο $\frac{a}{a+b}$. Υποθέτοντας τώρα τον βέλτιστο Discriminator, μπορούμε να δούμε ότι το κόστος του Generator είναι:

$$\begin{aligned} L(G) &= \mathbb{E}_{x\sim p_g}[\log(1 - D^*(x))] \\ &= \mathbb{E}_{x\sim p_g}\left[\log \frac{p_g(x)}{p_{data}(x) + p_g(x)}\right] \\ &= \int_x p_g(x)\log \frac{p_g(x)}{p_{data}(x) + p_g(x)}dx \\ &= KL\ \left(p_g\ ||\ \frac{p_{data} + p_g}{2}\right) - \log(2) \end{aligned} \tag{5.1.5}$$

όπου $KL(P\ ||\ Q)$ είναι η απόσταση Kullback-Leibler μεταξύ των κατανομών P, Q. Επειδή είναι γνωστό ότι η απόσταση Kullback-Leibler είναι πάντα μη αρνητική, και μηδέν μόνο στην περίπτωση που οι δύο κατανομές είναι ίσες παντού, μπορούμε να δούμε ότι η ελάχιστη τιμή του κόστους του Generator είναι όταν $p_g = \frac{p_{data}+p_g}{2} \Rightarrow p_g = p_{data}$. Μπορούμε δηλαδή να δούμε ότι όταν η εκπαίδευση είναι πετυχημένη, ο Generator μαθαίνει να εκτιμά την πραγματική κατανομή των δεδομένων.

Στην πράξη, η εκπαίδευση των GAN είναι αρκετά δύσκολη και πολύ συχνά αποτυχαίνει. Ένα πολύ συχνό φαινόμενο αποτυχίας είναι το mode collapse το οποίο συμβαίνει όταν όλες οι έξοδοι του Generator έχουν πολύ μικρή διακύμανση και μοιάζουν με ένα συγκεκριμένο πραγματικό δεδομένο. Αυτό μπορεί να συμβεί αν ο Discriminator στην αρχή της εκπαίδευσης είναι πολύ σίγουρος ότι αυτό το δεδομένο είναι πραγματικό και ο Generator ξεκίνησε σαν απάντηση να το βγάζει συνέχεια στην έξοδό του γιατί αντιστοιχεί σε μικρή συνάρτηση σφάλματος. Στη συνέχεια όμως, ο Discriminator μαθαίνει ότι το δεδομένο είναι ψεύτικο με μεγάλη πιθανότητα και έτσι καταφέρνει να αποκτά ποσοστό επιτυχίας σχεδόν 100%. Σαν αποτέλεσμα, ο Discriminator καταφέρνει να απορρίπτει όλες τις εξόδους του Generator με μεγάλη πιθανότητα και δεν προσφέρει κάποιο χρήσιμο gradient στον Generator για να βελτιωθεί με συνέπεια η εκπαίδευση να κολλάει. Μάλιστα, στην περίπτωση που η είσοδος του θορύβου έχει μικρότερες διαστάσεις και κατά συνέπεια μέτρο μηδέν στο χώρο των διαστάσεων των δεδομένων (όπως ισχύει στην συντριπτική πλειοψηφία των περιπτώσεων), έχει αποδειχθεί ότι υπάρχει ένας βέλτιστος Discriminator που έχει ποσοστό επιτυχίας όσο κοντά στο 100% απαιτείται, με την κλίση της συνάρτησης σφάλματος να είναι 0 σχεδόν παντού στον χώρο των δεδομένων [67].

Μεγάλο μέρος της επιτυχίας των GAN έγκειται στο γεγονός ότι δεν βασίζονται πάνω σε μία δεδομένη συνάρτηση ομοιότητας για την δημιουργία των δεδομένων και αν ο Discriminator είναι αρκετά ισχυρός, η συνάρτηση που ορίζει το πόσο ρεαλιστικά είναι τα δεδομένα μπορεί να είναι όσο περίπλοκη χρειάζεται. Για παράδειγμα, οι Variational Autoencoders συχνά βασίζονται στην L2 νόρμα για παράξουν ρεαλιστικά δεδομένα. Στην περίπτωση

των εικόνων, αυτό οδηγεί στο να τιμωρούνται περισσότερο τιμές που απέχουν αρκετά από την μέση τιμή και οι εικόνες που δημιουργούνται φαίνονται θολές γιατί οι τιμές των pixels δεν έχουν μεγάλη διακύμανση όπως έχουν οι πραγματικές εικόνες. Στα GAN αντίθετα, πρέπει να φαίνονται ρεαλιστικές σε ένα μοντέλο μεγάλης πολυπλοκότητας το οποίο σύντομα θα μάθει να απορρίπτει θολές εικόνες.

Η δομή των νευρωνικών δικτύων δεν είναι συγκεκριμένη και αλλάζει ανάλογα με την εφαρμογή. Παρ' όλα αυτά, μια πολύ συχνή επιλογή είναι τα συνελικτικά νευρωνικά δίκτυα τόσο για τον Discriminator όσο και για τον Generator [68].

5.2 Συναρτήσεις Σφάλματος

Διάφορες παραλλαγές της συνάρτησης σφάλματος των GAN έχουν προταθεί. Ο σκοπός τους είναι κυρίως η αποφυγή προβλημάτων κατά την εκπαίδευση, όπως το πρόβλημα mode collapse, ή η ταχύτητα σύγκλισης.

5.2.1 Λογαριθμικό Σφάλμα

Στην αρχική δημοσίευση [65], οι συγγραφείς παρατήρησαν ότι στην πράξη, η αρχική συνάρτηση σφάλματος δεν έδινε το απαραίτητο gradient για να μάθει ο Generator ικανοποιητικά. Στην αρχή της εκπαίδευσης, που ο Generator δεν είναι καλός, ο Discriminator απορρίπτει δείγματα με μεγάλη βεβαιότητα αφού είναι πολύ διαφορετικά από τα δεδομένα εκπαίδευσης. Σε αυτή την περίπτωση, το $\log(1 - D(G(z)))$ έχει μικρή τιμή. Η προτεινόμενη λύση είναι αντί να ελαχιστοποιείται το $\log(1-D(G(z)))$ ο Generator εκπαιδεύεται να μεγιστοποιεί το $\log D(G(z))$. Αυτή η αντικειμενική συνάρτηση οδηγεί στο ίδιο σταθερό σημείο των μοντέλων αλλά έχει πολύ πιο ισχυρές παραγώγους στην αρχή της εκπαίδευσης.

5.2.2 LSGAN

Το σφάλμα αυτής της παραλλαγής είναι η L2 νόρμα με τιμή 1 για τα πραγματικά δεδομένα και 0 για τα ψεύτικα [69]. Είναι δηλαδή:

$$\begin{aligned} L(D) &= \mathbb{E}_{x\sim p_{data}}[(D(x)-1)^2] + \mathbb{E}_{z\sim p_z}[(D(G(z)))^2] \\ L(G) &= \mathbb{E}_{z\sim p_z}[(D(G(z))-1)^2] \end{aligned} \tag{5.2.1}$$

Όταν ελαχιστοποιείται αυτό το σφάλμα, ελαχιστοποιείται η απόκλιση χ^2 του Pearson μεταξύ των κατανομών των πραγματικών δεδομένων και των δημιουργημένων δεδομένων.

5.2.3 Wasserstein GAN

Το Wasserstein GAN [70], έχει ως σκοπό την ελαχιστοποίηση της απόστασης Wasserstein-1 ή Earth Mover distance που ορίζεται στις κατανομές των δεδομένων και του Generator ως εξής:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma\in\Pi(\mathbb{P}_r,\mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}[||x-y||_1] \tag{5.2.2}$$

όπου $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ αντιστοιχεί στο σύνολο όλων των joint κατανομών $\gamma(x, y)$ που έχουν ως marginals τις κατανομές $\mathbb{P}_r$ και $\mathbb{P}_g$ (δηλαδή ισχύει $\int_{y=-\infty}^{=\infty} \gamma(x, y) = \mathbb{P}_r(x)$ και αντίστοιχα για το y). Διαισθητικά, το $\gamma(x, y)$ δείχνει πόση μάζα πρέπει να μεταφερθεί από το x στο y για να μετασχηματιστεί η κατανομή $\mathbb{P}_r$ στην κατανομή $\mathbb{P}_g$. Η Earth Mover distance αντιστοιχεί στο ελάχιστο κόστος της βέλτιστης μεταφοράς.

Η απόσταση Wasserstein έχει το πλεονέκτημα ότι αν η συνάρτηση που παρεμετροποιεί τον Generator είναι Lipschitz συνεχής, τότε η απόσταση Wasserstein είναι παντού συνεχής και σχεδόν παντού παραγωγίσιμη ως προς τις παραμέτρους του Generator. Αυτό δεν ισχύει για άλλες συναρτήσεις αποστάσεων κατανομών όπως η Kullback-Leibler.

Για την προσέγγιση της απόστασης Wasserstein, χρησιμοποιείται η δυϊκότητα Kantorovich-Rubinstein η οποία δίνει ότι:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \sup_{||f||_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_g}[f(x)] \tag{5.2.3}$$

όπου το supremum είναι πάνω σε όλες τις συναρτήσεις που είναι 1-Lipschitz συνεχείς.

Η συνάρτηση σφάλματος που προκύπτει είναι:

$$\begin{aligned} L(D) &= \mathbb{E}_{x \sim p_{data}}[D(x)] - \mathbb{E}_{z \sim p_z}[D(G(z))] \\ L(G) &= \mathbb{E}_{z \sim p_z}[D(G(z))] \end{aligned} \tag{5.2.4}$$

Ο Discriminator χρησιμοποιείται για την προσέγγιση της συνάρτησης f στην εξίσωση 5.2.3. Βελτιώνοντας τον, προσεγγίζουμε το supremum και για να εξασφαλίσουμε την Lipschitz συνέχεια, περιορίζουμε τα βάρη του σε ένα διάστημα $[-c, c]$ αποκόβοντας τις τιμές που βγαίνουν εκτός διαστήματος. Αυτό εξασφαλίζει την Lipschitz συνέχεια αλλά όχι την 1-Lipschitz συνέχεια. Προκύπτει όμως ότι όταν έχουμε K-Lipschitz συνέχεια προσεγγίζουμε την απόσταση Wasserstein επί K, η οποία έχει ελάχιστο στα ίδια σημεία. Επειδή όσο περισσότερο εκπαιδεύουμε τον Discriminator τόσο καλύτερη προσέγγιση της απόστασης έχουμε, και επειδή η απόσταση Wasserstein είναι παραγωγίσιμη σχεδόν παντού, κάνουμε μερικά βήματα ανανέωσης των βαρών του Discriminator πριν κάνουμε ένα βήμα του Generator, δηλαδή δεν εκπαιδεύουμε τα δίκτυα εναλλασσόμενα. Στη συνέχεια, η ανανέωση του Generator γίνεται έτσι ώστε να μειώσουμε την απόσταση και κατά συνέπεια να φέρουμε τις κατανομές πιο κοντά. Συνήθως κάνουμε 5 βήματα ανανέωσης των βαρών του Discriminator για ένα βήμα του Generator αλλά ο αριθμός αυτός αποτελεί υπερπαράμετρο που μπορεί να βελτιστοποιηθεί.

5.2.4 WGAN-GP

Επειδή το Wasserstein GAN έχει ακόμη κάποια προβλήματα όσον αφορά τον περιορισμό των βαρών του δικτύου ώστε να εξασφαλιστεί η Lipschitz συνέχεια, όπως το ότι περιορίζεται η πολυπλοκότητα των συναρτήσεων που μπορούν να προσεγγιστούν, ένας επιπλέον όρος προστίθεται στην συνάρτηση σφάλματος ο οποίος περιορίζει το gradient του Discriminator να είναι κοντά στο 1. Αυτό συμβαίνει επειδή στο Wasserstein GAN, ο βέλτιστος Discriminator έχει gradient παντού κοντά στο 1, και μάλιστα περιέχει ευθείες γραμμές μεταξύ δειγμάτων από τις δύο κατανομές οι οποίες έχουν κλίση 1. Αυτό το επιπλέον σφάλμα ονομάζεται gradient penalty [71] και υλοποιείται με το τετράγωνο της L2 νόρμας:

$$L_{GP} = \lambda \, \mathbb{E}_{\hat{x} \sim p_{\hat{x}}}(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2 \tag{5.2.5}$$

Το λ είναι μια υπερπαράμετρος με ενδεικτική τιμή 10. Η κατανομή $\hat{x}$ προκύπτει παίρνοντας στην τύχη ένα σημείο κατά μήκος της ευθείας γραμμής μεταξύ κάποιου δείγματος της πραγματικής κατανομής και ενός δείγματος του Generator, χρησιμοποιώντας σαν κίνητρο την βέλτιστη μορφή του Discriminator. Οι υπόλοιποι όροι του σφάλματος είναι ίδιοι με το Wasserstein GAN.

5.3 Conditional GAN

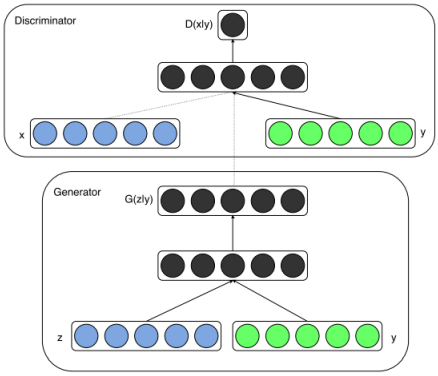
Τα GAN στην απλή τους μορφή κάνουν δειγματοληψία από την ζητούμενη κατανομή πιθανότητας χωρίς να υπάρχει κάποιος έλεγχος πάνω στο ποιο δείγμα θα προκύψει. Στην πράξη όμως τα δεδομένα πολύ συχνά είναι πολυτροπικά και χωρίζονται σε έναν διακριτό αριθμό από κλάσεις. Για παράδειγμα, όταν παράγουμε εικόνες από ψηφία όπως από τον dataset MNIST [72], τα ψηφία χωρίζονται από μόνα τους σε 10 διακριτές κλάσεις. Στην ιδανική περίπτωση θα θέλαμε να ελέγχουμε το ποιο ψηφίο δημιουργείται σε κάθε περίπτωση αλλά δεν υπάρχει κάποιος άμεσος τρόπος να επιλέξουμε το κατάλληλο δείγμα θορύβου το οποίο θα δημιουργήσει το ζητούμενο ψηφίο.

Η λύση σε αυτό το πρόβλημα, δίνεται από τα Conditional GAN [73], τα οποία παίρνουν σαν είσοδο, εκτός από τον θόρυβο που παίρνουν τα απλά GAN, και ένα επιπλέον διάνυσμα που αντιστοιχεί στην κλάση του ζητούμενου δεδομένου.

Η αντικειμενική συνάρτηση που βελτιστοποιείται σε αυτή την περίπτωση (υποθέτοντας κλασικό GAN) είναι:

$$\begin{aligned} L(D) &= -\mathbb{E}_{x \sim p_{data}(x)}[\log D(x|y)] - \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z|y)))] \\ L(G) &= \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z|y)))] \end{aligned} \tag{5.3.1}$$

όπου το y αντιστοιχεί στην κλάση του δείγματος x και δίνεται σε μορφή one hot vector.



Σχήμα 5.3.1: Σχηματική αναπαράσταση ενός Conditional GAN [73]

Ο τρόπος που δουλεύουν τα Conditional GAN είναι ότι ο Discriminator, βλέποντας κατά την διάρκεια της εκπαίδευσης όλα τα πραγματικά δείγματα μαζί με την αναπαράσταση της κλάσης τους, θεωρεί την κλάση σαν χαρακτηριστικό των δεδομένων. Έτσι, αν ο Generator δημιουργήσει ένα ρεαλιστικό δείγμα αλλά χωρίς την ζητούμενη κλάση, ο Discriminator μπορεί να το ξεχωρίσει γιατί δεν θα έχει δει ποτέ κατά την διάρκεια της εκπαίδευσης ένα τέτοιο δείγμα. Κατά συνέπεια, ο Generator αναγκάζεται να δημιουργήσει τα δείγματα με την σωστή κλάση και μπορούμε να αξιοποιήσουμε αυτό το γεγονός και να ελέγξουμε την έξοδο του.

Αρκετές παραλλαγές και προεκτάσεις του βασικού Conditional GAN έχουν προταθεί. Για παράδειγμα, σαν condition μπορεί να χρησιμοποιηθεί επεξεργασμένο κείμενο αντί για την κλάση του δεδομένου, κάτι που οδηγεί σε ένα μοντέλο που παράγει εικόνες που αντιστοιχούν σε ένα δοσμένο κείμενο [74].

5.4 CycleGAN

Το CycleGAN [**zhu2017unpaired**] αποτελεί ένα μοντέλο που μαθαίνει την απεικόνιση δεδομένων από ένα πεδίο (domain), σε κάποιο άλλο πεδίο. Ουσιαστικά, ο σκοπός του είναι η "μετάφραση" των δεδομένων, με τέτοιο τρόπο ώστε τα δεδομένα και από τα δύο πεδία να είναι ρεαλιστικά, χωρίς να επηρεάζεται η σωστή αντιστοίχιση των δεδομένων (η οποία εξαρτάται από τα συγκεκριμένα πεδία που χρησιμοποιούνται στα δεδομένα). Παραδείγματα τέτοιων "μεταφράσεων" είναι η αλλαγή μιας φωτογραφίας τοπίου από καλοκαίρι σε χειμώνα χωρίς την αλλαγή του σκηνικού, και αντίστροφα, ή η αλλαγή ενός πίνακα ζωγραφικής σε φωτογραφία τοπίου και αντίστροφα. Είναι πολύ παρόμοιο σε αρχιτεκτονική και τρόπο εκπαίδευσης με το DiscoGAN [75], το οποίο δημοσιεύτηκε ανεξάρτητα την ίδια περίοδο.

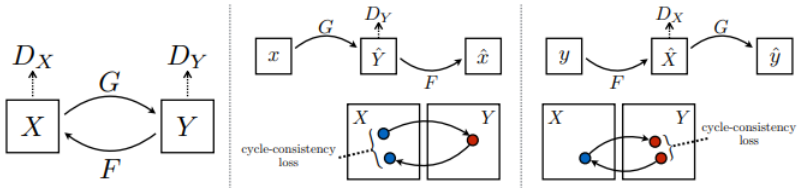
Μοντέλα αντιστοίχισης δεδομένων από ένα πεδίο σε άλλο προϋπήρχαν [76] αλλά υπήρχε ανάγκη από dataset το οποίο περιλαμβάνει ταιριασμένα ζευγάρια δεδομένων από τα δύο πεδία. Η συνεισφορά του CycleGAN είναι ότι δεν υπάρχει τέτοια ανάγκη και αρκεί ένα dataset το οποίο περιλαμβάνει δεδομένα και από τα δύο πεδία χωρίς κανενός είδους αντιστοίχιση. Αυτό έχει μεγάλη σημασία καθώς πολλές φορές είναι πρακτικά αδύνατον να βρεθεί η να δημιουργηθεί τέτοιο dataset ή μπορεί να γίνει μόνο με μεγάλο κόστος.

Η δομή του μοντέλου φαίνεται στο Σχήμα 5.4.1. Χρησιμοποιούνται δύο Generators, G και F και δύο Discriminators, D_X και D_Y. Ο Generator G, μαθαίνει να απεικονίζει τα δεδομένα από το πεδίο X στο πεδίο Y και ο F μαθαίνει την αντίστροφη απεικόνιση. Ο Discriminator D_X μαθαίνει να αναγνωρίζει αν τα δεδομένα προέρχονται από το πεδίο X ή είναι έξοδοι του F και αντίστοιχα, ο D_Y αναγνωρίζει αν προέρχονται από το πεδίο Y ή από τον G. Για την απεικόνιση $X \to Y$, η συνάρτηση σφάλματος είναι:

$$L_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[\log(D_Y(y))] + \mathbb{E}_{x \sim p_x}[\log(1 - D_Y(G(x)))] \tag{5.4.1}$$

Ο G προσπαθεί να ελαχιστοποιήσει αυτό το σφάλμα ενώ ο D_Y να το μεγιστοποιήσει. Αντίστοιχη συνάρτηση σφάλματος υπάρχει και για την απεικόνιση $Y \to X$.

Παρατηρούμε ότι δεν υπάρχει θόρυβος σαν είσοδος στα μοντέλα, και κατά συνέπεια οι απεικονίσεις είναι ντετερμινιστικές.



Σχήμα 5.4.1: Η δομή ενός CycleGAN [**zhu2017unpaired**].

Οι συναρτήσεις σφάλματος όπως δόθηκαν, θα δημιουργήσουν ρεαλιστικά δεδομένα και από τα δύο πεδία, αλλά δεν υπάρχει κανένας λόγος να κρατήσουν κάποια σημασιολογική αντιστοιχία μεταξύ των ζευγαριών των δειγμάτων. Μπορούν να θεωρήσουν την κατανομή εισόδου σαν απλά μια κατανομή θορύβου και να λειτουργήσουν ως απλά GAN χωρίς να κρατάνε κάποια πληροφορία από το συγκεκριμένο δείγμα που βάλαμε ως είσοδο. Για να αντιμετωπιστεί αυτό το πρόβλημα, τοποθετείται ένα επιπλέον σφάλμα, που αποκαλείται cycle consistency loss. Διαισθητικά, όταν πηγαίνουμε από μία εικόνα του πρώτου πεδίου στο δεύτερο και στη συνέχεια από το δεύτερο στο πρώτο πρέπει να έχουμε καταλήξει στην αρχική εικόνα, δηλαδή $x \to G(x) \to F(G(X)) \approx x$. Το cycle consistency loss αντιστοιχεί στο σφάλμα ανακατασκευής της εικόνας. Αυτό το σφάλμα, αναγκάζει το δίκτυο να κρατήσει την πληροφορία της αρχικής εικόνας στο δεύτερο πεδίο, κάτι που σημαίνει ότι θα κρατήσει την σημασιολογική αντιστοιχία μεταξύ των εικόνων.

Συγκεκριμένα, χρησιμοποιείται η L1 νόρμα:

$$L_{cyc}(G,F) = \mathbb{E}_{x\sim p_{data}(x)}||F(G(x)) - x||_1 + \mathbb{E}_{y\sim p_{data}(y)}||G(F(y)) - y||_1 \tag{5.4.2}$$

Η συνολική συνάρτηση σφάλματος είναι:

$$\begin{aligned} L(G,F,D_X,D_Y) = & L_{GAN}(G,D_Y,X,Y) \\ & + L_{GAN}(F,D_X,X,Y) \\ & + \lambda L_{cyc}(G,F) \end{aligned} \tag{5.4.3}$$

όπου το λ είναι μια υπερπαράμετρος που συσχετίζει τα δύο κριτήρια (π.χ. $\lambda = 10$).

Η εκπαίδευση του μοντέλου, μπορεί να θεωρηθεί σαν την εκμάθηση δύο διαφορετικών Autoencoders, ο ένας να μαθαίνει την απεικόνιση από το X στο X, με ενδιάμεση αναπαράσταση τον χώρο Y, και ο άλλος αντίστροφα.

Σε ορισμένες περιπτώσεις, χρησιμοποιείται ένα επιπλέον σφάλμα, που ονομάζεται identity loss. Αυτό το σφάλμα, είναι το εξής:

$$L_{id}(G,F) = \mathbb{E}_{x\sim p_{data}(x)}||F(x) - x||_1 + \mathbb{E}_{y\sim p_{data}(y)}||G(y) - y||_1 \tag{5.4.4}$$

δηλαδή αναγκάζει την είσοδο να ισούται με την έξοδο όταν η είσοδος είναι από το πεδίο της εξόδου. Ο λόγος που υπάρχει αυτό το επιπλέον σφάλμα είναι επειδή πολλές φορές το δίκτυο θα μάθει να αλλάζει ένα χαρακτηριστικό της εικόνας (π.χ. χρώμα) όταν πηγαίνει από το ένα πεδίο στο άλλο, και να το αλλάζει πάλι στο αρχικό όταν γυρίζει στο πρώτο πεδίο. Αυτό το σφάλμα, παροτρύνει το δίκτυο να μην αλλάζει τα χαρακτηριστικά τα οποία υπάρχουν σε δεδομένα και στα δύο πεδία.

5.5 Star GAN

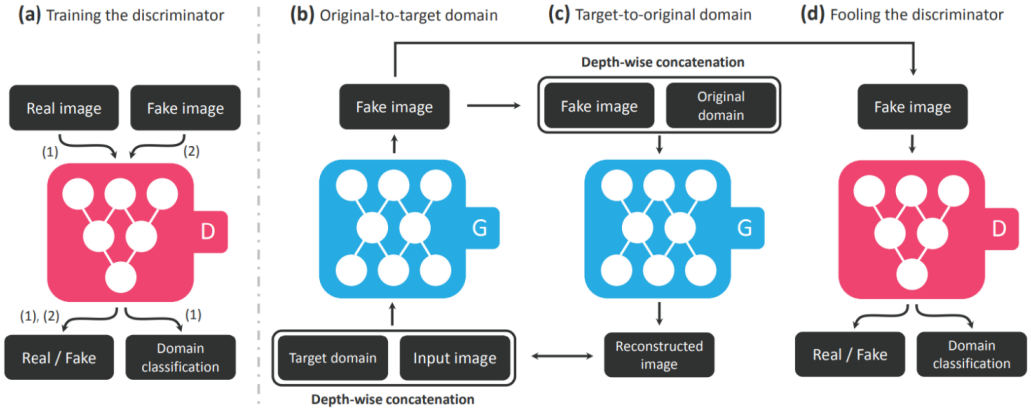
Το CycleGAN, παρά την επιτυχία του, έχει το μειονέκτημα ότι μπορεί να μετασχηματίσει δεδομένα μόνο μεταξύ δύο πεδίων. Σε πολλές περιπτώσεις όμως, όπως και στην περίπτωση των συναισθημάτων, υπάρχουν περισσότερα από δύο πεδία . Επίσης, για την εκπαίδευσή του χρειάζεται την εκπαίδευση τεσσάρων μοντέλων (2 Discriminators και 2 Generators) το οποίο το κάνει ακριβό υπολογιστικά.

Μία προφανής γενίκευση του CycleGAN είναι να μάθουμε τον μετασχηματισμό μεταξύ δύο πεδίων, για κάθε δυνατό ζευγάρι από πεδία. Αν εφαρμόζαμε την ίδια αρχιτεκτονική σε κάθε δυνατό ζευγάρι από πεδία, θα είχαμε συνολικά, $2N(N-1)$ διαφορετικά μοντέλα που θα έπρεπε να εκπαιδευτούν, για N πεδία. Αυτή η δομή επομένως δεν είναι αποδοτική. Επίσης, το κάθε μοντέλο δεν θα μπορούσε να αξιοποιήσει την γνώση που θα είχαν αποκτήσει τα άλλα μοντέλα, και κάθε ένα θα έπρεπε να μάθει από την αρχή τον μετασχηματισμό. Είναι προφανές ότι μπορεί να υπάρξει βελτίωση συνδυάζοντας μερικά από αυτά τα μοντέλα και γλιτώνοντας τόσο σε παραμέτρους όσο και σε απαιτούμενη υπολογιστική επεξεργασία.

Η αρχιτεκτονική του StarGAN [77], η οποία φαίνεται στο Σχήμα 5.5.1, συνδυάζει όλα τα μοντέλα σε μόνο δύο, έναν Discriminator και έναν Generator. Η δουλειά του Generator είναι να δημιουργεί ένα δείγμα από κάποιο δοσμένο πεδίο, παίρνοντας στην είσοδο ένα δείγμα από οποιοδήποτε άλλο πεδίο, και την κλάση εισόδου σαν condition, σε μορφή one hot vector. Ο Generator δηλαδή, είναι όπως ο Generator ενός Conditional GAN με την διαφορά ότι αντί να παίρνει σαν είσοδο θόρυβο, παίρνει ένα δεδομένο από κάποιο πεδίο. Αυτή η μορφή τον διευκολύνει να χρησιμοποιεί τα χαρακτηριστικά που έχει μάθει να χρησιμοποιεί για κάποιο πεδίο, όταν δημιουργεί και όλα τα υπόλοιπα. Ο Discriminator από την άλλη, παίρνει σαν είσοδο ένα δείγμα, και εκτός από το να προβλέπει αν είναι αληθινό ή όχι, κάνει και ταξινόμηση σε ποιο πεδίο πιστεύει ότι ανήκει.

Ο Discriminator αποτελείται, εκτός από ένα νευρωνικό δίκτυο το οποίο επεξεργάζεται το δεδομένο, και από δύο επιπλέον πλήρως συνεκτικά επίπεδα. Το ένα χρησιμοποιείται για να γίνει η πρόβλεψή του αν το δεδομένο είναι πραγματικό ή όχι και το δεύτερο για να γίνει η ταξινόμηση στις κλάσεις.

Η εκπαίδευση του Generator γίνεται αυξάνοντας και τα δύο σφάλματα, το σφάλμα ταξινόμησης σε πραγματικά ή ψεύτικα δεδομένα για τους ίδιους λόγους με το κανονικό GAN, και το σφάλμα ταξινόμησης σε πεδίο με σκοπό να καταφέρνει ο Generator να δημιουργεί ένα δείγμα από το ζητούμενο πεδίο. Αυτό γίνεται αυξάνοντας το σφάλμα ταξινόμησης του Discriminator με το target πεδίο, αυξάνοντας δηλαδή την πιθανότητα που νομίζει ο Discriminator ότι το συνθετικό δεδομένο ανήκει στο ζητούμενο πεδίο.



Σχήμα 5.5.1: Η δομή ενός StarGAN [77]

5.6 Domain Adversarial Training

Μία επιπλέον σημαντική τεχνική που έχει πολλές εννοιολογικές ομοιότητες με τα GAN, είναι η ιδέα του Domain Adversarial Training (ανταγωνιστική εκπαίδευση πεδίου) [78]. Στόχος αυτής της διαδικασίας εκπαίδευσης είναι να μάθει αναπαραστάσεις από δυο διαφορετικά αλλά παρόμοια πεδία εισόδου, οι οποίες είναι ανεξάρτητες από το ποιο πεδίο προέρχονται τα δεδομένα.

Συγκεκριμένα, το πρόβλημα που προσπαθεί να λυθεί είναι το εξής: Έστω ένα πρόβλημα ταξινόμησης στο οποίο X είναι ο χώρος εισόδου και $Y = \{0, 1, \ldots, L-1\}$ είναι τα L διαφορετικά labels. Επιπρόσθετα, έχουμε δύο διαφορετικές κατανομές πάνω στο $X \times Y$, τις οποίες αποκαλούμε πεδίο source $\mathcal{D}_S$ και πεδίο target $\mathcal{D}_T$. Σαν δεδομένα εκπαίδευσης δίνονται *i.i.d.* δείγματα (ανεξάρτητα και πανομοιότυπα κατανεμημένα) από την κατανομή $\mathcal{D}_S$ και *i.i.d.* δείγματα από την κατανομή $\mathcal{D}_T^X$, όπου $\mathcal{D}_T^X$ είναι η περιθωριακή (marginal) κατανομή της $\mathcal{D}_T$ πάνω στο X:

$$S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \sim (\mathcal{D}_S)^n, \; T = \{\mathbf{x}_i\}_{i=n+1}^{N} \sim (D_T^X)^{n'}, \tag{5.6.1}$$

όπου $N = n + n'$ είναι ο συνολικος αριθμός των δειγμάτων. Ο στόχος του αλγορίθμου είναι να φτιαχτεί ένας ταξινομητής $\eta : X \to Y$ με ελάχιστη πιθανότητα σφάλματος στην target κατανομή (target risk):

$$R_{\mathcal{D}_T}(\eta) = \Pr_{(\mathbf{x},y)\sim\mathcal{D}_T}\big(\eta(\mathbf{x}) \neq y\big) \tag{5.6.2}$$

ενώ τα labels της target κατανομής δεν δίνονται κατά την εκπαίδευση.

Σε προηγούμενη έρευνα [79] έχει δειχθεί ότι το σφάλμα στο target πεδίο είναι φραγμένο από το σφάλμα στο source πεδίο συν κάποια απόσταση μεταξύ των κατανομών πάνω στα δύο πεδία. Αυτό βγάζει νόημα γατί η συμπεριφορά του ταξινομητή στο source πεδίο είναι καλή μετρική και για το target πεδίο όταν τα πεδία είναι κοντινά μεταξύ τους. Η απόσταση που χρησιμοποιείται στην θεωρητική ανάλυση είναι η $\mathcal{H}$-απόκλιση:

$$d_{\mathcal{H}}(\mathcal{D}_S^X, \mathcal{D}_T^X) = 2 \sup_{\eta\in\mathcal{H}} \left| \Pr_{\mathbf{x}\sim\mathcal{D}_S^X}[\eta(\mathbf{x}) = 1] - \Pr_{\mathbf{x}\sim\mathcal{D}_T^X}[\eta(\mathbf{x}) = 1] \right| \tag{5.6.3}$$

όπου $\mathcal{H}$ είναι το σύνολο όλων των δυαδικών ταξινομητών $\eta : X \to \{0, 1\}$.

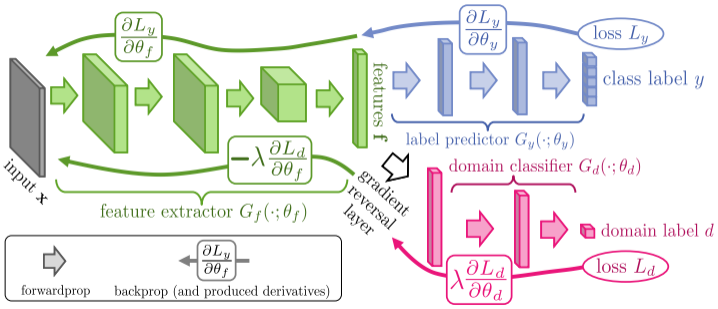
Η λύση που δίνει η Adversarial ταξινόμηση είναι να γίνει αναπαράσταση όλων των δεδομένων σε έναν χώρο τέτοιο ώστε να υπάρχουν όλα τα απαραίτητα στοιχεία έτσι ώστε να μπορεί να γίνει η ταξινόμηση ως προς τα labels αλλά ώστε ταυτόχρονα η $\mathcal{H}$-απόκλιση των αναπαραστάσεων να είναι όσο το δυνατόν πιο μικρή.

Επειδή από τον ορισμό της η $\mathcal{H}$-απόκλιση δεν είναι εύκολα προσεγγίσιμη χρησιμοποιείται ένας επιπλέον ταξινομητής ο οποίος προσπαθεί να διαχωρίσει τα δείγματα από τα source και target πεδία. Στην πράξη χρησιμοποιείται ένα νευρωνικό δίκτυο το οποίο εκπαιδεύεται στο νέο εικονικό dataset:

$$\mathcal{U} = \{(\mathbf{x}_i, 0)\}_{i=1}^{n} \cup \{(\mathbf{x}_i, 1)\}_{i=n+1}^{N} \tag{5.6.4}$$

όπου τα source δεδομένα έχουν πάρει το label 0 και τα target δεδομένα το label 1. Με την εκπαίδευση αυτού του δικτύου με Back Propagation ουσιαστικά προσεγγίζουμε το supremum της εξίσωσης 5.6.3.

Στην συνέχεια, το δίκτυο που εξάγει τα χαρακτηριστικά που θα χρησιμοποιηθούν για την ταξινόμηση, εκπαιδεύεται με σκοπό να μειώσει αυτή την προσέγγιση της $\mathcal{H}$-απόκλισης. Αυτό γίνεται με την βοήθεια του gradient reversal layer, (επιπέδου αντιστροφής της κλίσης), το οποίο αλλάζει το πρόσημο στις παραγώγους οι οποίες περνάνε από αυτό κατά το Back Propagation. Το δίκτυο που θα βγάλει αυτά τα χαρακτηριστικά, εκπαιδεύεται δηλαδή με σκοπό να μεγιστοποιήσει το σφάλμα του ταξινομητή πεδίου. Επειδή υποθέτουμε ότι ο ταξινομητής έχει εκπαιδευτεί αρκετά ώστε να προσεγγίζει την συνάρτηση απόκλισης, τα χαρακτηριστικά των ενδιάμεσων πεδίων εκπαιδεύονται με σκοπό να έχουν όσο λιγότερο πληροφορία γίνεται για το πεδίο από το οποίο προέρχονται, και κατά συνέπεια οι κατανομές των ενδιάμεσων χαρακτηριστικών των δύο πεδίων έρχονται πιο κοντά μεταξύ τους. Η λογική αυτής της εκπαίδευσης φαίνεται στο Σχήμα 5.6.1.



Σχήμα 5.6.1: Αρχιτεκτονική της Adversarial Ταξινόμησης

Παρόλο που αυτή η διαδικασία αναπτύχθηκε για δύο μόνο πεδία, η γενίκευση της σε περισσότερα είναι αρκετά εύκολη. Η μόνη ουσιαστική αλλαγή είναι ότι η συνάρτηση σφάλματος αλλάζει από binary cross entropy σε multi-class cross entropy, και αντί να προβλέπεται μία τιμή που αντικατοπτρίζει την πιθανότητα να έρχεται το δεδομένο από το ένα πεδίο, προβλέπεται μια τιμή για κάθε πεδίο με την βοήθεια της συνάρτησης softmax.

Πολλές εφαρμογές Text to Speech έχουν αξιοποιήσει την Adversarial ταξινόμηση, κυρίως σε περιπτώσεις όπου υπάρχουν πολλοί ομιλητές στο dataset και κάποια δεδομένα πρέπει να είναι ανεξάρτητα του ομιλητή. Παραδείγματα είναι όταν χρησιμοποιείται ένας Encoder πάνω στο κείμενο ο οποίος δεν πρέπει να έχει την πληροφορία του ομιλητή (πάνω σε dataset πολλών γλωσσών που κάποιες γλώσσες έχουν λίγους ομιλητές και κατά συνέπεια είναι εύκολο να προβλέψεις κάποιους ομιλητές μόνο από το ποια γλώσσα μιλάνε) [80] ή όταν χρησιμοποιείται ένας Encoder ο οποίος πρέπει να έχει πληροφορία μόνο για το αν υπάρχει θόρυβος στην ομιλία ή όχι [81].

5.7 Επεξεργασία Φωνής με GAN

5.7.1 Γενικά

Η βασική δομή των GAN περιλαμβάνει την μη επιβλεπόμενη δημιουργία δεδομένων από μία δοσμένη κατανομή. Η μορφή των παραγμένων δειγμάτων είναι ως επί το πλείστον ένα διάνυσμα σταθερού μεγέθους. Επειδή ο ήχος στην γενική περίπτωση είναι μεταβλητού μεγέθους, η βασική αρχιτεκτονική δεν συστήνεται για απλή δημιουργία μη επιβλεπόμενων δεδομένων ήχου. Για αυτό τον λόγο, ο μεγαλύτερος αριθμός των εφαρμογών των GAN στον ήχο έχει να κάνει με την μετατροπή φωνής από πεδίο σε πεδίο.

Μία εξαίρεση σε αυτό τον κανόνα αποτελεί το WaveGAN [82] το οποίο απλά δημιουργεί σήματα ήχου σταθερού μεγέθους. Για να δουλέψει αυτή η στρατηγική, πρέπει και τα δοσμένα δεδομένα να έχουν σταθερό μέγεθος, κάτι που επιτυγχάνεται προσθέτοντας τον απαραίτητο αριθμό από μηδενικά στο τέλος του σήματος (padding). Η αρχιτεκτονική του μοντέλου αποτελείται από ένα πλήρως συνεκτικό επίπεδο και στη συνέχεια μία σειρά από συνελίξεις, με σκοπό να αξιοποιηθεί ο περιοδικός χαρακτήρας των σημάτων φωνής. Αυτή η τεχνική δουλεύει κυρίως σε σχετικά απλά σήματα ήχου (συγκεκριμένα σε εκφωνήσεις των ψηφίων από το μηδέν μέχρι το εννιά) καθώς σε πιο περίπλοκα σήματα υπάρχουν συσχετίσεις μεταξύ δειγμάτων που απέχουν μεγάλη χρονική διάρκεια μεταξύ τους, κάτι που δεν μοντελοποιείται εύκολα με την δεδομένη αρχιτεκτονική. Επίσης, όταν τα δεδομένα έχουν μεγάλη διαφορά στην διάρκεια, το padding που απαιτείται αποτελεί μεγάλο μέρος του σήματος κάτι που δυσκολεύει αρκετά τον Generator να δημιουργήσει λόγω της συνελικτικής του δομής.

Ένα επιπλέον πρόβλημα που δημιουργείται με την απλοϊκή στρατηγική του WaveGAN είναι ότι λόγω των deconvolutional δικτύων που χρησιμοποιούνται στον Generator για το upsampling του σήματος από θόρυβο σε ομιλία, πολλές φορές δημιουργούνται artifacts σε σχήμα σκακιέρας τα οποία προκύπτουν όταν το stride της συνέλιξης δεν διαιρεί ακέραια το μέγεθος του φίλτρου, με αποτέλεσμα να υπάρχει άνιση επικάλυψη μεταξύ διαδοχικών εφαρμογών του φίλτρου σε κάθε σημείο του σήματος [83]. Αυτό είναι ένα γενικό πρόβλημα των GAN που υπάρχει και στην περίπτωση των εικόνων.

Η διαφορά που υπάρχει στα σήματα ήχου είναι ότι η διαδικασία του deconvolution δημιουργεί υψηλές συχνότητες οι οποίες υπάρχουν και στα πραγματικά δεδομένα, σε αντίθεση με τις εικόνες που γενικά δεν έχουν πολλές περιοχές με περιοδικότητα υψηλών συχνοτήτων. Με την αρχιτεκτονική του WaveGAN όμως, οι συχνότητες αυτές δημιουργούνται πάντα σε σταθερά σημεία του σήματος, κάτι που δεν ισχύει στα πραγματικά δεδομένα. Για να αντιμετωπιστεί αυτό το πρόβλημα, χρησιμοποιείται ένα phase suffling που σε κάθε επίπεδο του δικτύου που αποτελείται από μία κυκλική μετάθεση κατά έναν μικρό τυχαίο αριθμό. Αυτό αλλάζει σε κάθε επίπεδο το σημείο που δημιουργούνται οι υψηλές συχνότητες και έτσι δεν μπορεί ο Discriminator να μάθει ποια δεδομένα είναι ψεύτικα απλά ανιχνεύοντας το σημείο όπου βρίσκονται αυτές οι συχνότητες.

Ένα πεδίο εφαρμογής των GAN στην μετατροπή φωνής είναι στο Speech enhancement, σκοπός του οποίου είναι να μετασχηματίσει σήμα φωνής που έχει θόρυβο σε καθαρό σήμα. Στο SEGAN [84] για παράδειγμα, χρησιμοποιούν τον Discriminator για να ξεχωρίσουν ζευγάρια από ένα καθαρό σήμα και ένα θορυβώδες. Επειδή τέτοια δεδομένα είναι δύσκολο να βρεθούν, χρησιμοποιούν σαν dataset καθαρά σήματα φωνής που τους προσθέτουν τεχνητό θόρυβο. Η δομή του Generator είναι παρόμοια με αυτή ενός Autoencoder κάνει δηλαδή υποδειγματοληψία, χρησιμοποιώντας συνελικτικά δίκτυα και μετά υπερδειγματοληψία ώστε να καταλήξουμε σε μέγεθος με ίδιο σχήμα με το αρχικό. Κάθε ενδιάμεσο στάδιο κατά την υποδειγματοληψία συνδέεται με το αντίστοιχο στάδιο της υπερδειγματοληψίας με skip συνδέσεις, ώστε να μην χαθούν οι πληροφορίες χαμηλού επιπέδου περνώντας από πολλά στάδια. Επίσης, χρησιμοποιείται ένας επιπλέον θόρυβος σαν είσοδος στην ενδιάμεση αναπαράσταση, παίζοντας τον ρόλο της latent κατανομής των κλασικών GAN. Παρόμοια μοντέλα έχουν χρησιμοποιηθεί για τον ίδιο σκοπό όπως [85, 86].

Ένα ακόμα πεδίο εφαρμογής των GAN σε ήχο είναι σαν Vocoders καθώς ο μετασχηματισμός από Mel

Spectrograms σε αρχείο ήχου έχει σταθερή σχέση μεγέθους. Για παράδειγμα ένα δείγμα από Mel Spectrogram αντιστοιχεί πάντα σε τόσα δείγματα ήχου όσο το παράθυρο που χρησιμοποιήθηκε. Μάλιστα, τα GAN έχουν πλεονέκτημα σε αυτή την κατηγορία γιατί δεν είναι Auto-regressive με αποτέλεσμα να γίνεται η σύνθεση με μεγάλη ταχύτητα. Ένα τέτοιο παράδειγμα είναι το MelGAN [87] το οποίο χρησιμοποιεί έναν απλό συνελικτικό Generator παρόμοιο με του WaveGAN με την διαφορά ότι κάνει και υπερδειγματοληψία, και χρησιμοποιεί διαφορετικούς Discriminator οι οποίοι λειτουργούν σε διαφορετικά επίπεδα κλίμακας. Για παράδειγμα, ένας Discriminator κοιτάει αν το σήμα είναι ρεαλιστικό σε π.χ. ένα χρονικό διάστημα μερικών millisecond ενώ ένας άλλος κοιτάει στην κλίμακα των δευτερολέπτων. Ένα ακόμα τέτοιο μοντέλο το οποίο είναι βασισμένο στο MelGAN αλλά αποτελεί μια από τις πιο συνηθισμένες επιλογές για Vocoders αυτήν την εποχή είναι το HifiGAN [88].

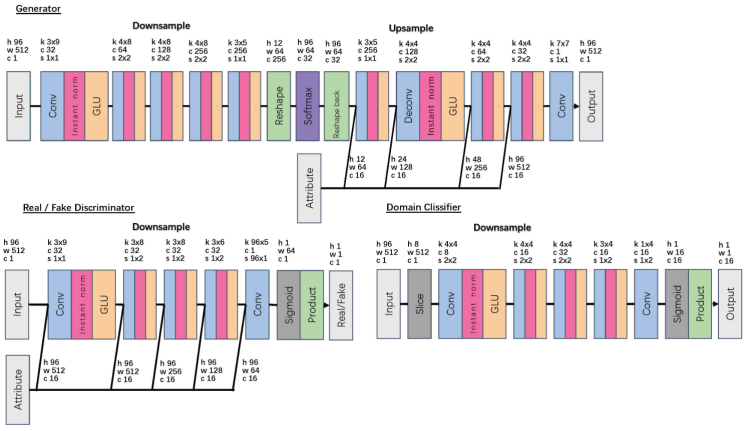
Ένα πρόσφατο πεδίο εφαρμογής των GAN είναι στο Text to Speech [89, 90]. Η διάρκεια του ζητούμενου σήματος υπολογίζεται είτε με ένα ξεχωριστό νευρωνικό δίκτυο, είτε θεωρείται κάποιο συγκεκριμένο πολλαπλάσιο του μεγέθους της εισόδου και επιτυγχάνεται με υπερδειγματοληψία. Έτσι ο Generator ουσιαστικά μετατρέπει την γλωσσολογική πληροφορία σε σήμα ήχου. Χρησιμοποιούνται πάλι πολλοί Discriminators σε διαφορετικές κλίμακες.

5.7.2 StarGAN-VC

Η διαδικασία εκπαίδευσης του CycleGAN, όπως περιγράφηκε στην Ενότητα 5.4 έχει εφαρμοστεί και στο πεδίο της ομιλίας στο Voice Conversion [91, 92, 93]. Το Voice Conversion έχει σκοπό την μετατροπή μιας πρότασης εκφωνημένη από έναν δοσμένο ομιλητή (source) σε έναν άλλο (target), αλλάζοντας μόνο τα χαρακτηριστικά της ομιλίας που αφορούν την ταυτότητα του ομιλητή και κρατώντας το κείμενο άλλα και όλη την υπόλοιπη προσωδία σταθερά. Όπως και το αρχικό CycleGAN, το CycleGAN-VC δουλεύει μεταξύ μόνο 2 ομιλητών.

Στην περίπτωση του προβλήματος του Voice Conversion με πολλούς ομιλητές, υπάρχει το StarGAN-VC [94] το οποίο έχει αποτελέσει ως η βάση για όλες τα πειράματα μας. Όπως και το κανονικό StarGAN, έχει ένα επιπλέον μοντέλο που κάνει ταξινόμηση στις κλάσεις, με σκοπό να χρησιμοποιηθεί το σφάλμα ταξινόμησης από τον Generator ώστε να γίνει η μετατροπή στα δεδομένα από την σωστή κλάση.

Για να δουλέψει το σύστημα στα δεδομένα του ήχου, χρειάστηκαν να γίνουν ορισμένες σχεδιαστικές επιλογές που καλύτερα αξιοποιούν τα χαρακτηριστικά του ήχου. Η αρχική προ-επεξεργασία των δεδομένων γίνεται με τον WORLD Vocoder και για κάθε πλαίσιο ομιλίας εξάγουμε 3 χαρακτηριστικά, τα Mel-Cepstral coefficients χαρακτηριστικά, τον λογάριθμο της θεμελιώδης συχνότητας, και τα απεριοδικά χαρακτηριστικά (aperiodicities). Προηγούμενη έρευνα [95] έδειξε ότι τα απεριοδικά χαρακτηριστικά δεν επηρεάζουν σε μεγάλο βαθμό την ποιότητα του σήματος, οπότε δεν μεταβάλλονται κατά την μετατροπή.



Σχήμα 5.7.1: Αρχιτεκτονική του StarGAN-VC. Τα c,h,w αντιστοιχούν στις διαστάσεις των δεδομένων (κανάλια, ύψος, βάθος), και τα k, s στα kernel size, stride των συνελίξεων.

Για την μετατροπή της θεμελιώδης συχνότητας γίνεται η υπόθεση ότι η κατανομή που ακολουθούν είναι λογαριθμική κανονική κατανομή [96]. Αρχικά η μέση τιμή και η τυπική απόκλιση των $\log(F_0)$ των δύο ομιλητών υπολογίζεται. Στην συνέχεια, για την μετατροπή, χρησιμοποιείται ο τύπος:

$$F_{0_t} = \exp\left(\frac{(\log(F_{0_s}) - \mu_s)}{\sigma_s} * \sigma_t + \mu_t\right) \tag{5.7.1}$$

όπου μ και σ είναι η μέση τιμή και η τυπική απόκλιση αντίστοιχα, και οι δείκτες s και t δείχνουν τους source και target ομιλητές.

Ο τύπος αυτός είναι ισοδύναμος με το να γίνεται κανονικοποίηση στον λογάριθμο των δεδομένων της F_0 και στην συνέχεια αποκανονικοποίηση στην ζητούμενη κατανομή του target. Έτσι, η συνθετική ομιλία θα έχει τα σωστά στατιστικά χαρακτηριστικά και ένα frame που στον source ομιλητή θα έχει F_0 π.χ. τιμή 3 τυπικών αποκλίσεων μακριά από την μέση τιμή του, η τιμή μετά την μετατροπή θα έχει τιμή 3 αποκλίσεων μακριά από την μέση τιμή και στην κατανομή του target ομιλητή.

Για την μετατροπή των spectral χαρακτηριστικών, καθώς και για τον Discriminator χρησιμοποιείται ένα διδιάστατο συνελικτικό νευρωνικό δίκτυο, η πλήρης αρχιτεκτονική του οποίου φαίνεται στο Σχήμα 5.7.1. Χρησιμοποιούνται residual connections μεταξύ των επιπέδων (όπως αναλύονται στην Ενότητα 3.4.3), και κάποιες επιπλέον μονάδες:

- Batch Normalization
  Σε κάθε συνελικτικό επίπεδο γίνεται κανονικοποίηση των δεδομένων σε κάθε batch [97]. Συγκεκριμένα, ο τύπος της κανονικοποίησης είναι:

$$\begin{aligned} \sigma_c^2 &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{ict} - \mu_c)^2 \\ \hat{x} &= \frac{x - \mu_c}{\sigma_c} \end{aligned} \tag{5.7.2}$$

  όπου n,c,t είναι οι δείκτες του δεδομένου στο batch, στο κανάλι και στον χρόνο αντίστοιχα, T η συνολική διάρκεια του σήματος και N ο αριθμός των batches.

- Gated Linear Unit
  Μια εναλλακτική μη γραμμικότητα αντί για τις κλασικές activation functions [98]. Το αποτέλεσμα μιας συνέλιξης χωρίζεται σε δύο μέρη και το ένα χρησιμοποιείται για να επιλέξει τις τιμές που θα περάσουν από το δεύτερο με την βοήθεια μιας σιγμοειδής συνάρτησης. Με τον τρόπο αυτό, διαφορετικά σημεία του σήματος περνάνε στο επόμενο επίπεδο, ανάλογα με τα δεδομένα.

  Για το επίπεδο l ο τύπος δίνεται από

$$H_{l+1} = (H_l * W_l + b_l) \otimes \sigma(H_l * V_l + c_l) \tag{5.7.3}$$

  όπου H_l είναι η έξοδος της συνέλιξης του προηγούμενου επιπέδου, b, c είναι biases, και W, H πίνακες που χρησιμοποιούνται για τον μετασχηματισμό και σ είναι η σιγμοειδής συνάρτηση.

Για να γίνει η μετατροπή, χρησιμοποιείται και το label του target ομιλητή, σε αναπαράσταση one hot vector. Για να διευκολυνθούν τα δίκτυα, αυτή η αναπαράσταση επαναλαμβάνεται κατά το μήκος του σήματος, όσες φορές χρειάζεται για να αποκτήσει το ίδιο μέγεθος και προστίθεται σαν επιπλέον κανάλι. Αυτή η διαδικασία γίνεται σε πολλά επίπεδα ενός Conditional GAN.

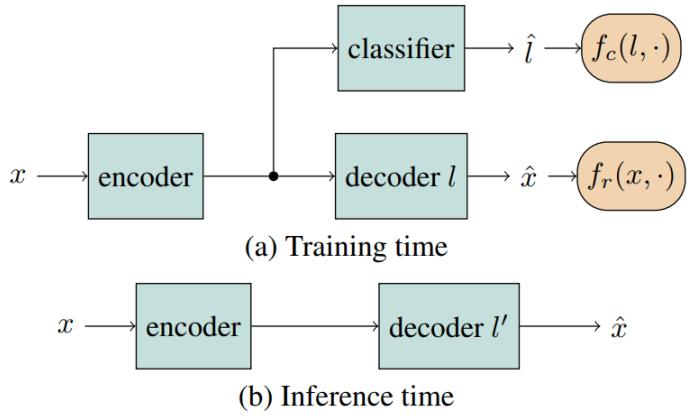
Αξίζει να σημειωθεί ότι όλη η αρχιτεκτονική είναι πλήρως συνελικτική, χωρίς καθόλου πλήρως συνεκτικά επίπεδα, κάτι που επιτρέπει την μετατροπή ολόκληρων ακολουθιών με αυθαίρετο μήκος. Σαν αποτέλεσμα, ο Classifier δεν δουλεύει στο επίπεδο όλης της πρότασης αλλά για κάθε κομμάτι που αντιστοιχεί στο receptive field του, κάνει ξεχωριστή ταξινόμηση. Το ίδιο ισχύει και για τον Discriminator.

5.7.3 Adversarially Trained Autoencoders

Μια επιπλέον αρχιτεκτονική που χρησιμοποιήθηκε στο πρόβλημα του Voice Conversion είναι οι Adversarially Trained Autoencoders [99]. Η βάση αυτής της αρχιτεκτονικής είναι η Adversarial ταξινόμηση που αναλύθηκε στην Ενότητα 5.6 και οι Autoencoders που αναλύθηκαν στην Ενότητα 3.4.4.

Η δομή φαίνεται στην Εικόνα 5.7.2 και αποτελείται από έναν Encoder και πολλούς Decoders. Υπάρχει ένας Decoder για κάθε πεδίο (ομιλητή στην περίπτωση του Voice Conversion). Ο σκοπός του Encoder είναι να

κάνει απεικόνιση την ομιλία από έναν ομιλητή σε έναν ενδιάμεσο χώρο ο οποίος δεν περιέχει πληροφορία για την ταυτότητα του ομιλητή και ο σκοπός του Decoder l είναι να κάνει σύνθεση με την φωνή του ομιλητή l. Κατά την σύνθεση, όταν θέλουμε να μετασχηματίσουμε μια πρόταση από έναν ομιλητή l σε ομιλητή l' χρησιμοποιούμε τον Encoder που είναι ο ίδιος σε όλους τους ομιλητές και τον Decoder l'.



Σχήμα 5.7.2: Adversarially trained Autoencoder

Κατά την διάρκεια της εκπαίδευσης χρησιμοποιείται ένας ταξινομητής που προσπαθεί να προβλέψει τον ομιλητή από την έξοδο του Encoder. Από αυτόν προκύπτει ένα σφάλμα ταξινόμησης f_c το οποίο ο ταξινομητής εκπαιδεύεται ώστε να ελαχιστοποιήσει. Το σφάλμα που χρησιμοποιείται είναι binary cross entropy. Ταυτόχρονα, ο Encoder εκπαιδεύεται έτσι ώστε να μεγιστοποιήσει το σφάλμα του ταξινομητή κάτι που μπορεί να γίνει μόνο αν μάθει να αφαιρεί όλη την πληροφορία για την ταυτότητα του ομιλητή από την είσοδό του, κάνοντας την κατανομή της εισόδου ανεξάρτητη από τον ομιλητή.

Όταν εκπαιδεύεται ένα δείγμα από τον ομιλητή l χρησιμοποιείται ο αντίστοιχος Decoder l με σκοπό να κάνει ανακατασκευή του σήματος (reconstruction), με την βοήθεια ενός σφάλματος f_r, το οποίο επιλέγεται να είναι η 1-νόρμα L_1.

Παρόλο που το μοντέλο εκπαιδεύεται μόνο από το σφάλμα ανακατασκευής, και κάθε δείγμα απεικονίζεται στον εαυτό του, δεν χρειάζονται παράλληλα δείγματα από διαφορετικά πεδία. Αυτό επιτυγχάνεται μέσω της ενδιάμεσης κατανομής η οποία, αν η εκπαίδευση πετύχει, είναι η ίδια σε όλα τα δείγματα. Έτσι, κατά την σύνθεση, μπορούμε να κάνουμε μετατροπή από τον ένα ομιλητή στον άλλο χωρίς το μοντέλο να έχει κάνει ποτέ ξανά μετατροπή κατά την εκπαίδευση.

Η διαφορά αυτού του μοντέλου με το StarGAN είναι ότι το StarGAN χρησιμοποιεί ένα μόνο μοντέλο για την απεικόνιση των δεδομένων από το ένα πεδίο στο άλλο και η μεταφορά γίνεται χρησιμοποιώντας τα labels του κάθε ομιλητή. Από αυτή την άποψη, το StarGAN είναι πιο αποτελεσματικό. Ο Autoencoder από την άλλη έχει το πλεονέκτημα ότι τα σφάλματα που χρησιμοποιούνται είναι μόνο σφάλματα ανακατασκευής καθώς και το γεγονός ότι κάθε μοντέλο μπορεί πιο εύκολα να εξειδικευτεί στην σύνθεση του ομιλητή που του αντιστοιχεί.

Κεφάλαιο 6

Μη παράλληλη, από Πολλά σε Πολλά, Μετατροπή Συναισθηματικής Ομιλίας

6.1 Σχετική Έρευνα

Το μεγαλύτερο μέρος της σχετικής έρευνας σε αυτόν τον τομέα επικεντρώνεται στην μετατροπή του ομιλητή, αντί για την μετατροπή του συναισθήματος, αλλά πέρα από κάποιες διαφορές οι οποίες θα αναλυθούν στην συνέχεια, οι μαθηματική μοντελοποίηση των προβλημάτων είναι η ίδια. Έχουμε πάλι την διαφοροποίηση σε παράλληλη και μη παράλληλη μετατροπή μεταξύ ομιλητών.

Στην περίπτωση των παράλληλων δεδομένων, υπάρχουν αρχιτεκτονικές που βασίζονται σε τεχνικές sequence to sequence (ακολουθία σε ακολουθία) όπως αυτές που χρησιμοποιούνται στο Tacotron ή σε επεξεργασία κείμενο όπως στους Transformers [40]. Τέτοια παραδείγματα είναι τα [100] και [101] στα οποία η αντιστοίχιση μεταξύ των παράλληλων δεδομένων μαθαίνεται από ένα δίκτυο attention (προσοχής). Στο [102] η αντιστοίχιση γίνεται επίσης με attention αλλά όλα τα δίκτυα που χρησιμοποιούνται είναι συνελικτικά. Επιπλέον στόχοι μπορούν να ανατεθούν στα μοντέλα, όπως στο [103] που από τις ενδιάμεσες αναπαραστάσεις γίνεται πρόβλεψη γλωσσολογικών χαρακτηριστικών με σκοπό να ενθαρρύνουν τις ενδιάμεσες αναπαραστάσεις να διατηρήσουν την αντίστοιχη πληροφορία. Σε ορισμένες δουλειές όπως στο [104] χρησιμοποιούνται GANs ώστε να αντικαταστήσουν τις συναρτήσεις σφάλματος όπως το L2 οι οποίες τείνουν να εξομαλύνουν υπερβολικά τα spectrograms.

Στην περίπτωση των μη παράλληλων δεδομένων, οι περισσότερες προσεγγίσεις βασίζονται είτε πάνω στους Variational Autoencoders είτε στα GAN. Στην περίπτωση των VAE, συνήθως η μετατροπή γίνεται σε επίπεδο πλαισίου όπως στο [105] που το label της εισόδου παρατίθεται μαζί με το διάνυσμα του ενδιάμεσου χώρου του VAE πριν πάει στον Decoder. Μια παρόμοια ιδέα είναι το [106], με την διαφορά ότι χρησιμοποιεί δύο ειδών ακουστικά χαρακτηριστικά, mfcc και χαρακτηριστικά από τον Vocoder STRAIGHT [107]. Χρησιμοποιεί δύο Encoders, έναν για κάθε χαρακτηριστικό και δύο αντίστοιχους Decoders, με την διαφορά ότι κάνει και ανακατασκευή από τον ένα Decoder με είσοδο τα χαρακτηριστικά του άλλου, καθώς και βάζει και ένα επιπλέον σφάλμα με σκοπό οι ενδιάμεσες αναπαραστάσεις να είναι κοντά μεταξύ τους για τα δύο χαρακτηριστικά. Άλλες προσεγγίσεις είναι η [108] που μεγιστοποιεί την κοινή πληροφορία μεταξύ του ανακατασκευασμένου σήματος και του label της κλάσης και η [109] που χρησιμοποιεί τα Phonetic Posterior-Grams (PPGs) με σκοπό να διατηρήσει την γλωσσική πληροφορία.

Οι GAN αρχιτεκτονικές βασίζονται συνήθως στο CycleGAN, όπως το [110] και το [91] που χρησιμοποιούνται για την μετατροπή μεταξύ δύο ομιλητών, με την διαφορά ότι το πρώτο χρησιμοποιεί κανονικά νευρωνικά δίκτυα και το δεύτερο συνελικτικά. Αντίστοιχο είναι και το [111] που χρησιμοποιεί το DiscoGAN (ανάλογο με το CycleGAN) για την μετατροπή του ομιλητή, προσθέτοντας και μερικά επιπλέον σφάλματα για την καλύτερη διατήρηση της γλωσσολογικής πληροφορίας, καθώς και για την μετατροπή του στυλ ομιλίας του ομιλητή. Επίσης, στο [112] χρησιμοποιούν πάλι CycleGAN αλλά με τα PPGs σαν χαρακτηριστικά εισόδου τα οποία μετά μετασχηματίζονται πίσω σε ομιλία με ένα ειδικό μοντέλο σύνθεσης. Φυσικά σε αυτήν την κατηγορία ανήκει και το StarGAN-VC [94] που αντικαθιστά το CycleGAN με το StarGAN, όπως εξηγήθηκε στην Ενότητα 5.7.2. Ένα επιπλέον μοντέλο που χρησιμοποιεί GAN είναι το VAW-GAN [113] το οποίο συνδυάζει ένα Conditional VAE με ένα WGAN. Τέλος, υπάρχει και το [114] το οποίο εκπαιδεύει πρώτα έναν Autoencoder με σκοπό να μετατρέψει την είσοδο σε έναν χώρο ανεξαρτήτου ομιλητή και στην συνέχεια εκπαιδεύει ένα GAN να κάνει την μετατροπή.

Όσον αφορά την μετατροπή συναισθήματος, έχει γίνει κάποια σχετική δουλειά η οποία κατά το μεγαλύτερο μέρος της έγινε κατά την διάρκεια εκπόνησης αυτής της εργασίας. Στο [115], μια δουλειά που είναι βασισμένη σε Autoencoders, χρησιμοποιούν έναν ενδιάμεσο χώρο που περιέχει όλη την πληροφορία που είναι ανεξάρτητη συναισθήματος και δύο χώρους οι οποίοι περιέχουν την πληροφορία για τα δύο συναισθήματα που χρησιμοποιούνται στην εκπαίδευση (ασχολείται με ένα σε ένα μετατροπή). Στην περίπτωση των ένα σε ένα μετατροπών, υπάρχει επίσης και το [116] που βασίζεται επίσης στο CycleGAN με τον μετασχηματισμό του F0 να γίνεται με έναν continuous wavelet transform (CWT). Επιπλέον, υπάρχουν δουλειές βασισμένες πάνω στο VAW-GAN, [117, 118], η πρώτη εκ των οποίων είναι απλή εφαρμογή του VAW-GAN για τον μετασχηματισμό των cepstral χαρακτηριστικών και του CWT μετασχηματισμό για το F0, και η δεύτερη χρησιμοποιεί έναν επιπλέον προεκπαιδευμένο ταξινομητή συναισθήματος του οποίου τα ενδιάμεσα χαρακτηριστικά δίνονται σαν είσοδος στο μοντέλο σύνθεσης. Η άμεση μεταφορά του StarGAN-VC στο πεδίο της μετατροπής της συναισθηματικής ομιλίας έχει γίνει από το [119], το οποίο μετέπειτα χρησιμοποιεί τα συνθετικά δεδομένα για να κάνει εκπαίδευση καλύτερου ταξινομητή συναισθημάτων. Το [120] χρησιμοποιεί έναν VAE με ιεραρχική δομή σε διαφορετικές χρονικές κλίμακες οι οποίες είναι conditioned στο επισημειωμένο. Τέλος, το [121] κάνει μετατροπή συναισθημάτων με GAN με την διαφορά ότι ο Discriminator παίρνει σαν είσοδο ζευγάρια από

ζεύγη A,B και ταξινομεί το αν αντιστοιχούν στον μετασχηματισμό $A \rightarrow B$ ή στον αντίστροφο, όπου A είναι ο ζητούμενος μετασχηματισμός συναισθημάτων και B είναι το παραγμένο σήμα. Επιπλέον χρησιμοποιούν έναν αντιστρέψιμο μετασχηματισμό για τα F0 χαρακτηριστικά.

6.2 Περιγραφή Βάσης Δεδομένων

Τα μοντέλα εκπαιδεύτηκαν σε δύο βάσεις δεδομένων συναισθηματικής ομιλίας, το CVSP-Expressive Audio-Visual Speech Corpus (CVSP_EAV) [6] και το Toronto Emotional Speech (TESS) [7].

- Το CVSP_EAV είναι μία βάση δεδομένων Ελληνικής ομιλίας, εκφωνημένο από έναν ηθοποιό, που περιέχει προτάσεις σε τέσσερα συναισθήματα (θυμός, χαρά, λύπη, ουδέτερο) με 899 προτάσεις για κάθε συναίσθημα. Αυτό το dataset έχει το πλεονέκτημα ότι έχει σχετικά καλό αριθμό από δεδομένα, αλλά αφού αποτελείται από μόνο τέσσερα συναισθήματα, είναι πιο δύσκολο να γίνει αξιολόγηση του κατά πόσο το μοντέλο μπορεί να χειριστεί μετατροπή μεταξύ πολλών πεδίων. Η συνολική διάρκεια της βάσης δεδομένων είναι 4 ώρες και 29 λεπτά, για τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση.
- Το TESS είναι μία Αγγλική βάση δεδομένων εκφωνημένη από δύο ηθοποιούς που περιέχει επτά συναισθήματα (θυμός, χαρά, λύπη, ουδέτερο, απέχθεια, ευχάριστη έκπληξη, φόβος) με 200 προτάσεις για κάθε συναίσθημα και ομιλητή, καταλήγοντας σε 2800 συνολικές προτάσεις. Παρόλο που ο μεγαλύτερος αριθμός των συναισθημάτων είναι χρήσιμος για την έρευνά μας, όλες οι ηχογραφημένες προτάσεις έχουν την μορφή "Say the word 'x' " ("Πες την λέξη 'x'"), που κάνει το μοντέλο επιρρεπές στο overfitting. Αφού το μοντέλο πρέπει να μάθει τις απεικονίσεις μόνο από 200 λέξεις, συν την αρχή της πρότασης που είναι κοινή σε όλα τα δεδομένα, αποτυγχάνει πιο εύκολα στις λέξεις που δεν έχει ξαναδεί στην εκπαίδευση. Η συνολική διάρκεια της βάσης δεδομένων είναι 1 ώρα και 26 λεπτά, για τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση, σημαντικά μικρότερη από το CVSP_EAV.

Και οι δύο βάσεις δεδομένων, έχουν δειγματοληφθεί σε συχνότητα 16000 Hz.

Αν και η μέθοδος είναι μη παράλληλη, και τα δυο datasets είναι παράλληλα, δηλαδή περιέχουν τις ίδιες προτάσεις σε διαφορετικά συναισθήματα. Είναι σημαντικό να τονίσουμε ότι οι μέθοδοι που χρησιμοποιούνται σε αυτή την εργασία δεν αξιοποιούν κάπως αυτό το γεγονός και αντιμετωπίζουν κάθε πρόταση σαν να υπάρχει μόνο σε ένα συναίσθημα. Η μη παραλληλότητα των βάσεων χρησιμοποιείται μόνο κατά την αξιολόγηση των μοντέλων.

6.3 Περιγραφή Μοντέλων

6.3.1 Baseline

Σαν baseline χρησιμοποιούμε το μοντέλο StarGAN-VC με την διαφορά ότι το χρησιμοποιούμε για την μετατροπή συναισθήματος αντί για ομιλητή. Ακολουθούμε δηλαδή την διαδικασία από το [119] (το οποίο δημοσιεύτηκε κατά την διάρκεια εκπόνησης αυτής της εργασίας). Χρησιμοποιούμε το Wasserstein-GP σαν σφάλμα, (συναρτήσεις σφάλματος 5.2.4 και 5.2.5) με υπερπαράμετρο λ 10. Οι παράμετροι του Generator ενημερώνεται μια φορά για κάθε 5 επαναλήψεις του Discriminator και του Classifier.

Τα χαρακτηριστικά είναι όπως και στο κανονικό μοντέλα τα χαρακτηριστικά του WORLD Vocoder με τα απεριοδικά χαρακτηριστικά να μένουν αμετάβλητα, τα Cepstral χαρακτηριστικά να μετασχηματίζονται από τα νευρωνικά δίκτυα και την θεμελιώδη συχνότητα να μετασχηματίζεται από τον τύπο:

$$F_{0_t} = \exp\left(\frac{(\log(F_{0_s}) - \mu_s)}{\sigma_s} * \sigma_t + \mu_t\right) \tag{6.3.1}$$

Η οποία είναι η ίδια εξίσωση με την εξίσωση που χρησιμοποιείται και από το StarGAN-VC 5.7.2, με την διαφορά ότι οι μέσες τιμές και οι τυπικές αποκλίσεις υπολογίζονται για κάθε συναίσθημα.

6.3.2 Προτεινόμενο Μοντέλο

Το μοντέλο που προτείνεται σε αυτήν την εργασία είναι ένας συνδυασμός του StarGAN και των Adversarially Trained Autoencoders όπως περιγράφηκαν στις Ενότητες 5.7.2 και 5.7.3 αντίστοιχα.

Ένα πρόβλημα που έχει η αρχική αρχιτεκτονική του StarGAN είναι ότι ο Generator πρέπει να μάθει $n*(n-1)$ απεικονίσεις όπου n είναι ο αριθμός των κλάσεων των δεδομένων. Αυτό συμβαίνει επειδή χρησιμοποιείται το ίδιο μοντέλο για όλα τα πιθανά ζευγάρια εισόδου εξόδου. Στο μοντέλο των Adversarially Trained Autoencoders χρησιμοποιείται ένας ενδιάμεσος χώρος που δεν έχει πληροφορία από τον ομιλητή της εισόδου. Το αρνητικό αυτού του μοντέλου είναι ότι χρειάζεται έναν ξεχωριστό Decoder για κάθε πεδίο, κάτι που κάνει το συνολικό μοντέλο απαγορευτικά μεγάλο, ιδιαίτερα όταν έχουμε μεγάλο αριθμό κλάσεων.

Συνδυάζοντας τις δύο προσεγγίσεις, μπορούμε να έχουμε έναν Decoder για όλα τα πεδία, ο οποίος θα παίζει τον ρόλο του Generator στο StarGAN. Χρησιμοποιούμε σαν condition την ζητούμενη κλάση, και εκπαιδεύουμε το μοντέλο με τα σφάλματα του StarGAN. Χρησιμοποιείται ένας Encoder και ένας Classifier με τον ίδιο τρόπο όπως στο μοντέλο των Adversarially Trained Autoencoders, δηλαδή με τον Classifier να προσπαθεί να ταξινομήσει τα δείγματα στην σωστή κλάση (συναισθήματα στην περίπτωσή μας) και τον Encoder να προσπαθεί να αυξήσει αυτό το σφάλμα. Αυτό θα οδηγήσει σε έναν ενδιάμεσο χώρο που ιδανικά δεν θα διατηρεί καθόλου πληροφορία από το συναίσθημα της εισόδου, αλλά θα διατηρεί όλα τα υπόλοιπα χαρακτηριστικά όπως κείμενο, προσωδία, ταυτότητα ομιλητή (για dataset με πολλούς ομιλητές).

Αυτή η αρχιτεκτονική μειώνει τον αριθμό των απεικονίσεων που θα πρέπει να μάθει το μοντέλο από $n*(n-1)$ σε n για τον Encoder (από κάθε κλάση στον ενδιάμεσο ανεξάρτητο χώρο) και n για τον Decoder (από τον χώρο ανεξάρτητου συναισθήματος σε κάθε ξεχωριστό συναίσθημα). Αυτό, αν όλα πάνε καλά, κάνει πιο εύκολη την εκπαίδευση, καθώς απλοποιεί το πρόβλημα που έχει να λύσει κάθε δίκτυο. Επίσης, είναι πιθανόν να μειώσει την επιρροή του συναισθήματος εισόδου στο συναίσθημα στόχου, καθώς η πληροφορία της εισόδου αφαιρείται πριν γίνει η τελική μετατροπή.

Τα features που χρησιμοποιούνται είναι όπως και στο baseline, τα features του WORLD Vocoder, με τα απεριοδικά χαρακτηριστικά να μένουν αμετάβλητα, το F0 να μετατρέπεται με τον ίδιο τρόπο όπως το baseline 6.3.1 και τα Cepstral χαρακτηριστικά να μοντελοποιούνται από τα νευρωνικά δίκτυα.

Σε αρχικά πειράματα παρατηρήθηκε ότι ο Encoder μπορεί να κάνει καλύτερη ταξινόμηση όταν σαν είσοδος τοποθετηθεί και το label του συναισθήματος της εισόδου, δηλαδή πρακτικά, είναι πιο εύκολο για τον Encoder να αφαιρέσει την πληροφορία του χώρου όταν ξέρει ποιος είναι αυτός ο χώρος. Αυτό είναι λίγο περιοριστικό γιατί κατά το inference, χρειάζεται να γνωρίζουμε το συναίσθημα της εισόδου. Είναι όμως μια υπόθεση η οποία ισχύει στην κατηγορία μετατροπής από πολλά σε πολλά καθώς αν αφαιρέσουμε αυτήν την υπόθεση, τότε η μετατροπή ονομάζεται από οποιοδήποτε σε πολλά (any to many). Επιπλέον, επειδή η ταξινόμηση συναισθήματος είναι γενικά πιο εύκολο πρόβλημα από την σύνθεση συναισθήματος, μπορούμε να υποθέσουμε ότι έχουμε στην διάθεσή μας ένα ταξινομητή συναισθήματος ο οποίος προβλέπει το συναίσθημα εισόδου, κάτι που δεν πρέπει να επηρεάσει δραματικά την απόδοση των μοντέλων.

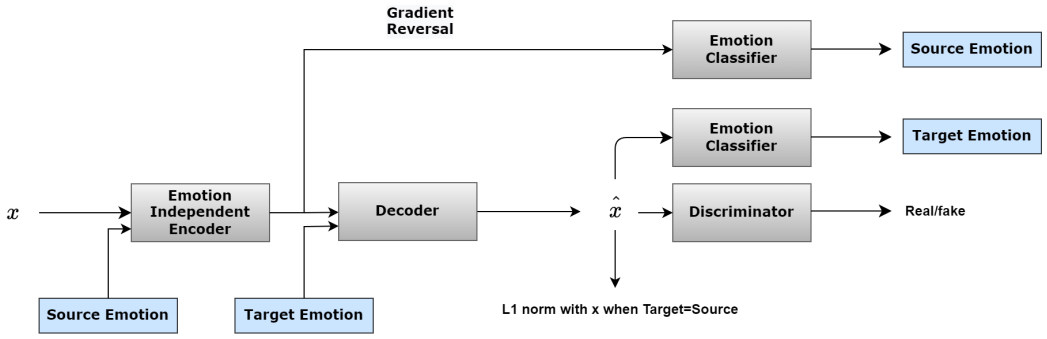
Αρκετές σχεδιαστικές επιλογές παρουσιάζονται όσο αφορά τις αρχιτεκτονικές των μοντέλων. Για παράδειγμα, έχουμε μια ταξινόμηση στον χώρο του συναισθήματος με σκοπό να αφαιρεθεί η πληροφορία του συναισθήματος από την αναπαράσταση του Encoder και μια ακόμα ταξινόμηση με σκοπό να πάρουμε το σφάλμα ταξινόμησης για να εκπαιδεύσουμε τον Generator. Κατά συνέπεια, μπορούμε να χρησιμοποιήσουμε τον ίδιο ταξινομητή και για τις δύο διαδικασίες, ή να συγχωνεύσουμε τον ένα απο τους δύο με τον Discriminator (όπως γίνεται στο StarGAN).

Αρχικά πειράματα έδειξαν ότι η συγχώνευση του Discriminator με τον Classifier που κάνει την ταξινόμηση στον τελικό χώρο, δίνει καλύτερα αποτελέσματα, μειώνοντας παράλληλα των αριθμό των παραμέτρων των μοντέλων. Αυτό γίνεται χρησιμοποιώντας το ίδιο δίκτυο με ένα διαφορετικό γραμμικό επίπεδο να κάνει στο τέλος την ταξινόμηση σε κλάσεις συναισθήματος και στο αν είναι πραγματικό ή όχι το δεδομένο. Έτσι οι ενδιάμεσες αναπαραστάσεις χρησιμοποιούνται και στα δύο προβλήματα. Έγιναν δοκιμές με εισαγωγή βαθύτερων δικτύων στα τελικά επίπεδα που διαχωρίζουν τον Discriminator με τον Classifier αλλά τα αποτελέσματα ήταν χειρότερα.

Αντίθετα, η συγχώνευση του ταξινομητή που λειτουργεί στον ενδιάμεσο χώρο με αυτόν που λειτουργεί στον τελικό χώρο οδηγεί σε χειρότερα αποτελέσματα. Αυτό κατά πάσα πιθανότητα συμβαίνει επειδή ο Encoder στην περίπτωση που οι ταξινομητές είναι διαφορετικοί δεν έχει τον επιπλέον περιορισμό η έξοδος του να βρίσκεται στον ίδιο χώρο της εισόδου και μπορεί να μάθει καλύτερη κωδικοποίηση (πάντα χωρίς να χάνεται το περιεχόμενο της πρότασης βέβαια). Αντίστοιχα αποτελέσματα έδωσε και η εισαγωγή ενός επιπλέον σφάλματος από τον Discriminator με είσοδο την έξοδο του Encoder, κάτι που είχε σκοπό να κάνει την ενδιάμεση αναπαράσταση να βρίσκεται στον ίδιο χώρο (να είναι δηλαδή σήμα ήχου) με την τελική, χωρίς την πληροφορία συναισθήματος προφανώς.

Με αυτές τις σχεδιαστικές επιλογές το μοντέλο που προκύπτει είναι αρκετά παρόμοιο με την αρχιτεκτονική του [122] που χρησιμοποιείται για μετατροπή εικόνων. Σε αυτή την δουλειά, είχε χρησιμοποιηθεί επίσης ένας Encoder που πήγαινε τις εικόνες σε έναν χώρος ανεξάρτητο της κλάσης, και ένας Discriminator που έκανε επίσης ταξινόμηση στις κλάσεις της τελικής εικόνας. Αντίστοιχο μοντέλο για μετατροπή φωνής είναι και το [114], με την διαφορά ότι εκπαιδεύεται σε δύο στάδια.

Το μοντέλο φαίνεται στο Σχήμα 6.3.1.



Σχήμα 6.3.1: Προτεινόμενη Αρχιτεκτονική

Οι συναρτήσεις σφάλματος που χρησιμοποιούνται στην εκμάθηση όλων των μοντέλων είναι οι εξής:

$$L_{rec}(E,G) = \mathbb{E}_{x\sim\mathcal{X},c\sim\mathcal{C}}||G(E(x,c),c) - x||_1 \tag{6.3.2a}$$
$$L_{class_1}(C,E) = \mathbb{E}_{x\sim\mathcal{X},c\sim\mathcal{C}}\left[\log p_C(c|E(x))\right] \tag{6.3.2b}$$
$$L_{adv}(D) = -\mathbb{E}_{x\sim\mathcal{X}}\left[\log(D(x))\right] - \mathbb{E}_{x\sim\mathcal{X},(c,c')\sim\mathcal{C}}\left[\log(1 - D\big(G(E(x),c),c'\big))\right] \tag{6.3.2c}$$
$$L_{adv}(G) = -\mathbb{E}_{x\sim\mathcal{X},(c,c')\sim\mathcal{C}}\left[\log(D\big(G(E(x),c),c'\big))\right] \tag{6.3.2d}$$
$$L_{class_2}(D) = \mathbb{E}_{x\sim\mathcal{X},c\sim\mathcal{C}}\left[\log p_D(c|x)\right] + \mathbb{E}_{x\sim\mathcal{X},(c,c')\sim\mathcal{C}}\left[\log p_D(c|D\big(G(E(x),c),c'\big))\right] \tag{6.3.2e}$$
$$L_{class_2}(G) = \mathbb{E}_{x\sim\mathcal{X},(c,c')\sim\mathcal{C}}\left[\log p_D(c'|D\big(G(E(x),c),c'\big))\right] \tag{6.3.2f}$$

όπου είναι $\mathbb{X}$ και $\mathbb{C}$ είναι η κατανομή των δεδομένων και των labels των συναισθημάτων αντίστοιχα, C το μοντέλο του ταξινομητή συναισθήματος, D είναι ο Discriminator και E και G είναι ο Encoder και ο Decoder αντίστοιχα. Τα σφάλματα των ταξινομήσεων είναι όλα cross entropy με βάση τις πιθανότητες που προκύπτουν για κάθε κλάση από τα μοντέλα D και C, στα οποία το χρησιμοποιείται η συνάρτηση softmax στο τελευταίο επίπεδο.

Τα σφάλματα L_{adv} είναι τα σφάλματα που αντιστοιχούν στην κανονική εκπαίδευση των GANs και το L_{rec} είναι αντίστοιχο με το cycle consistency loss από το CycleGAN. Το σφάλμα L_{class_1} είναι του ενδιάμεσου ταξινομητή ενώ το σφάλμα L_{class_2} είναι το σφάλμα ταξινόμησης του Discriminator όπως και στο StarGAN.

Τα συνολικά σφάλμα που ελαχιστοποιούνται σε κάθε μοντέλο ξεχωριστά είναι τα εξής:

$$\mathcal{L}(E) = L_{rec}(E,G) - L_{class1}(C,E) \tag{6.3.3a}$$
$$\mathcal{L}(C) = L_{class1}(C,E) \tag{6.3.3b}$$
$$\mathcal{L}(D) = L_{adv}(D) + \lambda_{cls}L_{class2}(D) \tag{6.3.3c}$$
$$\mathcal{L}(G) = L_{adv}(G) + \lambda_{cls}L_{class2}(G) + \lambda_{rec}L_{rec}(E,G) \tag{6.3.3d}$$

Μπορούμε να δούμε ότι ο Encoder εκπαιδεύεται με σκοπό να αυξήσει το σφάλμα του Classifier, ο οποίος προσπαθεί να ταξινομήσει της εξόδους του Encoder στα σωστά συναισθήματα. Η διαδικασία είναι ανάλογη του Domain Adversarial Training όπως εξηγήθηκε στην Ενότητα 5.6.

Τα σφάλματα των Decoder, Discriminator, είναι τα ίδια όπως και στο κανονικό StarGAN, με τον Decoder να παίζει τον ρόλο του Generator (ο Encoder δεν εκπαιδεύεται από αυτό το σφάλμα).

Τα λ_{cls} και λ_{rec} είναι υπερπαράμετροι που καθορίζουν το βάρος που θα έχει το κάθε σφάλμα στο τελικό. Οι τιμές που χρησιμοποιήθηκαν στα πειράματα είναι 10 και για τις δύο υπερπαραμέτρους, μετά από αρχικό fine-tuning.

Το cycle consistency loss από το CycleGAN και το StarGAN έχουν αντικατασταθεί από το σφάλμα ανακατασκευής του Decoder L_{rec}. Αυτό συμβαίνει επειδή ουσιαστικά ο Generator των GAN έχει σπάσει σε 2 μοντέλα που αντιστοιχούν στον Encoder και Decoder των Autoencoders. Κατά συνέπεια, δεν χρειάζεται να κάνουμε μετατροπή σε κάποιο άλλο πεδίο για εισάγουμε ένα σφάλμα που θα εξασφαλίζει ότι κρατάμε την γλωσσολογική πληροφορία αμετάβλητη. Το σφάλμα ανακατασκευής των Autoencoders αρκεί, και τα μοντέλα G και E εκπαιδεύονται με αυτό το σφάλμα.

Παρ' όλ' αυτά, έγιναν πειράματα όπου το σφάλμα ανακατασκευής ήταν μεταξύ του $G(E(y, c'), c)$ όπου $y = G(E(x), c), c')$ όπως θα ήταν στο κανονικό CycleGAN (με την διαφορά ότι τα G, E είναι ένα μοντέλο). Αυτό οδήγησε σε πολύ μικρή διαφορά στην ποιότητα των αποτελεσμάτων προς το χειρότερο και κατά συνέπεια το σφάλμα επιλέχθηκε να είναι αυτό που φαίνεται στην εξίσωση.

Στην πράξη, το σφάλμα που χρησιμοποιήθηκε για το κομμάτι των GAN δεν είναι το κανονικό σφάλμα που φαίνεται στις εξισώσεις αλλά έγινε δοκιμή με τα σφάλματα Wasserstein-GP (WGAN-GP) που αναλύθηκε στην Ενότητα 5.2.4 και LSGAN που αναλύθηκε στην Ενότητα 5.2.2, καθώς αρχικά πειράματα δείχνουν ότι το κανονικό σφάλμα των GAN δεν έχει την απαραίτητη σταθερότητα στην εκπαίδευση. Οι συναρτήσεις σφάλματος αντίστοιχα αλλάζουν στις 5.2.4 και 5.2.5 με υπερπαράμετρο λ επίσης 10 για την περίπτωση του WGAN-GP και στις 5.2.1 για την περίπτωση του LSGAN. Επίσης, όταν εκπαιδεύουμε με το σφάλμα WGAN-GP, οι παράμετροι του Generator (που στο προτεινόμενο μοντέλο αντιστοιχεί στο ζευγάρι Encoder, Decoder) ενημερώνεται μια φορά για κάθε 5 επαναλήψεις του Discriminator και του Classifier.

Η αρχιτεκτονική του κάθε επιμέρους μοντέλου είναι ως εξής:

1. **Encoder**, **Decoder**

   Ο Encoder και ο Decoder είναι βασισμένοι πάνω στον Generator του StarGAN-VC (Σχήμα 5.7.1). Αποτελούνται από μία σειρά από δισδιάστατα συνελικτικά δίκτυα τα οποία κάνουν down-sampling το σήμα σε μικρότερες διαστάσεις (τόσο στο πεδίο του χρόνου όσο και της συχνότητας), μια σειρά από residual blocks και στην συνέχεια πάλι μια σειρά από συνελικτικά δίκτυα (de-convolutional) για να επαναφέρουν το σήμα στις αρχικές του διαστάσεις με upsampling.

   Τα downsampling συνελικτικά δίκτυα αποτελούνται από μια δισδιάστατη συνέλιξη με kernel $(4, 8)$ και stride 2 για το down-sampling, ένα επίπεδο Instance Normalization και ένα ReLU. Σε κάθε επίπεδο τα κανάλια της συνέλιξης διπλασιάζονται (ενώ οι διαστάσεις του σήματος πέφτουν στη μέση λόγω του stride). Συνολικά χρησιμοποιούνται 3 τέτοια επίπεδα.

   Τα ενδιάμεσα επίπεδα αποτελούνται 5 Residual blocks. Το κάθε ένα από αυτά τα blocks αποτελείται από την συνελικτική δομή που περιγράφηκε πάνω (συνέλιξη, Instance Normalization, ReLU) δύο φορές, συν ένα residual connection από την αρχή στην έξοδο. Για να δουλεύει το residual connection χρειάζεται τα κανάλια και το μέγεθος της εξόδου και και της εισόδου να είναι ίδια, κάτι που σημαίνει ότι οι συνελίξεις δεν αλλάζουν αριθμό καναλιών, και έχουν stride 1.

   Τα upsampling συνελικτικά δίκτυα έχουν την ίδια δομή με τα downsampling, με την διαφορά ότι είναι de-concolutional και το μέγεθος του kernel είναι $(4, 4)$. Χρησιμοποιούνται πάλι 3 τέτοια επίπεδα.

   Το label του συναισθήματος εισόδου δίνεται στην είσοδο του Encoder με one-hot αναπαράσταση η οποία γίνεται concatenate με το σήμα σαν επιπλέον κανάλια (όσα και ο αριθμός των συναισθημάτων). Αυτό δημιουργεί έναν πλεονασμό στα kernel αυτού του επιπέδου, καθώς τα σημεία του kernel που αντιστοιχούν σε αυτά τα κανάλια δεν βλέπουν καμμία διαφορά στο πεδίο του χρόνου ή της συχνότητας, αλλά έγινε η επιλογή να διατηρηθεί ίδιο με το StarGAN-VC.

   Αντίστοιχα, στον Decoder, το label της επιθυμητής κλάσης της εξόδου τοποθετείται με τον ίδιο τρόπο στην είσοδο.

   Στην έξοδο χρησιμοποιείται ένα επιπλέον συνελικτικό επίπεδο με kernel 7.

Η τελική έξοδος των μοντέλων έχει ακριβώς την ίδια διάσταση με το σήμα της εισόδου.

2. **Classifier**, **Discriminator**

   Ο ταξινομητής και ο Discriminator αποτελούνται από 5 επίπεδα δισδιάστατων συνελικτικών δικτύων με kernel 4 και στις δύο διαστάσεις, με leaky RELU με παράμετρο 0.01 σαν συνάρτηση ενεργοποίησης, και ένα επιπλέον συνελικτικό δίκτυο με kernel διάστασης 1 στο πεδίο της συχνότητας και 8 στο πεδίο του χρόνου. Τα πρώτα επίπεδα έχουν stride 2 ενώ το τελευταίο stride 1.

   Η επιλογή αυτών των τιμών έχει γίνει έτσι ώστε με το downsampling που γίνεται λόγω του stride στα πρώτα επίπεδα, να καταλήγουμε σε ένα σήμα μήκους 8 (πάντα με είσοδο μεγέθους 128) και με τον kernel διάστασης 8 στην τελευταία διάσταση να εξαφανίζεται τελείως το πεδίο του χρόνου έτσι ώστε να καταλήγουμε σε έναν μονοδιάστατο αριθμό.

   Στην περίπτωση του Discriminator, το τελευταίο επίπεδο είναι διπλό, ένα ώστε να κάνει ταξινόμηση στο πεδίο των συναισθημάτων με cross entropy (ο αριθμός των καναλιών είναι όσα και τα συναισθήματα), και ένα ώστε να προσπαθεί να ξεχωρίσει τα πραγματικά από τα ψεύτικα δείγματα.

Όλα τα μοντέλα εκπαιδεύτηκαν με batch size 32, για 200 χιλιάδες επαναλήψεις, με τον αλγόριθμο Adam [25]. Οι παράμετροι του Adam είναι $(\beta_1, \beta_2) = (0.5, 0.99)$ τόσο στο baseline όσο και στα υπόλοιπα μοντέλα. Οι ρυθμοί μάθησης είναι 0.0001 για όλα τα μοντέλα και εφαρμόζεται weight decay στις τελευταίες 100 χιλιάδες επαναλήψεις.

6.3.3 Μοντελοποίηση F0

Το μοντέλο αυτό έχει προκύψει από προσαρμογή μοντέλου για μετατροπή χροιάς ομιλητή, στο πρόβλημα της μετατροπής συναισθήματος. Είναι γνωστό όμως [17] πως η θεμελιώδης συχνότητα F0, είναι πολύ σημαντική για την αναγνώριση του συναισθήματος. Υποπτευόμαστε λοιπόν πως ένας πολύ απλός μετασχηματισμός όπως ο 5.7.2 δεν αρκεί για να προσεγγίσει όλη την περιπλοκότητα της διαφοράς του F0 μεταξύ διαφορετικών συναισθημάτων, παρόλο που φέρνει τα εύρη της θεμελιώδης συχνότητας στα σωστά σημεία για το κάθε στοχευμένο συναίσθημα. Η μοντελοποίηση του F0 με νευρωνικά δίκτυα είναι πιθανόν να δώσει καλύτερες ιδιότητες συναισθηματικής μετατροπής, από το απλό μοντέλο.

Ενώ υπάρχουν προσεγγίσεις που προβλέπουν το F0 όπως το Fastspeech 2 [47], συνήθως βασίζονται σε Auto-regressive μοντέλα, ή σε παράλληλες βάσεις δεδομένων. Στην περίπτωση της μη παράλληλης βάσης δεδομένων, δεν έχει επιχειρηθεί εν γνώση μας μοντελοποίηση του F0.

Η πρώτη προφανής ιδέα είναι να χρησιμοποιηθεί το F0 σαν ένα έξτρα χαρακτηριστικό των δεδομένων, απλά ενώνοντάς το με τα spectral χαρακτηριστικά. Οι εξισώσεις των μοντέλων γίνονται:

$$L_{rec}(E,G) = \mathbb{E}_{x,f_0\sim\mathcal{X},c\sim\mathcal{C}}||G(E(x \parallel f_0, c), c) - x \parallel f_0||_1 \tag{6.3.4a}$$

$$L_{class_1}(C,E) = \mathbb{E}_{x,f_0\sim\mathcal{X},c\sim\mathcal{C}}\left[\log p_C(c|E(x \parallel f_0))\right] \tag{6.3.4b}$$

$$L_{adv}(D) = -\mathbb{E}_{x,f_0\sim\mathcal{X}}\left[\log(D(x \parallel f_0))\right] - \mathbb{E}_{x,f_0\sim\mathcal{X},(c,c')\sim\mathcal{C}}\left[\log(1 - D\big(G(E(x \parallel f_0), c), c')\big)\right] \tag{6.3.4c}$$

$$L_{adv}(G) = -\mathbb{E}_{x,f_0\sim\mathcal{X},(c,c')\sim\mathcal{C}}\left[\log(D\big(G(E(x \parallel f_0), c), c')\big)\right] \tag{6.3.4d}$$

$$L_{class_2}(D) = \mathbb{E}_{x,f_0\sim\mathcal{X},c\sim\mathcal{C}}\left[\log p_D(c|x \parallel f_0)\right] + \mathbb{E}_{x\sim\mathcal{X},(c,c')\sim\mathcal{C}}\left[\log p_D(c|D\big(G(E(\|f_0), c), c')\big)\right] \tag{6.3.4e}$$

$$L_{class_2}(G) = \mathbb{E}_{x,f_0\sim\mathcal{X},(c,c')\sim\mathcal{C}}\left[\log p_D(c'|D\big(G(E(x \parallel f_0), c), c')\big)\right] \tag{6.3.4f}$$

όπου το σύμβολο $\parallel$ υποδηλώνει την παράθεση, και κάναμε την υπόθεση ότι το F0 προκύπτει από την κατανομή των δεδομένων ($\mathcal{X}$) ή οποία σε αυτή την περίπτωση έχει απλά μία διάσταση περισσότερο.

Το πρόβλημα με αυτού του είδους την αντιμετώπιση είναι ότι αφενός το F0 έχει τελείως διαφορετική τάξη μεγέθους από τα υπόλοιπα χαρακτηριστικά, καθώς και πολλά σημεία στα οποία είναι 0, αλλά και ότι είναι γενικά δύσκολο να μοντελοποιηθεί. Σε αρχικά πειράματα, παρατηρήθηκε ότι ο Discriminator μπορούσε πολύ εύκολα να ξεχωρίσει τα δείγματα με το "ψεύτικο" F0 και πολύ νωρίς κατά την διάρκεια της εκπαίδευσης μάθαινε να αγνοεί τα δείγματα που προκύπτουν από τον Generator. Αυτό οδηγούσε είτε σε τεράστια σφάλματα στην περίπτωση του WGAN-GP σφάλματος, είτε σε mode collapse στην περίπτωση άλλου σφάλματος. Για να αντιμετωπιστεί αυτό, ο αριθμός των εκπαιδεύσεων του Discriminator σε σχέση με τον Generator μειώθηκε από 5 σε 1.

Για την διευκόλυνση των μοντέλων, πριν περάσουμε τα δεδομένα από κάποιο μοντέλο, εξάγουμε τα frames τα οποία είναι μηδέν (άφωνοι ήχοι) και στην έξοδο του μοντέλου αναγκάζουμε τα αντίστοιχα frames να είναι μηδέν (πολλαπλασιάζοντας με 0). Αυτό σημαίνει ότι το μοντέλο δεν μπορεί να μετατρέψει έναν άφωνο ήχο σε έμφωνο, κάτι που όμως δεν επηρεάζει την απόδοση του μοντέλου, καθώς σε όλα τα μοντέλα έχει γίνει η υπόθεση ότι η διάρκεια του κάθε ήχου παραμένει σταθερή.

Παρ' όλα αυτά, εξαιτίας των προαναφερθέντων προβλημάτων (και όπως θα αναλυθεί στα αποτελέσματα στην συνέχεια) η απλή μοντελοποίηση του F0 οδηγεί σε αρκετά προβλήματα στην ποιότητα των ακουστικών σημάτων. Για την διευκόλυνση των μοντέλων, αντί να τεθεί ο στόχος του μοντέλου η εκμάθηση του F0 κατευθείαν, έγινε μια επιπλέον παραλλαγή της αρχιτεκτονικής, κατά την οποία χρησιμοποιείται η εξίσωση 6.3.1, σαν πρώτη προσέγγιση του F0 του ζητούμενου συναισθήματος, και το νευρωνικό δίκτυο μαθαίνει την διαφορά (residual) από την ζητούμενη F0.

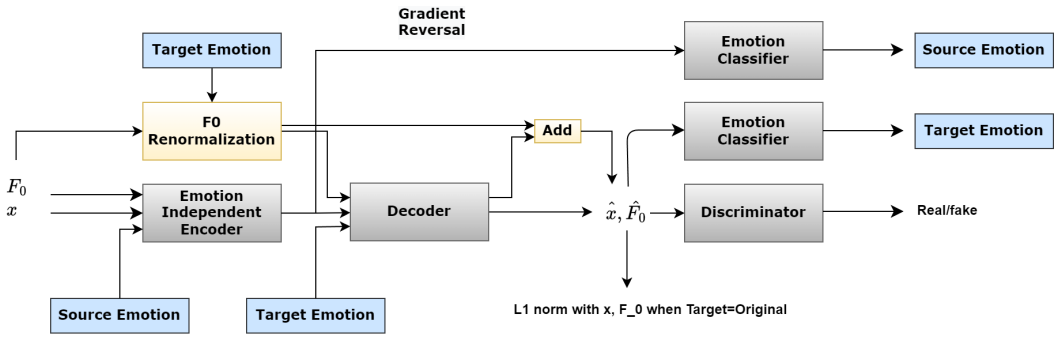
Πιο συγκεκριμένα, το μετετραμμένο F0 προκύπτει από τον εξής τύπο:

$$\hat{F}_{0,t} = G_{F_0}(F_{0,approx,t}, E(x \parallel F_0, s, s), t) + F_{0,approx,t}, \tag{6.3.5a}$$

$$F_{0,approx,t} = \exp\left(\frac{(\log(F_{0_s}) - \mu_s)}{\sigma_s} * \sigma_t + \mu_t\right) \tag{6.3.5b}$$

όπου $F_{0,approx,t}$ δηλώνει την προσεγγιστική F0 όπως προκύπτει από τον τύπο 6.3.1, s, t είναι τα (source, target) συναισθήματα εισόδου εξόδου. Το G_{F_0} υποδηλώνει ότι παίρνουμε το κομμάτι της εξόδου του G που αντιστοιχεί στο F0.

Η αρχιτεκτονική αυτή φαίνεται στο Σχήμα 6.3.2.



Σχήμα 6.3.2: Προτεινόμενη αλλαγή στην μοντελοποίηση του F0

6.4 Αξιολόγηση

Η αξιολόγηση μοντέλων σύνθεσης φωνής γίνεται με αντικειμενικές μετρικές, αλλά και χρησιμοποιώντας ακροατές οι οποίοι αξιολογούν τα επιθυμητά χαρακτηριστικά της συντεθειμένης ομιλίας με υποκειμενικό τρόπο. Χρησιμοποιούμε και τους δύο τρόπους αξιολόγησης για τα μοντέλα μας.

6.4.1 Υποκειμενική Αξιολόγηση

Για την υποκειμενική αξιολόγηση χρησιμοποιήθηκαν 25 ακροατές που αξιολόγησαν την ποιότητα των συνθετικών προτάσεων καθώς και το αντιληπτό συναίσθημα. Οι ακροατές είναι και των δύο φύλων και διαφορετικών ηλικιακών ομάδων και έχουν όλοι μητρική γλώσσα τα ελληνικά. Η υποκειμενική αξιολόγηση έγινε μόνο σε δείγματα από μοντέλα εκπαιδευμένα στο CVSP_EAV dataset καθώς περιέχει τις ελληνικές προτάσεις, και είναι και η μεγαλύτερη εκ των βάσεων δεδομένων που χρησιμοποιούμε.

Η αξιολόγηση της ποιότητας γίνεται σε κλίμακα από το 1 μέχρι το 5 όπου το 1 υποδεικνύει "όχι ομιλία" και το 5 "απόλυτα φυσική ομιλία". Αυτού του είδους αξιολόγηση αναφέρεται σαν Mean Opinion Score (MOS) (Μέση βαθμολογία γνώμης) και είναι η πλέον χρησιμοποιούμενη μετρική για την αξιολόγηση συνθετικής ομιλίας [123].

Η αξιολόγηση της ικανότητας των μοντέλων να εκτελέσουν μετατροπή του συναισθήματος της ομιλίας γίνεται απλά ρωτώντας τους ακροατές να εκτιμήσουν με ποιο συναίσθημα είναι ειπωμένη η ομιλία. Η επιλογή του συναισθήματος έγινε από μια λίστα προκαθορισμένων συναισθημάτων για απλοποίηση της διαδικασίας (θυμός, χαρά, λύπη, ουδέτερο). Αυτά τα συναισθήματα είναι τα συναισθήματα τα οποία αποτελούν την βάση δεδομένων που εκπαιδεύτηκε το μοντέλο. Στη συνέχεια, υπολογίζουμε το ποσοστό των ακροατών που επέλεξαν το συναίσθημα που δόθηκε σαν στόχος στο μοντέλο (target) σαν μετρική αξιολόγησης του συστήματος μετατροπής. Υποθέτουμε δηλαδή ότι ένα μοντέλο με υψηλή ικανότητα μετατροπής θα ωθήσει τους ακροατές να επιλέξουν το συναίσθημα που του τέθηκε σαν στόχος.

6.4.2 Αντικειμενική Αξιολόγηση

Μετρικές Αξιολόγησης

Οι μετρικές που επιλέχθηκαν για την αξιολόγηση της ικανότητας μετατροπής συναισθήματος είναι το Mel Cepstral Distortion (MCD) [124] καθώς και τα Voice Decision Error, Gross Pitch Error, F0 Frame Error για το F0 [125, 126]. Όλες οι μετρικές είναι μετρικές απόκλισης μεταξύ δύο δοσμένων ομιλιών και λειτουργούν σε επίπεδο παραθύρων.

Χρησιμοποιούμε αυτές τις μετρικές με δύο τρόπους, με σκοπό να αξιολογήσουμε την ικανότητα μετατροπής και την ικανότητα ανακατασκευής.

Η ικανότητα ανακατασκευής αξιολογείται απλά κάνοντας επανασύνθεση της ομιλίας με το ίδιο συναίσθημα σαν στόχο. Αυτό, σύμφωνα με την εκπαίδευση του μοντέλου, θα έπρεπε να οδηγεί στο ίδιο ακριβώς σήμα ήχου. Παρ' όλα αυτά, επειδή τα μοντέλα δεν είναι το ίδιο ικανά να κάνουν την ανακατασκευή, μπορούμε να αξιολογήσουμε την ποιότητα του μοντέλου χρησιμοποιώντας αυτές τις μετρικές.

Ο τρόπος που χρησιμοποιούνται αυτές οι μετρικές για την αξιολόγηση της μετατροπής είναι βρίσκοντας την απόκλιση μιας μετατρεμμένης ομιλίας σε ένα συναίσθημα με την πρόταση ειπωμένη με το ζητούμενο συναίσθημα από τον ομιλητή (πραγματική ομιλία). Αυτό μπορεί να συμβεί επειδή τα dataset που χρησιμοποιήθηκαν έχουν παράλληλα δεδομένα, ασχέτως που δεν χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων. Στην πράξη, αυτή η μετρική έχει το μειονέκτημα ότι υπάρχουν πολλοί τρόποι να ειπωθεί μια πρόταση με ένα συναίσθημα, ενώ αυτός ο τρόπος αξιολόγησης υποθέτει ότι υπάρχει μόνο ένας σωστός τρόπος. Παρ' όλα αυτά, με αρχικές αξιολογήσεις είδαμε ότι μπορεί να προσφέρει χρήσιμη πληροφορία για την ικανότητα μετατροπής του μοντέλου, όπως θα δούμε και στη συνέχεια.

Επειδή στην γενική περίπτωση τα μήκη των δύο ομιλιών διαφέρουν (καθώς το σήμα που χρησιμοποιείται ως στόχος έχει επίσης ειπωθεί από τον ομιλητή), αρχικά χρησιμοποιείται ο αλγόριθμος Dynamic Time Warping (DTW) για να ευθυγραμμίσει τις δύο ομιλίες.

Ο αλγόριθμος αυτός βρίσκει το βέλτιστο ταίριασμα μεταξύ δύο ακολουθιών ως προς μια μετρική απόστασης, η οποία συνήθως επιλέγεται να είναι η νόρμα L2 πάνω στα cepstrum χαρακτηριστικά.

Χρησιμοποιεί δυναμικό προγραμματισμό για να βρει την απόσταση με το μικρότερο δυνατό κόστος με βάση τις εξής προϋποθέσεις:

1. Η ευθυγράμμιση ξεκινάει από το πρώτο στοιχείο και των δύο προτάσεων και τελειώνει με το τελευταίο στοιχείο και των δύο προτάσεων
2. Κάθε στοιχείο της μιας ακολουθίας θα ευθυγραμμιστεί με ένα ή περισσότερα στοιχεία της άλλης και το αντίστροφο
3. Η ευθυγράμμιση είναι μονοτονική, δηλαδή σε κάθε χρονική στιγμή μπορούμε να ευθυγραμμίσουμε το επόμενο στοιχείο της μίας ακολουθίας με το προηγούμενα ευθυγραμμισμένο στοιχείο της άλλης, ή τα επόμενα στοιχεία και των δύο ακολουθιών μεταξύ τους.
4. Το συνολικό κόστος είναι το άθροισμα των αποστάσεων των ευθυγραμμισμένων σημείων.

Οι μετρικές που χρησιμοποιούμε υπολογίζονται πάνω στην ευθυγράμμιση που προκύπτει από το DTW, επαναλαμβάνοντας τα στοιχεία που αντιστοιχούν σε πάνω από ένα στοιχείο στην άλλη ακολουθία μέχρι που οι ακολουθίες να έχουν το ίδιο μήκος.

Το MCD ορίζεται ως το μέσο Root Mean Square Error των cepstral χαρακτηριστικών:

$$MCD = \frac{10\sqrt{2}}{\ln 10}\frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{i}\left(C_{ti}-\hat{C}_{ti}\right)^2} \tag{6.4.1}$$

όπου C_{ti} και $\hat{C}_{ti}$ είναι τα i-οστά cepstral χαρακτηριστικά την χρονική στιγμή t (μετά την ευθυγράμμιση) των δύο ομιλιών. Ο παράγοντας στην αρχή είναι απλά παράγοντας κανονικοποίησης.

Οι μετρικές Voice Decision Error, Gross Pitch Error, F0 Frame Error χρησιμοποιούνται για την αξιολόγηση της μετατροπής του F0 και ορίζονται ως το ποσοστό των παραθύρων στα οποία γίνεται ένα σφάλμα. Πιο συγκεκριμένα, έχουμε ότι VDE είναι:

$$VDE = \frac{N_{V\to U}+N_{U\to V}}{N}\times 100\% \tag{6.4.2}$$

όπου N είναι ο συνολικός αριθμός παραθύρων, $N_{V\to U}$ είναι ο αριθμός των παραθύρων που είναι έμφωνα στο πρώτο σήμα και άφωνα στο δεύτερο και $N_{U\to V}$ το αντίθετο. Το VDE είναι δηλαδή, το ποσοστό των παραθύρων στα οποία η συντεθειμένη ομιλία έχει έμφωνους ήχους ενώ η πραγματική ακολουθία έχει άφωνους ή το αντίστροφο.

Το GPE ορίζεται ως:

$$GPE = \frac{N_{F_0E}}{N_{VV}}\times 100\% \tag{6.4.3}$$

όπου N_{VV} είναι ο συνολικός αριθμός των έμφωνων παραθύρων:

$$|\frac{F_{0i,estimated}}{F_{0i,reference}}| - 1 > \delta\%| \tag{6.4.4}$$

όπου τα $F_{0i,estimated}$, $F_{0i,reference}$ είναι τα F0 των δύο ομιλιών. Το δ είναι ένα κατώφλι που καθορίζει την ανοχή στο λάθος και συνήθως επιλέγεται να είναι 20, τιμή που χρησιμοποιείται και σε αυτή την εργασία. Το GPE είναι δηλαδή το ποσοστό των έμφωνων παραθύρων στα οποία έχει γίνει σφάλμα στην τιμή του F0, μεγαλύτερο από 20%.

Τέλος, το FFE είναι συνδυασμός των δύο προηγούμενων μετρικών και ορίζεται απλά ως:

$$\begin{aligned} FFE &= \frac{\#\text{πλαισίων με σφάλμα}}{\#\text{συνολικών πλαισίων}} \\ &= \frac{N_{V\to U}+N_{U\to V}+N_{F_0E}}{N}\times 100\% \end{aligned} \tag{6.4.5}$$

Επειδή η μοντελοποίηση του F0 γίνεται άμεσα από τα μοντέλα μας, είτε με νευρωνικά δίκτυα είτε με τον τύπο της κανονικοποίησης, η εξαγωγή του F0 για αυτές τις μετρικές γίνεται ξανά από την κυματομορφή, (μάλιστα γίνεται με άλλο εργαλείο, χρησιμοποιώντας το Praat[18] αντί για τον WORLD Vocoder). Επειδή η εξαγωγή αυτή δεν είναι χωρίς σφάλμα και δεν είναι πλήρως αναστρέψιμη, εξηγείται πως, για παράδειγμα το VDE δεν είναι ταυτοτικά 0 αφού σε όλες τις προσεγγίσεις τα μηδενικά του F0 έχουν μείνει αμετάβλητα (φυσικά αυτό αφορά μόνο την περίπτωση της ανακατασκευής, στην περίπτωση που χρησιμοποιείται διαφορετική ομιλία του ομιλητή σαν στόχος, δεν υπάρχει κανένας λόγος να είναι 0 το VDE).

Αξιολόγηση με την βοήθεια Νευρωνικών δικτύων

Επειδή η υποκειμενική αξιολόγηση πολλών μοντέλων χρησιμοποιώντας ακροατές είναι δύσκολη καθώς απαιτεί την εύρεση αρκετών ανθρώπων να αξιολογήσουν τα δείγματα ώστε να βρεθούν στατιστικά σημαντικά αποτελέσματα, τα τελευταία χρόνια έχει γίνει μια προσπάθεια αυτοματοποίησης αυτής της αξιολόγησης από νευρωνικά δίκτυα.

Η διαδικασία εκπαίδευσης τέτοιων νευρωνικών δικτύων έχει ως εξής: Αρχικά γίνεται αξιολόγηση πολλών δειγμάτων συνθετικών ή μη, από μεγάλο αριθμό ατόμων [127, 128], με MOS. Αυτές οι αξιολογήσεις χρησιμοποιούνται ως βάση δεδομένων για να εκπαιδευτεί ένα νευρωνικό μοντέλο το οποίο προβλέπει την αξιολόγηση από το 1 μέχρι το 5 με είσοδο την ακουστική ακολουθία [129, 130].

Για την αξιολόγηση αυτής της εργασίας, χρησιμοποιήθηκε το προ-εκπαιδευμένο μοντέλο UTMOS [131] το οποίο χρησιμοποιεί πολλές διαφορετικές μεθόδους ταξινόμησης κλασικών αλγορίθμων τεχνητής νοημοσύνης (π.χ. SVM) και διαφορετικά μοντέλα νευρωνικών δικτύων τα οποία συναθροίζει στην τελική πρόβλεψη του MOS. Στα νευρωνικά δίκτυα, η πρόβλεψη γίνεται σε κάθε πλαίσιο της ακουστικής ακολουθίας χρησιμοποιώντας χαρακτηριστικά από self-supervised learning μοντέλα [132] τα οποία περνάνε από ένα LSTM. Η μέση τιμή των πλαισίων, αποτελεί την τελική πρόβλεψη του νευρωνικού δικτύου. Αντίστοιχα οι προβλέψεις των κλασικών νευρωνικών αλγορίθμων είναι πάνω σε χαρακτηριστικά όλης της πρότασης. Οι προβλέψεις των διαφορετικών μοντέλων συνδυάζονται με την βοήθεια κλασικών αλγορίθμων απόφασης για την τελική πρόβλεψη.

Ένα πρόβλημα της χρήσης προ-εκπαιδευμένων δικτύων για την αξιολόγηση των μοντέλων, είναι η διαφορά μεταξύ των δεδομένων στα οποία εκπαιδεύτηκαν και στον δεδομένων στα οποία εφαρμόζονται. Στην προκειμένη περίπτωση, υπάρχει μεγάλη διαφορά μεταξύ των δύο πεδίων, καθώς τα δεδομένα στα οποία εφαρμόζονται είναι συναισθηματικά φορτισμένη ομιλία, και στην περίπτωση του CVSP_EAV είναι σε διαφορετική γλώσσα, καθώς το UTMOS εκπαιδεύτηκε στην αγγλική γλώσσα.

Για την αξιολόγηση της ικανότητας συναισθηματικής μετατροπής, χρησιμοποιήθηκε ένας προ-εκπαιδευμένος ταξινομητής ομιλίας. Ο ταξινομητής που χρησιμοποιήθηκε είναι ο ταξινομητής συναισθημάτων από το εργαλείο "Speechbrain" [133]. Ο ταξινομητής αυτός είναι επίσης βασισμένος πάνω σε self supervised χαρακτηριστικά (wav2vec [132]). Η προεκπαίδευση αυτού του μοντέλου έγινε στο σύνολο δεδομένων IEMOCAP [5], το οποίο περιέχει διαλόγους μεταξύ ηθοποιών οι ομιλίες των οποίων ταξινομήθηκαν μετά την ηχογράφηση τους ανάλογα με το συναίσθημα τους από ακροατές. Τα δείγματα που επιλέχθηκαν για την εκπαίδευση του μοντέλου είναι αυτά που αντιστοιχούσαν στα συναισθήματα θυμός, χαρά, λύπη, ουδέτερο, στα ίδια δηλαδή με αυτά που υπάρχουν στο CVSP_EAV. Και σε αυτή την περίπτωση, για την αξιολόγηση δειγμάτων από το CVSP_EAV υπάρχει το πρόβλημα του ότι έχει εκπαιδευτεί σε αγγλική βάση δεδομένων. Αντίθετα, στην περίπτωση του TESS, υπάρχει το πρόβλημα ότι το TESS έχει περισσότερες κατηγορίες συναισθημάτων από τα δεδομένα που εκπαιδεύτηκε ο ταξινομητής. Η μετρική που επιλέχθηκε από αυτόν τον ταξινομητή είναι απλά το accuracy του ταξινομητή (ακρίβεια), θεωρώντας σαν σωστό συναίσθημα το συναίσθημα στόχου που δόθηκε στο μοντέλο. Θεωρούμε δηλαδή πως ένα μοντέλο που κάνει καλά την μετατροπή θα κάνει το μοντέλο πρόβλεψης συναισθημάτων να προβλέπει το συναίσθημα στόχου με μεγαλύτερο βαθμό, σε αντιστοιχία με τον τρόπο αξιολόγησης με ακροατές.

6.5 Αποτελέσματα

6.5.1 Αντικειμενική Αξιολόγηση

CVSP_EAV

Χρησιμοποιήσαμε 128 προτάσεις από το test set για την αξιολόγηση κάθε μοντέλου (8 για κάθε δυνατή μετατροπή). Η αξιολόγηση των μοντέλων έγινε αρχικά στο σύνολο δεδομένων CVSP_EAV και στην συνέχεια με κριτήρια που θα επεξηγηθούν στην συνέχεια, επιλέχθηκαν 3 μοντέλα τα οποία εκπαιδεύτηκαν και στο TESS, συμπεριλαμβανομένου και του baseline.

Αρχικά συγκρίνουμε το μοντέλο μας με το baseline στις αντικειμενικές μετρικές και σύμφωνα με την αξιολόγηση των νευρωνικών δικτύων. Έχουμε συμπεριλάβει το προτεινόμενο μοντέλο με δύο συναρτήσεις σφάλματος, το WGAN-GP και το LSGAN (Η απλή συνάρτηση σφάλματος των GAN είναι πολύ δύσκολο να βγάλει λογικά αποτελέσματα στα δεδομένα μας). Στον Πίνακα 6.1 φαίνονται τα αποτελέσματα, καθώς και οι τιμές της ακρίβειας της ταξινόμησης και η προβλεπόμενη τιμή του MOS από τα νευρωνικά μοντέλα για τις πραγματικές προτάσεις. Οι αντικειμενικές μετρικές αντιστοιχούν στις μετρικές μετατροπής οπως εξηγήθηκαν στο Κεφάλαιο 6.4.2 (δηλαδή έχουν σαν στόχο τις πραγματικές προτάσεις ειπωμένες με το συναίσθημα στόχου).

Αρχικά βλέπουμε ότι οι τιμές του UTMOS είναι αρκετά μικρές για όλα τα μοντέλα, ακόμα και για τις πραγματικές προτάσεις. Αυτό συμβαίνει πιθανότατα, επειδή το μοντέλο έχει εκπαιδευτεί σε βάση δεδομένων που διαφέρει αρκετά από την βάση που χρησιμοποιούμε εμείς, και επειδή είναι στην αγγλική γλώσσα αλλά και επειδή οι προτάσεις έχουν έντονο συναισθηματικό περιεχόμενο. Παρ' όλα αυτά, οι πραγματικές προτάσεις παίρνουν την μεγαλύτερη βαθμολογία και ακούγοντας τα δείγματα φαίνεται πως υπάρχει συσχέτιση μεταξύ της ποιότητας των δειγμάτων και της εξόδου του μοντέλου οπότε συμπεριλαμβάνουμε και αυτήν την μετρική.

Όσον αφορά την συνάρτηση σφάλματος, η WGAN-GP υπερισχύει σε όλες τις μετρικές την LSGAN, και σαν συνέπεια επιλέχθηκε για την εκπαίδευση των υπόλοιπων μοντέλων, και θα αναφερόμαστε στο μοντέλο proposed

	Objective Metrics				NN Based Evaluation.	
Model	MCD↓	FFE↓	VDE ↓	GPE ↓	MOS ↑	Accuracy ↑
Ground Truth	-	-	-	-	2.60	0.52
Baseline	**2.89**	**22.7**	**7.4**	**20.8**	**2.13**	0.42
Proposed WGAN-GP	3.14	23.5	7.9	21.4	2.00	**0.44**
Proposed LSGAN	5.68	55.7	31.0	38.1	1.26	0.43

Πίνακας 6.1: Αξιολόγηση της συνάρτησης σφάλματος.

WGAN-GP απλά σαν proposed ή προτεινόμενο.

Συγκρίνοντας το baseline με το προτεινόμενο μοντέλο, μπορούμε να δούμε ότι υπερισχύει σε όλες τις μετρικές εκτός από την ακρίβεια της ταξινόμησης, στην οποία είναι καλύτερο το προτεινόμενο αν και με μικρή διαφορά. Σε γενικές γραμμές όμως οι διαφορές είναι αρκετά μικρές και όπως θα φανεί στην υποκειμενική αξιολόγηση μάλλον δεν αντιστοιχούν σε αντιληπτή διαφορά ποιότητας.

Στην συνέχεια αξιολογούμε την παραλλαγή του προτεινόμενου μοντέλου, που χρησιμοποιεί και το F0 σαν χαρακτηριστικό κατασκευής. Συγκρίνουμε 3 τρόπους αξιοποίησης του F0: 1) απευθείας σαν χαρακτηριστικό των δεδομένων σε γραμμική κλίμακα, 2) χρησιμοποιώντας την residual σύνδεση όπως εξηγήθηκε στο Κεφάλαιο 6.3.3 και 3) χρησιμοποιώντας την residual σύνδεση αλλά με το F0 να έχει μετατραπεί πρώτα σε λογαριθμική κλίμακα. Αυτή η τελευταία δοκιμή, έγινε για να μην είναι το χαρακτηριστικό του F0 σε άλλη τάξη μεγέθους από τα υπόλοιπα χαρακτηριστικά των δεδομένων, καθώς και επειδή σε γενικές γραμμές οι άνθρωποι αντιλαμβάνονται την συχνότητα σε λογαριθμική κλίμακα. Τα αποτελέσματα φαίνονται στον Πίνακα 6.2.

	Objective Metrics				NN Based Evaluation	
Model	MCD ↓	FFE ↓	VDE ↓	GPE ↓	MOS ↑	Accuracy ↑
Ground Truth	-	-	-	-	2.60	52.9 %
Baseline	**2.89**	**22.7**	**7.4**	**20.8**	**2.13**	42.0 %
Proposed	3.14	23.5	7.9	21.4	2.00	44.0 %
Proposed with F0	3.55	43.1	8.9	47.4	1.85	35.0 %
Proposed with F0 Residual	3.25	25.9	8.1	25.3	1.77	**51.2** %
Proposed with log F0 Residual	3.74	31.3	10.69	31.34	1.82	34.0 %

Πίνακας 6.2: Αξιολόγηση της μοντελοποίησης του F0.

Μία ενδιαφέρουσα παρατήρηση είναι ότι οι μετρικές που αφορούν το F0 (FFE, VDE, GPE) έχουν σημαντικά χειρότερες τιμές σε σχέση με το baseline. Αυτό δείχνει ότι η μοντελοποίηση του F0 είναι αρκετά δύσκολο πρόβλημα. Επειδή η τροχιά του F0 μεταβάλλεται σημαντικά σε βάθος χρόνου είναι πιθανό να μην αρκεί το receptive field των μοντέλων μας. Επίσης, σε κάποια βαθμό, επειδή η συναισθηματική ομιλία δεν είναι μονοσήμαντο πρόβλημα, (υπάρχουν αρκετοί "έγκυροι" τρόποι να ειπωθεί μια πρόταση με δεδομένο συναίσθημα) η διαφορά των τιμών των F0 μετρικών μπορεί και να οφείλεται στο ότι το μοντέλο κάνει σύνθεση την πρόταση με έναν διαφορετικό αλλά εξίσου έγκυρο τρόπο.

Το μόνο μοντέλο που έχει σχετικά κοντινές τιμές με το baseline στις αντικειμενικές μετρικές είναι αυτό με το residual F0, κάτι που δικαιολογεί την χρήση residual ενώσεων αντί για την μοντελοποίηση το F0 κατευθείαν. Επίσης, βλέπουμε ότι το ίδιο μοντέλο έχει με διαφορά την καλύτερη ταξινόμηση συναισθημάτων, η οποία μάλιστα είναι σχεδόν ίδια με των πραγματικών προτάσεων. Το γεγονός ότι αυτό το μοντέλο υπερτερεί του μοντέλου με το residual connection σε λογαριθμική κλίμακα είναι λίγο απροσδόκητο.

Όσο αφορά το προβλεπόμενο MOS, βλέπουμε ότι το baseline έχει πολύ καλύτερο αποτέλεσμα από όλα τα υπόλοιπα μοντέλα. Μάλιστα, βλέπουμε ότι το μοντέλο με την καλύτερη απόδοση στις υπόλοιπες μετρικές έχει το χειρότερο σκορ. Δεν είναι προφανές πως να αξιολογήσουμε την σημασία των διαφορετικών μετρικών, αλλά δεδομένου ότι η διαφορά είναι μικρή και το μοντέλο με το residual F0 πετυχαίνει την καλύτερη ακρίβεια στην ταξινόμηση, αποφασίστηκε να επιλεχθεί αυτό το μοντέλο ως το πιο υποσχόμενο τις κατηγορίας του.

Αξίζει να δούμε τις ίδιες μετρικές, αλλά στην περίπτωση της ανακατασκευής αυτή την φορά, δηλαδή με στόχο

Objective Reconstruction Metrics				
Model	MCD ↓	FFE ↓	VDE ↓	GPE↓
Baseline	**1.80**	8.80	4.10	6.70
Proposed	1.84	7.65	4.60	4.20
Proposed with F0	1.97	6.90	4.38	3.52
Proposed with F0 residual	1.86	**5.95**	**4.05**	**2.75**

Πίνακας 6.3: Αποτελέσματα αντικειμενικής αξιολόγησης, ανακατασκευής.

το ίδιο το συναίσθημα της πρότασης, κάτι που σύμφωνα με τις υποθέσεις μας θα έπρεπε να οδηγήσει στην ίδια ακριβώς πρόταση (δεδομένου ότι το μοντέλο μας μπορεί να θεωρηθεί και σαν Autoencoder). Τα αποτελέσματα είναι στον Πίνακα 6.3.

Προφανώς τα νούμερα είναι για όλα τα μοντέλα αρκετά μικρότερα, κάτι που είναι αναμενόμενο καθώς η ανακατασκευή είναι αρκετά πιο εύκολη από την μετατροπή, καθώς και επειδή η ανακατασκευή είναι άμεσος στόχος βελτιστοποίησης στις συναρτήσεις σφάλματος. Επίσης, παρατηρούμε ότι τα μοντέλα που μοντελοποιούν το F0 έχουν γενικά, αρκετά καλύτερες μετρικές που αφορούν το F0 από τα υπόλοιπα, κάτι που δείχνει να ενισχύει την υπόθεση ότι η χαμηλή ποιότητα στον προηγούμενο πίνακα οφείλεται στο ότι το πρόβλημα δεν είναι μονοσήμαντο, παρά σε κακή μοντελοποίηση του F0. Βέβαια, η μετρική του MCD είναι πάλι χειρότερη στα μοντέλα με το F0 οπότε, πάλι αναμένουμε ότι η ποιότητα θα είναι λίγο χειρότερη. Η παρατήρηση ότι το μοντέλο με το residual F0 έχει καλύτερες μετρικές από τη απλή παραγωγή του F0 φαίνεται πως συνεχίζει να ισχύει.

TESS

Σύμφωνα με τα αποτελέσματα της αξιολόγησης των μοντέλων στο CVSP_EAV, τα μοντέλα που επιλέχθηκαν για το TESS, είναι το προτεινόμενο και το μοντέλο με το residual F0. Η εκπαίδευση των μοντέλων έγινε με τον ίδιο ακριβώς τρόπο και δεδομένου ότι οι ομιλητές ήταν μόνο 2, δεν προστέθηκε κάποιο speaker label ή embedding. Ουσιαστικά είναι σαν να θεωρείται η ταυτότητα του ομιλητή σαν αμετάβλητο χαρακτηριστικό της ομιλίας, ανεξάρτητο από το συναίσθημα.

Η αξιολόγηση έγινε επίσης με τον ίδιο τρόπο, με την διαφορά ότι τα συναισθήματα σε αυτήν την περίπτωση είναι 7 και κατά συνέπεια, δεν μπορούμε να χρησιμοποιήσουμε το μοντέλο UTMOS όπως είναι. Για να αποφύγουμε αυτό το θέμα, απλά χρησιμοποιήσαμε σαν στόχους τα 4 συναισθήματα στα οποία έχει εκπαιδευτεί το UTMOS, τα οποία είναι υποσύνολο των συναισθημάτων του TESS. Σαν συναίσθημα εισόδου, χρησιμοποιήσαμε οποιοδήποτε συναίσθημα από το TESS, καθώς αυτό θεωρητικά δεν θα έπρεπε να επηρεάζει το μοντέλο. Τα αποτελέσματα όλων των μετρικών φαίνονται στον Πίνακα 6.4.

	Objective Metrics				NN Based Evaluation	
Model	MCD ↓	FFE ↓	VDE ↓	GPE ↓	MOS ↑	Accuracy ↑
Ground Truth	-	-	-	-	2.43	56.3 %
Baseline	**3.59**	**34.3**	**4.33**	**34.9**	**1.84**	26.3 %
Proposed	4.42	39.1	6.89	38.0	1.75	**33.6** %
Proposed with F0 residual	10.6	52.9	18.7	46.1	1.27	31.6 %

Πίνακας 6.4: Αντικειμενική αξιολόγηση στο σύνολο δεδομένων TESS.

Αρχικά βλέπουμε ότι παρόλο που η βάση των δεδομένων είναι στα αγγλικά, οι πραγματικές προτάσεις παίρνουν περίπου την ίδια βαθμολογία από το UTMOS, και περίπου την ίδια ακρίβεια στην ταξινόμηση από το speechbrain το οποίο είναι λίγο απροσδόκητο. Σε κάποιο βαθμό αυτό μπορεί να οφείλεται στην χαμηλότερη ποιότητα των ηχογραφήσεων.

Επίσης, βλέπουμε γενικά ότι οι μετρικές για όλα τα μοντέλα είναι αρκετά χειρότερες. Αυτό είναι αναμενόμενο, καθώς η βάση δεδομένων είναι σημαντικά μικρότερη, με λιγότερη ποικιλία περιεχομένου και με δυο ομιλητές. Παρ' όλα αυτά, βλέπουμε ότι το baseline έχει καλύτερες μετρικές από το προτεινόμενο για όλες τις μετρικές εκτός από την ακρίβεια της ταξινόμησης. Στην περίπτωση της ακρίβειας της ταξινόμησης βέβαια, βλέπουμε πως το baseline έχει σχεδόν όση ακρίβεια θα περίμενε κανείς από έναν τυχαίο ταξινομητή για τέσσερις κλάσεις.

Το μοντέλο με το F0 φαίνεται να αποτυγχάνει τελείως σε αυτό το σύνολο δεδομένων, με τεράστιες διαφορές σε όλες τις μετρικές εκτός από την ταξινόμηση που έχει λίγο καλύτερο αποτέλεσμα από τον τυχαίο ταξινομητή. Ένας πιθανός λόγος για αυτό το φαινόμενο είναι ότι στην μοντελοποίηση του F0, δεν υπολογίστηκε καθόλου ο ρόλος του ομιλητή. Αυτό όμως είναι ανεπαρκές, καθώς υπάρχουν αρκετές διαφορές μεταξύ των διαφορετικών F0 διαφορετικών ομιλητών.

Αξίζει εδώ να σημειωθεί πως μπορούμε να πετύχουμε σχεδόν την ίδια ακρίβεια της ταξινόμησης με το baseline, απλά κάνοντας αντιγραφή της εισόδου στην έξοδο. Αυτό θα έδινε πολύ καλά νούμερα στις υπόλοιπες μετρικές αφού θα ήταν πραγματική ομιλία, αλλά φυσικά δεν θα αποτελούσε μοντέλο μετατροπής συναισθηματικής ομιλίας (στην πράξη, σε αυτό το επιχείρημα πρέπει να αναλογιστούμε και το σφάλμα του ταξινομητή στον υπολογισμό της ακρίβειας, οπότε μάλλον το baseline κάνει καλύτερη μετατροπή από το μοντέλο που απλά αντιγράφει την είσοδο).

Μπορούμε να πάρουμε μια πρόχειρη αξιολόγηση για το κατά πόσο ισχύει αυτή η παρατήρηση κοιτώντας το πόσο συχνά ο ταξινομητής ταξινομεί την πρόταση στο συναίσθημα εισόδου αντί για εξόδου (Πίνακας 6.5).

Accuracy of TESS data with regards to source emotion	
Model	Classification Accuracy
Baseline	34.8%
Proposed	25.2%
Proposed with residual F0	31.5%

Πίνακας 6.5: Ποσοστό του TESS που ταξινομήθηκε ως το συναίσθημα εισόδου από το μοντέλο speechbrain.

Παρατηρούμε ότι σε αυτήν την περίπτωση, το baseline έχει αρκετά καλύτερο ποσοστό σωστής ταξινόμησης. Αντίθετα, το προτεινόμενο μοντέλο δίνει τελείως τυχαίο αποτέλεσμα σε αυτήν την περίπτωση. Συμπεραίνουμε λοιπόν ότι το συναίσθημα εισόδου επηρεάζει το baseline μοντέλο σε μεγαλύτερο βαθμό απ ότι το προτεινόμενο. Αυτό δικαιολογεί την προσθήκη του adversarial classifier, καθώς φαίνεται πως το συναίσθημα εισόδου φαίνεται να εξαφανίζεται από την αναπαράσταση του Encoder, όπως ήταν το ζητούμενο.

6.5.2 Υποκειμενική Αξιολόγηση

Για την υποκειμενική αξιολόγηση χρησιμοποιήθηκαν 64 δείγματα από κάθε μοντέλο προς αξιολόγηση. Τα δείγματα είναι όλα από το test set της βάσης δεδομένων, που σημαίνει ότι το μοντέλο δεν τα έχει δει κατά την εκπαίδευση, και επιλέχθηκαν ως εξής: έγινε μια τυχαία επιλογή από το test set 4 προτάσεων από το κάθε συναίσθημα, οι οποίες μετατράπηκαν και στα 4 πιθανά συναισθήματα (συμπεριλαμβανομένου και του αρχικού). Για τα πραγματικά δείγματα, απλά επιλέχθηκαν 16 τυχαίες προτάσεις από κάθε συναίσθημα. Κάθε ακροατής άκουσε ένα τυχαίο δείγμα από 32 προτάσεις, το οποίο όμως περιείχε υποχρεωτικά 8 δείγματα από κάθε μοντέλο (και 8 από τις πραγματικές προτάσεις).

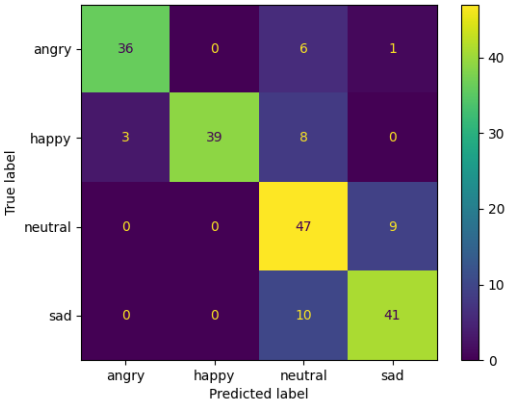
Η αξιολόγηση, με βάση τα αποτελέσματα της αντικειμενικής αξιολόγησης έγινε (εκτός από τις πραγματικές προτάσεις) σε 3 μοντέλα, στο baseline, στο προτεινόμενο μοντέλο, και στο προτεινόμενο μοντέλο με την residual σύνδεση στο F0.

Παρακάτω, στον Πίνακα 6.6, βλέπουμε τα αποτελέσματα της υποκειμενικής αξιολόγησης. Παρατίθενται οι μέσες τιμές των αξιολογήσεων κάθε μοντέλου, καθώς και τα διαστήματα εμπιστοσύνης 95% για την περίπτωση της ποιότητας.

MOS and Rater Accuracy		
Model	MOS ↑	Rater Accuracy ↑
Ground Truth	4.48 ±0.09	81.5%
Baseline	**3.84 ±0.13**	38.5%
Proposed	**3.83 ±0.14**	45.0%
Proposed with residual F0	2.99 ±0.18	**59.0%**

Πίνακας 6.6: Αποτελέσματα υποκειμενικής αξιολόγησης.

Αρχικά βλέπουμε ότι, όπως είναι αναμενόμενο, οι αληθινές προτάσεις παίρνουν σημαντικά μεγαλύτερη βαθμολογία και στην ποιότητα αλλά και στην ταξινόμηση του συναισθήματος. Η τιμή του 81% για την κατηγοριοποίηση του συναισθήματος, είναι σχετικά λογική καθώς η υποκειμενικότητα στην έκφραση κάθε συναισθήματος αποτρέπει υψηλότερες τιμές. Επίσης, όπως μπορεί να φανεί κοιτώντας τον πίνακα σύγχυσης (confusion matrix) για την ταξινόμηση των πραγματικών προτάσεων (Πίνακας 6.5.1), οι περισσότερες "λάθος" ταξινομήσεις έγιναν προβλέποντας το ουδέτερο συναίσθημα (neutral) σε προτάσεις στις οποίες είχε δοθεί ένα διαφορετικό συναίσθημα σαν οδηγία στην ηθοποιό. Αυτό είναι πιο λογικό, καθώς προτάσεις ειπωμένες με ένα συναίσθημα αλλά σχετικά μικρή ένταση είναι πιθανόν να ακουστούν σαν ουδέτερες.



Σχήμα 6.5.1: Πίνακας σύγχυσης για την ταξινόμηση των πραγματικών προτάσεων.

Στην συνέχεια, βλέπουμε ότι ενώ η ποιότητα του baseline και του προτεινόμενου μοντέλου είναι πολύ κοντά μεταξύ τους (μέσα στο ίδιο διάστημα εμπιστοσύνης), οι ακροατές ταξινόμησαν τις προτάσεις σαν το ζητούμενο συναίσθημα με αρκετά μεγαλύτερη συχνότητα στο προτεινόμενο μοντέλο σε σχέση με το baseline. Ισχυριζόμαστε με βάση αυτό το αποτέλεσμα ότι η προτεινόμενη μέθοδος εκπαίδευσης του μοντέλου βοηθάει στην ικανότητα μετατροπής του συναισθήματος, χωρίς επιδείνωση της ποιότητας.

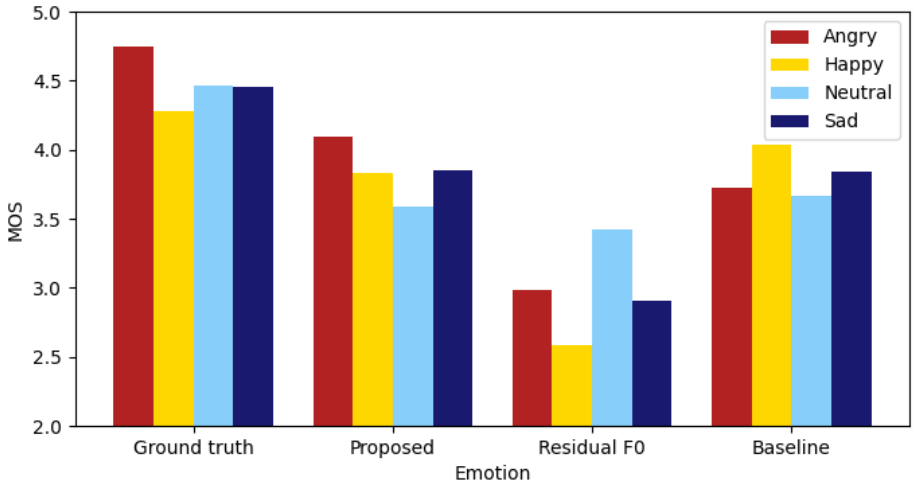
Τέλος, παρατηρούμε ότι η ποιότητα του προτεινόμενου μοντέλου με την προσθήκη του F0 πέφτει σημαντικά. Ταυτόχρονα όμως, η ικανότητα μετατροπής του συναισθήματος αυξάνεται επίσης σημαντικά. Αυτό μπορεί να εξηγηθεί αν αναλογιστούμε πως το F0 είναι αρκετά σημαντικό για την αναγνώριση συναισθήματος, αλλά είναι αρκετά δύσκολο να μοντελοποιηθεί με την βοήθεια νευρωνικών δικτύων (με μη αναδρομικά μοντέλα όπως τα συνελικτικά που χρησιμοποιούμε σε αυτή την εργασία). Επίσης, αυτό το αποτέλεσμα υποδεικνύει ότι η μετατροπή του F0 είναι μάλλον απαραίτητη για το πρόβλημα της μετατροπής συναισθήματος, και μάλλον υπάρχει κάποιο όριο στο πόσο αποτελεσματική μπορεί να είναι μια μέθοδος που βασίζεται σε απλοϊκές μετατροπές F0.

Ένας πιθανός λόγος που η μοντελοποίηση του F0 είναι δύσκολη να γίνει με τα συνελικτικά μοντέλα αυτής της εργασίας είναι ότι το receptive field όλων των μοντέλων είναι αρκετά μικρό και κατά συνέπεια είναι δύσκολο να συνθέσει τροχιές του F0 που να έχουν τις σωστές χρονικές εξαρτήσεις.

Συγκρίνοντας τις αντικειμενικές μετρικές με την υποκειμενική αξιολόγηση, μπορούμε να δούμε ότι σε γενικές γραμμές τα αποτελέσματα είναι συμβατά κάτι που δικαιολογεί σε κάποιο βαθμό την επιλογή των μετρικών αξιολόγησης. Το προτεινόμενο μοντέλο έχει όντως καλύτερη ικανότητα μετατροπής από το baseline, και το μοντέλο με το F0 ακόμα καλύτερη, αλλά με σημαντικό κόστος στην ποιότητα των προτάσεων. Τα μόνα συμπεράσματα τα οποία δεν φαίνεται να επαληθεύονται είναι ότι το μοντέλο με το F0 φαίνεται να δημιουργεί δείγματα τα οποία μοιάζουν με το ζητούμενο συναίσθημα όσο και τα αληθινά, όπως φαίνεται στον Πίνακα 6.2, και ότι γενικά το προτεινόμενο μοντέλο έχει σχετικά χαμηλότερη ποιότητα από το baseline (Πίνακας 6.1).

Στη συνέχεια, μπορούμε να παρατηρήσουμε την διαφορά ποιότητας κάθε μοντέλου ανάλογα με το συναίσθημα στόχου. Αυτό φαίνεται στο Σχήμα 6.5.2, όπου για κάθε μοντέλο και κάθε συναίσθημα, φαίνεται η μέση αξιολόγηση των ακροατών.

Αρχικά, παρατηρούμε ότι τα μοντέλα διαφέρουν στην διαφορά ποιότητας μεταξύ των διαφορετικών συναισθημάτων στόχου. Πιο συγκεκριμένα, τόσο το baseline μοντέλο όσο και το προτεινόμενο, παράγουν γενικά δείγματα



Σχήμα 6.5.2: MOS για κάθε ξεχωριστό συναίσθημα σαν στόχο, για τα μοντέλα, ή πραγματικό συναίσθημα για τις πραγματικές προτάσεις.

σταθερής ποιότητας, ανεξαρτήτως συναισθήματος στόχου (το baseline ίσως περισσότερο από το προτεινόμενο). Αντίθετα, με την προσθήκη του F0, παρατηρούμε ότι υπάρχει διαφορά σχεδόν ενός βαθμού από το χειρότερο συναίσθημα (χαρά) στο καλύτερο (ουδέτερο). Αυτό μπορεί μάλλον να εξηγηθεί από τις έντονες μεταβολές που υπάρχουν στο F0 στις χαρούμενες προτάσεις, οι οποίες είναι δύσκολο να μοντελοποιηθούν σωστά. Αντίθετα, οι προτάσεις με το ουδέτερο συναίσθημα, δεν έχουν ιδιαίτερα μεγάλες διακυμάνσεις στο F0 και είναι πολύ πιο εύκολες να μοντελοποιηθούν, με αποτέλεσμα το μοντέλο να έχει βαθμό σχεδόν όσο καλό όσο και τα άλλα μοντέλα.

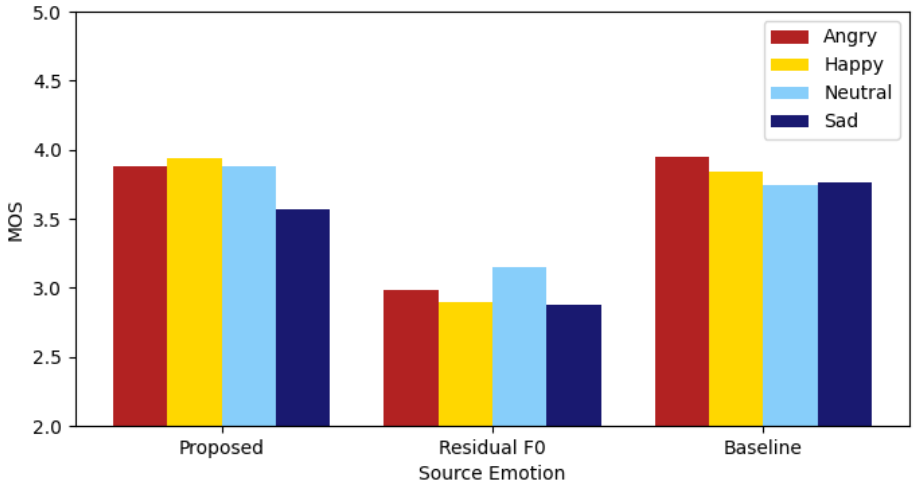
Επίσης, δεν φαίνεται κάποια γενική σχέση ποιότητας-συναισθήματος, που να ισχύει ανεξαρτήτως μοντέλου. Τα μοντέλα έχουν την καλύτερη απόδοση σε διαφορετικά συναισθήματα μεταξύ τους (baseline στη χαρά, προτεινόμενο στον θυμό και το μοντέλο με το F0 στο ουδέτερο συναίσθημα), και δεν φαίνεται κάποια προφανής εξήγηση για αυτό, πέρα από το σχόλιο της προηγούμενης παραγράφου για το F0. Στην περίπτωση των πραγματικών προτάσεων, οι διακυμάνσεις είναι αρκετά μικρές, με εξαίρεση τις προτάσεις που αντιστοιχούν στον θυμό που για κάποιον λόγο έχουν αρκετά μεγαλύτερη ποιότητα (πιθανώς λόγω στατιστικού σφάλματος).

Αντίστοιχα, η διαφορά μεταξύ της ποιότητας των προτάσεων ανάλογα με το συναίσθημα της εισόδου, φαίνεται να είναι αρκετά μικρότερη, σε όλα τα μοντέλα, όπως φαίνεται στο Σχήμα 6.5.3. Μοναδικές αξιόλογες παρατηρήσεις είναι ότι το προτεινόμενο μοντέλο οδηγεί σε χειρότερη ποιότητα με είσοδο προτάσεις λύπης και το μοντέλο με το F0 οδηγεί σε καλύτερη ποιότητα με είσοδο ουδέτερες προτάσεις.

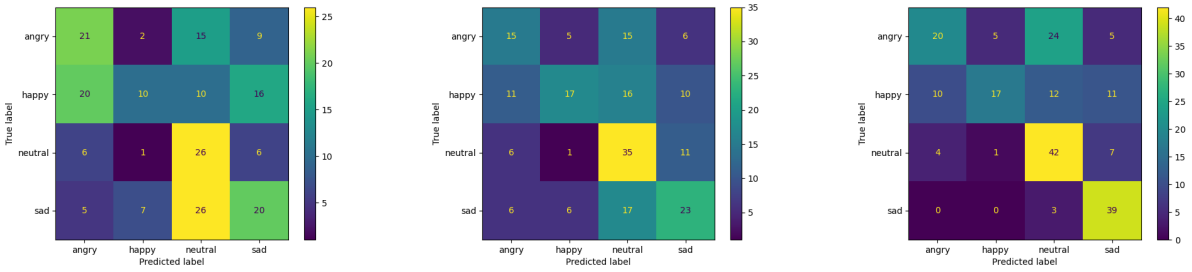
Στην συνέχεια αξιολογούμε την ικανότητα μετατροπής του κάθε μοντέλου, ανάλογα με το συναίσθημα της εισόδου και της εξόδου. Στο Σχήμα 6.5.4, βλέπουμε τον πίνακα σύγχυσης της ταξινόμησης των δειγμάτων του κάθε μοντέλου. Και στα 3 μοντέλα, υπάρχει μια τάση οι ακροατές να προβλέπουν το ουδέτερο συναίσθημα με μεγαλύτερο βαθμό, ανεξαρτήτως συναισθήματος στόχου. Αυτό είναι αναμενόμενο, και δεδομένης της ίδιας παρατήρησης για τα πραγματικά δεδομένα (Πίνακας 6.5.1) αλλά και αν σκεφτούμε ότι στις περιπτώσεις όπου το δείγμα δεν είχε κάποιο προφανές συναίσθημα, η επιλογή του ουδέτερου συναισθήματος είναι η πιο λογική.

Αντίστοιχα, και στα 3 μοντέλα οι ακροατές τείνουν να μην επιλέγουν το χαρούμενο συναίσθημα, ούτε και για "χαρούμενα" δείγματα. Στην περίπτωση του baseline, μάλιστα, επιλέγουν το χαρούμενο συναίσθημα, αρκετά λιγότερο συχνά απ ότι θα περίμενε κανείς αν επέλεγαν τυχαία, κάτι που δεν συμβαίνει με τα προτεινόμενα μοντέλα. Το χαρούμενο συναίσθημα είναι επίσης το συναίσθημα που ταξινομούν συστηματικά περισσότερο λάθος, είναι δηλαδή το συναίσθημα που είναι πιο δύσκολο να μοντελοποιηθεί από τα μοντέλα. Το πόσο αυτή η παρατήρηση είναι ιδιότητα του συναισθήματος ή της συγκεκριμένης βάσης δεδομένων είναι μια δύσκολη ερώτηση να απαντηθεί.

Μια επιπλέον παρατήρηση από τον Πίνακα 6.5.4c, είναι ότι εκτός από τα ουδέτερα συναισθήματα, το μοντέλο



Σχήμα 6.5.3: MOS για κάθε ξεχωριστό συναίσθημα σαν είσοδο.



(a) Baseline μοντέλο. (b) Προτεινόμενο μοντέλο. (c) Προτεινόμενο μοντέλο με F0.

Σχήμα 6.5.4: Πίνακας σύγχυσης για την ταξινόμηση των προτάσεων κάθε μοντέλου.

με το residual F0 έχει πολύ καλή δυνατότητα μετατροπής και στα λυπημένα συναισθήματα. Αυτό ισχύει παρά του ότι τα δείγματα αυτά έχουν σχετικά χαμηλή αξιολόγηση ποιότητας.

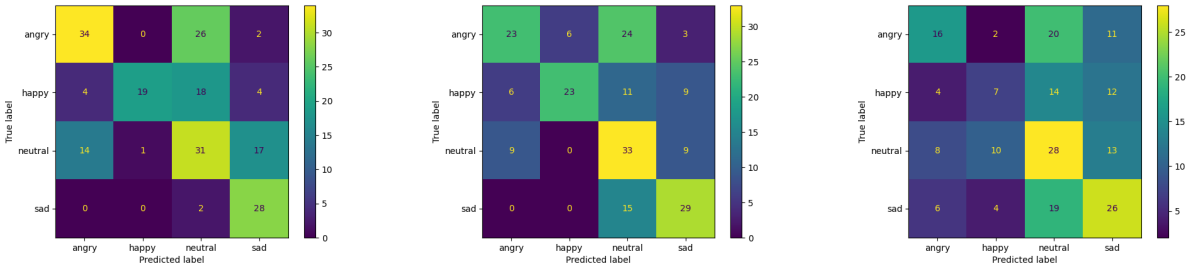
Τέλος, αξίζει να αξιολογήσουμε το βαθμό στον οποίο το συναίσθημα της πρότασης εισόδου του μοντέλου, επηρεάζει την ταξινόμηση των ακροατών (ουσιαστικά πόσο συναίσθημα από την είσοδο "περνάει" στην έξοδο). Αυτό μπορούμε να το δούμε άμεσα από τον βαθμό στον οποίο οι ακροατές επιλέξανε το συναίσθημα εισόδου, αντί για το ζητούμενο συναίσθημα στόχου, που φαίνεται στον Πίνακα 6.7, ο οποίος είναι αντίστοιχος του 6.5 αλλά για την ταξινόμηση των ακροατών.

Παρατηρούμε, ότι και το baseline και το προτεινόμενο μοντέλο διατηρούν το συναίσθημα εισόδου πολύ πιο συχνά απ ότι θα ήταν αναμενόμενο από την τύχη και μάλιστα και στις δύο περιπτώσεις, η έξοδος του μοντέλου ταξινομείται πιο συχνά σαν το συναίσθημα της εισόδου παρά το συναίσθημα στόχου (Παρατήρηση: στο ένα τέταρτο των δειγμάτων το συναίσθημα εισόδου είναι το ίδιο με το συναίσθημα εξόδου οπότε το άθροισμα των δυο αριθμών δεν είναι απαραίτητα μικρότερο από 100). Στην περίπτωση του προτεινόμενου μοντέλου τα νούμερα είναι κοντά. Αντίθετα, το συναίσθημα εισόδου φαίνεται να έχει πολύ μικρότερη επιρροή στο μοντέλο με το residual F0, αν και πάλι περισσότερο από το τυχαίο.

Τα προηγούμενα αποτελέσματα φαίνονται και αν δούμε ξανά τους πίνακες σύγχυσης των μοντέλων, αλλά θεωρώντας αυτή την φορά το συναίσθημα εισόδου σαν στόχο της ταξινόμησης (Σχήμα 6.5.5). Κοιτάμε δηλαδή πόσο συχνά οι ακροατές ταξινόμησαν το ηχητικό στο συναίσθημα εισόδου αντί για εξόδου, για κάθε διαφορετικό συναίσθημα. Ενώ ο ιδανικός πίνακας του 6.5.4 θα ήταν μια διαγώνιος (οι ακροατές διαλέγουν πάντα το συναίσθημα στόχου), ο ιδανικός πίνακας σε αυτή την περίπτωση θα ήταν ομοιόμορφος (οι ακροατές διαλέγουν πάλι το συναίσθημα στόχου, χωρίς καμμία επιρροή από το συναίσθημα εισόδου).

Accuracy of CVSP_EAV data with regards to source emotion	
Model	Rater Accuracy
Baseline	56.0%
Proposed	54.0%
Proposed with residual F0	38.5%

Πίνακας 6.7: Ποσοστό που ταξινομήθηκε σαν το συναίσθημα εισόδου.



(a) Baseline μοντέλο. (b) Προτεινόμενο μοντέλο. (c) Προτεινόμενο μοντέλο με F0.

Σχήμα 6.5.5: Πίνακας σύγχυσης για την ταξινόμηση των προτάσεων κάθε μοντέλου, ως προς την είσοδο του μοντέλου.

Στις περιπτώσεις του baseline και του προτεινόμενου μοντέλου χωρίς το F0, βλέπουμε ότι σε γενικές γραμμές ο πίνακας είναι πιο κοντά σε διαγώνιος απ ότι στο προηγούμενο σχήμα, το οποίο επιβεβαιώνει την προηγούμενη παρατήρηση.

Κεφάλαιο 7

Επίλογος

7.1 Σύνοψη και Συμπεράσματα

Στην παρούσα διπλωματική εργασία μελετήθηκε το πρόβλημα της μετατροπής συναισθηματικής ομιλίας με την χρήση νευρωνικών δικτύων. Αξιοποιήθηκαν μοντέλα που χρησιμοποιούνται στην μετατροπή ομιλητή και στην συνέχεια προτάθηκαν κάποιες επιπλέον αλλαγές στην αρχιτεκτονική και στον τρόπο εκπαίδευσης των δικτύων.

Αρχικά προτάθηκε το σπάσιμο του Generator σε δύο δίκτυα Encoder και Decoder ο πρώτος εκ των οποίων, με την βοήθεια της adversarial εκπαίδευσης, μετατρέπει την ομιλία σε ένα χώρο ο οποίος δεν περιέχει πληροφορία για το συναίσθημα, αλλά διατηρεί όλη την γλωσσολογική πληροφορία και γενικότερα όλη την πληροφορία που είναι ανεξάρτητη με το συναίσθημα εκφοράς και ο δεύτερος κάνει την μετατροπή στο ζητούμενο συναίσθημα. Με βάση την υποκειμενική και αντικειμενική αξιολόγηση που έγινε, καταλήγουμε στο ότι αυτή η διαδικασία εκπαίδευσης έχει μικρή επιρροή στην ποιότητα του παραγόμενου σήματος, αλλά οδηγεί σε βελτίωση στην ικανότητα μετατροπής του συναισθήματος. Επίσης, αυτή η αρχιτεκτονική φαίνεται πως βοηθάει στο να είναι το παραγόμενο σήμα ανεξάρτητο από το συναίσθημα εισόδου, καθώς ο Encoder βοηθάει στο να "ξεχαστεί" αυτή η πληροφορία.

Η επόμενη αλλαγή που προτάθηκε είναι στην μοντελοποίηση του F0. Είδαμε ότι η μοντελοποίηση του F0 είναι γενικά μια αρκετά δύσκολη διαδικασία, η οποία όμως φαίνεται να είναι απαραίτητη για την σωστή μετατροπή του συναισθήματος. Αυτό προκύπτει επειδή η ποιότητα των μοντέλων που χρησιμοποιούν το F0 είναι συστηματικά χειρότερη, αλλά με την χρήση residual δικτύων με είσοδο το F0 επανα-κανονικοποιημένο στο συναίσθημα στόχου, γίνεται σημαντικά μεγαλύτερη βελτίωση στην μετατροπή του συναισθήματος.

7.2 Μελλοντικές Επεκτάσεις

Αρκετές επεκτάσεις της εργασίας παρουσιάζονται, αν αναλογιστούμε τα δεδομένα που χρησιμοποιήσαμε αλλά και τις μειονεκτήματα των προτάσεων αυτής της εργασίας.

Αρχικά φαίνεται πως μια βελτίωση της μοντελοποίησης του F0 είναι απαραίτητη. Διαφορετικοί τρόποι να γίνει αυτό είναι χρησιμοποιώντας μεγαλύτερο receptive field σε όλα τα δίκτυα ώστε να μοντελοποιηθούν καλύτερα μεγαλύτερες χρονικά εξαρτήσεις μεταξύ των διαφορετικών τιμών του F0. Επίσης, στην περίπτωση των βάσεων δεδομένων με πολλούς ομιλητές, μάλλον είναι πιθανόν το μοντέλο που παράγει το F0 να χρειάζεται και σαν είσοδο την ταυτότητα του ομιλητή, δεδομένου ότι τα αποτελέσματα στην βάση TESS ήταν πολύ κακά.

Δεδομένου ότι η βάση δεδομένων που χρησιμοποιήθηκε είναι γενικά αρκετά μικρή, θα ήταν χρήσιμο να επεκταθούν τα πειράματα σε μεγαλύτερη βάση. Επειδή αυτό μάλλον θα έχει σαν συνέπεια την χρήση αρκετών ομιλητών, η μοντελοποίηση του ομιλητή δίνει αρκετές επιπλέον σχεδιαστικές επιλογές για το δίκτυο, καθώς θα είναι ενδιαφέρον να δούμε ένα μοντέλο που θα μοντελοποιεί ταυτόχρονα τους ομιλητές και το συναίσθημα, και ενδεχομένως να κάνει μετατροπή και στα δυο.

Σε όλα τα μοντέλα που χρησιμοποιήσαμε, μια υπόθεση που έχει γίνει έμμεσα με την χρήση συνελικτικών δικτύων, είναι ότι η διάρκεια της κάθε πρότασης (και κάθε ήχου γενικά) θα παραμείνει σταθερή. Αυτή η υπόθεση όμως

είναι προφανώς λάθος σε κάποιο βαθμό, αφού διαφορετικά συναισθήματα έχουν διαφορετικό ρυθμό ομιλίας. Μια προφανής βελτίωση του μοντέλου είναι να χρησιμοποιηθεί κάποιου είδος μοντελοποίηση της διάρκειας της πρότασης σαν είσοδο του μοντέλου. Αυτό δεν είναι προφανές πως μπορεί να γίνει και μάλλον απαιτεί την χρήση κειμένου ώστε να μπορούμε να το ευθυγραμμίσουμε με την ακουστική ακολουθία και να πάρουμε τις διάρκειες κάθε λέξης ή φωνήματος. Βέβαια, δεν είναι προφανές πως μπορεί αυτό να χρησιμοποιηθεί σε μη παράλληλα δεδομένα, και μάλλον θα απαιτήσει κάποια σημαντική αλλαγή στο μοντέλο.

Επίσης, σε όλα τα μοντέλα κρατήσαμε τα απεριοδικά χαρακτηριστικά σταθερά. Αυτό είναι πιο εύλογη υπόθεση από την διάρκεια, αλλά σε περιπτώσεις όπως για παράδειγμα το συναίσθημα της λύπης, ενδεχομένως να δημιουργεί πρόβλημα, καθώς σε αυτό το συναίσθημα μπορεί να υπάρχουν ήχοι αναφιλητών ή κλάματος. Αυτοί οι ήχοι είναι σε μεγάλο βαθμό άφωνοι και πιθανόν η ποιότητα των μετετραμμένων προτάσεων να επηρεάζεται άμεσα από τέτοιους ήχους. Γενικότερα, αξίζει τον κόπο να δοκιμαστούν διαφορετικά χαρακτηριστικά, όπως π.χ. από άλλους Vocoders, νευρωνικούς ή μη, ή και από self-supervised μοντέλα.

Τέλος, μια επιπλέον πιθανή προέκταση της εργασίας είναι σε μετατροπή συναισθημάτων σε μη διακριτούς χώρους, όπως π.χ. χρησιμοποιώντας το valence και το arousal. Αυτό δημιουργεί και το θέμα της μοντελοποίησης αυτών των χαρακτηριστικών το οποίο δεν μπορεί να αντιμετωπιστεί με τα μοντέλα μας χωρίς αλλαγή καθώς η υπόθεση μας σε αυτή την εργασία είναι των διακριτών πεδίων, αλλά και το θέμα της βάσης δεδομένων που πρέπει να έχει πληροφορία για τα συνεχή χαρακτηριστικά. Αποτελεί όμως υποσχόμενη κατεύθυνση καθώς η ταξινόμηση συναισθημάτων με βάση μη διακριτές κατηγορίες είναι πιο κοντά στην πραγματικότητα.

Βιβλιογραφία

[1] Ekman, P. "An argument for basic emotions". In: *Cognition & Emotion* (1992).

[2] Plutchik, R. "Emotions in the practice of psychotherapy: Clinical implications of affect theories." In: *American Psychological Association* (2012).

[3] Schacter, D. et al. *Psychology: Second European edition*. 2016.

[4] Laukka, P. et al. "Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations". In: *Frontiers in Psychology* (2013).

[5] Busso, C. et al. "IEMOCAP: interactive emotional dyadic motion capture database". In: *Language Resources and Evaluation* (2008).

[6] Filntisis, P. et al. "Video-realistic expressive audio-visual speech synthesis for the Greek Language". In: *Speech Communication* (2017).

[7] Pichora-Fuller, M. K. and Dupuis, K. *Toronto emotional speech set (TESS)*. 2020.

[8] Brownell, W. E. "How the ear works - Nature's solutions for listening". In: *The Volta review* (1997).

[9] Rabiner, L. R. and Juang, B.-H. *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[10] Fant, G. *Acoustic Theory Of Speech Production*. De Gruyter, 1960.

[11] Denes, P. and Pinson, E. N. *The speech chain: the physics and biology of spoken language*. W.H. Freeman, 1963.

[12] Shannon, C. E. "A mathematical theory of communication". In: *The Bell System Technical Journal* (1948).

[13] Oppenheim, A. V. and Schafer, R. W. *Discrete-Time Signal Processing*. Prentice Hall Press, 2009.

[14] Bäckström, T. et al. *"Introduction to Speech Processing"*.

[15] Rabiner, L. R. and Schafer, R. W. *Introduction to Digital Speech Processing*. Now Publishers Inc., 2007.

[16] Griffin, D. and Lim, J. "Signal estimation from modified short-time Fourier transform". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* (1984).

[17] Paeschke, A. and Sendlmeier, W. "Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements". In: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion* (2000).

[18] Boersma, P. and Weenink, D. *Praat: Doing phonetics by computer*. Miscellaneous. 2010.

[19] Morise, M., Yokomori, F., and Ozawa, K. "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications". In: *IEICE Trans. Inf. Syst.* (2016).

[20] Kawahara, H. et al. "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2008).

[21] Morise, M., Kawahara, H., and Katayose, H. "Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech". In: *journal of the audio engineering society* (2009).

[22] Morise, M. "CheapTrick, a spectral envelope estimator for high-quality speech synthesis". In: *Speech Communication* (2015).

[23] Chen, J. and Miller, D. "Pitch-Synchronous Analysis of Human Voice". In: *Journal of Voice* (2019).

[24] Morise, M. "PLATINUM: A method to extract excitation signals for voice synthesis system". In: *Acoustical Science and Technology* (2012).

[25] Kingma, D. P. and Ba, J. *Adam: A Method for Stochastic Optimization*. 2014.

[26] Karpathy, A. *"Convolutional neural networks for visual recognition."*
[27] Tibshirani, R. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. series$_B$(Methodological)* (1996).
[28] Srivastava, N. et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research (JMLR)* (2014).
[29] He, K. et al. *Deep Residual Learning for Image Recognition*. 2015.
[30] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.
[31] Hinton, G. and Salakhutdinov, R. "Reducing the Dimensionality of Data with Neural Networks". In: *Science (New York, N.Y.)* (2006).
[32] Hinton, G. E. and Zemel, R. S. "Autoencoders, Minimum Description Length and Helmholtz Free Energy". In: *Conference on Neural Information Processing Systems (NIPS)*. 1993.
[33] Kingma, D. P. and Welling, M. *Auto-Encoding Variational Bayes*. 2013.
[34] Lecun, Y. et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* (1998).
[35] Deng, J. et al. "ImageNet: A Large-Scale Hierarchical Image Database". In: *Computer Vision and Pattern Recognition Conference (CVPR)*. 2009.
[36] Abdel-Hamid, O. et al. "Convolutional Neural Networks for Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2014).
[37] Scherer, D., Müller, A. C., and Behnke, S. "Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition". In: *International Conference on Artificial Neural Networks (ICANN)*. 2010.
[38] Olah, C. et al. "Zoom In: An Introduction to Circuits". In: *Distill* (2020).
[39] Oord, A. van den and Kalchbrenner, N. "Pixel RNN". In: *International Conference on Machine Learning (ICML)*. 2016.
[40] Vaswani, A. et al. "Attention is All you Need". In: *Conference on Neural Information Processing Systems (NIPS)*. 2017.
[41] Graves, A. "Generating Sequences With Recurrent Neural Networks". In: *ArXiv* abs/1308.0850 (2013).
[42] Hochreiter, S. and Schmidhuber, J. "Long Short-term Memory". In: *Neural computation* (1997).
[43] Sutskever, I., Vinyals, O., and Le, Q. V. "Sequence to Sequence Learning with Neural Networks". In: *ArXiv* abs/1409.3215 (2014).
[44] Bahdanau, D., Cho, K., and Bengio, Y. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *ArXiv* (2014).
[45] Battenberg, E. et al. "location_*RelativeAttentionMechanismsforRobustLong* − *FormSpeechSynthesis*". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019).
[46] He, M., Deng, Y., and He, L. "Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS". In: *ArXiv* (2019).
[47] Ren, Y. et al. "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech". In: *ArXiv* (2020).
[48] Shen, J. et al. "Non-Attentive Tacotron: Robust and Controllable Neural TTS Synthesis Including Unsupervised Duration Modeling". In: *ArXiv* (2020).
[49] McAuliffe, M. et al. "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi". In: *Interspeech*. 2017.
[50] Zeng, Z. et al. "Alignments: Efficient Feed-Forward Text-to-Speech System Without Explicit Alignment". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020).
[51] Wang, Y. et al. "Tacotron: Towards End-to-End Speech Synthesis". In: *Interspeech*. 2017.
[52] Skerry-Ryan, R. J. et al. "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron". In: *ArXiv* (2018).
[53] Wang, Y. et al. "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis". In: *International Conference on Machine Learning (ICML)*. 2018.
[54] Hsu, W.-N. et al. "Hierarchical Generative Modeling for Controllable Speech Synthesis". In: *ArXiv* abs/1810.07217 (2018).
[55] Sun, G. et al. "Generating Diverse and Natural Text-to-Speech Samples Using a Quantized Fine-Grained VAE and Autoregressive Prosody Prior". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020).

[56] Rezende, D. J. and Mohamed, S. "Variational Inference with Normalizing Flows". In: *International Conference on Machine Learning (ICML)* (2015).
[57] Kim, J. et al. "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search". In: *Conference on Neural Information Processing Systems (NIPS)* (2020).
[58] Kim, J., Kong, J., and Son, J. "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech". In: *International Conference on Machine Learning (ICML)* (2021).
[59] Oord, A. van den et al. "WaveNet: A Generative Model for Raw Audio". In: *Speech Synthesis Workshop (SSW)* (2016).
[60] Oord, A. van den et al. "Parallel WaveNet: Fast High-Fidelity Speech Synthesis". In: *International Conference on Machine Learning*. 2017.
[61] Kingma, D. P., Salimans, T., and Welling, M. "Improving Variational Inference with Inverse Autoregressive Flow". In: *Conference on Neural Information Processing Systems (NIPS)* (2016).
[62] Kalchbrenner, N. et al. "Efficient Neural Audio Synthesis". In: *International Conference on Machine Learning (ICML)* (2018).
[63] Prenger, R. J., Valle, R., and Catanzaro, B. "Waveglow: A Flow-based Generative Network for Speech Synthesis". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018).
[64] Ping, W. et al. "WaveFlow: A Compact Flow-based Model for Raw Audio". In: *International Conference on Machine Learning (ICML)* (2019).
[65] Goodfellow, I. J. et al. "Generative Adversarial Nets". In: *Conference on Neural Information Processing Systems (NIPS)*. 2014.
[66] Al, G. *"Generative Adversarial Networks – Hot Topic in Machine Learning"*.
[67] Arjovsky, M. and Bottou, L. "Towards Principled Methods for Training Generative Adversarial Networks". In: *International Conference on Learning Representations (ICLR)* (2017).
[68] Radford, A., Metz, L., and Chintala, S. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2015.
[69] Mao, X. et al. "Least Squares Generative Adversarial Networks". In: *IEEE International Conference on Computer Vision (ICCV)* (2016).
[70] Arjovsky, M., Chintala, S., and Bottou, L. "Wasserstein GAN". In: *ArXiv* abs/1701.07875 (2017).
[71] Gulrajani, I. et al. "Improved Training of Wasserstein GANs". In: *Conference on Neural Information Processing Systems (NIPS)*. 2017.
[72] LeCun, Y., Cortes, C., and Burges, C. "MNIST handwritten digit database". In: *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* (2010).
[73] Mirza, M. and Osindero, S. "Conditional Generative Adversarial Nets". In: *ArXiv* abs/1411.1784 (2014).
[74] Reed, S. E. et al. "Generative Adversarial Text to Image Synthesis". In: *International Conference on Machine Learning (ICML)*. 2016.
[75] Kim, T. et al. "Learning to Discover Cross-Domain Relations with Generative Adversarial Networks". In: *International Conference on Machine Learning (ICML)*. 2017.
[76] Isola, P. et al. "Image-to-Image Translation with Conditional Adversarial Networks". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
[77] Choi, Y. et al. "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017).
[78] Ganin, Y. et al. "Domain-Adversarial Training of Neural Networks". In: *Journal of machine learning research (JMLR)*. 2015.
[79] Ben-David, S. et al. "Analysis of Representations for Domain Adaptation". In: *Conference on Neural Information Processing Systems (NIPS)*. 2006.
[80] Zhang, Y. et al. "Learning to Speak Fluently in a Foreign language: Multilingual Speech Synthesis and Cross-language Voice Cloning". In: *Interspeech*. 2019.
[81] Hsu, W.-N. et al. "Disentangling Correlated Speaker and Noise for Speech Synthesis via Data Augmentation and Adversarial Factorization". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019).
[82] Donahue, C., McAuley, J., and Puckette, M. "Synthesizing Audio with GANs". In: *International Conference on Learning Representations (ICLR)*. 2018.
[83] Odena, A., Dumoulin, V., and Olah, C. "Deconvolution and Checkerboard Artifacts". In: *Distill* (2016).

[84] Pascual, S., Bonafonte, A., and Serrà, J. "SEGAN: Speech Enhancement Generative Adversarial Network". In: *Interspeech*. 2017.
[85] Baby, D. and Verhulst, S. "Sergan: Speech Enhancement Using Relativistic Generative Adversarial Networks with Gradient Penalty". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.
[86] Meng, Z. et al. "Cycle-Consistent Speech Enhancement". In: *Interspeech* (2018).
[87] Kumar, K. et al. "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis". In: *Advances in Neural Information Processing Systems (NIPS)*. 2019.
[88] Su, J., Jin, Z., and Finkelstein, A. "HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks". In: *Interspeech* (2020).
[89] Binkowski, M. et al. "High Fidelity Speech Synthesis with Adversarial Networks". In: *International Conference on Learning Representations (ICLR)* (2019).
[90] Donahue, J. et al. "End-to-End Adversarial Text-to-Speech". In: *International Conference on Learning Representations (ICLR)* (2020).
[91] Kaneko, T. and Kameoka, H. "Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks". In: *ArXiv* abs/1711.11293 (2017).
[92] Kaneko, T. et al. "Cyclegan-VC2: Improved Cyclegan-based Non-parallel Voice Conversion". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.
[93] Kaneko, T. et al. "CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-spectrogram Conversion". In: *Proceedings of the Annual Conference of the International Speech Communication Association*. 2020.
[94] Kameoka, H. et al. "StarGAN-VC: non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks". In: *IEEE Spoken language Technology Workshop (SLT)* (2018).
[95] Ohtani, Y. et al. "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation". In: *Interspeech*. 2006.
[96] Liu, K., Zhang, J., and Yan, Y. "High Quality Voice Conversion through Phoneme-Based Linear Mapping Functions with STRAIGHT for Mandarin". In: *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. 2007.
[97] Ioffe, S. and Szegedy, C. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *International Conference on Machine Learning (ICML)*. 2015.
[98] Dauphin, Y. et al. "Language Modeling with Gated Convolutional Networks". In: *International Conference on Machine Learning (ICML)*. 2017.
[99] Ocal, O. et al. "Adversarially Trained Autoencoders for Parallel-data-free Voice Conversion". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019).
[100] Zhang, J. et al. "Sequence-to-sequence acoustic modeling for voice conversion". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2019).
[101] Tanaka, K. et al. "ATTS2S-VC: Sequence-to-sequence Voice Conversion with Attention and Context Preservation Mechanisms". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018).
[102] Kameoka, H. et al. "ConvS2S-VC: Fully convolutional sequence-to-sequence voice conversion". In: *ArXiv* abs/1811.01609 (2018).
[103] Zhang, J.-X. et al. "Improving Sequence-to-sequence Voice Conversion by Adding Text-supervision". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018).
[104] Kaneko, T. et al. "Sequence-to-Sequence Voice Conversion with Similarity Metric Learned Using Generative Adversarial Networks". In: *Interspeech*. 2017.
[105] Hsu, C.-C. et al. "Voice conversion from non-parallel corpora using variational auto-encoder". In: *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE. 2016.
[106] Huang, W.-C. et al. "Voice Conversion Based on Cross-Domain Features Using Variational Auto Encoders". In: *International Symposium on Chinese Spoken Language Processing (ISCSLP)* (2018), pp. 51–55. URL:
[107] Kawahara, H. "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds". In: *Acoust Sci Technol* (2006).
[108] Kameoka, H. et al. "ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder". In: *ArXiv* abs/1808.05092 (2018).

[109] Saito, Y. et al. "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018.

[110] Fang, F. et al. "High-quality nonparallel voice conversion based on cycle-consistent adversarial network". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018.

[111] Gao, Y., Singh, R., and Raj, B. "Voice impersonation using generative adversarial networks". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018.

[112] Yeh, C.-c. et al. "Rhythm-Flexible Voice Conversion Without Parallel Data Using Cycle-GAN Over Phoneme Posteriorgram Sequences". In: *2018 IEEE Spoken Language Technology Workshop (SLT)* (2018).

[113] Hsu, C.-C. et al. "Voice Conversion from Unaligned Corpora Using Variational Autoencoding Wasserstein Generative Adversarial Networks". In: *Interspeech* (2017).

[114] Chou, J.-C. et al. "Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations". In: *Interspeech*. 2018.

[115] Gao, J. et al. "Nonparallel Emotional Speech Conversion". In: *Interspeech*. 2019.

[116] Zhou, K., Sisman, B., and Li, H. "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data". In: *The Speaker and Language Recognition Workshop*. 2020. URL:

[117] Zhou, K., Sisman, B., and Li, H. "Vaw-Gan For Disentanglement And Recomposition Of Emotional Elements In Speech". In: *Spoken Language Technology Workshop*. 2020.

[118] Zhou, K. et al. "Seen and Unseen Emotional Style Transfer for Voice Conversion with A New Emotional Speech Dataset". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020).

[119] Rizos, G. et al. "Stargan for Emotional Speech Conversion: Validated by Data Augmentation of End-To-End Emotion Recognition". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020).

[120] Choi, H. and Hahn, M. "Sequence-to-Sequence Emotional Voice Conversion With Strength Control". In: *IEEE Access* (2021).

[121] Shankar, R., Sager, J., and Venkataraman, A. "Non-parallel Emotion Conversion using a Deep-Generative Hybrid Network and an Adversarial Pair Discriminator". In: *Interspeech*. 2020.

[122] Liu, A. H. et al. "A Unified Feature Disentangler for Multi-Domain Image Translation and Manipulation". In: *Advances in Neural Information Processing Systems (NIPS)*. 2018.

[123] Black, A. W. and Tokuda, K. "The blizzard challenge - 2005: evaluating corpus-based speech synthesis on common datasets". In: *Interspeech*. 2005.

[124] Kubichek, R. "Mel-cepstral distance measure for objective speech quality assessment". In: *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing* (1993).

[125] Chu, W. and Alwan, A. "Reducing F0 Frame Error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009.

[126] Nakatani, T. et al. "A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments". In: *Speech Commun.* (2008).

[127] Lorenzo-Trueba, J. et al. "The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods". In: *The Speaker and Language Recognition Workshop* (2018).

[128] Maniati, G. et al. "SOMOS: The Samsung Open MOS Dataset for the Evaluation of Neural Text-to-Speech Synthesis". In: *Interspeech*. 2022.

[129] Lo, C.-C. et al. "MOSNet: Deep Learning based Objective Assessment for Voice Conversion". In: *Interspeech* (2019).

[130] Cooper, E. et al. "Generalization Ability of MOS Prediction Networks". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021).

[131] Saeki, T. et al. "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022". In: *Interspeech* (2022).

[132] Baevski, A. et al. "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *Conference on Neural Information Processing Systems (NIPS)*. Curran Associates Inc., 2020.

[133] Ravanelli, M. et al. *SpeechBrain: A General-Purpose Speech Toolkit*. 2021.