ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ
ΥΠΟΛΟΓΙΣΤΩΝ

Αυτόματη Μετάφραση Μουσικής με χρήση Νευρωνικών Δικτύων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παναγιώτης Ι. Αποστολίδης

Επιβλέπων : Γεώργιος Στάμου
Αν. Καθηγητής Ε.Μ.Π

Αθήνα, Οκτώβρης 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ
ΥΠΟΛΟΓΙΣΤΩΝ

Αυτόματη Μετάφραση Μουσικής με χρήση Νευρωνικών Δικτύων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παναγιώτης Ι. Αποστολίδης

Επιβλέπων : Γεώργιος Στάμου
Αν. Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 27η Οκτώβρη 2021.

.....	.....	.....
Γεώργιος Στάμου	Α. – Γ. Σταφυλοπάτης	Σ. Κόλλιας
Αν. Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π

Αθήνα, Οκτώβρης 2021

.....

Παναγιώτης Ι. Αποστολίδης

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

Περίληψη

Η Αυτόματη Σύνθεση Μουσικής αποτελεί ένα από τα πλέον κομβικά, αλλά και δύσκολα έργα στον τομέα της ανάκτησης μουσικής πληροφορίας. Ειδικότερα, η Αυτόματη Μετάφραση Μουσικής, η οποία υπάγεται στο πρόβλημα της Αυτόματης Σύνθεσης Μουσικής, παρουσιάζει αρκετό ενδιαφέρον, καθώς είναι ένα έργο που δεν έχει διερευνηθεί ακόμα σε ικανοποιητικό βαθμό.

Ως όρος μπορεί να ερμηνευθεί με δύο διαφορετικούς τρόπους: Ως μετάφραση από ένα είδος μουσικής σε κάποιο άλλο (π.χ. από κλασική μουσική σε Τζαζ), ή ως μετάφραση μουσικής από ένα μουσικό όργανο ή μια ομάδα οργάνων σε κάποιο άλλο μουσικό όργανο. Στην παρούσα διπλωματική εργασία θα μελετήσουμε τη δεύτερη ερμηνεία, δηλαδή την αυτόματη μετάφραση μουσικής από ένα μουσικό όργανο σε κάποιο άλλο.

Κατά μια έννοια, η μεταφορά μουσικής μπορεί να θεωρηθεί ως το μουσικό αντίστοιχο της μετάφρασης ενός λογοτεχνικού έργου. Οι διάφορες φωνές και τα όργανα αποτελούν την "γλώσσα", με την οποία ο συνθέτης εξωτερικεύει τις σκέψεις και τα συναισθήματα του. Συνεπώς, σκοπός της μεταφοράς μουσικής είναι να γίνει κατανοητό ένα μουσικό έργο σε κάποια διαφορετική "γλώσσα". Από την πλευρά του συνθέτη η διαδικασία αυτή απαιτεί ένα είδος προσαρμογής/διασκευής της αρχικής σύνθεσης, ώστε να ταιριάζει στο νέο μέσο, δηλαδή στα μουσικά όργανα απ' τα οποία θα εκτελεστεί. Για το λόγο αυτό, του δίνεται η ελευθερία να διαφοροποιήσει την εναρμόνιση των φωνών, την ενορχήστρωση, αλλά και την ανάπτυξη του μουσικού θέματος σε σχέση με την πρωτότυπη μουσική σύνθεση.

Από την πλευρά των υπολογιστών, η Αυτόματη Μετάφραση Μουσικής αποτελεί ένα εξίσου απαιτητικό πρόβλημα. Τα τελευταία χρόνια, με την ανάπτυξη νέων βαθιών νευρωνικών δικτύων και τεχνικών εκπαίδευσης, έχουν αναπτυχθεί ορισμένα μοντέλα με υποσχόμενα αποτελέσματα. Τα χαρακτηριστικότερα εξ’ αυτών αποτελούν το “ A Universal Music Translation Network” και το “MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis”.

Στο πειραματικό μέρος της εργασίας δημιουργήθηκε και εκπαιδεύθηκε ένα μοντέλο που προέκυψε από έναν συνδυασμό των προαναφερθέντων δικτύων. Ως σύνολο δεδομένων χρησιμοποιήθηκε το MusicNet και πιο συγκεκριμένα όσα δεδομένα του συνόλου αυτού αποτελούν μουσικές συνθέσεις του Μπετόβεν για πιάνο

Λέξεις Κλειδιά: Τεχνητά Νευρωνικά Δίκτυα, Βαθιά Μάθηση, Συνελικτικά Νευρωνικά Δίκτυα Δίκτυο WaveNet, Παραγωγικά Αντιπαραθετικά Δίκτυα, Αρχιτεκτονική Κωδικοποιητή – Αποκωδικοποιητή, Μουσικά Όργανα, Σπεκτρόγραμμα

Abstract

Automatic Music Generation is one of the most crucial, but also demanding tasks in the field of Music Information Retrieval. More specifically, Automatic Music Translation is a promising sub-task of Automatic Music Generation that has not yet been developed to a satisfactory level.

The term of Music Translation can be interpreted either as a Music Transfer from one music genre to another, or as a Music Transfer from one set of musical instruments to another. In this project we will delve into the second interpretation, meaning the task of translating music from one set of musical instruments to another.

In a sense, the art of Music Arrangement can be considered as the equivalent of the translation of a literary piece. Composers use voices and instruments as their "language", in order to communicate their emotions to the rest of the world. Thus, the purpose of arrangement is to make that which was written in one musical language intelligible in another. On the composer's side, this procedure requires some sort of adaptation of the original piece so as to fit in the new medium, meaning the musical instruments for which it is arranged. Hence, the composer is given the freedom to modify the voicings and harmonization, as well as the development of the musical theme, compared to the original.

On the other hand, the task of Automatic Translation is equally challenging for computers. However, there has been great progress in the past years, due to the growth of new deep neural networks and training methods. As a result, certain Machine Learning models with promising results have been introduced. Two of the most noteworthy networks are "A Universal Music Translation Network" and "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis".

For the experimental part of this thesis, a model combining the aforementioned networks has been created and trained. The Dataset "MusicNet" was selected for the training of the model. More precisely, the network was trained only using MusicNet data that are included in the domain "Beethoven Solo Piano".

Keywords: Artificial Neural Networks, Deep Learning, Convolutional Neural Networks, WaveNet Network, Generative Adversarial Networks, Encoder – Decoder Architecture, Musical Instruments, Spectrogram

Πίνακας Περιεχομένων

1 Εισαγωγή

Η Αυτόματη Μετάφραση Μουσικής (Music Translation) είναι ένα έργο που δεν έχει διερευνηθεί ακόμα αρκετά στον τομέα της ανάκτησης πληροφορίας από μουσική (Music Information Retrieval). Ως όρος μπορεί να διερμηνευθεί με δύο διαφορετικούς τρόπους: Ο πρώτος αποτελεί μετάφραση από ένα είδος μουσικής σε κάποιο άλλο , όπως για παράδειγμα από κλασική μουσική σε Τζαζ, ενώ ο δεύτερος αποτελεί μετάφραση μουσικής από ένα μουσικό όργανο ή μια ομάδα οργάνων σε κάποιο άλλο μουσικό όργανο. Χαρακτηριστικό παράδειγμα αποτελεί η αυτόματη μετάφραση μουσικής από πιάνο σε ορχήστρα. Σε αυτή την διπλωματική εργασία θα ασχοληθούμε με την αυτόματη μετάφραση μουσικής από ένα μουσικό όργανο σε κάποιο άλλο.

1.1 Σκοπός της Εργασίας

Παραδοσιακά, τα μοντέλα αυτόματης μετάφρασης μουσικής σε πρώτο στάδιο πραγματοποιούν επεξεργασία του σήματος εισόδου με σκοπό να εξάγουν χαρακτηριστικά τα οποία θα διευκολύνουν την εκπαίδευση του μοντέλου ή δέχονται ως είσοδο midi αρχεία τα οποία περιλαμβάνουν τέτοια χαρακτηριστικά. Ωστόσο τα τελευταία χρόνια, έχει γίνει προσπάθεια να αντικατασταθούν τα στάδια επεξεργασίας σήματος και εξαγωγής χαρακτηριστικών με αρχιτεκτονικές βαθιάς μηχανικής μάθησης. Έτσι λοιπόν υπάρχουν δίκτυα που δέχονται στην είσοδο τους απευθείας την ακολουθία εισόδου και δίκτυα που πραγματοποιούν απλούς μετασχηματισμούς όπως είναι ο μετασχηματισμός Fourier μικρού χρόνου (Short Time Fourier Transform). Στην παρούσα εργασία θα εξετάσουμε και θα προσπαθήσουμε να επεκτείνουμε μοντέλα τέτοιου τύπου.

Σε ότι αφορά το πειραματικό μέρος της εργασίας, σκοπός είναι η δημιουργία και η εκπαίδευση ενός δικτύου που πραγματοποιεί αυτόματη μετάφραση μουσικής από ένα Μουσικό Domain σε κάποιο άλλο. Για τους σκοπούς της εργασίας, ως μουσικό Domain ορίζουμε το σύνολο των δεδομένων που αποτελούνται από μουσικές συνθέσεις ενός συνθέτη και ενός συγκεκριμένου συνόλου οργάνων. Η αξιολόγηση των παραγόμενων δεδομένων πραγματοποιήθηκε τόσο ποιοτικά από περιορισμένο αριθμό χρηστών, όσο και ποσοτικά, με την καταγραφή του σφάλματος ανακατασκευής ως μετρική αξιολόγησης.

1.2 Δομής της Εργασίας

Η παρούσα εργασία είναι δομημένη σε 9 κεφάλαια. Στο Κεφάλαιο 2 παρουσιάζεται η τέχνη της Μετάφρασης Μουσικής από ένα σύνολο οργάνων σε κάποιο άλλο. Ταυτόχρονα αναλύονται οι δυσκολίες που συναντάει ένας συνθέτης όταν επιχειρεί να πραγματοποιήσει ένα τέτοιο έργο.

Το κεφάλαιο 3 αναφέρεται στους διαφορετικούς τρόπους αναπαράστασης δεδομένων από Παραγωγικά Νευρωνικά Δίκτυα. Πιο συγκεκριμένα γίνεται αναφορά στους παραδοσιακούς τρόπους αναπαράστασης των δεδομένων και ανάπτυξη μεθόδων που αντικαθιστούν τα πρώτα στάδια προ-επεξεργασίας ενός σήματος ήχου. Χαρακτηριστικές τέτοιες μέθοδοι αποτελούν τα σπεκτρογράμματα τα οποία είναι αναπαραστάσεις που προέρχονται από Μετασχηματισμό Fourier του αρχικού σήματος.

Στο κεφάλαιο 4 αναλύονται ορισμένα βασικά μοντέλα νευρωνικών δικτύων, τα οποία χρησιμοποιούνται ως υποσυστήματα των συνθετότερων μοντέλων που έχουν αναπτυχθεί τα τελευταία χρόνια.

Στο κεφάλαιο 5 παρουσιάζονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση των δικτύων στα οποία βασίστηκε αυτή η εργασία. Παράλληλα, γίνεται και παρουσίαση των βασικών περιεχομένων αλλά και χαρακτηριστικών των συνόλων δεδομένων αυτών.

Στο κεφάλαιο 6 παρουσιάζονται αναλυτικά τα δίκτυα στα οποία βασίστηκε αυτή η εργασία. Πιο συγκεκριμένα, αυτά είναι το “A Universal Music Translation Network” και το “MelGAN”.

Το κεφάλαιο 7 εστιάζει στην δική μας υλοποίηση που αποτελεί έναν συνδυασμό των δύο προαναφερθέντων μοντέλων. Ταυτόχρονα, εξηγείται η διαδικασία εκπαίδευσης του τελικού μοντέλου.

Στο κεφάλαιο 8 παρουσιάζονται οι διαφορετικές μέθοδοι αξιολόγησης των Παραγωγικών Αντιπαραθετικών Δικτύων και αναλύονται αυτές που επιλέχθηκαν για το δικό μας δίκτυο. Επιπλέον, γίνεται μια μικρή παρουσίαση της ποιότητας των αποτελεσμάτων.

Το κεφάλαιο 9 περιέχει την βιβλιογραφία στην οποία βασίστηκε η παρούσα εργασία.

2 Θεωρία Μουσικής

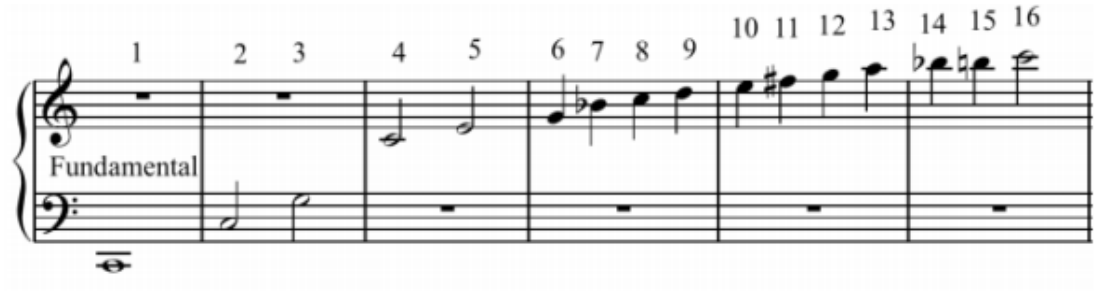
2.1 Ορισμός Μουσικής

Η μουσική είναι μια οργανωμένη συλλογή από ήχους. Η σύνθεση μουσικής είναι η διαδικασία με την οποία συνδυάζουμε ήχους και τόνους με σκοπό να δημιουργήσουμε ένα ενοποιημένο έργο. Η οργάνωση αυτή διαφορετικών ήχων έχει ως σκοπό την δημιουργία μιας συγκεκριμένης ατμόσφαιρας ή την έκφραση ιδεών και συναισθημάτων του συνθέτη. Η μουσική αποτελείται λοιπόν από ήχους, δονήσεις και παύσεις και εκτελείται από ένα ευρύ φάσμα μουσικών οργάνων ή/και την ανθρώπινη φωνή. Μάλιστα, υπάρχουν μουσικά έργα τα οποία εκτελούνται μόνο από μουσικά όργανα (instrumental) και άλλα που είναι αποκλειστικά φωνητικά (a Capella). Σύμφωνα με την κλασική οργανολογία τα μουσικά όργανα μπορούν να κατηγοριοποιηθούν σε χάλκινα πνευστά, ξύλινα πνευστά, κρουστά, έγχορδα και πληκτροφόρα.

2.2 Ο ήχος που παράγουν τα μουσικά όργανα

Κάθε μουσικό έργο εκτελείται από ένα υποσύνολο μουσικών οργάνων ή ακόμα και από μόνο ένα μουσικό όργανο. Η επιλογή της ενορχήστρωσης καθορίζει σε πολύ μεγάλο βαθμό την συνολική αίσθηση και ατμόσφαιρα που αφήνει το μουσικό έργο. Για παράδειγμα, αν ένα κοντραμπάσο παίξει μια νότα Μι θα έχει μια σοβαρή αίσθηση, ενώ αν παίξει ακριβώς την ίδια νότα ένα φλάουτο θα ακουστεί πολύ πιο εύθυμη. Αυτό συμβαίνει επειδή τα μουσικά όργανα, ακόμα και όταν παίζουν τον ίδιο τόνο, παράγουν διαφορετικές αρμονικές σε διαφορετικές εντάσεις και έχουν διαφορετικά ηχοχρώματα. Σημαντικό ρόλο παίζει και το υλικό απ' το οποίο είναι κατασκευασμένο το μουσικό όργανο αφού ο ήχος διαδίδεται μέσα από αυτό.

Πιο συγκεκριμένα, μια χορδή που πάλλεται δεν παράγει μια μοναδική συχνότητα, αλλά έναν συνδυασμό από θεμελιώδεις συχνότητες και αρμονικές. Αυτό δεν σημαίνει ότι ο άνθρωπος έχει την δυνατότητα να διαχωρίσει την κάθε αρμονική που παράγεται καθώς αυτές "αναμειγνύονται" με αποτέλεσμα να παράγεται ο ήχος που τελικά ακούμε εμείς. Αν κάθε μουσικό όργανο είχε την δυνατότητα να παράγει μόνο την θεμελιώδη συχνότητα που αντιστοιχεί σε μία νότα, τότε κάθε όργανο θα είχε ακριβώς τον ίδιο μουντό ήχο. Παρακάτω παρουσιάζονται όλες οι αρμονικές συχνότητες που εμπλέκονται όταν μια νότα ΛΑ παίζεται σε βιολί, οι οποίες λέμε ότι αποτελούν την αρμονική σκάλα Λα του βιολιού



Σχήμα 1: Αρμονική σκάλα Λα βιολιού

Ο πρώτος που μελέτησε τον ρόλο των διαφορετικών συχνοτήτων που εμπεριέχονται σε έναν ήχο ήταν ο Πυθαγόρας ο οποίος παρατήρησε ότι όταν σταματούσε την δόνηση μιας χορδής στο μέσο του μήκους της. τότε ο τόνος υψωνόταν κατά μια οκτάβα. Έτσι επανέλαβε το πείραμα αυτό σε διαφορετικά σημεία της χορδής μελετώντας τις αρμονικές που προκύπταν κάθε φορά ανάλογα με το σημείο στο οποίο σταματούσε την χορδή. Μάλιστα υπάρχει μια συγκεκριμένη σχέση που συνδέει το μήκος κύματος με τις αρμονικές που παράγονται.

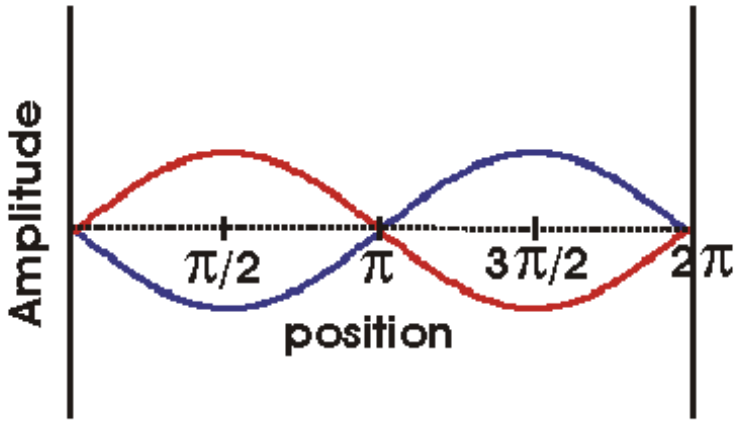
Όταν πάλουμε μια χορδή δημιουργείται στάσιμο κύμα. Πιο συγκεκριμένα με το χέρι μας, με μια πένα ή με ένα δοξάρι μεταφέρουμε ενέργεια στην χορδή η οποία είναι στηριγμένη σε δύο σημεία. Με αυτό τον τρόπο το κύμα που αρχικά έχει δημιουργηθεί ανακλάται στα δύο σημεία που στηρίζεται η χορδή με αποτέλεσμα να δημιουργηθούν δύο κύματα τα οποία όταν συναντώνται συμβάλλουν υπερθετικά. Ωστόσο τα στάσιμα κύματα προκύπτουν μόνο όταν η χορδή πάλλεται σε συγκεκριμένες συχνότητες. Σε αυτό το σημείο θα θεωρήσουμε ότι **L** είναι το μήκος της χορδής, **λ** το μήκος κύματος, **F** η συχνότητα με την οποία πάλλεται η χορδή και **V** η ταχύτητα που μεταφέρεται το κύμα στην χορδή η οποία είναι και σταθερή. Τότε η συχνότητα **F** είναι αντιστρόφως ανάλογη του μήκους κύματος της χορδής **λ**. Όπως βλέπουμε στο σχήμα 2, στην θεμελιώδη συχνότητα έχουμε τις εξής σχέσεις:

$$\lambda_0 = 2L \qquad (1)$$

$$F_0 = \frac{v}{\lambda_0} \qquad (2)$$

Σχήμα 2 : Στιγμιότυπο παλμού χορδής στην θεμελιώδη συχνότητα (πρώτη αρμονική)

Όταν ο παλμός της χορδής διακόπτεται σε συγκεκριμένα σημεία προκύπτουν δεσμοί, δηλαδή σημεία με μέγιστο πλάτος, και κοιλίες, δηλαδή σημεία που τα δύο κύματα αλληλοαναιρούνται. Οι δεσμοί έχουν διαφορά φάσης π ενώ οι κοιλίες διαφορά φάσης 2π όπως φαίνεται και στο παρακάτω σχήμα.

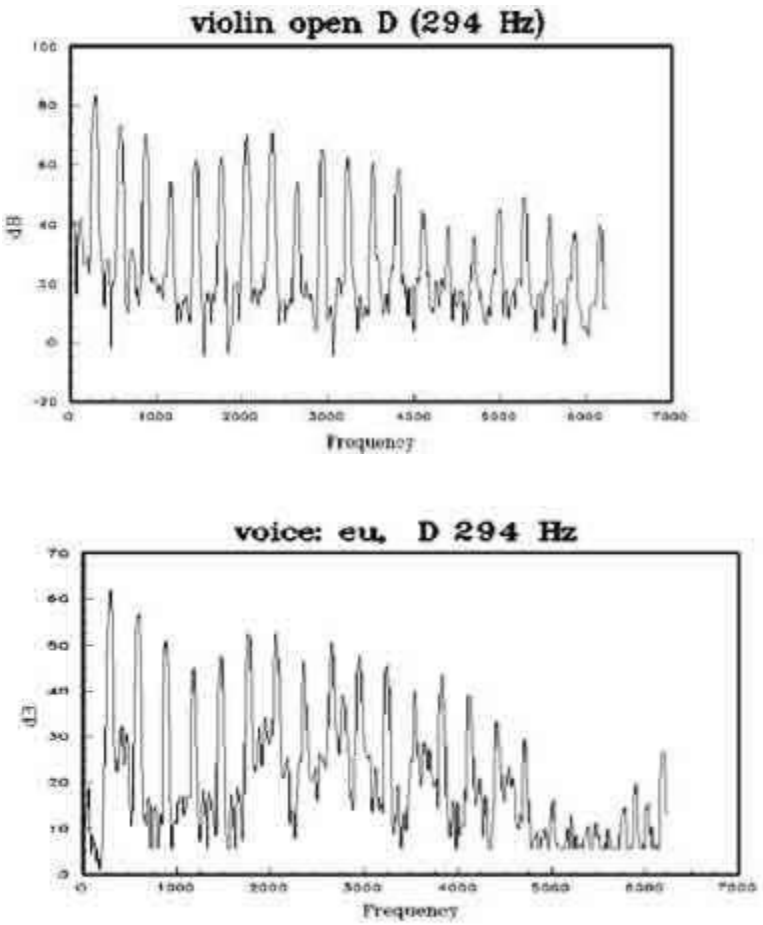


Σχήμα 3: Εμφάνιση δεσμών και κοιλιών όταν $\lambda_1 = L$ και $F_1 = \frac{v}{\lambda_1}$

Τελικά η n-οστή αρμονική δίνεται από τον τύπο:

$$\lambda_n = \frac{2 * L}{n} \qquad (3)$$

Για να βρούμε την μορφή που θα έχει το κύμα που παράγεται όταν παίζουμε μια νότα θα πρέπει να αθροίσουμε τα πλάτη όλων των αρμονικών που εμφανίζονται (mixtures of harmonics) Τα πλάτη των αρμονικών για μια νότα δεν είναι όμως ίδια σε κάθε όργανο με αποτέλεσμα να προκύπτουν διαφορετικά ηχοχρώματα. Για παράδειγμα, το κλαρινέτο έχει πολύ δυνατές αρμονικές περιττού αριθμού και πολύ πιο αδύναμες άρτιου αριθμού, ενώ για το φλάουτο ισχύει ακριβώς το αντίθετο. Παρακάτω παρουσιάζουμε όλες τις αρμονικές που παρουσιάζονται όταν κάποιος παίζει τη νότα Ρε στο βιολί αλλά και όταν κάποιος τραγουδάει την ίδια νότα. Παρατηρούμε ότι υπάρχουν ομοιότητες στα δύο διαγράμματα, αλλά και σημαντικές διαφορές που καθορίζουν την διαφορά στους ήχους που προκύπτουν.



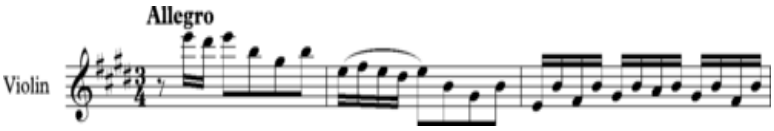
Σχήμα 4 : Τα πλάτη των αρμονικών όταν παίζεται η νότα Ρε (συχνότητας 294 Hz) στο βιολί και από φωνή αντίστοιχα.

2.3 Μεταφορά μουσικής

Η μεταφορά μουσικής από ένα όργανο ή σύνολο οργάνων σε κάποιο άλλο δεν είναι καθόλου απλή υπόθεση αν λάβουμε υπόψιν όσα αναφέραμε στο προηγούμενο κεφάλαιο για τις ιδιαιτερότητες κάθε οργάνου. Ο συνθέτης πρέπει να λάβει υπόψιν τις δυνατότητες ενός οργάνου, όπως το αν είναι πολυφωνικό ή μονοφωνικό και τις τονικές του δυνατότητες, αλλά και την απαραίτητη ενορχήστρωση, Για παράδειγμα, αν ένας συνθέτης θέλει να μεταφέρει ένα μουσικό έργο πιάνου για ένα βιολί συνοδευόμενο από άλλα έγχορδα, θα πρέπει να επιλέξει ανάλογα με τις δυνατότητες του κάθε οργάνου, ποια μέρη της μελωδίας θα πάρει το βιολί και ποια μέρη θα πάρουν τα υπόλοιπα έγχορδα. Με αυτή την έννοια η μεταφορά της μουσικής (στα αγγλικά χρησιμοποιείται ο όρος arrangement) απαιτεί ένα είδος προσαρμογής/διασκευής της αρχικής σύνθεσης ώστε να ταιριάζει στο μέσο, δηλαδή στα μουσικά όργανα, απ' το οποίο θα παιχτεί. Είναι συνεπώς κατανοητό ότι δίνεται η ελευθερία στον συνθέτη να διαφοροποιήσει την εναρμόνιση των φωνών, την ενορχήστρωση αλλά και την ανάπτυξη του μουσικού θέματος σε σχέση με την αυθεντική μουσική σύνθεση.

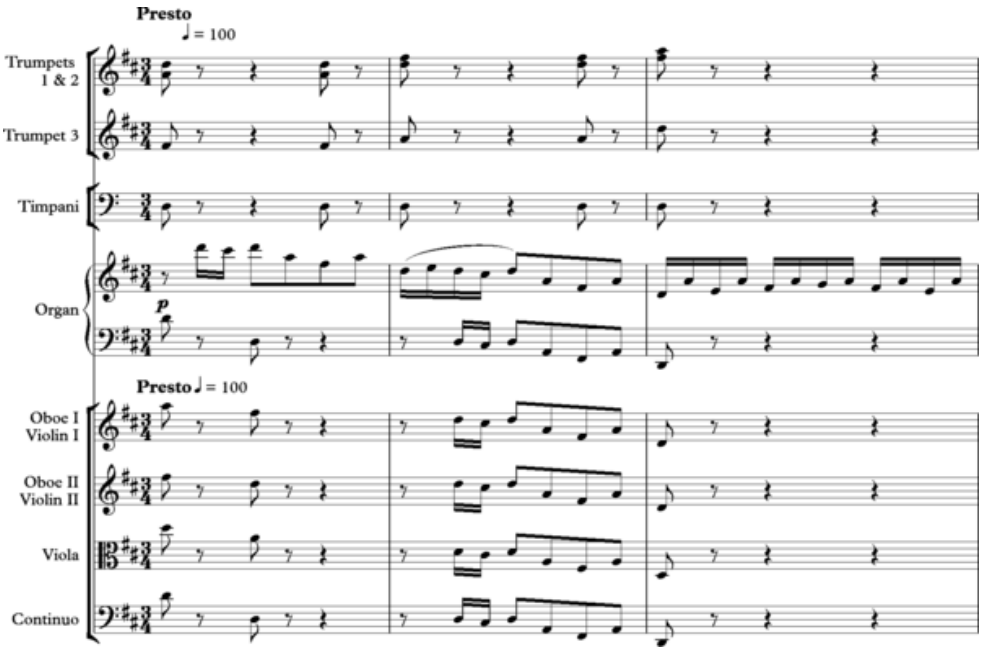
2.3.1 Ιστορικά Στοιχεία

Η μεταφορά της μουσικής σε διαφορετικά όργανα παρατηρείται ήδη απ' τα τέλη του Μεσαίωνα όπου και χρησιμοποιείται για πρώτη φορά το πολυφωνικό σύστημα. Στην Αναγέννηση, όπου και εμφανίζονται πολλά νέα όργανα η ανάγκη αυτή αυξάνεται αφού τα περισσότερα φωνητικά έργα πλέον εκτελούνται με την συνοδεία τσέμπαλου ή λαούτου. Στην περίοδο του Μπαρόκ το ενδιαφέρον για την μεταφορά μουσικής χάνεται. Εξαίρεση αποτελεί ο Johann Sebastian Bach ο οποίος είναι και ο πρώτος που ασχολήθηκε συστηματικά με την μεταφορά μουσικής, την δημιουργία νέων εκτελέσεων και την ενορχήστρωση τους. Έτσι ο Bach, μετέφερε πολλά κονσέρτα βιολιού του Antonio Vivaldi για τσέμπαλο και εκκλησιαστικό όργανο. Ταυτόχρονα πραγματοποίησε και διαφορετικές εκτελέσεις δικών του έργων. Χαρακτηριστικό παράδειγμα αποτελεί η Παρτίτα νούμερο 3 για σόλο βιολί (BWV 1006), της οποίας ένα μικρό μέρος παρουσιάζεται παρακάτω



Σχήμα 5: Παρτίτα νούμερο 3 για βιολί του Μπαχ, Πρελούδιο.

Ο Μπαχ μετέτρεψε το σολιστικό αυτό έργο σε μια ορχηστρική συμφωνία ως εισαγωγή για μια Καντάτα του. Μάλιστα σύμφωνα με τον J. Mincham [4] έγινε με τέτοιο τρόπο που κανείς δεν μπορεί να καταλάβει ότι η βασική μελωδία προϋπήρχε της ενορχήστρωσης. Η σχετική παρτιτούρα παρουσιάζεται παρακάτω:



Σχήμα 6: Μπαχ Καντάτα 29

Κατά τον 19ο αιώνα η μεταφορά μουσικής έγινε πάλι διάσημη και χρησιμοποιήθηκε σε πολλά είδη όπως στην όπερα, σε μπαλέτα και σολιστικές σπουδές οργάνων. Στην σύγχρονη pop μουσική η μεταφορά μουσικής εμφανίζεται κυρίως ως διασκευές στις οποίες μπορεί να έχουν γίνει αλλαγές στον ρυθμό, την τονικότητα και το μέτρο. Θα μπορούσαμε να πούμε ότι ακόμα και τα remixes αποτελούν ένα είδος μεταφοράς μουσικής.

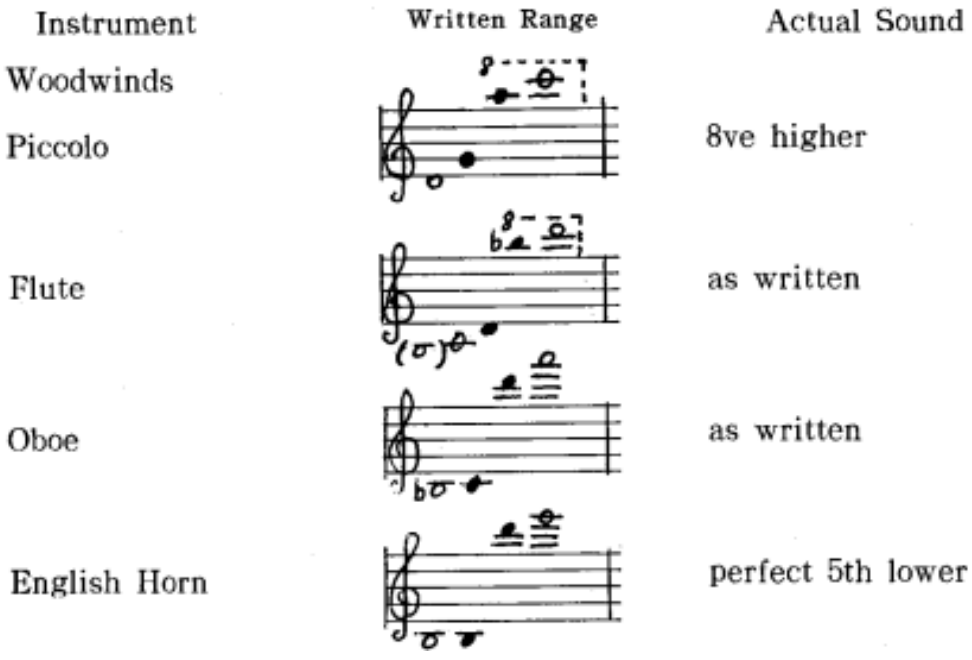
2.3.2 Η μεταφορά μουσικής ως τέχνη

Κατά μια έννοια η μεταφορά μουσικής μπορεί να θεωρηθεί ως το μουσικό αντίστοιχο της μετάφρασης ενός λογοτεχνικού έργου. Οι διάφορες φωνές και τα όργανα αποτελούν την "γλώσσα" με την οποία ο συνθέτης εξωτερικεύει τις σκέψεις και τα συναισθήματα του. Έτσι σκοπός της μεταφοράς μουσικής είναι να γίνει κατανοητό ένα μουσικό έργο σε κάποια διαφορετική "γλώσσα".

Σύμφωνα με το [6] υπάρχουν δύο διαφορετικά είδη μεταφοράς μουσικής. Το πρώτο έχει ως σκοπό την εξοικείωση με την τέχνη της μεταφοράς και καθαρά εκπαιδευτικό σκοπό. Γι' αυτό και είναι πολύ αυστηρός καθώς δεν επιτρέπεται να προστεθεί κάτι διαφορετικό στην ήδη υπάρχουσα σύνθεση. Το δεύτερο είδος μεταφοράς μουσικής προορίζεται για δημόσια παρουσίαση. Σε αυτή την περίπτωση το νέο μουσικό έργο έχει την ίδια υπόσταση με το αρχικό αλλά εκφράζεται μέσω ενός διαφορετικού συνόλου οργάνων. Επιπλέον, η μεταφορά αυτή είναι πολύ πιο ελεύθερη καθώς επιτρέπονται αλλαγές στα μουσικά θέματα.

Κατά την διαδικασία της μεταφοράς μουσικής θα πρέπει να δοθεί ιδιαίτερη προσοχή σε μία σειρά από παράγοντες ώστε να προκύψει ένα ικανοποιητικό αποτέλεσμα. Σημαντικοί τέτοιοι παράγοντες είναι η έκταση των διαφόρων φωνών και οργάνων, ο χρωματισμός των διαφόρων φωνών αλλά και η επιλογή των συγχορδιών και γενικότερα της εναρμόνισης. Όσον αφορά τις εκτάσεις των φωνών θα πρέπει να υπάρχει ιδιαίτερη προσοχή ως προς τη μίξη τους, δηλαδή το αν θα βρίσκονται σε διαφορετικές θέσεις ή αν θα επικαλύπτονται. Όσον αφορά τον τονικό χρωματισμό θα πρέπει να δοθεί προσοχή στην επιλογή των ηχοχρωμάτων, των φωνητικών και μελωδικών γραμμών αλλά και της δυναμικής κάθε οργάνου και φωνής ξεχωριστά. Προφανώς, θεωρητικές γνώσεις μουσικής είναι ως ένα βαθμό απαραίτητες ώστε να γίνει σωστά η παραπάνω διαδικασία. Πιο συγκεκριμένα, ο συνθέτης θα πρέπει να έχει βασικές γνώσεις αρμονίας, αντίστιξης και προφανώς ενορχήστρωσης ώστε να πραγματοποιεί τις σωστές επιλογές των φωνών και των οργάνων.

Κάθε όργανο, πέρα από την μουσική του έκταση, διαθέτει ένα υποσύνολο αυτής της έκτασης στο οποίο αποδίδει καλύτερα. Αυτό είναι πιο εμφανές στα πνευστά αλλά και στις ανθρώπινες φωνές που δεν έχουν την δυνατότητα να παίζουν πολύ υψηλές ή χαμηλές νότες της έκτασης τους για πολύ ώρα και με μεγάλη ένταση. Έτσι λοιπόν, κατά τη μεταφορά μουσικής, ο συνθέτης θα πρέπει να λάβει υπόψιν αυτά τα χαρακτηριστικά και να επιλέξει τις κατάλληλες στιγμές, πιθανώς κάποιο κρεσέντο, ώστε να υπάρξει κάποια μορφή κορύφωσης της σύνθεσης. Παρακάτω παρουσιάζεται ένα μικρό διάγραμμα με τις εκτάσεις των φωνών ορισμένων πνευστών οργάνων αλλά και των πραγματικών τους εκτάσεων, μέσα στις οποίες αποδίδουν καλύτερα.



Σχήμα 7: Μουσικές εκτάσεις του Πίκολο (μικρό φλάουτο), Φλάουτο, Όμποε και Αγγλικής κόρνας

Ως επόμενο βήμα, θα πρέπει να γίνει η μεταφορά σε κάθε όργανο ώστε να υπάρχει μια ροή και μια συνοχή στις μελωδίες. Ιδιαίτερα, στην κλασική μουσική τα συνοδευτικά όργανα θα πρέπει να ακολουθούν μικρές κινήσεις, οι μελωδίες τους δηλαδή να κινούνται είτε διατονικά είτε με διαστήματα τρίτης. Μεγαλύτερες μεταβολές συνήθως ορίζουν την αρχή μιας νέας μελωδικής γραμμής ή φράσης. Στην συνέχεια θα πρέπει να καθοριστούν οι αποστάσεις μεταξύ των φωνών και των οργάνων. Σε αυτή την περίπτωση, ακόμα και στην κλασική μουσική οι κανόνες είναι πολύ χαλαροί, ενώ μεγάλη αυστηρότητα υπάρχει στις αποστάσεις των φωνών σε ένα φωνητικό έργο ή σε μια χορωδία, καθώς δεν πρέπει να υπάρχει απόσταση μεγαλύτερη της οκτάβας μεταξύ δύο διαδοχικών φωνών, ενώ ταυτόχρονα οι φωνές απαγορεύεται να επικαλύπτονται. Προφανώς, όλοι οι παραπάνω κανόνες και συστάσεις στην σύγχρονη μουσική και ιδιαίτερα στην Jazz σχεδόν αγνοούνται.

Όσον αφορά τις τεχνικές σύνθεσης, ο συνθέτης μπορεί να επιλέξει είτε αρμονική γραφή είτε συνεχόμενα παράλληλα διαστήματα. Στην πρώτη περίπτωση κάθε όργανο και φωνή ακολουθεί μια διαφορετική μελωδία η οποία στην κλασική μουσική πρέπει να ακολουθεί τις κανόνες που αναφέρθηκαν παραπάνω. Όλες οι μελωδίες βασίζονται πάνω σε κοινές συγχορδίες ώστε να υπάρχει μια αρμονική συνοχή. Στην δεύτερη περίπτωση οι φωνές "ντουμπλάρονται" και ακολουθούν παρόμοια μελωδική γραμμή με ίσες αποστάσεις. Η πιο κοινή είναι η ταυτοφωνία η οποία χρησιμοποιείται για να δοθεί έμφαση στη μελωδία αλλά και για να αποδοθούν διαφορετικοί χρωματισμοί. Άλλα κοινά διαστήματα είναι τα τρίτης, πέμπτης, έκτης , οκτάβες και δεκάτης. Ιδιαίτερο ρόλο στη μορφή του νέου έργου που θα προκύψει με την μεταφορά παίζει η επιλογή της συνοδείας (accompaniment), καθώς καθορίζει τη ρυθμική βάση του έργου και ολοκληρώνει την αρμονική μορφή του. Επιπλέον η συνοδεία μπορεί να προσδώσει μια ξεχωριστή αίσθηση και χρωματισμούς στην μελωδία τονίζοντας συγκεκριμένα σημεία της.

3 Αναπαράσταση Δεδομένων σε Παραγωγικά Δίκτυα

Μια από τις σημαντικότερες αποφάσεις που πρέπει να λάβει κανείς όταν σχεδιάζει ένα νευρωνικό δίκτυο, ανεξαρτήτου εφαρμογής, είναι το πως θα αναπαρίστανται τα διαθέσιμα δεδομένα, ώστε να τροφοδοτηθούν στο δίκτυο και πιθανώς να παραχθούν αντίστοιχα δεδομένα από αυτό. Σε ό,τι αφορά τις ηχητικές αναπαραστάσεις, η ιδανικότερη επιλογή δεν είναι τόσο προφανής όσο είναι στην επεξεργασία εικόνας. Γι' αυτό τον λόγο, έχουν δοκιμαστεί πολλές διαφορετικές αναπαραστάσεις δεδομένων εισόδου για διαφορετικές εφαρμογές. Στην πραγματικότητα η "καλύτερη" αναπαράσταση εξαρτάται από το πρόβλημα και από τους διαθέσιμους πόρους (υπολογιστικές ικανότητες, διαθέσιμες GPUs και χρόνος). Όταν λοιπόν σχεδιάζουμε ένα νευρωνικό δίκτυο για επεξεργασία ήχου, καλούμαστε να επιλέξουμε μεταξύ μιας πληθώρας αναπαραστάσεων των δεδομένων, όπως είναι το απευθείας σήμα ήχου χωρίς επεξεργασία (raw audio), χαρακτηριστικά προσημασμένα με το χέρι, χαρακτηριστικά που έχουν εξαχθεί από κάποιο άλλο δίκτυο, Mel Frequency Cepstral Coefficients (MfCC), συμβολικές αναπαραστάσεις (π.χ. φωτογραφίες από παρτιτούρες ή MIDI αρχεία), αλλά και πολλούς φασματικούς μετασχηματισμούς.

Γενικά πρέπει να έχουμε υπόψιν ότι τα ακουστικά σήματα αποτελούνται από μεγάλο όγκο δεδομένων, όπου η σχετική πληροφορία με το πρόβλημα που αντιμετωπίζουμε είναι συνήθως κρυμμένη, και εξαπλωμένη σε μεγάλα χρονικά διαστήματα. Με αυτή την έννοια τα νευρωνικά δίκτυα μπορούν να επωφεληθούν εάν τροφοδοτηθούν με αραιές αναπαραστάσεις δεδομένων, στις οποίες ορισμένοι συντελεστές αποκαλύπτουν την απαραίτητη πληροφορία. Εάν λοιπόν το δίκτυο μας τροφοδοτηθεί με τέτοιου είδους πληροφορίες η εκπαίδευση θα γίνει γρηγορότερα και ταυτόχρονα θα απαιτείται λιγότερο πολύπλοκο δίκτυο.

Για πολλά χρόνια, ο σχεδιασμός, η επιλογή και η εξαγωγή των κατάλληλων χαρακτηριστικών θεωρούταν απαραίτητο βήμα για πολλά προβλήματα επεξεργασίας ήχου. Συνήθη τέτοια χαρακτηριστικά είναι το κέντρο βάρους του φάσματος του σήματος (spectral centroid) και στατιστικά υψηλότερης τάξης, ο ρυθμός διέλευσης από το μηδέν (zero crossing rate), η αρμονικότητα - δηλαδή ένα μέτρο της έντασης του περιοδικού σήματος ως προς τον θόρυβο σε dB (Harmonics-to-noise-ratio (HNR)) - , η θεμελιώδης συχνότητα και χρονικές περιγραφές του φάσματος.

Αντίθετα, τα τελευταία χρόνια, κυρίως λόγω της επικράτησης της βαθιάς μάθησης, η γενικότερη κατεύθυνση είναι να αφήσουμε το δίκτυο να επιλέξει τα κατάλληλα χαρακτηριστικά ώστε να αντιμετωπίσει το πρόβλημα τροφοδοτώντας στην είσοδο απευθείας το σήμα ήχου ή με μια ενδιάμεση αναπαράσταση, όπως είναι το σπεκτρόγραμμα. Αυτό μας έδωσε την δυνατότητα να σχεδιάζουμε μοντέλα που απαιτούν λιγότερη πρότερη γνώση, πάντα με κόστος δεδομένων, υπολογιστικών απαιτήσεων και χρόνου εκπαίδευσης.

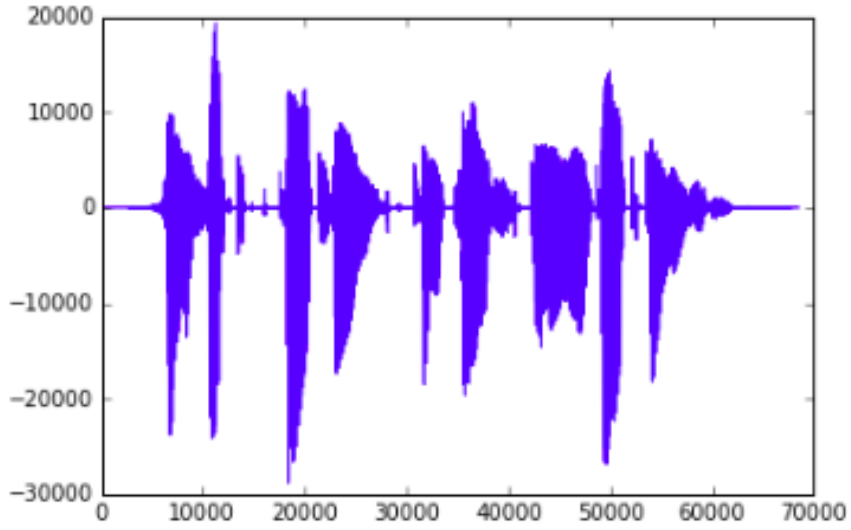
Το πρόβλημα που αντιμετωπίζουμε μπορεί να θεωρηθεί παραπλήσιο της αυτόματης σύνθεσης μουσικής από ένα νευρωνικό δίκτυο. Εξάλλου ένα μέρος του δικτύου συντίθεται από ένα

παραγωγικό δίκτυο, και πιο συγκεκριμένα από ένα παραγωγικό αντιπαραθετικό δίκτυο (Generative Adversarial Network). Είναι λοιπόν χρήσιμο να μελετήσουμε τις αναπαραστάσεις δεδομένων που χρησιμοποιούνται σε τέτοια δίκτυα για την παραγωγή μουσικής. Είναι προφανές ότι η αναπαράσταση που θα επιλεχθεί πρέπει να είναι ικανή να συνθέσει ήχο υψηλής ποιότητας. Συνεπώς, πρέπει να αποκλεισθούν όλες τις αναπαραστάσεις που έχουν απώλειες πληροφορίας, όπως είναι οι συντελεστές Mel-Frequency Cepstrum (MFCCs), και οι περισσότερες τεχνικές οι οποίες εξάγουν "χειροποίητα" προσημασμένα χαρακτηριστικά. Εντούτοις, πολλές διαφορετικές αναπαραστάσεις παραμένουν διαθέσιμες.

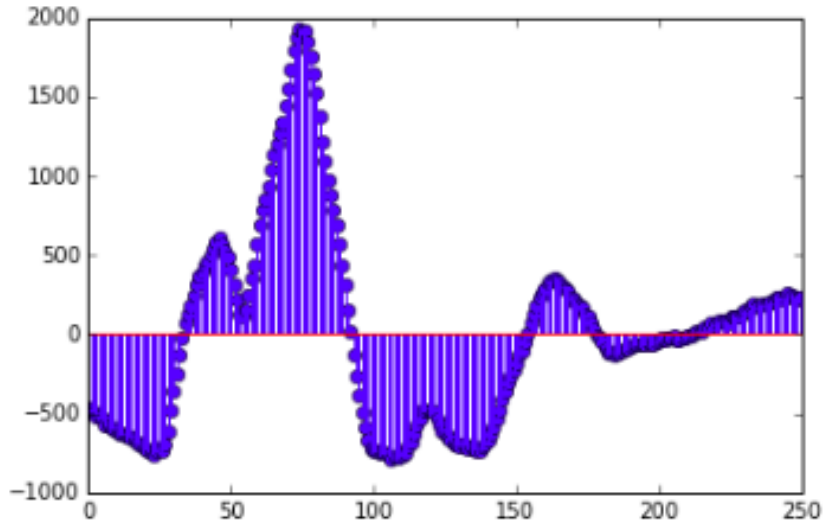
3.1 Κυματομορφή του σήματος ήχου (Raw Audio Waveform)

Μια αρκετά συνηθισμένη επιλογή είναι να τροφοδοτήσουμε το δίκτυο απευθείας με το σήμα ήχου. (raw audio). Αυτό αποτελείται από μια κυματομορφή σε διακριτή μορφή, δηλαδή από μια ακολουθία αριθμητικών δειγμάτων που καθορίζουν τις αριθμητικές τιμές του σήματος κατά βήματα t. Παρακάτω παρουσιάζεται ένα σήμα φωνής με ρυθμό δειγματοληψίας 16 kHz και μια μεγέθυνση του ίδιου σήματος ώστε να φαίνονται μόλις 200 δείγματα.

Η αναπαράσταση αυτή είναι αρκετά απαιτητική για τα νευρωνικά δίκτυα, ιδίως για το πρόβλημα της παραγωγής μουσικής. Σε αυτά τα προβλήματα λοιπόν, τα νευρωνικά δίκτυα τείνουν να μαθαίνουν χρονικά τοπικές δομές του σήματος και να αγνοούν συσχετίσεις μακράς διαρκείας οι οποίες είναι απαραίτητες για να υπάρχει μια δομή στον παραγόμενο ήχο. Εάν αντιμετωπίσουμε το προαναφερθέν πρόβλημα έχουμε την δυνατότητα να χειριστούμε απευθείας τα δεδομένα μας χωρίς καμία απώλεια, δίνοντας έτσι την δυνατότητα στο μοντέλο μας να εξάγει όλες τις απαραίτητες πληροφορίες που θα χρειαστεί για να παράξει το σήμα ήχου που θέλουμε. [9]



Time-domain speech signal sampled at 16 kHz.



Σχήμα 7: Διακριτό σήμα φωνής και μεγέθυνση του

Το πιο διαδεδομένο δίκτυο για την παραγωγή μουσικής που δέχεται στην είσοδο του το ακουστικό σήμα είναι το WaveNet. Το WaveNet είναι ένα συνελικτικό δίκτυο, το οποίο εκπαιδεύεται ώστε να προβλέψει το πιο πιθανό επόμενο δείγμα δοθείσας μιας ακολουθίας. Η βασική του διαφορά ως προς τα παραγωγικά αντιπαραθετικά δίκτυα (GAN) είναι ακριβώς αυτή η σειριακή λογική που ακολουθεί καθώς προβλέπει το ένα δείγμα μετά το άλλο, σε αντίθεση με τα παραγωγικά αντιπαραθετικά δίκτυα τα οποία προβλέπουν όλα τα δείγματα παράλληλα το ένα με το άλλο. Οι

δύο αυτοί τύποι δικτύου θα αναλυθούν εκτενέστερα στα δύο επόμενα κεφάλαια. Ωστόσο, πρέπει να αναφέρουμε ότι το βασικό μειονέκτημα του WaveNet είναι ότι οι πρακτικές του υλοποιήσεις χρησιμοποιούν δίκτυα με 60 ή και περισσότερα στρώματα, με αποτέλεσμα να απαιτείται πολύς χρόνος και υπολογιστική δύναμη για την εκπαίδευση τους. Παράλληλα, οι ρυθμοί δειγματοληψίας τους είναι από 16kHz έως 48kHz ανά δευτερόλεπτο, με αποτέλεσμα η σύνθεση του ήχου να είναι πολύ αργή, καθώς για την παραγωγή μόλις λίγων δευτερολέπτων απαιτείται επεξεργασία μερικών λεπτών.

3.2 Μετασχηματισμοί στο πεδίο της Συχνότητας

Η νέα αναπαράσταση που προκύπτει από τέτοιους μετασχηματισμούς είναι το σπεκτρόγραμμα. Το σπεκτρόγραμμα είναι μια "εικόνα" δύο διαστάσεων που αναπαριστά μια φασματική ακολουθία, με τον οριζόντιο άξονα να αντιπροσωπεύει τον χρόνο και τον κάθετο τις συχνότητες, ενώ η φωτεινότητα ή το χρώμα δηλώνουν την ένταση του φασματικού περιεχομένου σε μια δεδομένη χρονική στιγμή t.

Τέτοιοι μετασχηματισμοί χρησιμοποιούνται πλέον ευρέως για δύο βασικούς λόγους. Πρώτον, το υπολογιστικό κόστος είναι αρκετά μικρό, καθώς απαιτείται μια σειρά από μετασχηματισμούς Fourier Βραχέως Χρόνου (STFT) οι οποίοι ταυτόχρονα είναι και αντιστρέψιμοι, και δεύτερον το σπεκτρόγραμμα, δηλαδή ο τύπος δεδομένων που θα τροφοδοτήσουμε πλέον το μοντέλο μας στην είσοδο του, μπορεί να θεωρηθεί ως μια εικόνα. Με αυτή την έννοια έχουμε την δυνατότητα να εφαρμόσουμε αρχιτεκτονικές με συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks (CNNs)), τα οποία χρησιμοποιούνται ευρέως σε εφαρμογές επεξεργασίας εικόνας και όρασης υπολογιστών. Βασικοί τέτοιοι μετασχηματισμοί είναι το απλό Σπεκτρόγραμμα ή αλλιώς Γραμμικό σπεκτρόγραμμα, το σπεκτρόγραμμα σε λογαριθμική κλίμακα και το Mel σπεκτρόγραμμα.

3.2.1 Γραμμικό Σπεκτρόγραμμα

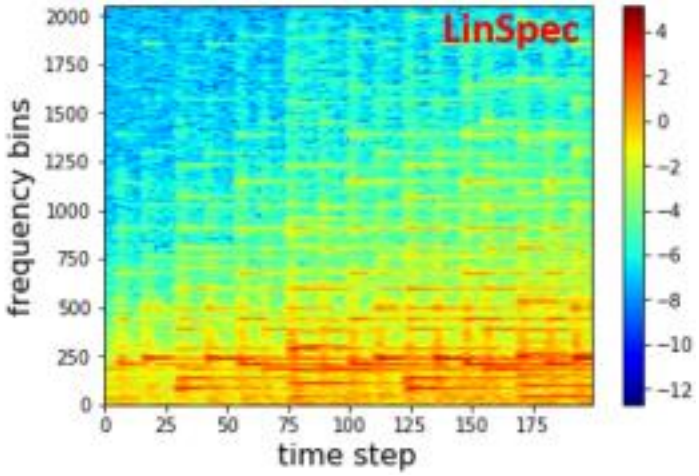
Το απλό σπεκτρόγραμμα είναι το αποτέλεσμα ενός μετασχηματισμού Fourier βραχέως χρόνου (STFT) ο οποίος αναλύει ένα σήμα ως ένα σταθμισμένο άθροισμα από μιγαδικά διανύσματα ημιτονικής βάσης. Οι κεντρικές συχνότητες του μετασχηματισμού είναι γραμμικά κατανεμημένες. Με αυτόν τον μετασχηματισμό ανακύπτει η δομή συχνότητας - χρόνου ενός ακουστικού σήματος. Πιο αναλυτικά ο μετασχηματισμός βραχέως χρόνου προκύπτει ως εξής:

$$X[k,\tau] = \sum_{n=0}^{N-1} x\,[n + ht] \cdot e^{-2\pi i k \frac{n}{N}} \qquad (4)$$

όπου k είναι οι κολόνες συχνότητας (frequency bins), τ το βήμα στον άξονα του χρόνου, N το μέγεθος του παραθύρου του μετασχηματισμού, n ένας δείκτης της κυματομορφής εισόδου και h το μέγεθος του βήματος. Ουσιαστικά κάθε κολόνα στο πεδίο της συχνότητας αντιστοιχεί σε μια πραγματική συχνότητα f μέσω της σχέσης:

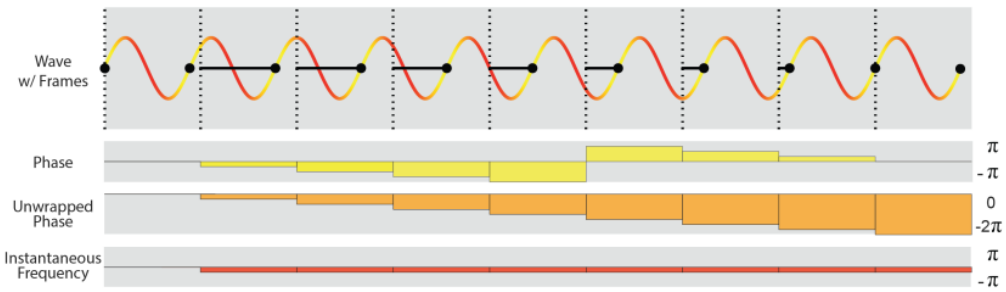
$$f = k\frac{s}{N} \qquad (5)$$

όπου s είναι ο ρυθμός δειγματοληψίας. Παρακάτω παρουσιάζεται το σπεκτρόγραμμα ενός σήματος μουσικής:



Σχήμα 8: Γραμμικό Σπεκτρόγραμμα Πρελούδιου του Bach

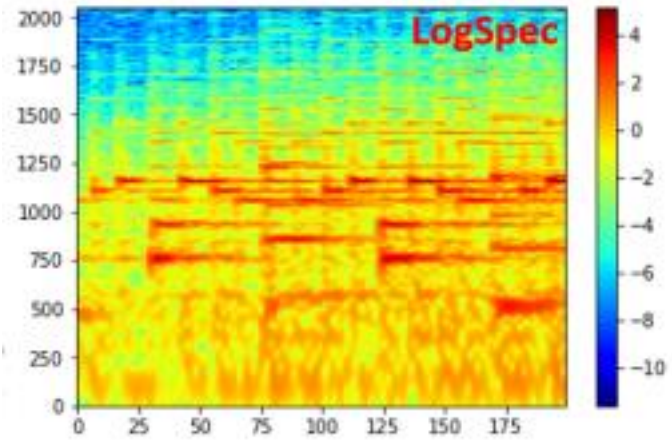
Από τον μετασχηματισμό Fourier προκύπτει και η συνιστώσα της φάσης, η οποία όμως εμφανίζει ασυνέχεια όταν υπερβαίνει το π ή πέφτει κάτω από αυτό**.** Τα δίκτυα σύνθεσης μουσικής πρέπει να διατηρούν μια συνέχεια στην έξοδο τους καθώς η ανθρώπινη αντίληψη είναι πολύ ευαίσθητη σε τέτοιες ασυνέχειες σε μικρές ή μεσαίες κλίμακες χρόνου, δηλαδή από 1- 100 ms. Όταν ο ρυθμός δειγματοληψίας δεν είναι ακριβώς ίσος με την περιοδικότητα του σήματος (κάτι που είναι αδύνατον να προβλέψουμε) έχουμε προήγηση φάσης. Μια τεχνική που χρησιμοποιείται ευρέως ([13], [14]) ώστε να λυθεί αυτό το πρόβλημα είναι η χρησιμοποίηση της ακαριαίας συχνότητας (instantaneous frequency (IF)). Σε αυτή την περίπτωση αναδιπλώνουμε την φάση προσθέτοντας 2π κάθε φορά που υπάρχει ασυνέχεια φάσης. Αυτό έχει ως αποτέλεσμα η φάση να αυξάνεται γραμμικά. Έτσι το παράγωγο της αναδιπλωμένης φάσης σε σχέση με τον χρόνο είναι ίσο με την κυκλική διαφορά μεταξύ του παραθύρου δειγματοληψίας και της περιοδικότητας του σήματος. Αυτό το μέγεθος ονομάστηκε ακαριαία συχνότητα καθώς μεταβάλλεται χρονικά ανάλογα με την αλλαγή της περιοδικότητας του σήματος (περιμένουμε ότι ένα σήμα φωνής δεν είναι ένας απλός τόνος με σταθερή περιοδικότητα αλλά ένα άθροισμα χρονικά μεταβαλλόμενων τόνων). Το παραπάνω φαινόμενο μπορεί να γίνει πολύ πιο κατανοητό απ' το παρακάτω σχήμα:



Σχήμα 9: Το πρόβλημα προήγησης φάσης και η ακαριαία συχνότητα στην απλή περίπτωση ενός μονοτονικού σήματος.

3.2.2 Σπεκτρόγραμμα Λογαριθμικής Κλίμακας

Σε αυτή την περίπτωση το σπεκτρόγραμμα προκύπτει πάλι απ΄ την εξίσωση (1) του Μετασχηματισμού Fourier Βραχέως Χρόνου. Η διαφορά τώρα είναι ότι η κατανομή των στηλών συχνότητας (frequency bins) είναι λογαριθμική αντί για γραμμική. Η βασική διαίσθηση μιας τέτοιας αλλαγής οφείλεται στην λογαριθμική σχέση των μουσικών νοτών. Περιμένουμε δηλαδή ότι η λογαριθμική κλίμακα θα λειτουργήσει καλά σε εφαρμογές σύνθεσης μουσικής αφού με τον κατάλληλο σχεδιασμό κάθε συχνοτική στήλη θα περιλαμβάνει μια μουσική νότα. Παρακάτω παρουσιάζουμε ένα αντίστοιχο σπεκτρόγραμμα λογαριθμικής κλίμακας. Σε σχέση με το γραμμικό σπεκτρόγραμμα μπορούμε να παρατηρήσουμε πόσο πιο απλωμένες είναι οι μικρές συχνότητες, φαινόμενο που συνεπάγεται την αύξηση της ανάλυσης στις χαμηλές συχνότητες.



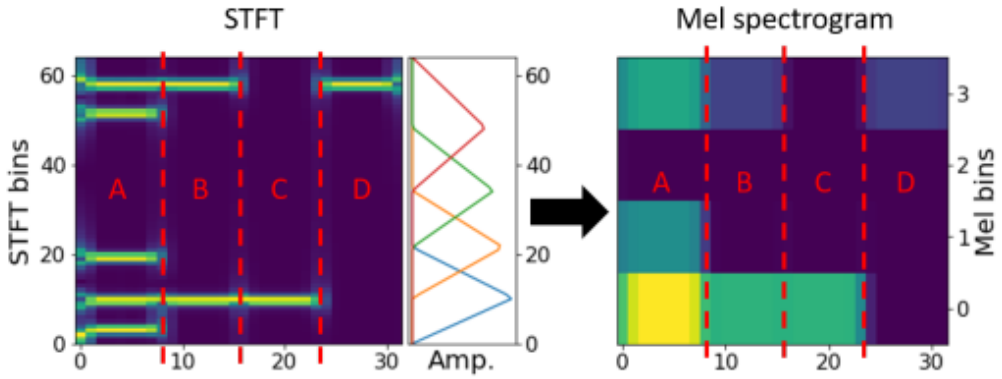
Σχήμα 10: Λογαριθμικό Σπεκτρόγραμμα Πρελούδιου του Bach.

Πράγματι, στο [10] έδειξαν ότι το σπεκτρόγραμμα λογαριθμικής κλίμακας έχει την δυνατότητα να διατηρήσει ένα πολύ μεγάλο ποσοστό χρήσιμης πληροφορίας με αρκετά μικρότερη ανάλυση

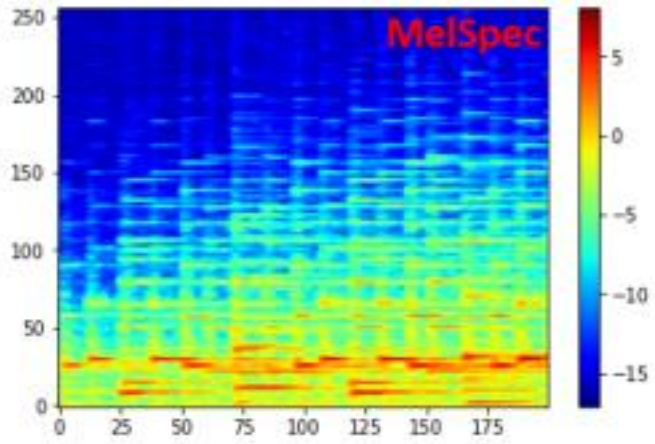
σε σχέση με το γραμμικό σπεκτρόγραμμα, δηλαδή με πολύ λιγότερες στήλες συχνότητας. Επιπλέον, εάν αυξήσουμε την ανάλυση στην συχνότητα, η βελτίωση που προκύπτει απ' το σπεκτρόγραμμα λογαριθμικής κλίμακας δεν είναι τόσο εμφανής, καθώς το μεγάλο πλήθος στηλών περιορίζει την χωρητική ικανότητα των στηλών.

3.2.3 Σπεκτρόγραμμα κλίμακας Mel

Η κλίμακα Mel είναι ένας μη γραμμικός μετασχηματισμός της κλίμακας των συχνοτήτων. Κατασκευάστηκε με τέτοιο τρόπο ώστε να προσομοιάζει την ανθρώπινη αντίληψη για την αλλαγή στις συχνότητες. Με αυτή την έννοια ήχοι με σταθερή απόσταση στην κλίμακα Mel, ακούγονται στους ανθρώπους ακριβώς στην ίδια απόσταση μεταξύ τους. Γνωρίζουμε δηλαδή ότι το ανθρώπινο ακουστικό μοντέλο αποτελείται από μια σειρά από μη γραμμικά φίλτρα, με αποτέλεσμα η διαφορά μεταξύ 500 και 1000 Hz να είναι πολύ εμφανής, ενώ η διαφορά μεταξύ 2500 και 3000 Hz να είναι σχεδόν μη αναγνωρίσιμη. Ο μετασχηματισμός λοιπόν της κλίμακας Mel προκύπτει με την εφαρμογή μιας τράπεζας ζωνοπερατών φίλτρων, το εύρος των οποίων μεγαλώνει, καθώς αυξάνεται η κεντρική συχνότητα. Ταυτόχρονα, σε αυτήν την περίπτωση παραβλέπουμε εντελώς την συνιστώσα της φάσης του σήματος. Γι' αυτό για την ανασκευή του αρχικού σήματος χρειάζεται να χρησιμοποιήσουμε αλγορίθμους βασισμένους στην μέθοδο των Griffin Lim (μέθοδος ανακατασκευής ενός σήματος από σπεκτρόγραμμα χωρίς την πληροφορία της ακαριαίας συχνότητας) ώστε να ανακτήσουμε την φάση του σήματος κατά την σύνθεση. Παρακάτω παρουσιάζουμε την μέθοδο εφαρμογής της τράπεζας των μη γραμμικών φίλτρων πάνω στις στήλες συχνότητας του απλού σπεκτρογράμματος αλλά και το σπεκτρόγραμμα που προκύπτει.



Σχήμα 11: Κάθε τράπεζα φίλτρων εφαρμόζεται σε πολλαπλές στήλες συχνότητας και τις μειώνει σε μόλις μια στήλη συχνότητας Mel.



Σχήμα 12: Mel Σπεκτρόγραμμα μετά την εφαρμογή της τράπεζας φίλτρων.

Όταν επιλέγουμε να αυξήσουμε την ανάλυση της κλίμακας Mel είναι πολύ πιθανό κάποια τράπεζα φίλτρων να είναι άδεια, δηλαδή να μην υπάρχει περιεχόμενο του σήματος. Αυτό έχει ως αποτέλεσμα μικρότερη αποτελεσματικότητα αυτής της αναπαράστασης. Εντούτοις, η τράπεζα φίλτρων Mel μπορεί να θεωρηθεί ως ένας πολύ αποτελεσματικός αλγόριθμος συμπίεσης για σπεκτρογράμματα, καθώς ενώ μειώνουμε το πλήθος των στηλών στον άξονα της συχνότητας, τα μοντέλα διατηρούν σε ένα μεγάλο βαθμό την ικανότητα τους.

4 Νευρωνικά Δίκτυα

Η έρευνα πάνω στα νευρωνικά δίκτυα είναι εμπνευσμένη από την δομή και την λειτουργία των νευρώνων του εγκεφάλου. Με αυτή την έννοια τα νευρωνικά δίκτυα αποτελούν υπολογιστικά μοντέλα τα οποία βασίζονται, ή τουλάχιστον σε αρχικό στάδιο βασίζονταν, σε μια δικτυακή μορφή παρόμοια με αυτή του εγκεφάλου. Το αντικείμενο των νευρωνικών δικτύων είναι η ανάπτυξη και η μελέτη μαθηματικών αλγορίθμων που έχουν την δυνατότητα να αντιμετωπίσουν ιδιαίτερα απαιτητικά προβλήματα, όπως είναι η αναγνώριση φυσικής γλώσσας, η αναγνώριση προσώπων και περιβάλλοντος, η επίτευξη αυτόματης πλοήγησης ενός ρομπότ σε περιβάλλον με φυσικά εμπόδια, η ανάπτυξη βέλτιστων στρατηγικών για ένα πρόβλημα, η εκτέλεση συλλογισμών που καταλήγουν σε λογικά συμπεράσματα και ιδανικά η προσαρμοστικότητα τους σε νέες καταστάσεις και σε άγνωστα περιβάλλοντα. Η βασική ιδέα για την αντιμετώπιση αυτών των προβλημάτων είναι η εκπαίδευση των νευρωνικών δικτύων και η μάθηση μέσα από τις εμπειρίες τους.

Πιο συγκεκριμένα, τα νευρωνικά δίκτυα πολλών επιπέδων (Multi-Layer neural networks) είναι μη γραμμικά μοντέλα τα οποία πραγματοποιούν μια αντιστοίχιση του διανύσματος εισόδου **x**, σε ένα διάνυσμα εξόδου **y**. Πολύ απλά το νευρωνικό δίκτυο μπορεί να περιγραφεί από μια συνάρτηση μεταφοράς της παρακάτω μορφής:

$$y = f(x; \Theta) \qquad (1)$$

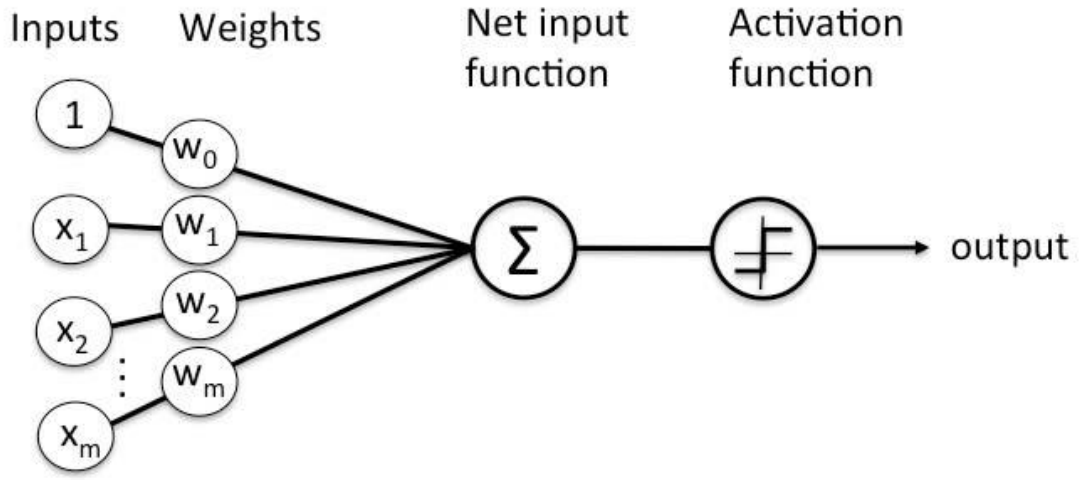
όπου Θ είναι το διάνυσμα των εκπαιδεύσιμων (trainable) παραμέτρων του συστήματος.

Η επιλογή των κατάλληλων παραμέτρων πραγματοποιείται κατά την διάρκεια της εκπαίδευσης, ώστε το δίκτυο να μάθει την βέλτιστη αναπαράσταση της εισόδου. Βέβαια, η επίτευξη της κατάλληλης αναπαράστασης εισόδου εξαρτάται πάντα τόσο από την τοπολογία του δικτύου όσο και από την επιλογή του αλγορίθμου εκπαίδευσης.

4.1 Feed Forward Neural Networks

4.1.1 Δίκτυο Perceptron

Το δίκτυο Perceptron είναι το θεμελιώδες στοιχείο ενός πλήρως συνδεδεμένου (fully connected) νευρωνικού δικτύου. Αποτελείται από μόλις ένα επίπεδο και έχει την δυνατότητα να χωρίζει τα δεδομένα σε δύο γραμμικά διαχωρίσιμες κλάσεις. Γι' αυτόν ακριβώς τον λόγο λέμε ότι είναι ένας Γραμμικός Δυαδικός Ταξινομητής (Linear Binary Classifier). Παρακάτω παρουσιάζεται ένα χαρακτηριστικό διάγραμμα ενός δικτύου Perceptron



Σχήμα 1: Δίκτυο Perceptron.

Η έξοδος του δικτύου **h**, υπολογίζεται ως το αποτέλεσμα μιας μη γραμμικής συνάρτησης **g**, η οποία ονομάζεται συνάρτηση ενεργοποίησης (Activation Function). Χαρακτηριστικές συναρτήσεις ενεργοποίησης αποτελούν η βηματική, η SoftMax αλλά και η σιγμοειδής, ενώ η επιλογή τους αποτελεί σημαντικό παράγοντα για την εκπαίδευση του δικτύου. Ο υπολογισμός της εξόδου δίνεται από τον παρακάτω τύπο:

$$\mathrm{h} = \mathrm{g}(\mathbf{x}^{\mathrm{T}}\mathbf{w} + w_0) = g(x;\theta) \qquad (2)$$

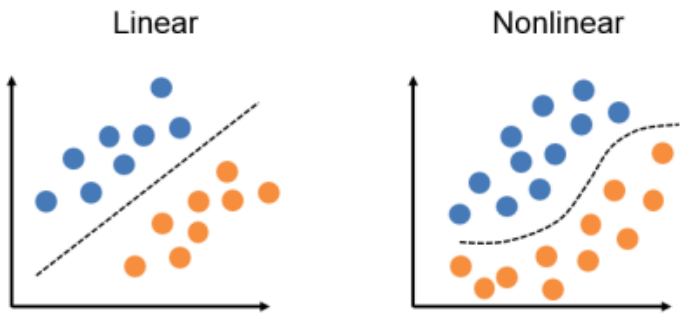
όπου **w** είναι το διάνυσμα βαρών ενώ w_0 η "μεροληψία" (bias) του δικτύου. Για διευκόλυνση της αναπαράστασης επαυξάνουμε το διάνυσμα **x** και ενσωματώνουμε το bias με τα βάρη του δικτύου στο διάνυσμα **Θ**.

Όπως ήδη αναφέρθηκε, ο νευρώνας Perceptron μπορεί να λύσει μόνο γραμμικά διαχωρίσιμα προβλήματα. Είναι λοιπόν κατανοητό ότι η υπολογιστική του ισχύς είναι αρκετά περιορισμένη. Για το λόγο αυτό, σχεδόν όλα τα σύγχρονα νευρωνικά δίκτυα που επιδιώκουν να λύσουν

περίπλοκα προβλήματα (τα οποία και είναι μη γραμμικώς διαχωρίσιμα) αποτελούνται από συνδέσεις νευρώνων.

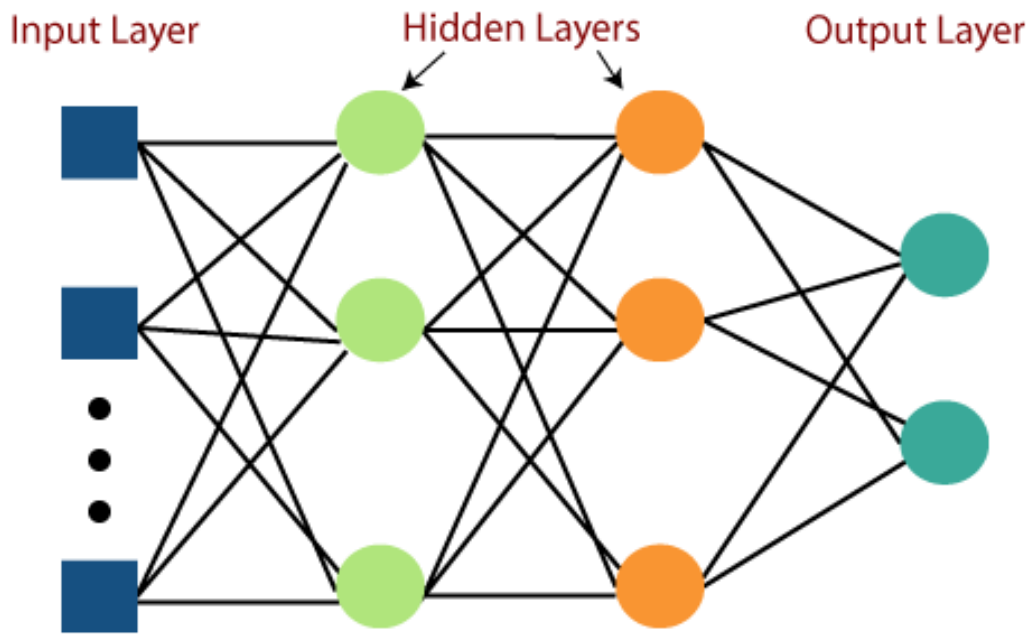
4.1.2. Δίκτυο Perceptron πολλών επιπέδων (Multi-Layer-Perceptron)

Το βασικό πρόβλημα ενός δικτύου Perceptron ενός νευρώνα είναι η αδυναμία του να λύσει γραμμικά διαχωρίσιμα σύνολα προτύπων. Γι' αυτόν ακριβώς τον λόγο, το 1986 έγινε η πρώτη εισαγωγή των Δικτύων Perceptron πολλών επιπέδων, αλλά και του βασικού αλγορίθμου back propagation. Σε μια τέτοια τοπολογία λοιπόν, οι διάφοροι νευρώνες Perceptron διαμερίζονται σε επίπεδα (layers). Τα γειτονικά επίπεδα συνδέονται με συνδέσεις μίας κατεύθυνσης (feed forward neural networks).



Σχήμα 2: Αριστερά, γραμμικά διαχωρίσιμο, και δεξιά, μη γραμμικά διαχωρίσιμο πρόβλημα. Στην πρώτη περίπτωση το όριο απόφασης είναι μια ευθεία γραμμή, ενώ στην δεύτερη περίπτωση όχι.

Το πρώτο επίπεδο, ονομάζεται επίπεδο εισόδου (input layer) και είναι αυτό το οποίο δέχεται το σήμα, ενώ το τελευταίο επίπεδο εξόδου (output layer) είναι αυτό το οποίο πραγματοποιεί την απόφαση ή την πρόβλεψη της εισόδου που δέχθηκε. Τα ενδιάμεσα επίπεδα, τα οποία έχουν αυθαίρετο πλήθος, ονομάζονται κρυφά επίπεδα (hidden layers) και αποτελούν την υπολογιστική ικανότητα του δικτύου. Ουσιαστικά, κάθε νευρώνας του **i**-οστού επιπέδου δέχεται ως είσοδο την έξοδο κάθε νευρώνα του **i-1** επιπέδου και με την σειρά του, τροφοδοτεί τους νευρώνες του επόμενου επιπέδου, όπως εξάλλου φαίνεται και στο παρακάτω σχήμα:



Σχήμα 3: Τοπολογία δικτύου Perceptron πολλών επιπέδων.

Εάν υποθέσουμε ότι το δίκτυο αποτελείται από **m** επίπεδα , τότε η έξοδος του **i**-οστού επιπέδου δίνεται από τον παρακάτω τύπο:

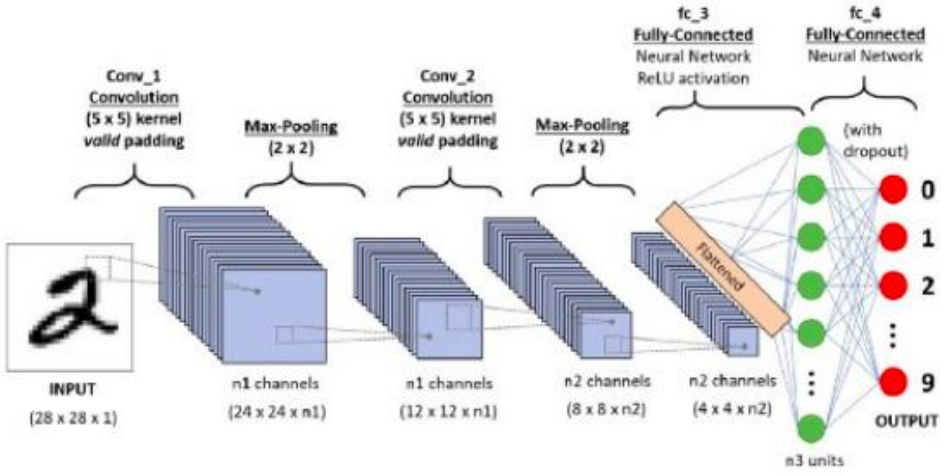
$$h^{(i)} = g^{(i)}\left(\boldsymbol{h}^{(i-1)T}\boldsymbol{W}(i) + \boldsymbol{w}_{i,0}\right) = g^{(i)}\left(\boldsymbol{h}^{(i-1)}; \boldsymbol{\Theta}^{(i)}\right) \qquad (3)$$

,όπου $h^{(i)}$η έξοδος του επιπέδου **i**, $g^{(i)}$ η συνάρτηση ενεργοποίησης των νευρνώνων του επιπέδου **i**, $W^{(i)}$ ο πίνακας των βαρών των συνδέσεων και $w_{i,0}$ η μεροληψία ή bias.

Κατά την εκπαίδευση του δικτύου, οι παράμετροι του ρυθμίζονται με σκοπό να ελαχιστοποιηθεί το σφάλμα μεταξύ της εξόδου του δικτύου και των πραγματικών δεδομένων. Αυτό γίνεται μέσω του αλγορίθμου backpropagation, ο οποίος συνδέει αυτό το σφάλμα με τις παραμέτρους του δικτύου. Για την μέτρηση του σφάλματος υπάρχουν πολλές επιλογές οι οποίες θα αναλυθούν στην συνέχεια.

4.2 Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks (CNNs))

Τα συνελικτικά δίκτυα είναι αυτά που έφεραν την μεγαλύτερη πρόοδο στον τομέα της όρασης υπολογιστών. Μάλιστα, έχουν γίνει πολλές προσπάθειες να χρησιμοποιηθούν οι δυνατότητες τους σε μοντέλα επεξεργασίας ήχου, συνήθως μεταφέροντας κάποιον επιτυχημένο αλγόριθμο επεξεργασίας εικόνας. Ωστόσο, είναι προφανές ότι ακόμα υπάρχει πολύ έδαφος που πρέπει να καλυφθεί ώστε να πλησιάσουν τα αντίστοιχα μοντέλα επεξεργασίας/ παραγωγής ήχου σε απόδοση αυτά της επεξεργασίας εικόνας. Στην περίπτωση των σπεκτρογραμμάτων η μεταφορά ενός αλγορίθμου επεξεργασίας εικόνας είναι πιο απλή καθώς και τα 2 σήματα εισόδου είναι 2 διαστάσεων. Πρέπει βέβαια να λάβουμε υπόψιν τις διαφορετικές συσχετίσεις μεταξύ των δύο διαστάσεων, καθώς στην περίπτωση της εικόνας μιλάμε για διαφορετικά pixel, ενώ στην περίπτωση ενός σπεκτρογράμματος οι δύο διαστάσεις αποτελούν τον άξονα των συχνοτήτων και τον άξονα του χρόνου. Στην συνέχεια λοιπόν θα αναλύσουμε τα συνελικτικά δίκτυα για επεξεργασία εικόνας. Οι βασικές ιδέες που έχουν αναπτυχθεί για την μεταφορά τους σε σπεκτρογράμματα θα αναλυθούν στην συνέχεια.



Σχήμα 4: Ένα συνελικτικό νευρωνικό δίκτυο που κατηγοριοποιεί χειρόγραφα ψηφία

Ένα από τα βασικά πλεονεκτήματα των CNNs είναι ότι δεν απαιτούν εκτενή προ-επεξεργασία των δεδομένων εισόδου. Πιο συγκεκριμένα, προγενέστερες μέθοδοι απαιτούσαν ορισμένα χαρακτηριστικά του σήματος να επιλεγούν και να εξαχθούν πριν τροφοδοτηθούν σε αυτά. Αντίθετα, εάν εκπαιδευτούν σωστά αλλά και για αρκετό χρονικό διάστημα, τα συνελικτικά δίκτυα έχουν την δυνατότητα να μάθουν από μόνα τους αυτά τα χαρακτηριστικά που θα τους είναι χρήσιμα για την επιτυχή αντιμετώπιση του προβλήματος.

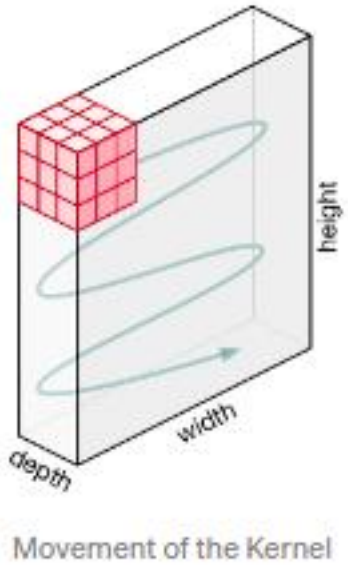
Τα δίκτυα Perceptron πολλών επιπέδων εξομαλύνουν την δισδιάστατη εικόνα εισόδου, σε ένα διάνυσμα μιας διάστασης. Ωστόσο, γειτονικά pixel μιας εικόνας αλληλοσχετίζονται. Αντίθετα, τα συνελικτικά δίκτυα έχουν την ικανότητα να εκμεταλλευτούν αυτές τις συσχετίσεις τόσο στο πεδίο του χώρου όσο και στο πεδίο του χρόνου (όταν μιλάμε για σήματα εισόδου άλλης μορφής όπως

μια αλληλουχία εικόνων (βίντεο)). Έτσι τα συνελικτικά δίκτυα επιτυγχάνουν παρόμοια αποτελέσματα με τα απλά νευρωνικά δίκτυα αλλά με πολύ μικρότερο πλήθος παραμέτρων.

Ένα άλλο βασικό πλεονέκτημα των συνελικτικών δικτύων είναι η επεκτασιμότητα τους σε πολύ μεγάλα Datasets. Στα περισσότερα προβλήματα, το νευρωνικό δίκτυο καλείται να επεξεργαστεί έναν πολύ μεγάλο όγκο δεδομένων, όπου κάθε δεδομένο μπορεί να αποτελείται από πολλές τιμές (π.χ. εικόνα 8K (7680 x 4320)). Η βασική ιδέα πίσω από τα συνελικτικά δίκτυα είναι να μειώσουν το μέγεθος των δεδομένων εισόδου εφαρμόζοντας διαδοχικά φίλτρα, τα οποία εξάγουν τα σημαντικά χαρακτηριστικά του σήματος.

4.2.1 Βασική Αρχιτεκτονική

Η λέξη συνελικτικό στην ονομασία του δικτύου υποδεικνύει ότι αξιοποιεί τη μαθηματική πράξη της συνέλιξης, η οποία αποτελεί και έναν γραμμικό υπολογισμό. Η πράξη της συνέλιξης πραγματοποιείται, στο πρώτο μέρος ενός συνελικτικού επιπέδου, που ονομάζεται φίλτρο ή πυρήνας. Το φίλτρο αυτό έχει βάθος όσο και το διάνυσμα εισόδου. Το φίλτρο που έχει επιλεγεί εφαρμόζεται πάνω στην είσοδο και μεταφέρεται πάνω σε αυτή με μια μετατόπιση (stride) που επιλέγει ο δημιουργός του μοντέλου. Σε κάθε θέση εφαρμογής του φίλτρου (kernel) πραγματοποιείται μια πράξη πολλαπλασιασμού πινάκων μεταξύ του φίλτρου και του μέρους της εισόδου πάνω στην οποία έχει εφαρμοστεί. Σχηματικά, η μετατόπιση αυτή παρουσιάζεται στο παρακάτω σχήμα:



Σχήμα 5: Μετατόπιση του φίλτρου πάνω στην είσοδο του συνελικτικού επιπέδου.

Σκοπός της παραπάνω διαδικασίας είναι να εξαχθούν τα high-level χαρακτηριστικά του σήματος εισόδου. Υποθετικά, όσο η είσοδος "προχωράει" απ' το ένα συνελικτικό επίπεδο στο επόμενο, τόσο πιο high level είναι τα χαρακτηριστικά που εξάγονται.

Μετά την εφαρμογή του φίλτρου ακολουθεί το γέμισμα (padding) της εικόνας. Σε αυτό το σημείο επιλέγεται η διάσταση της εξόδου απ' το συγκεκριμένο συνελικτικό επίπεδο. Ανάλογα με το padding που θα επιλεγεί (same padding ή valid padding), το μέγεθος του σήματος αυξάνεται, μειώνεται ή παραμένει το ίδιο.

Στην συνέχεια, ακολουθεί το pooling επίπεδο, το οποίο είναι υπεύθυνο για την μείωση του μεγέθους της εξόδου του συνελικτικού επιπέδου. Έτσι, μέσω του pooling επιπέδου επιτυγχάνεται η μείωση της απαιτούμενης υπολογιστικής ικανότητας του μοντέλου. Υπάρχουν κατά βάση δύο είδη pooling: το max pooling και το average pooling. Το πρώτο επιστρέφει την μέγιστη τιμή από το αποτέλεσμα της συνέλιξης του μεγίστου της εισόδου με το φίλτρο, ενώ το average pooling επιστρέφει τον μέσο όρο αυτών των τιμών. Μάλιστα, πολύ συχνά το max pooling χρησιμοποιείται και για αποθορυβοποίηση του σήματος εισόδου, καθώς αγνοεί εντελώς τον πρόσθετο θόρυβο. Γι' αυτόν τον λόγο το max pooling χρησιμοποιείται και συχνότερα.

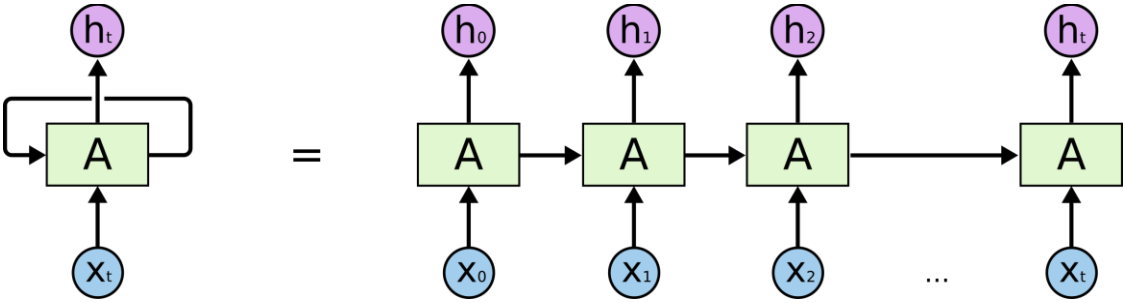
4.3 Αναδρομικά (Recurrent Neural Networks (RNNs))

Ένα κύριο πρόβλημα από τους τύπους νευρωνικών δικτύων που ήδη αναφέραμε είναι η έλλειψη μνήμης. Αυτή η έλλειψη είναι ιδιαίτερα εμφανής όταν τα δεδομένα εισόδου έχουν χρονικές συσχετίσεις, όπως σε ένα μουσικό μέρος ή σε μια αλληλουχία εικόνων ή βίντεο, καθώς τα δίκτυα που ήδη αναλύσαμε δεν έχουν την δυνατότητα να κρατάνε σημαντικές πληροφορίες από γεγονότα που προηγήθηκαν. Αυτή η βασική ιδέα, δηλαδή η διατήρηση μνήμης ή αλλιώς εσωτερικής κατάστασης ήταν το έναυσμα για την δημιουργία ενός νέου τύπου δικτύων, που ονομάστηκαν αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Networks (RNNs))

Ο τρόπος που τα αναδρομικά δίκτυα αντιμετωπίζουν το πρόβλημα της μνήμης είναι η δημιουργία ενός βρόχου (loop) που διατηρεί την πληροφορία στο δίκτυο σε διαφορετικές χρονικές στιγμές. Με βάση τα παραπάνω, τα RNNs φαίνονται ιδανικά για την αντιμετώπιση προβλημάτων που έχουν ως σύνολο δεδομένων ακολουθίες δεδομένων στις οποίες είναι έντονη η χρονική εξάρτηση.

4.3.1 Βασική Λειτουργία των RNNs

Η βασική λειτουργία των RNNs μπορεί να γίνει πιο κατανοητή εάν τα ξεδιπλώσουμε σε διαφορετικές χρονικές στιγμές, όπως φαίνεται στην παρακάτω εικόνα:



Σχήμα 6: Ένα αναδιπλωμένο αναδρομικό νευρωνικό δίκτυο

όπουx_t είναι ή είσοδος του δικτύου τη χρονική στιγμή t , και h_t είναι η αντίστοιχη έξοδος του.

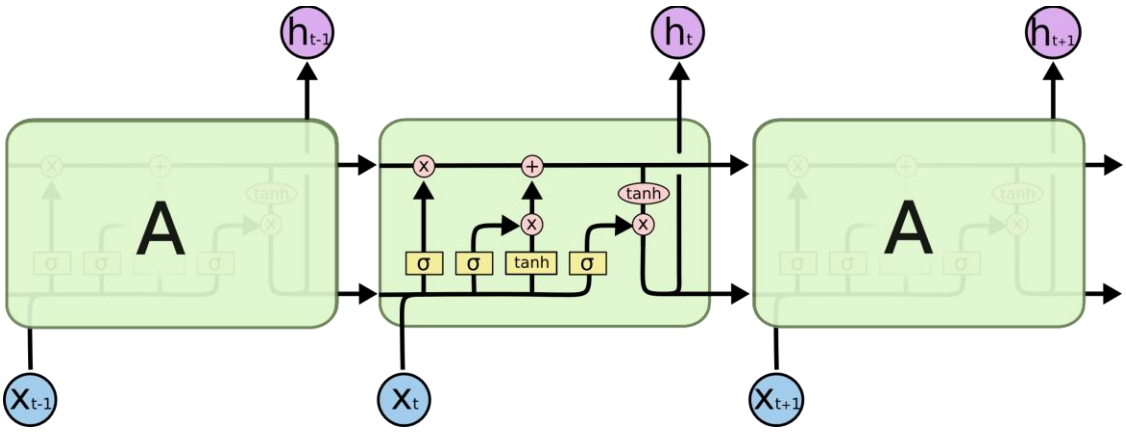
Με την αναδίπλωση αυτή μπορούμε να σκεφτούμε τα αναδρομικά δίκτυα ως πολλαπλές εκδοχές ενός δικτύου σε διαφορετικές χρονικές στιγμές. Επιπλέον, αυτό το χαρακτηριστικό κάνει εμφανή την σχέση μεταξύ των αναδρομικών δικτύων και των ακολουθιακών δεδομένων εισόδου. Γι' αυτόν ακριβώς τον λόγο έχουν χρησιμοποιηθεί εκτενώς σε προβλήματα αναγνώρισης φωνής, αναγνώρισης κειμένου αλλά και σε πιο σύνθετα όπως είναι η παραγωγή τέτοιου τύπου δεδομένων.

4.3.2 Long Short-Term Memory (LSTMs)

Ένα βασικό μειονέκτημα των RNNs είναι η αδυναμία τους να συνδέσουν πληροφορίες που εμφανίστηκαν αρκετά νωρίτερα από την δεδομένη χρονική στιγμή. Για να αυξηθεί η δυνατότητα τους να διατηρούν τέτοιες πληροφορίες πρέπει να αυξηθεί και η υπολογιστική πολυπλοκότητα. Έτσι, τα δίκτυα αυτά πρακτικά δεν είναι ιδιαίτερα χρήσιμα για μεγάλες ακολουθίες, ή για ακολουθίες που σημαντικά γεγονότα εμφανίζονται αραιά στο πεδίο του χρόνου. Θεωρητικά, μια πολύ προσεκτική επιλογή των παραμέτρων του δικτύου θα έδινε την δυνατότητα να διατηρούν πληροφορίες από μεγάλες ακολουθίες από δεδομένα. Ωστόσο, κατά την εκπαίδευση με τη μέθοδο back-propagation, τα διαφορικά, τα οποία διαδίδονται προς τις προηγούμενες χρονικές στιγμές, σε κάποια στιγμή θα μηδενιστούν ή θα απειριστούν. Το πρόβλημα αυτό ονομάζεται Vanishing/Exploding Gradient και είναι αυτό που καθιστά αδύνατη την εκπαίδευση του δικτύου σε μεγάλες ακολουθίες.

Μια λύση σε αυτό το πρόβλημα, δίνει μια συνθετότερη αρχιτεκτονική αναδρομικού δικτύου που ονομάζεται Long Short Term Memory (LSTM), τα οποία χρησιμοποιούνται ευρέως τα τελευταία χρόνια για την αντιμετώπιση προβλημάτων τέτοιου τύπου. Τα LSTMs μπορούν με τη σειρά τους να αναλυθούν ως πολλαπλές εκδοχές ενός δικτύου σε διαφορετικές χρονικές στιγμές. Ωστόσο, η

εσωτερική τους δομή είναι διαφορετική και πιο πολύπλοκη από αυτή των RNNs. Πιο συγκεκριμένα, ενώ το RNN αποτελείται από μόλις ένα επίπεδο νευρώνων, στο LSTM υπάρχουν τέσσερα διαφορετικά επίπεδα, όπως φαίνεται στο παρακάτω σχήμα.



Σχήμα 7: Ένα αναδιπλωμένο LSTM. Μέσα σε κίτρινα κουτάκια φαίνονται τα τέσσερα διαφορετικά επίπεδα του δικτύου.

5 Δεδομένα

Η εργασία αυτή βασίζεται σε δύο μοντέλα, το "MelGan" και το "A Universal Music Translation System", τα οποία και θα αναλυθούν στο επόμενο κεφάλαιο. Στην ενότητα αυτή θα εξετάσουμε τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση αυτών των δικτύων, αλλά και τα δεδομένα που επιλέξαμε εμείς για την εκπαίδευση του δικού μας μοντέλου, το οποίο εξάλλου αποτελεί και συνδυασμό των δύο προαναφερθέντων δικτύων. Πιο συγκεκριμένα, θα αναλύσουμε το NSynth Dataset, πάνω στο οποίο εκπαιδεύτηκε το "A Universal Music Translation System" αλλά και το "MusicNet" το οποίο αποτελεί ένα από τα Datasets πάνω στο οποίο εκπαιδεύτηκε το “A Universal Music Translation System” αλλά και το MelGan για ένα από τα subtasks του

5.1 NSynth

Το NSynth Dataset δημιουργήθηκε από την Google AI το 2017 [17] και είναι ίσως το πιο διαδεδομένο Dataset για την εκπαίδευση δικτύων παραγωγής και σύνθεσης μουσικής. καθώς διαθέτει μακράν τα περισσότερα δεδομένα για την αντιμετώπιση τέτοιων προβλημάτων (άλλα αντίστοιχα Datasets είναι των Goto et al. (2003), Romani et al.(2015)). Το NSynth αποτελείται από 306.043 νότες/ δείγματα, καθεμία με διαφορετικό τόνο, ηχόχρωμα ή/και περιβάλλουσα (Envelope). Κάθε δείγμα έχει συχνότητα 4 δευτερόλεπτα, είναι μονοφωνικό και έχει συχνότητα 16kHz. Στο σύνολο καλύφθηκε κάθε τόνος ενός Midi πιάνου (νότες 21-108) σε 5 διαφορετικές ταχύτητες. Η Midi ταχύτητα είναι σχεδόν συνώνυμη έννοια με την ένταση μιας νότας. Συνολικά χρησιμοποιήθηκαν οι ήχοι 1006 διαφορετικών μουσικών οργάνων. Προφανώς κάθε όργανο δεν έχει τις ίδιες δυνατότητες σε ότι αφορά το εύρος των νοτών που μπορεί να παράγει, γι’ αυτό και προέκυψε ένα σύνολο δεδομένων που αποτελείται από 306.043 δείγματα. Κάθε νότα έχει διάρκεια 3 δευτερόλεπτα και στο τελευταίο δευτερόλεπτο ο ήχος της σβήνει.

Ένα ακόμη πολύ σημαντικό χαρακτηριστικό του NSynth είναι ότι αποτελεί ένα προσημασμένο σύνολο δεδομένων. Αυτό σημαίνει πως σε κάθε δείγμα έχουν προστεθεί οι εξής τρεις πληροφορίες.

1) **Πηγή:** Ως επιλογές υπάρχουν "ακουστική", "ηλεκτρική, "συνθετική" ώστε να καθοριστεί ο τύπος του οργάνου που παρήγαγε τη νότα

2) **Οικογένεια:** Η οικογένεια οργάνων στην οποία ανήκει το όργανο που παρήγαγε τη νότα. (κάθε όργανο ανήκει ακριβώς σε μία ομάδα οργάνων)

3) **Ιδιότητες:** Προαιρετικό πεδίο. Περιλαμβάνει ξεχωριστά χαρακτηριστικά που μπορεί να έχει η νότα όπως το αν έχει distortion ή reverb.

5.2 MusicNet

Η βασική ιδέα που οδήγησε στο MusicNet [18] ήταν η δημιουργία ενός νέου συνόλου δεδομένων στον χώρο του Music Information Retrieval (MIR) που θα είναι αντίστοιχο με το ImageNet στον τομέα της επεξεργασίας εικόνας και όρασης υπολογιστών. Γι' αυτόν ακριβώς τον λόγο το MusicNet ενδείκνυται για μια ποικιλία προβλημάτων σχετικών με την εξαγωγή πληροφορίας από μουσική.

Το MusicNet λοιπόν περιλαμβάνει 330 ηχογραφήσεις κλασικής μουσικής, τα δικαιώματα των οποίων παρέχονται δωρεάν. Σε αυτές τις 330 ηχογραφήσεις υπάρχει μεγάλη ποικιλία από μουσικά όργανα και διαφορετικοί συνδυασμοί αυτών, αλλά και διαφορετικές ακουστικές συνθήκες και ηχογραφήσεις. Με βάση αυτά λοιπόν είναι κατανοητό το ότι οι συνθήκες του συγκεκριμένου συνόλου δεδομένων είναι πολύ πιο ελαστικές από το NSynth.

Κάθε ηχογράφηση περιλαμβάνει επιπλέον πληροφορίες (labeled dataset), οι οποίες είναι χωρισμένες σε 513 κλάσεις. Ως κλάση εννοούμε ένα διαφορετικό συνδυασμό νότας και μουσικού οργάνου. Στον παρακάτω πίνακα παρουσιάζονται συνοπτικά τα πιο σημαντικά χαρακτηριστικά του συνόλου δεδομένων.

MusicNet

Minutes	Labels	Recordings	Error Rate
2,048	1,299,329	330	4.0%

Ensemble	Minutes	Labels
Solo Piano	917	576,471
String Quartet	405	259,702
Accompanied Violin	148	124,886
Piano Quartet	73	60,362
Accompanied Cello	63	37,557
String Sextet	48	33,248
Piano Trio	46	28,873
Piano Quintet	25	27,545
Wind Quintet	43	24,820
Horn Piano Trio	30	18,799
Wind Octet	23	14,635
Clarinet-Cello-Piano Trio	25	13,447
Pairs Clarinet-Horn-Bassoon	24	12,218
Clarinet Quintet	26	11,184
Solo Cello	49	10,876
Accompanied Clarinet	20	10,049
Solo Violin	30	8,837
Violin and Harpsichord	16	7,469
Viola Quintet	15	4,156
Solo Flute	8	2,214

Composer	Minutes	Labels
Beethoven	1,085	736,072
Schubert	253	146,648
Brahms	192	133,109
Mozart	156	99,641
Bach	184	62,782
Dvorak	56	46,261
Cambini	43	24,820
Faure	33	22,349
Ravel	27	21,243
Haydn	15	6,404

Instrument	Minutes	Labels
Piano	1346	794,532
Violin	874	230,484
Viola	621	99,407
Cello	800	99,132
Clarinet	173	24,426
Bassoon	102	14,954
Horn	132	11,468
Oboe	66	8,696
Flute	69	8,310
Harpsichord	16	4,914
String Bass	38	3,006

	Piano	Violin	Cello	Viola	Clarinet	Bassoon	Horn	Oboe	Flute	Bass	Harpsichord
Notes	83	51	51	51	41	36	41	28	37	43	51

Σχήμα 1: Σύνοψη των στατιστικών του συνόλου δεδομένων MusicNet. Η τελευταία στήλη αντιστοιχεί στο σύνολο των νοτών που παίζονται για τον συγκεκριμένο συνδυασμό οργάνων.

Όπως μπορούμε να παρατηρήσουμε από το παραπάνω σχήμα, υπάρχει μεγάλη διασπορά σε ότι αφορά τα διαθέσιμα δεδομένα για κάθε συνδυασμό οργάνων αλλά και για κάθε καλλιτέχνη. Πιο συγκεκριμένα, υπάρχουν πολλά διαθέσιμα δεδομένα για έργα του Beethoven και για έργα πιάνου, ενώ αντίθετα δεν υπάρχει καλή εκπροσώπηση για όργανα όπως το όμποε ή το φλάουτο.

6 Βασικά Μοντέλα

Σε αυτό το κεφάλαιο θα αναλύσουμε τα δύο δίκτυα, "A Universal Music Translation System" και "MelGan" πάνω στα οποία θα βασιστεί το μοντέλο που θα αναπτύξουμε και θα εκπαιδεύσουμε στην συνέχεια. Η βασική δομή και των δύο μοντέλων είναι τα δίκτυα που αναλύσαμε στο κεφάλαιο 4.

6.1 A Universal Music Translation System

Το δίκτυο αυτό [18] αναπτύχθηκε το 2018 με σκοπό να μεταφράζει μουσικά έργα ή μέρη σε διαφορετικά μουσικά όργανα αλλά και είδη και στυλ μουσικής. Η βασική προσφορά του μοντέλου αυτού είναι ότι εκπαιδεύεται end-to-end, δηλαδή όλες του οι παράμετροι είναι εκπαιδεύσιμες, δεχόμενο ως είσοδο, απευθείας την κυματομορφή του ήχου. Η αρχιτεκτονική του δικτύου βασίζεται σε αυτοκωδικοποιητή (autoencoder) τύπου Wavenet, ο οποίος εκπαιδεύεται ταυτόχρονα σε πολλούς "τομείς" της εισόδου (multi-domain learning). Στην συγκεκριμένη περίπτωση, ως τομέα εισόδου εννοούμε τη μουσική που παράγει ξεχωριστά κάθε όργανο ή και το κάθε είδος μουσικής. Καταλαβαίνουμε λοιπόν ότι είναι απαραίτητο να ξεκινήσουμε την περιγραφή του δικτύου αυτού με την ανάλυση των Wavenets αλλά και των αυτοκωδικοποιητών.

6.1.1 Wavenet

Το Wavenet [12] είναι ένα βαθύ νευρωνικό δίκτυο που αναπτύχθηκε από την Deepmind το 2016, με σκοπό να αντιμετωπίσει το πρόβλημα της απευθείας σύνθεσης της κυματομορφής του ήχου (raw audio waveform generation). Το αρχικό πρόβλημα που επιχειρεί να λύσει το WaveNet είναι αυτό της παραγωγής ανθρώπινης ομιλίας για εφαρμογές text-to-speech. Ωστόσο, το μοντέλο επεκτείνεται με επιτυχία για την παραγωγή άλλων ηχητικών σημάτων, όπως είναι τα σήματα μουσικής. Για το πεδίο της παραγωγής ανθρώπινης ομιλίας λοιπόν, το Wavenet θεωρείται ως το πλέον σημαντικό μοντέλο και αποτελεί την βάση στην οποία στηρίχθηκαν επόμενα δίκτυα, καθώς η φωνή που αυτό παράγει είναι πολύ φυσική για το ανθρώπινο αυτί και προσομοιάζει σε πολύ μεγάλο βαθμό την ανθρώπινη ομιλία. Σε αυτό το κεφάλαιο θα κάνουμε μια σύντομη παρουσίαση αυτού του δικτύου.

Αυτοπαλινδρομικό Παραγωγικό Δίκτυο (Autoregressive Generative Model)

Τα παραγωγικά δίκτυα δέχονται ως είσοδο μη επισημειωμένα δεδομένα (unlabeled data) και επιχειρούν να "μάθουν" την κατανομή πιθανότητας που ακολουθούν αυτά τα δεδομένα. Στην συνέχεια, παράγουν νέα δεδομένα, τα οποία ιδανικά προσομοιάζουν τα δεδομένα εισόδου, με βάση αυτή την κατανομή πιθανότητας που έμαθαν. Για την εκμάθηση της κατανομής πιθανότητας των δεδομένων έχουν αναπτυχθεί κατά βάση δύο κατηγορίες μεθόδων. Στην πρώτη κατηγορία, η κατανομή πιθανότητας ορίζεται αναλυτικά και προσπαθεί να προσαρμοστεί στα μη επισημειωμένα δεδομένα εισόδου. Με αυτή την έννοια, οι παράμετροι της κατανομής πιθανότητας αποτελούν και εκπαιδεύσιμες παραμέτρους του μοντέλου. Στην δεύτερη μεθοδολογία, το μοντέλο επιχειρεί να μάθει έμμεσα την κατανομή των δεδομένων χωρίς αυτή να οριστεί εκ των προτέρων. Τα παραγωγικά αντιπαραθετικά δίκτυα (Generative Adversarial Networks (GANs)), τα οποία και θα παρουσιαστούν στην συνέχεια, αποτελούν χαρακτηριστικά παραγωγικά δίκτυα που κάνουν χρήση της δεύτερης μεθόδου. Αντίθετα, το Wavenet ανήκει στην πρώτη κατηγορία παραγωγικών δικτύων.

Πιο συγκεκριμένα, για μια ακολουθία εισόδου **X** , με **T** δείγματα, το Wavenet επιχειρεί να μοντελοποιήσει την από κοινού κατανομή πιθανότητας με την παρακάτω σχέση:

$$p(x) = \prod_{t=1}^{T} p\left(x_t | x_1, \dots, x_{t-1}\right) \qquad (1)$$

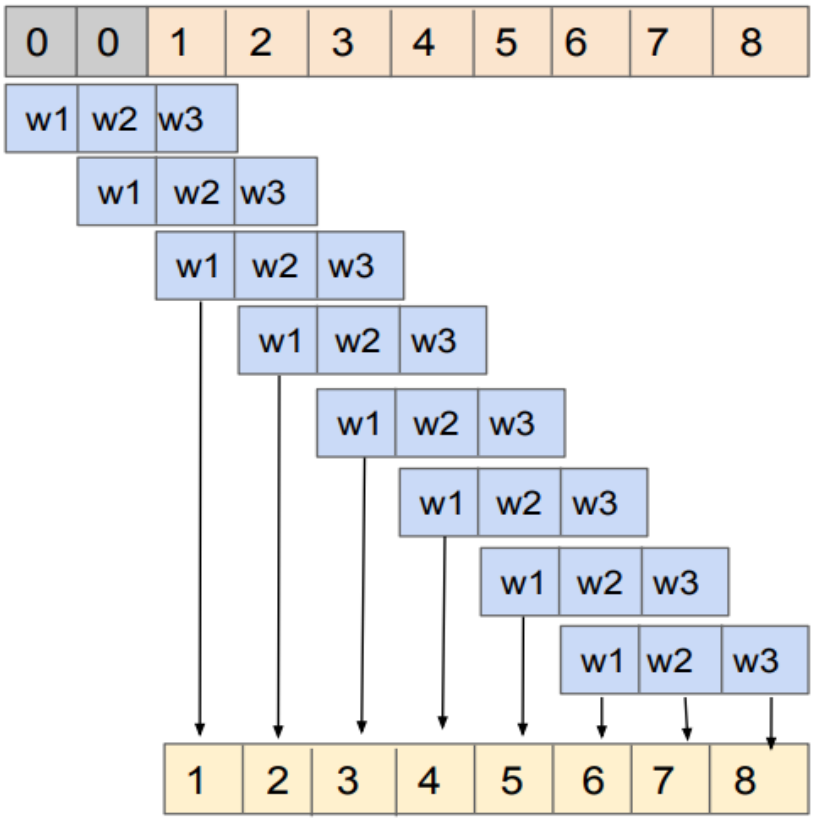
.δηλαδή ως το γινόμενο των εξαρτημένων πιθανοτήτων, για κάθε στοιχείο X_t στην ακολουθία. Με αυτή την έννοια, θεωρούμε ότι κάθε δείγμα εξαρτάται από όλα τα προηγούμενα. Διαφορετικά μπορούμε να πούμε ότι κάθε παρατήρηση εξαρτάται από τις παρατηρήσεις που προηγούνται. Ένα δίκτυο που δέχεται δεδομένα με αυτή την συσχέτιση λέμε ότι είναι "αυτοπαλινδρομικό" (autoregressive) (δεν χρησιμοποιείται ιδιαίτερα ο όρος στα ελληνικά)

Για την μοντελοποίηση κάθε όρου του παραπάνω γινομένου, δηλαδή για την μοντελοποίηση κάθε εξαρτημένης πιθανότητας, μια αρχική ιδέα θα ήταν να χρησιμοποιηθούν LSTMs. Αυτή είναι μια σχετικά προφανής επιλογή αφού τα LSTMs αποτελούν μη-γραμμικά ακολουθιακά δίκτυα. Μια σχετική υλοποίηση ακολουθεί το δίκτυο Pixel RNN [20], το οποίο χρησιμοποιείται για την παραγωγή συνθετικών εικόνων οι οποίες φαίνονται παρόμοιες με τις εικόνες εισόδου. Ωστόσο, το πρόβλημα με τα Wavenets είναι ότι δέχονται στην είσοδο τους μια κυματομορφή ήχου, η οποία έχει ρυθμό δειγματοληψίας τουλάχιστον 16 kHz. Αυτό σημαίνει ότι για 1 δευτερόλεπτο ηχητικού σήματος χρειάζονται 16.000 δείγματα. Όπως αναφέραμε και στο κεφάλαιο 5, το πρόβλημα των RNNs αλλά και των LSTMs σε μικρότερο βαθμό, είναι η μοντελοποίηση χρονικών εξαρτήσεων σε μεγάλη κλίμακα, καθώς όσο ο χρόνος αυξάνεται, το δίκτυο ξεχνάει την πληροφορία που εξήγαγε σε παλαιότερες χρονικές στιγμές. Ένα ακόμη πρόβλημα που προκύπτει στην χρήση

LSTM και RNN δικτύων είναι ο χρόνος εκπαίδευσης που απαιτείται, λόγω του ακολουθιακού τους χαρακτήρα.

Συνελικτικό Δίκτυο (CNN)

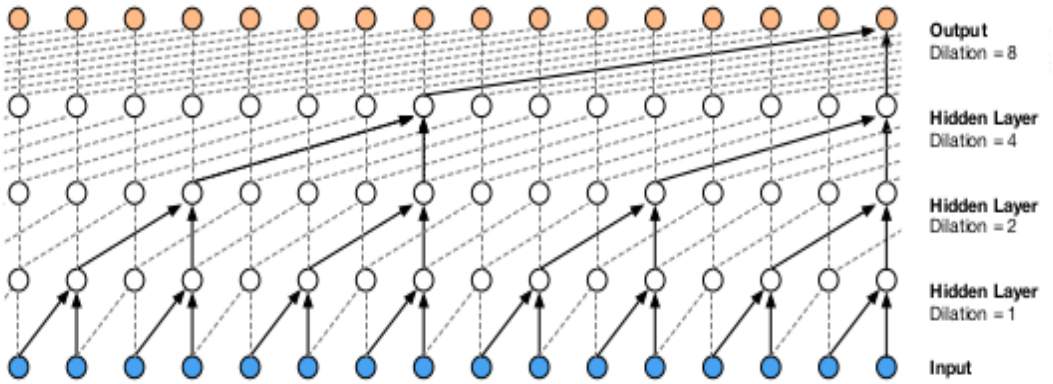
Για την αντιμετώπιση των παραπάνω προβλημάτων, το Wavenet αποτελείται από CNNs. Ο βασικός λόγος που αυτά επιλέχθηκαν είναι η ταχύτητα εκπαίδευσης τους, αφού είναι δυνατή η παράλληλη εκπαίδευση, σε αντίθεση με τα LSTMS τα οποία δέχονται μόνο ακολουθιακά δεδομένα. Για την αντιμετώπιση της εξάρτησης κάθε δείγματος από τα προηγούμενα (autoregressive property) χρησιμοποιήθηκε η τεχνική της αιτιατής συνελίξεις (Causal Convolution). Η βασική ιδέα για μονοδιάστατα δεδομένα είναι ότι εφαρμόζεται padding στην αρχή της ακολουθίας ώστε να μπορέσει να εφαρμοστεί το φίλτρο ή μάσκα του CNN. Το γέμισμα αυτό στην αρχή της ακολουθίας γίνεται με μηδενικά και μπορεί να γίνει κατανοητό με την παρακάτω εικόνα:



Σχήμα 1: Εφαρμογή του CNN φίλτρου με την μέθοδο της αιτιατής συνέλιξης (causal convolution)

Με βάση τα παραπάνω, καταλαβαίνουμε ότι η μέθοδος αυτή μας επιτρέπει να επιλέξουμε το πλήθος των προηγούμενων δειγμάτων από τα οποία θεωρούμε ότι εξαρτάται το τρέχον δείγμα. Αυτός ο όρος ονομάζεται look-back length. Ωστόσο, όπως ήδη είπαμε για 1 δευτερόλεπτο ήχου, χρειαζόμαστε τουλάχιστον 16.000 δείγματα. Είναι λοιπόν κατανοητό ότι το μοντέλο θα πρέπει να έχει look-back length τουλάχιστον 16.000 δειγμάτων. Η πιο απλή ιδέα θα ήταν το Wavenet να αποτελείται από συνελικτικά δίκτυα, τα οποία θα έχουν αντίστοιχο μήκος. Ωστόσο, με αυτή την λύση αυξάνεται σημαντικά το πλήθος των παραμέτρων του δικτύου και η πολυπλοκότητα του, κάτι που μεταφράζεται σε δυσκολότερη και πιο χρονοβόρα εκπαίδευση. Μια δεύτερη λύση θα ήταν να αυξηθούν τα επίπεδα του δικτύου. Ωστόσο προσθέτοντας επιπλέον επίπεδα, αυξάνεται η πολυπλοκότητα του δικτύου, δημιουργώντας ανάλογα προβλήματα με την πρώτη λύση. Ο βασικός λόγος που αυτό συμβαίνει είναι επειδή το δεκτικό πεδίο (receptive field) των συνελικτικών δικτύων, αλλά και το look-back length αυξάνονται γραμμικά με την αύξηση των επιπέδων του δικτύου. Σε αυτό το σημείο είναι χρήσιμο να αναφέρουμε ότι, ως δεκτικό πεδίο ορίζουμε το μέγεθος της περιοχής του πεδίου εισόδου ενός συνελικτικού δικτύου, που είναι υπεύθυνο για την παραγωγή ενός χαρακτηριστικού (feature). Στην πραγματικότητα, το δεκτικό πεδίο είναι μια συσχέτιση μεταξύ ενός χαρακτηριστικού εξόδου (σε οποιοδήποτε επίπεδο) με την περιοχή εισόδου.

Η μέθοδος που επιλέχθηκε ώστε να μην αυξηθούν τα επίπεδα του δικτύου αλλά και το μέγεθος των φίλτρων είναι αυτή των διασταλμένων συνελίξεων (dilated convolutions). Η βασική ιδέα αυτής της μεθόδου είναι η εφαρμογή του φίλτρου σε μια μεγαλύτερη περιοχή από το μέγεθος του, αγνοώντας ορισμένες τιμές εισόδου, με ένα σταθερό βήμα. Αυτό ισοδυναμεί με μια συνέλιξη με ένα μεγαλύτερο φίλτρο το οποίο προκύπτει από το αρχικό "διαστέλλοντας" το με μηδενικά. Στο Wavenet συγκεκριμένα, εφαρμόζονται πολλαπλές διασταλμένες συνελίξεις με διαφορετικό βήμα, ώστε να αυξηθεί το δεκτικό πεδίο του δικτύου. Στην πραγματικότητα, το Wavenet αποτελείται από πολλαπλά επίπεδα, με κάθε επίπεδο να έχει μεγαλύτερο "βήμα διαστολής" από το προηγούμενο. Η αύξηση αυτή του βήματος είναι μάλιστα εκθετική και ισοδυναμεί σε εκθετική αύξηση του δεκτικού πεδίου. Στο παρακάτω σχήμα γίνεται φανερή η υλοποίηση αυτής της μεθόδου:



Σχήμα 2: Διασταλμένες συνελίξεις στο Wavenet.

SoftMax Κατανομή

Για την μοντελοποίηση της εξαρτημένης πιθανότητας της σχέσης (1), εφαρμόζεται η SoftMax κατανομή, κυρίως επειδή δεν κάνει κάποια υπόθεση για την κατανομή των δεδομένων εισόδου. Με αυτό τον τρόπο, το μοντέλο έχει την δυνατότητα να μάθει αυθαίρετες κατανομές που τυχόν ακολουθούν τα δεδομένα εισόδου. Ωστόσο, το πρόβλημα που προκύπτει με την εφαρμογή ενός SoftMax επιπέδου είναι η αύξηση της υπολογιστικής πολυπλοκότητας του δικτύου. Πιο συγκεκριμένα, ένα ηχητικό σήμα συνήθως αναπαρίσταται ως μια ακολουθία από ακέραιες τιμές 16-bit (δηλαδή έχει εύρος [-32.768, 32.768]). Με αυτή την έννοια, σε κάθε χρονική στιγμή θα πρέπει το μοντέλο να υπολογίζει 65.535 τιμές. Γι' αυτό τον λόγο, πριν ακριβώς από το SoftMax επίπεδο μειώνεται το πλήθος των bit κάθε τιμής. Η βασική ιδέα για την μείωση των bit, είναι ότι σε χαμηλές τιμές απαιτείται μεγαλύτερη ακρίβεια. Γι' αυτό τον λόγο δεν εφαρμόζεται γραμμική μείωση των bit εισόδου, που χωρίζει τον χώρο σε σταθερά επίπεδα κβάντισης , αλλά μη γραμμική κβάντιση, γνωστή και ως μ-law companding. Ο τύπος αυτής της κβάντισης είναι ο παρακάτω:

$$f(x_t) = sign(x_t)\frac{ln(1+\mu|x_t|)}{ln(1+\mu)} \qquad (2)$$

, όπου ισχύει $-1 < X_t < 1$ όπου μ = 255. Με αυτό τον τρόπο, το δίκτυο έχει στην έξοδο του 256 τιμές, αντί για 65.536 που έχει στην είσοδο, για κάθε χρονική στιγμή μειώνοντας με αυτό τον τρόπο τον απαιτούμενο χρόνο εκπαίδευσης.

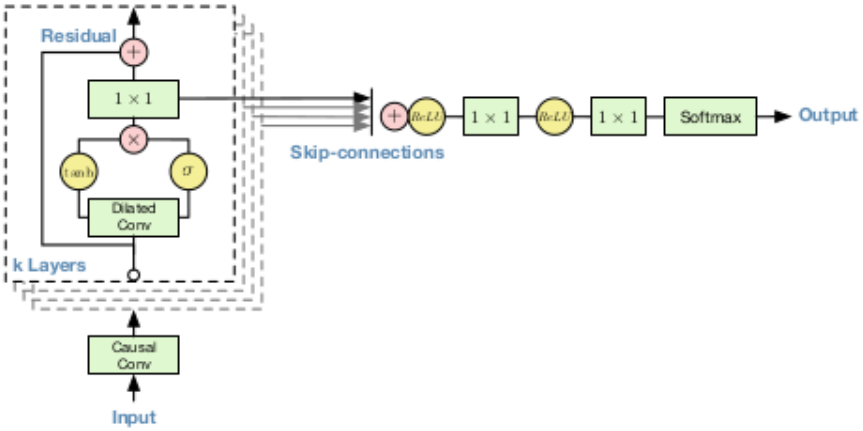
Συνάρτηση ενεργοποίησης, "residual" και "skip connections"

Αρχικά. για την εκπαίδευση του Wavenet επιλέχθηκε η ReLu συνάρτηση ενεργοποίησης. Ωστόσο, έπειτα από δοκιμές φάνηκε ότι μια μη-γραμμική "gated" συνάρτηση ενεργοποίησης παρουσιάζει καλύτερα αποτελέσματα. Η συνάρτηση αυτή χρησιμοποιείται και στο δίκτυο Pixel CNN και είναι η εξής:

$$\mathbf{z} = \tanh(\mathrm{W}_{\mathrm{f,k}} * \mathbf{x}) \odot \sigma(\mathrm{W}_{\mathrm{g,k}} * \mathbf{x}) \qquad (3)$$

, όπου **W** είναι το εκπαιδεύσιμο φίλτρο, * αναπαριστά την πράξη της συνέλιξης και το ⊙ αναπαριστά το εσωτερικό γινόμενο.

Επιπλέον, στο δίκτυο χρησιμοποιούνται και υπολειπόμενες συνδέσεις (residual connections) αλλά και παραλείψεις συνδέσεων (skip connections) [21]. Στην πρώτη περίπτωση στην έξοδο ενός επιπέδου, προστίθεται και η έξοδος του προηγούμενου ακριβώς επιπέδου, ενώ στην περίπτωση της παράλειψης σύνδεσης, στην έξοδο του δικτύου αυξάνεται απευθείας η έξοδος ενός αρχικού επιπέδου. Με αυτό τον τρόπο , μειώνεται ο χρόνος σύγκλισης, δηλαδή και ο χρόνος μάθησης του δικτύου. Παρακάτω παρατίθεται σχηματικά ένα residual block αλλά και εποπτικά το Wavenet δίκτυο:



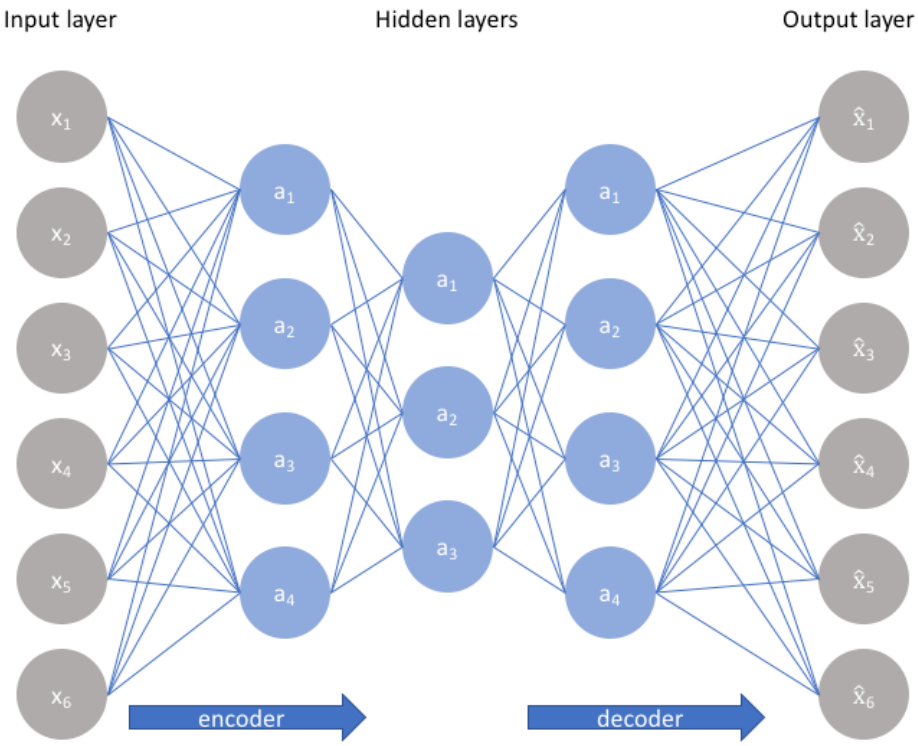
Σχήμα 3: Residual block και αρχιτεκτονική του Wavenet

6.1.2 Wavenet Αυτοκωδικοποιητής (Wavenet Autoencoder)

Το δίκτυο αυτό [22] δημιουργήθηκε το 2017 και αποτελεί επέκταση του δικτύου Wavenet που είδαμε και αναλύσαμε στο προηγούμενο κεφάλαιο. Βασικός του στόχος ήταν να επεκτείνει τη μελέτη στον τομέα της "Σύνθεσης Ήχου" (Audio synthesis), παράγοντας νέους τύπους εκφραστικών και ρεαλιστικών ήχων μουσικών οργάνων μέσα από νευρωνικά δίκτυα. Το απλό μοντέλο Wavenet έχει την δυνατότητα να μοντελοποιεί πολύ αποτελεσματικά ηχητικά σήματα μικρής και μεσαίας κλίμακας, δηλαδή γύρω στα 500 msec. Ωστόσο, η "συνεκτικότητα" και η πιστότητα του ήχου χάνεται για ηχητικά σήματα μεγαλύτερης διάρκειας. Μάλιστα, οι δημιουργοί του Wavenet, πρότειναν μια λύση στο συγκεκριμένο πρόβλημα, η οποία όμως βασίζεται σε εξωτερικές συνθήκες ("external conditioning"). Αντίθετα ο Wavenet αποκωδικοποιητής λύνει αυτό το πρόβλημα χωρίς κάποια εξωτερική συνθήκη, καθώς έχει την δυνατότητα να μάθει σημασιολογικά σημαντικές κρυφές αναπαραστάσεις τις οποίες και χρησιμοποιεί ως ένα σήμα ελέγχου ώστε να ελέγξει τον τόνο, την υφή αλλά και τις δυναμικές του ήχου κατά την παραγωγή του. Προκειμένου να γίνει κατανοητή η αρχιτεκτονική του συγκεκριμένου δικτύου, αρχικά, θα παρουσιάσουμε συνοπτικά τους αυτοκωδικοποιητές και την λειτουργία τους.

Αυτοκωδικοποιητές ("Autoencoders")

Ο αυτοκωδικοποιητής είναι ένα δίκτυο μη - επιβλεπόμενης μάθησης, το οποίο επιχειρεί να μάθει μια αναπαράσταση δεδομένων. Η βασική ιδέα είναι η συμπίεση των δεδομένων μέσα στο δίκτυο σε σημείο που επιβάλλεται ένα bottleneck, ώστε να προκύψει μια συμπιεσμένη και κρυμμένη αναπαράσταση των αρχικών δεδομένων. Στην συνέχεια ακολουθεί η ανακατασκευή των δεδομένων ώστε να δημιουργηθεί το σήμα εξόδου. Σε περίπτωση που τα δεδομένα εισόδου είναι ανεξάρτητα μεταξύ τους, τότε η υλοποίηση της παραπάνω ιδέας είναι αρκετά δύσκολη. Ωστόσο, υποθέτουμε ότι τα δεδομένα εισόδου ακολουθούν κάποια κατανομή ή κάπως συσχετίζονται. Με αυτή την συνθήκη, ο αυτοκωδικοποιητής μπορεί να μάθει αποτελεσματικά την δομή των δεδομένων και να την χρησιμοποιήσει για την εκπαίδευση του. Παρακάτω, παρουσιάζεται μια τυπική αρχιτεκτονική ενός δικτύου αυτοκωδικοποιητή:



Σχήμα 4: Αρχιτεκτονική αυτοκωδικοποιητή

Όπως γίνεται εμφανές και στο παραπάνω σχήμα, το μέρος του δικτύου που πραγματοποιείται η συμπίεση των δεδομένων εισόδου ονομάζεται κωδικοποιητής ("encoder"), ενώ το μέρος που πραγματοποιείται η αποσυμπίεση και ανακατασκευή των δεδομένων ονομάζεται αποκωδικοποιητής ("decoder"). Κατά την διαδικασία της εκπαίδευσης, το δίκτυο δέχεται ως είσοδο ένα μη επισημειωμένο (unlabeled) Dataset και επιχειρεί να ανακατασκευάσει τα δεδομένα εισόδου, δημιουργώντας τις εξόδους $\hat{x}$. Κατά την εκπαίδευση, βασικός στόχος είναι η ελαχιστοποίηση του σφάλματος ανακατασκευής ("Reconstruction Error") $L(x, \hat{x})$, το οποίο υπολογίζει τις διαφορές μεταξύ των δεδομένων εισόδου και των δεδομένων ανακατασκευής. Ο ρόλος του bottleneck στο δίκτυο είναι να αποτρέψει το δίκτυο από το να απομνημονεύει τις τιμές εισόδου, καθώς περιορίζει το πλήθος της πληροφορίας που μπορεί να διασχίσει μέσα απ' όλο το δίκτυο.

Ένας αποτελεσματικός αποκωδικοποιητής θα πρέπει να συνδυάζει τα παρακάτω δύο στοιχεία. Αρχικά πρέπει να είναι αρκετά ευαίσθητος στα δεδομένα εισόδου ώστε η ανακατασκευή τους να είναι αποτελεσματική. Από την άλλη, πρέπει ταυτόχρονα να μην είναι και υπερβολικά ευαίσθητος στα δεδομένα εισόδου ώστε να μην απομνημονεύει το σύνολο δεδομένων εισόδου. Με αυτή την έννοια, υπάρχει ένα αντιστάθμισμα (trade-off), μεταξύ της αποτελεσματικής ανακατασκευής των δεδομένων και της απομνημόνευσης τους ("overfitting").

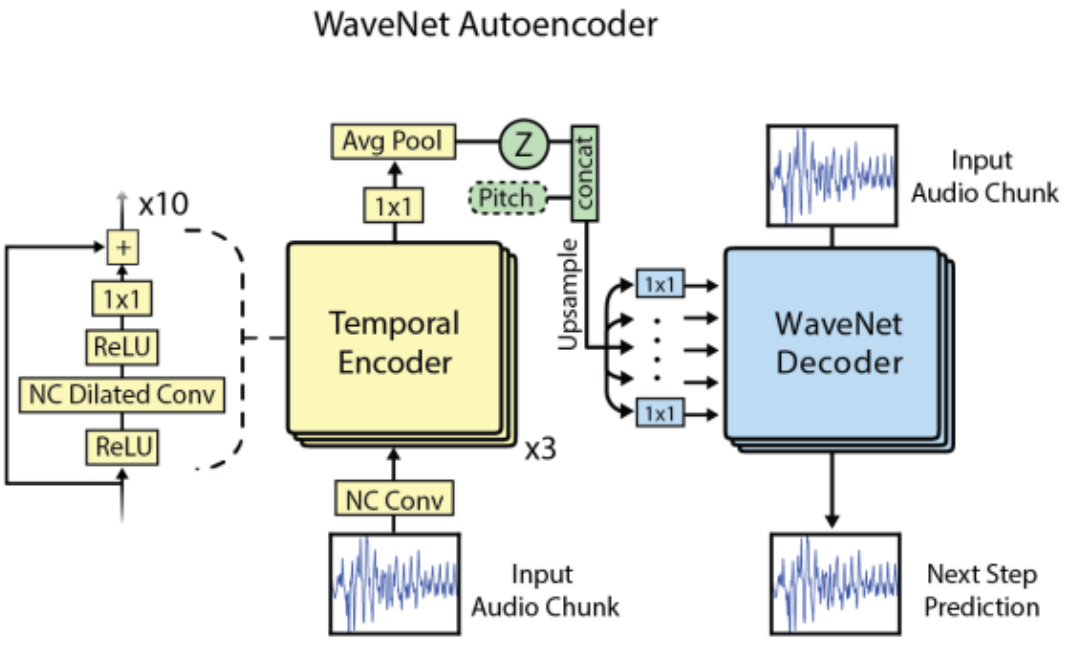
Wavenet Autoencoder

Ο Wavenet Autoencoder, όπως ήδη αναφέραμε, δέχεται στην είσοδο του απευθείας ηχητική κυματομορφή. Στην συνέχεια, ο κωδικοποιητής εξάγει ένα embedding ("εμφύτευση" (structure within a structure)) Z = f(x). Αυτό το embedding τροφοδοτείται στον αποκωδικοποιητή ο οποίος επιχειρεί να ανακατασκευάσει το σήμα εισόδου. Η από κοινού πιθανότητα του δικτύου Wavenet είναι λοιπόν τώρα η εξής:

$$p(x) = \prod_{t=1}^{T} p\left(x_t | x_1, \dots, x_{T-1}, f(x)\right) \qquad (4)$$

Ένα πιθανό πρόβλημα σε αυτή την περίπτωση είναι ο αποκωδικοποιητής να αγνοήσει πλήρως το embedding και να επιχειρεί να ανακατασκευάσει το δείγμα εισόδου μόνο με βάση τα προηγούμενα δείγματα. Ωστόσο, στην πράξη κάτι τέτοιο δεν συμβαίνει, καθώς το embedding είναι ιδιαίτερα σημαντικό κατά την διαδικασία της εκπαίδευσης του δικτύου, οπότε και ο αποκωδικοποιητής μαθαίνει να το χρησιμοποιεί.

Κατά την διαδικασίας παραγωγής νέων δεδομένων, ο αποκωδικοποιητής αυτοπαλινδρομικά παράγει σειριακά δείγματα εξόδου, βασισμένος σε ένα embedding και σε μια αρχική ακολουθία από μηδενικά δείγματα εισόδου. Παρακάτω παρουσιάζουμε την αρχιτεκτονική του δικτύου:



Σχήμα 5: Wavenet Autoencoder

Ο κωδικοποιητής είναι ένα δίκτυο 30 επιπέδων με υπολειπόμενες συνδέσεις (residual connections), που πραγματοποιεί διασταλμένες συνελίξεις, όπως ακριβώς και το απλό δίκτυο Wavenet, ακολουθούμενες από συνελίξεις με 1x1 φίλτρα. Κάθε CNN φίλτρο έχει 128 κανάλια και ως συνάρτηση ενεργοποίησης την ReLu (Rectified Linear Unit). Το embedding προκύπτει από άλλο ένα 1x1 CNN φίλτρο, ακολουθούμενο από ένα επίπεδο που πραγματοποιείται average pooling. Το embedding λέμε ότι είναι χρονικό, με την έννοια ότι το αποτέλεσμα του κωδικοποιητή είναι μια ακολουθία από κρυφούς κωδικούς με άξονες τον χρόνο και τα διαφορετικά κανάλια. Η χρονική ανάλυση εξαρτάται από την μετατόπιση (stride) του pooling. Στην πράξη, επιλέχθηκε σταθερό το μέγεθος του embedding και εφαρμόστηκε περίπου 32 φορές συμπύκνωση των δεδομένων εισόδου.

Ο αποκωδικοποιητής του δικτύου είναι εντελώς παρόμοιος με το απλό δίκτυο Wavenet που αναλύσαμε στο προηγούμενο κεφάλαιο. Η αύξηση των δειγμάτων του embedding σήματος, ώστε να έχει το ίδιο μέγεθος με το δείγμα εισόδου, πραγματοποιείται με τη μέθοδο της παρεμβολής του κοντινότερου γείτονα (nearest neighbor interpolation), δηλαδή κάθε τιμή αντιστοιχεί στην κοντινότερη της από τις διαθέσιμες τιμές με υψηλότερο ρυθμό δειγματοληψίας.

6.1.3 A Universal Music Translation Network

Εποπτικά, το δίκτυο αποτελείται από πολλαπλούς Wavenet αυτοκωδικοποιητές, όπως αυτόν που είδαμε στο προηγούμενο κεφάλαιο. Πιο συγκεκριμένα, το μοντέλο αυτό αποτελείται από έναν κοινό, καθολικό κωδικοποιητή (universal encoder), ο οποίος τροφοδοτείται με όλα τα διαφορετικά δεδομένα εισόδου, ανεξάρτητα από τον τομέα (domain) που προέρχονται. Πέρα από το προφανές πλεονέκτημα της εκπαίδευσης λιγότερων δικτύων, ο κοινός κωδικοποιητής επιτρέπει τη μεταφορά μουσικής από διαφορετικούς τομείς εισόδου, ακόμα και αν αυτοί δεν είχαν εμφανιστεί κατά την εκπαίδευση του μοντέλου. Επιπλέον, το δίκτυο αποτελείται από πολλαπλούς αποκωδικοποιητές, έναν για κάθε τομέα εισόδου (π.χ. έναν για κάθε όργανο, σύνολο οργάνων ή είδος μουσικής)

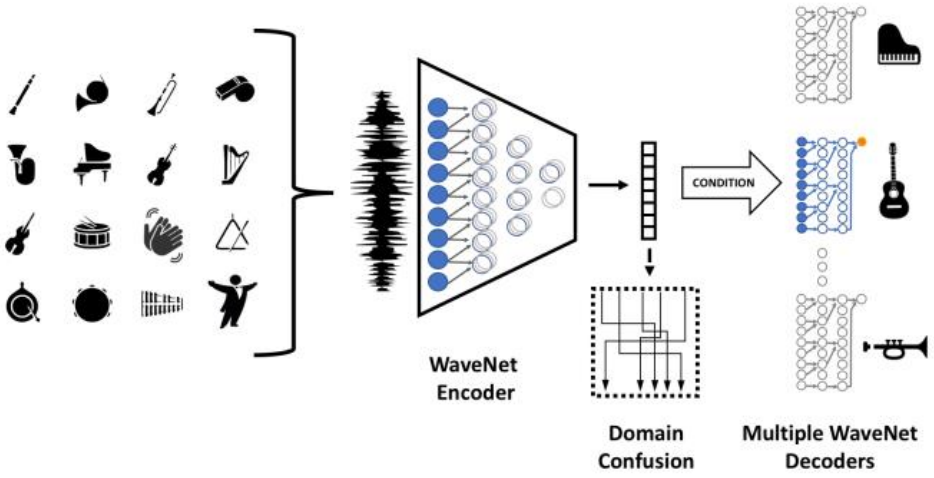
Το κλειδί για την εκπαίδευση ενός δικτύου με αυτή την αρχιτεκτονική είναι να διασφαλίσουμε ότι οποιαδήποτε πληροφορία που είναι σχετική με τον συγκεκριμένο τομέα εισόδου δεν κωδικοποιείται. Στο συγκεκριμένο δίκτυο, αυτό επιτυγχάνεται με την χρήση ενός δικτύου που μπερδεύει τον τομέα εισόδου, το οποίο ονομάστηκε Domain Confusion Network. Σκοπός του δικτύου αυτού είναι να στείλει ένα αντιπαραθετικό σήμα ("adversarial signal") στον κοινό κωδικοποιητή.

Επιπλέον, προκειμένου ο κωδικοποιητής να μην απομνημονεύσει απλά τα δεδομένα εισόδου, αλλά να κωδικοποιήσει σημασιολογική πληροφορία από αυτά, το σήμα εισόδου δέχεται μια παραμόρφωση, με μια τυχαία διαμόρφωση της τονικότητας, Έτσι, κατά την διαδικασία της εκπαίδευσης, το μοντέλο εκπαιδεύεται ως ένας αυτοκωδικοποιητής αποθορυβοποίησης, που επιχειρεί δηλαδή να ανακατασκευάσει την μη παραμορφωμένη μορφή της αρχικής εισόδου. Από την στιγμή βέβαια που το σήμα που δέχεται στην είσοδο του το δίκτυο είναι η παραμορφωμένη

είσοδος, και ταυτόχρονα ο αποκωδικοποιητής επιχειρεί να ανακατασκευάσει το αρχικό σήμα εισόδου, το μοντέλο μαθαίνει να προβάλλει δεδομένα που δεν ανήκουν στον τομέα του αντίστοιχου αποκωδικοποιητή. Επιπρόσθετα, με αυτή την μέθοδο, το δίκτυο δεν επωφελείται απομνημονεύοντας απλά τα δεδομένα εισόδου και εφαρμόζει μια κωδικοποίηση υψηλότερου επιπέδου.

Αρχιτεκτονική Δικτύου

Σε ένα αρκετά υψηλό επίπεδο, η αρχιτεκτονική του δικτύου παρουσιάζεται στο παρακάτω σχήμα.



Σχήμα 6: Η βασική αρχιτεκτονική του δικτύου

Η βασική αρχιτεκτονική, τόσο του κωδικοποιητή, όσο και των πολλαπλών αποκωδικοποιητών, στηρίζεται στον Wavenet Autoencoder που παρουσιάσαμε στην προηγούμενη ενότητα. Σε ό,τι αφορά τον Wavenet κωδικοποιητή , αυτός είναι ένα πλήρως συνελικτικό δίκτυο, το οποίο μπορεί να δεχθεί στην είσοδο του μια ακολουθία τυχαίας διάρκειας. Πιο συγκεκριμένα, το δίκτυο έχει ακριβώς την ίδια αρχιτεκτονική που ορίστηκε στον Wavenet Autoencoder του σχήματος 5. Σε ότι αφορά το average pooling layer, που εφαρμόζεται στην έξοδο του κωδικοποιητή, επιλέχθηκε σταθερό φίλτρο μήκους 50 milliseconds (800 δειγμάτων), με σκοπό το embedding που προκύπτει να είναι 64 διαστάσεων. Με αυτό τον τρόπο, εφαρμόζεται περίπου 12.5 φορές συμπύκνωση των δεδομένων εισόδου.

Αντίστοιχα, οι διαφορετικοί αποκωδικοποιητές έχουν την ίδια μορφή με το βασικό Wavenet Autoencoder, που αναλύσαμε προηγουμένως. Πιο συγκεκριμένα, κάθε αποκωδικοποιητής αποτελείται από 4 κομμάτια (blocks), καθένα από τα οποία αποτελείται από 10 υπολειπόμενα επίπεδα (residual layers). Έτσι, κάθε αποκωδικοποιητής έχει ένα δεκτικό πεδίο μεγέθους 250 milliseconds (ή 4.093 δείγματα).

Εκπαίδευση Δικτύου και συναρτήσεις απώλειας (Loss Functions)

Κατά την διαδικασία της εκπαίδευσης, το δίκτυο δέχεται στην είσοδο του δεδομένα διάρκειας ενός δευτερολέπτου. Επιπλέον, πραγματοποιείται μια αύξηση του συνόλου δεδομένων εκπαίδευσης. Πιο συγκεκριμένα, επιλέγεται ένα μέρος κάθε δεδομένου εκπαίδευσης, διάρκειας 0.25 έως 0.5 δευτερόλεπτα και ο τόνος του μετατοπίζεται τυχαία κατά -0.5 μέχρι +0.5 απ' τον αρχικό τόνο.

Προκειμένου να ορίσουμε μαθηματικά της συναρτήσεις απώλειας του δικτύου, θα θεωρήσουμε ως s^j ένα δείγμα εισόδου απ'τον τομέα (domain) j = 1,2...,k, όπου k το πλήθος των domains που επιλέχθηκαν κατά την εκπαίδευση. Επιπλέον, ορίζουμε ως E τον καθολικό κωδικοποιητή και ως D^j τον Wavenet αποκωδικοποιητή για τον τομέα j. Ως C θεωρούμε ένα δίκτυο που κατηγοριοποιεί τα δεδομένα ανάλογα με τον τομέα (domain) στον οποίο ανήκουν (Domain Confusion Network) και ως O(s,r) ορίζουμε την τυχαία επαύξηση που εφαρμόζεται σε ένα δείγμα s με τυχαίο seed r.

Το δίκτυο C που αναφέραμε προβλέπει τον τομέα στον οποίο ανήκει το κάθε δείγμα εισόδου, δεχόμενο στην είσοδο του το διάνυσμα εξόδου του Wavenet κωδικοποιητή. Το δίκτυο αποτελείται από τρία μονοδιάστατα συνελικτικά επίπεδα και ένα ELU ("exponential linear unit") επίπεδο. Το ELU επίπεδο ορίστηκε στο [23] και έχει ως στόχο την αύξηση της ταχύτητας εκπαίδευσης ενός βαθιού νευρωνικού δικτύου, αλλά και την αύξηση της απόδοσης του. Σε αντίθεση με τις ReLUs συναρτήσεις ενεργοποίησης, η ELU δέχεται και αρνητικές τιμές, με αποτέλεσμα οι μέσες συναρτήσεις ενεργοποίησης του δικτύου να είναι κοντά στο 0. Η ELU συνάρτηση ενεργοποίησης ορίζεται λοιπόν ως παρακάτω:

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(exp(x) - 1) & \text{if } x \leq 0 \end{cases} \qquad (5)$$

, όπου το α>0 αποτελεί μια υπερπαράμετρο που ελέγχει την τιμή κορεσμού (saturation value) της συνάρτησης για αρνητικές τιμές εισόδου. Το βασικό πλεονέκτημα της ύπαρξης της υπερπαραμέτρου α είναι η προστασία από το πρόβλημα των εξαφανιζόμενων κλίσεων (Vanishing Gradient Problem).

Οι αυτοκωδικοποιητές εκπαιδεύονται με βάση την παρακάτω συνάρτηση απώλειας:

$$\sum_n \sum_{s^j} \mathop{\mathbb{E}}_r \mathcal{L}\left(D^j\left(\mathbb{E}\left(O(s^j, r)\right)\right), s^j\right) - \lambda \mathcal{L}\left(C\left(\mathbb{E}\left(O(s^j, r)\right)\right), j\right) \qquad (6)$$

, όπου $\mathcal{L}(o, y)$ είναι η διασταυρώμενη εντροπία (cross entropy) συνάρτηση απώλειας που εφαρμόζεται σε κάθε στοιχείο της εξόδου o και του αντίστοιχου στιχείου-στόχου y.

Η συνάρτηση απώλειας Cross Entropy υπολογίζει την απόσταση της τιμής που προβλέπει ένα δίκτυο με την πραγματική επισημειωμένη τιμή. Όσο μεγαλύτερη είναι αυτή η απόσταση, τόσο αυξάνεται και η απώλεια διασταυρωμένης εντροπίας. Το penalty που θέτει η συνάρτηση αυτή είναι λογαριθμικό, το οποίο σημαίνει ότι η απώλεια παίρνει υψηλές τιμές για διαφορές κοντά στο 1 και αρκετά μικρές για διαφορές κοντά στο 0. (τόσο τα στοιχεία εξόδου, όσο και τα επισημειωμένα στοιχεία ανήκουν στο πεδίο [0,1]). Πρωταρχικός στόχος είναι η μείωση της συνάρτησης απώλειας, ώστε να αυξηθεί η απόδοση του δικτύου. Μαθηματικά, η συνάρτηση διασταυρωμένης εντροπίας ορίζεται ως παρακάτω:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{n} t_i \, log(p_i), \qquad (7)$$

, όπου n το πλήθος των κλάσεων, t_i το επισημειωμένο στοιχείο και p_i η πιθανότητα για την i-οστή κλάση.

Επιστρέφοντας στην σχέση (6), γίνεται εμφανές ότι ο αποκωδικοποιητής είναι ένα Wavenet μοντέλο το οποίο υπόκειται στην έξοδο του κωδικοποιητή E. Κατά την εκπαίδευση, το δίκτυο δέχεται στην είσοδο του την επισημειωμένη έξοδο s^{j-1} , αντί για την έξοδο που παράγει το ίδιο το δίκτυο. Αυτή η μέθοδος ονομάζεται teacher forcing και έχει ως στόχο γρηγορότερη σύγκλιση και σταθερότητα του δικτύου.

Τέλος, το domain confusion δίκτυο C εκπαιδεύεται με στόχο την ελαχιστοποίηση της παρακάτω συνάρτησης κόστους , που αποτελεί και το κόστος κατηγοριοποίησης:

$$\sum_{n} \sum_{s^j} \mathop{\mathbb{E}}_{r} \mathcal{L}\left(C\left(\mathbb{E}\left(O\left(s^j, r\right)\right)\right), s^j\right) \qquad (8)$$

Παραγωγή νέων δειγμάτων

Αφού ολοκληρωθεί η εκπαίδευση, το δίκτυο έχει την δυνατότητα παραγωγής δειγμάτων ακόμα και από δεδομένα εισόδου που προέρχονται από κάποιον μουσικό τομέα πάνω στον οποίο δεν έχει εκπαιδευτεί. Πιο συγκεκριμένα, έστω ότι το δίκτυο δέχεται στην είσοδο του ένα δείγμα s, ανεξαρτήτως τομέα εισόδου (musical domain) και ως επιθυμητό τομέα εξόδου, τον τομέα j. Σε αυτή την περίπτωση το δείγμα s περνάει από τον αυτοκωδικοποιητή χωρίς την εφαρμογή παραμόρφωσης. Με αυτή την έννοια, το νέο δείγμα δίνεται από τον παρακάτω τύπο:

$$\hat{s}^j = D^j\big(E(s)\big) \qquad (9)$$

6.2 MelGAN: Conditional Waveform Synthesis

Το μοντέλο αυτό [23] αναπτύχθηκε το 2019, με σκοπό την παραγωγή συνεκτικής ακουστικής κυματομορφής με την χρήση Παραγωγικών Αντιπαραθετικών Δικτύων (Generative Adversarial Networks (GANs)). Η βασική προσφορά του μοντέλου είναι η εκπαίδευση ενός GAN δικτύου το οποίο έχει την δυνατότητα να παράγει ηχητικά σήματα υψηλής ποιότητας τα οποία ταυτόχρονα έχουν συνοχή. Μάλιστα, αυτό επιτυγχάνεται με την εισαγωγή ορισμένων αλλαγών στην αρχιτεκτονική του GAN δικτύου και απλών τεχνικών εκπαίδευσης, που θα αναλυθούν στην συνέχεια. Το MelGAN μπορεί να εκπαιδευτεί για την αντιμετώπιση πολλαπλών προβλημάτων παραγωγής ηχητικών σημάτων, όπως είναι η παραγωγή φωνής, η μεταφορά ενός μουσικού μέρους σε άλλο μουσικό τομέα, αλλά και η απευθείας σύνθεση μουσικής (unconditional music synthesis).

Επιπλέον, σε σχέση με μοντέλα τύπου WaveNet, παρουσιάζει πλεονεκτήματα στην εκπαίδευση, καθώς το πλήθος των εκπαιδεύσιμων παραμέτρων είναι σημαντικά μικρότερο, με αποτέλεσμα ο χρόνος εκπαίδευσης που απαιτείται να είναι κι αυτός με την σειρά του σαφώς μικρότερος. Βέβαια, το βασικό χαρακτηριστικό των GAN δικτύων είναι η παραλληλοποίηση, δηλαδή η ικανότητα τους να δέχονται και να εκπαιδεύονται παράλληλα σε πολλαπλά δείγματα εισόδου. Με αυτό τον τρόπο, μη αυτοπαλινδρομικά (non autoregressive) μοντέλα εκμεταλλεύονται πλήρως τις δυνατότητες που παρέχουν οι GPUs για την εκπαίδευση δικτύων μηχανικής μάθησης.

Προκειμένου να γίνει κατανοητή η αρχιτεκτονική του δικτύου MelGAN, θα ξεκινήσουμε με μια σύντομη παρουσίαση των Παραγωγικών Αντιπαραθετικών Δικτύων (Generative Adversarial Networks) και της λειτουργίας τους.

6.2.1 Παραγωγικά Αντιπαραθετικά Δίκτυα (Generative Adversarial Networks)

Τα Παραγωγικά Αντιπαραθετικά Δίκτυα αναπτύχθηκαν το 2014 από τον Ian Goodfellow [25]. Έκτοτε έχει γίνει ευρεία χρήση τους σε εφαρμογές παραγωγής εικόνας και βίντεο, καθώς έχουν την δυνατότητα να παράγουν πολύ πιστό και ταυτόχρονα νέο υλικό. Ωστόσο η χρήση τους είναι περιορισμένη σε εφαρμογές ηχητικών σημάτων, καθώς δεν παρουσιάζουν εξίσου καλά αποτελέσματα. Με αυτή την έννοια, υπάρχουν ακόμα μεγάλα περιθώρια εξέλιξης για την εφαρμογή τέτοιων μοντέλων σε προβλήματα παραγωγής φωνής και μουσικής.

Παραγωγικά Δίκτυα (Generative Networks) και η χρήση νευρωνικών δικτύων

Η βασική αρχή τέτοιων δικτύων, ανεξαρτήτως εφαρμογής αλλά και τύπου δεδομένων (εικόνα, κείμενο, ηχητικό σήμα ή άλλο) είναι η υπόθεση ότι το σήμα εισόδου ακολουθεί μια κατανομή. Ωστόσο, αυτή η κατανομή πιθανότητας είναι σύνθετη και ορίζεται σε έναν πολυδιάστατο χώρο. Έτσι, ακόμα και αν το σήμα εισόδου διέπεται από μια κατανομή πιθανότητας, είναι αδύνατον να την εκφράσουμε αναλυτικά. Σε διαφορετική περίπτωση, το πρόβλημα θα ήταν αρκετά απλό, καθώς το ζητούμενο θα ήταν η παραγωγής μιας τυχαίας μεταβλητής από μια γνωστή και συγκεκριμένη κατανομή πιθανότητας.

Μια αρχική ιδέα για την επίλυση του προβλήματος που περιγράψαμε παραπάνω είναι η μέθοδος της Μέγιστης Πιθανοφάνειας (Maximum Likelihood estimation). Σύμφωνα με αυτή τη μέθοδο εκτιμάται η κατανομή πιθανότητας των δεδομένων εισόδου, με την προϋπόθεση ότι είναι διαθέσιμα δεδομένα εισόδου. Αυτό επιτυγχάνεται μεγιστοποιώντας μια συνάρτηση πιθανότητας, πάνω στην οποία ταιριάζουν τα δεδομένα εισόδου. Παραδοσιακά παραγωγικά δίκτυα, όπως είναι τα Restricted Boltzmann Machines (RBMs), Gaussian Mixture Models (GMMs), Απλοϊκά μοντέλα Bayes (Naïve Bayes Models) αλλά και τα κρυφά μοντέλα Markov (Hidden Markov Models), στηρίζονται κατά βάση στην μέθοδο της Μέγιστης Πιθανοφάνειας.

Ωστόσο, είναι πολύ πιθανό, ειδικά σε πιο περίπλοκα προβλήματα, η εκτίμηση μέγιστης πιθανοφάνειας να μην ανταποκρίνεται στην πολυπλοκότητα της πραγματικής κατανομής πιθανότητας που ακολουθούν τα δεδομένα εισόδου και ως αποτέλεσμα, η παραπάνω μέθοδος να μην είναι ικανή να μάθει πλήρως αυτή την κατανομή.

Μια διαφορετική προσέγγιση στο πρόβλημα αυτό είναι η χρήση νευρωνικών δικτύων για την παραγωγή νέων δειγμάτων. Πιο συγκεκριμένα, το νευρωνικό δίκτυο πρέπει να μάθει μια συνάρτηση μεταφοράς, σύμφωνα με την οποία, δέχεται στην είσοδο του μια N-διάστατη τυχαία μεταβλητή και επιστρέφει στην έξοδο, μια διαφορετική N-διάστατη τυχαία μεταβλητή που ακολουθεί την κατανομή πιθανότητας των δεδομένων εισόδου. Ιδιαίτερα σημαντική είναι η ανάπτυξη της κατάλληλης τεχνικής εκπαίδευσης ενός τέτοιου δικτύου.

Εκπαίδευση Παραγωγικών Νευρωνικών Δικτύων

Για την εκπαίδευση παραγωγικών νευρωνικών δικτύων έχουν προταθεί δύο διαφορετικές μέθοδοι.

Στην πρώτη μέθοδο (άμεση εκπαίδευση) συγκρίνονται απευθείας η πραγματική και η παραγόμενη κατανομή πιθανότητας. Στην συνέχεια, μέσω της μεθόδου της οπισθοδρόμησης (back propagation) το σφάλμα μεταφέρεται προς τα πίσω στο δίκτυο ώστε να προσαρμοστούν οι εκπαιδεύσιμες παράμετροι του. Το πρόβλημα στην συγκεκριμένη περίπτωση, όπως εξάλλου ήδη αναφέραμε, είναι ότι η πραγματική κατανομή πιθανότητας που ακολουθούν τα δεδομένα είναι άγνωστη, αλλά και δύσκολο να υπολογιστεί. Γι' αυτόν τον λόγο, το σφάλμα υπολογίζεται μέσω κάποιας μετρικής που συγκρίνει δύο "κατανομές" βασισμένες στα δείγματα από τα οποία αποτελούνται. Μια τέτοια τεχνική είναι η μέγιστη μέση απόκλιση (Maximum Mean Discrepancy), σύμφωνα με την οποία ορίζεται μια απόσταση μεταξύ δύο κατανομών πιθανότητας οι οποίες έχουν υπολογιστεί από τα δείγματα που αποτελούνται.

Στην δεύτερη μέθοδο (έμμεση εκπαίδευση), δεν συγκρίνουμε άμεσα τις δύο αυτές κατανομές πιθανότητας. Αντίθετα, το δίκτυο εκπαιδεύεται πάνω σε ένα παρόμοιο πρόβλημα, το οποίο ωθεί τις δύο κατανομές να είναι όσο το δυνατόν πιο κοντά. Πάνω σε αυτή την ιδέα βασίστηκαν τα παραγωγικά αντιπαραθετικά δίκτυα.

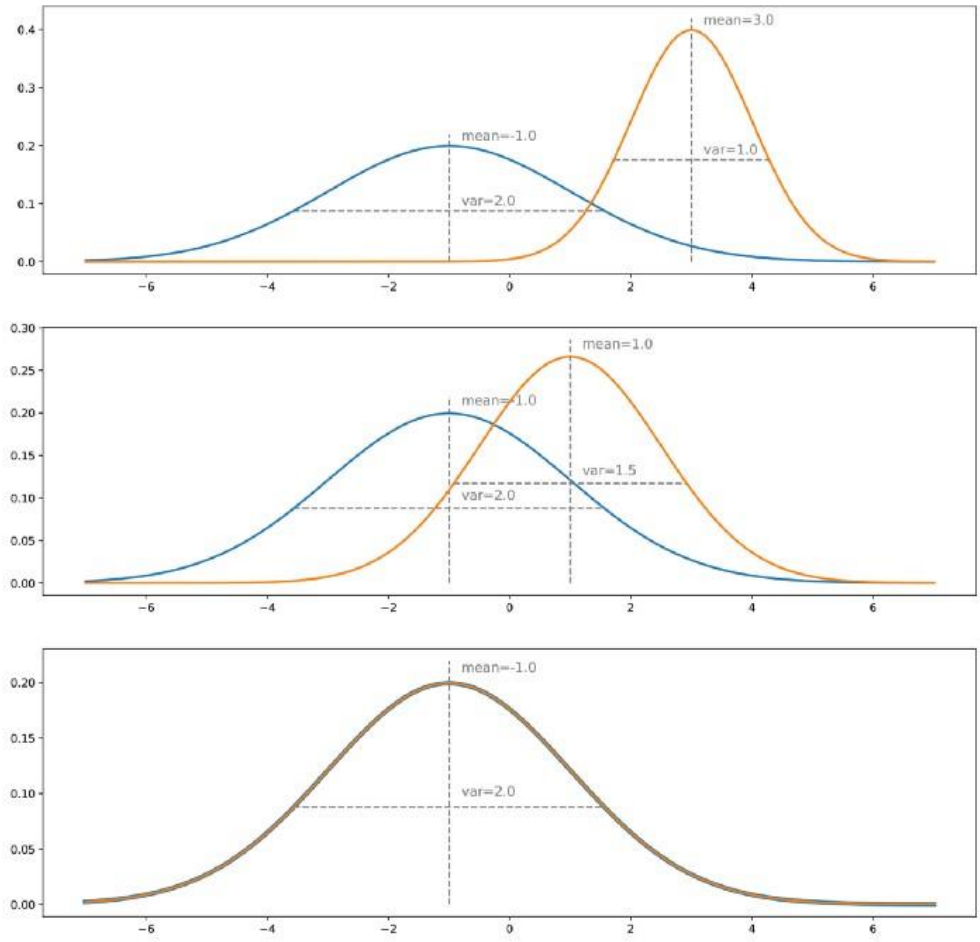
Παραγωγικά Αντιπαραθετικά Δίκτυα (Generative Adversarial Networks)

Το πρόβλημα που έχουν να αντιμετωπίσουν τα Παραγωγικά Αντιπαραθετικά Δίκτυα, ώστε η κατανομή των παραγόμενων δεδομένων να είναι πολύ κοντά στην κατανομή των πραγματικών δεδομένων, είναι το πρόβλημα του διαχωρισμού ("discrimination") των δεδομένων σε πραγματικά και παραγόμενα δεδομένα. Στην πραγματικότητα θέλουμε αυτός ο διαχωρισμός να αποτυγχάνει όσο το δυνατόν περισσότερο.

Με βάση την υλοποίηση της παραπάνω ιδέας προκύπτει η αρχιτεκτονική ενός Παραγωγικού Αντιπαραθετικού Δικτύου. Ένα GAN μοντέλο λοιπόν αποτελείται από ένα παραγωγικό δίκτυο ("Generator") και ένα διαχωριστικό δίκτυο ("Discriminator"). Ο Discriminator λαμβάνει στην είσοδο του δείγματα από πραγματικά δεδομένα και από δεδομένα που παράγει ο Generator και επιχειρεί να τα κατηγοριοποιήσει σε πραγματικά ή παραγόμενα. Απ' την άλλη, ο Generator επιχειρεί να εξαπατήσει τον Discriminator και μάλιστα εκπαιδεύεται ακριβώς γι' αυτόν τον σκοπό.

Σε ένα θεωρητικό πλαίσιο, η εκπαίδευση ενός τέτοιου δικτύου θα έχει ως αποτέλεσμα ένα ιδανικό παραγωγικό δίκτυο. Ας δούμε λοιπόν την ιδανική περίπτωση, στην οποία έχουμε έναν ιδανικό Generator και Discriminator. Ως ιδανικά δίκτυα εννοούμε δίκτυα χωρίς περιορισμούς ως προς την χωρητικότητα και δίκτυα που δεν περιορίζονται από κάποιο άλλο παραμετρικό μοντέλο. Γι' αυτό τον λόγο, προς το παρόν θα αγνοήσουμε την αρχιτεκτονική και των δύο δικτύων και θα τα θεωρήσουμε ως δύο ασαφείς οντότητες.

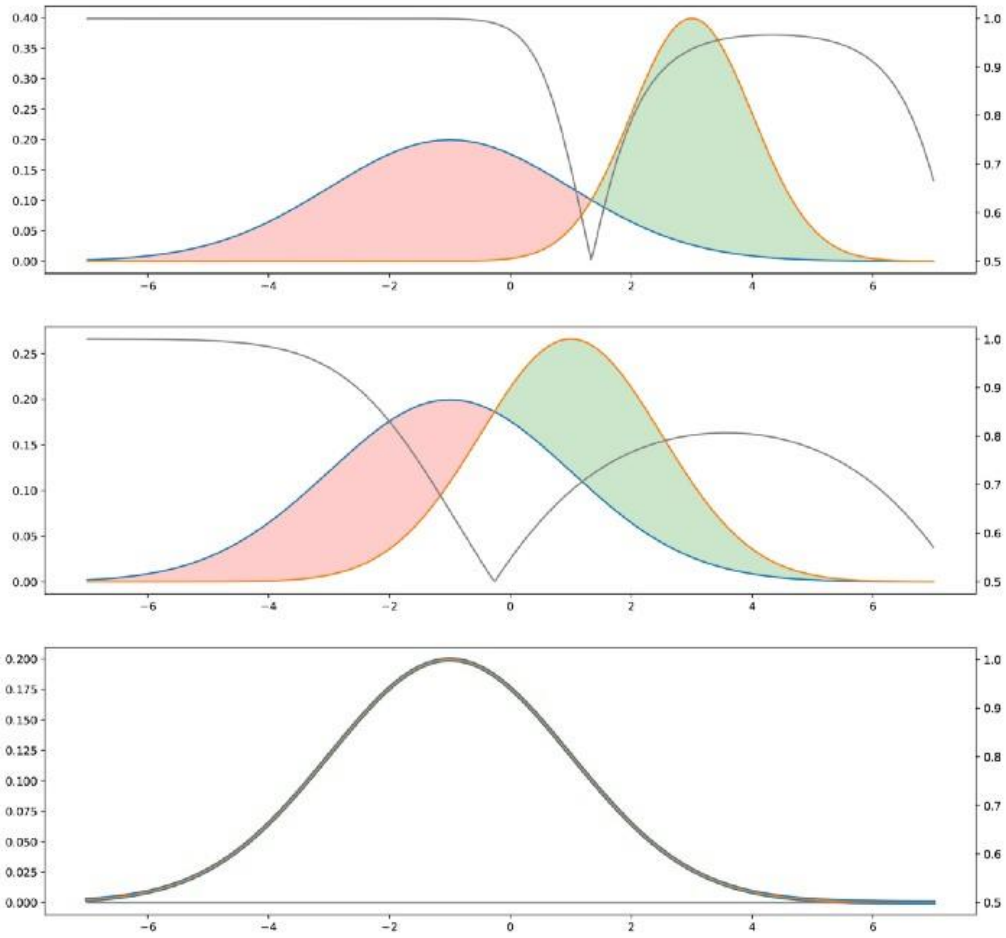
Ας υποθέσουμε ότι έχουμε μια πραγματική κατανομή, όπως για παράδειγμα μια μονοδιάστατη κανονική κατανομή. Στην περίπτωση της «άμεσης» εκπαίδευσης (που αναφέραμε προηγουμένως, το παραγωγικό δίκτυο θα προσαρμοζόταν επαναλαμβανόμενα κατά την εκπαίδευση, ώστε να διορθώσει την διαφορά ή τη μέτρηση του σφάλματος μεταξύ της πραγματικής και παραγόμενης κατανομής, μέχρις ότου, ιδανικά, η παραγόμενη κατανομή να ταιριάζει πλήρως με την πραγματική (Σχήμα 7).



Σχήμα 7: Απεικόνιση της άμεσης μεθόδου εκπαίδευσης. Η μπλε καμπύλη αναπαριστά την πραγματική κατανομή, ενώ η πορτοκαλί την παραγόμενη κατανομή του Generator.

Στην περίπτωση του Παραγωγικού Αντιπαραθετικού Δικτύου, ας υποθέσουμε προς το παρόν ότι ο Discriminator έχει πλήρη γνώση της πραγματικής κατανομής. Με αυτή την έννοια, έχει την δυνατότητα να κάνει τέλειες προβλέψεις ως προς το αν ένα δείγμα είναι πραγματικό ή παραγόμενο. Εάν οι δύο κατανομές έχουν μεγάλη απόσταση, τότε ο Discriminator θα κατηγοριοποιεί με πολύ μεγάλη ευκολία αλλά και υψηλό επίπεδο εμπιστοσύνης σε ποια κατανομή ανήκει ένα δείγμα. Για να εξαπατηθεί ο Discriminator, θα πρέπει οι δύο κατανομές να είναι πολύ κοντινές. Μάλιστα, εάν οι δύο κατανομές ταιριάξουν πλήρως, τότε ο Discriminator δεν θα έχει τη

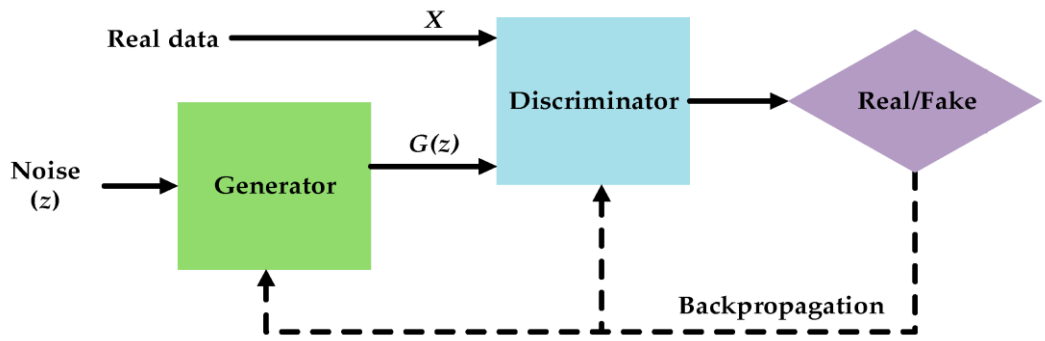
δυνατότητα να κατηγοριοποιήσει τα δείγματα καθώς θα υπάρχουν ίσες πιθανότητες αυτά να ανήκουν στην πραγματική ή παραγόμενη κατανομή. Η παραπάνω συμπεριφορά γίνεται εμφανής στο επόμενο σχήμα.



Σχήμα 8: Σχηματικά η διαισθητική ιδέα της αντιπαραθετικής μεθόδου. Η μπλε καμπύλη αντιστοιχεί στην πραγματική κατανομή, ενώ η πορτοκαλί στην παραγόμενη κατανομή. Σε γκρι, παρουσιάζεται η πιθανότητα να προβλέψει σωστά ο Discriminator (με τον y-άξονα στο δεξί μέρος του σχήματος). Κατά την εκπαίδευση, στόχος είναι να μειωθούν οι πράσινες και κόκκινες περιοχές.

Αρχιτεκτονική Δικτύου

Παρακάτω, παρουσιάζεται σχηματικά η βασική αρχιτεκτονική ενός Παραγωγικού Αντιπαραθετικού Δικτύου.



Σχήμα 9: Αρχιτεκτονική ενός GAN δικτύου.

Ο Generator είναι ένα νευρωνικό δίκτυο το οποίο μοντελοποιεί μια συνάρτηση μεταφοράς, σύμφωνα με την οποία προκύπτουν τα νέα δεδομένα. Στην είσοδο του λαμβάνει μια τυχαία παράμετρο με την μορφή τυχαίου θορύβου, όπως φαίνεται στο παραπάνω σχήμα, και πρέπει να επιστρέψει, αφού εκπαιδευτεί, μια τυχαία μεταβλητή η οποία ακολουθεί την κατανομή των δεδομένων εισόδου. Όπως είναι κατανοητό από το παραπάνω σχήμα, ο Discriminator είναι επίσης ένα νευρωνικό δίκτυο, καθώς η συνάρτηση διαχωρισμού είναι άγνωστη και πιθανότατα περίπλοκη. Ο Discriminator λοιπόν δέχεται ένα δείγμα στην είσοδο του και επιστρέφει στην έξοδο του την πιθανότητα αυτό το δείγμα να είναι πραγματικό.

Από στην στιγμή που ορίσαμε τα μοντέλα του Generator και του Discriminator ως δύο νευρωνικά δίκτυα, πλέον δεν ισχύει πλήρως η ιδανική περίπτωση που υποθέσαμε προηγουμένως. Ο λόγος είναι ότι τα δύο δίκτυα αποτελούνται από εκπαιδεύσιμες παραμέτρους και πλέον, η κατανομή των δεδομένων που παράγει ο Generator εξαρτάται από την χωρητικότητα και την ακρίβεια των δύο αυτών παραμετρικών μοντέλων.

Κατά την εκπαίδευση, τα δύο αυτά δίκτυα εκπαιδεύονται ταυτόχρονα. Το παραγωγικό δίκτυο έχει ως στόχο να εξαπατήσει το Διαχωριστικό δίκτυο. Με αυτή την έννοια ο Generator εκπαιδεύεται έτσι ώστε να μεγιστοποιήσει τον τελικό σφάλμα κατηγοριοποίησης, μεταξύ πραγματικών και παραγόμενων δεδομένων. Από την άλλη, το διαχωριστικό δίκτυο έχει ως στόχο να εντοπίζει τα “ψεύτικα” δεδομένα που παράγει ο Generator. Έτσι, ο Discriminator εκπαιδεύεται έτσι ώστε να ελαχιστοποιήσει το προαναφερθέν σφάλμα.

Γίνεται λοιπόν κατανοητό, ότι σε κάθε επανάληψη της διαδικασίας εκπαίδευσης, τα βάρη του παραγωγικού δικτύου ενημερώνονται έτσι ώστε να αυξηθεί το σφάλμα κατηγοριοποίησης, ενώ τα βάρη του διαχωριστικού δικτύου ενημερώνονται έτσι ώστε να μειωθεί αυτό το σφάλμα. Αυτοί οι

δύο αντικρουόμενοι στόχοι εξηγούν πλήρως και την χρήση του όρου “Αντιπαραθετικό Δίκτυο”, καθώς τα δύο δίκτυα επιχειρούν να υπερνικήσουν και με αυτό τον τρόπο, κατά την εκπαίδευση, βελτιώνονται και τα δύο ταυτόχρονα.

Ως μια άλλη εξήγηση, σύμφωνα με την θεωρία παιγνίων, θα μπορούσαμε να θεωρήσουμε την λειτουργία ενός τέτοιου δικτύου ως ένα minimax παίγνιο μηδενικού αθροίσματος μεταξύ δύο παικτών , δηλαδή ένα παιχνίδι στο οποίο ο ένας παίκτης επιχειρεί να μεγιστοποιήσει και ο άλλος να ελαχιστοποιήσει τον ίδιο στόχο. Σε μια τέτοια περίπτωση, το σημείο ισορροπίας αντιστοιχεί στην περίπτωση που ο Generator παράγει δείγματα που ανήκουν στην πραγματική κατανομή των δεδομένων εισόδου και ο Discriminator προβλέπει τα δείγματα αυτά με πιθανότητα ½ αν είναι πραγματικά ή παραγόμενα.

Μαθηματική Μοντελοποίηση

Όπως αναφέραμε, ένα Παραγωγικό Αντιπαραθετικό Μοντέλο αποτελείται από ένα παραγωγικό δίκτυο G(.) το οποίο δέχεται μια τυχαία είσοδο z με πυκνότητα πιθανότητας p_z και επιστρέφει την έξοδο $X_g = G(z)$, η οποία στην ιδανική περίπτωση ακολουθεί την κατανομή πιθανότητας των πραγματικών δεδομένων. Επιπλέον, το μοντέλο αποτελείται από το διαχωριστικό δίκτυο D(.) το οποίο δέχεται μια είσοδο x που μπορεί να είναι ένα πραγματικό x_t ή ένα παραγόμενο x_g δείγμα και επιστρέφει την πιθανότητα D(x), το x να είναι πραγματικό δεδομένο.

Ας θεωρήσουμε ότι ο Discriminator δέχεται το ίδιο πλήθος από πραγματικά και παραγόμενα δεδομένα. Τότε , το απόλυτο σφάλμα του Discriminator είναι το παρακάτω:

$$E(G,D) = \frac{1}{2}\mathbb{E}_{x\sim p_t}[1 - D(x)] + \frac{1}{2}\mathbb{E}_{z\sim p_z}[D(G(z)] \qquad (10)$$

Βέβαια, από την πλευρά του Generator θέλουμε αυτό το σφάλμα να μεγιστοποιηθεί. Έτσι, συνολικά έχουμε το παρακάτω σφάλμα:

$$Total\ Error = \max_G \left(\min_D E(G,D) \right) \qquad (11)$$

Από τις παραπάνω σχέσεις αποδεικνύεται ότι βέλτιστος Generator πληροί την παρακάτω συνθήκη:

$$p_g(x) = p_t(x) \qquad (12)$$

,δηλαδή στην ιδανική περίπτωση, μπορεί να επιβεβαιωθεί και μαθηματικά ότι ο Generator παράγει δεδομένα ακριβώς την ίδια πυκνότητα πιθανότητας με αυτή που ακολουθούν τα πραγματικά δεδομένα

6.2.2 MelGAN: Conditional Waveform Synthesis

Πρωταρχικός στόχος του δικτύου αυτού, όπως εξάλλου ήδη αναφέραμε, είναι η απευθείας σύνθεση ενός ηχητικού σήματος. Γι' αυτό τον λόγο, η ανάλυση του δικτύου που ακολουθεί, επικεντρώνεται στην επίτευξη αυτού του στόχου. Η χρήση ενός τέτοιου δικτύου για την αντιμετώπιση του προβλήματος της εργασίας αυτής, δηλαδή για την αντιμετώπιση της μεταφοράς μουσικής από ένα σύνολο μουσικών οργάνων σε κάποιο άλλο, θα γίνει ξεκάθαρη στο επόμενο κεφάλαιο.

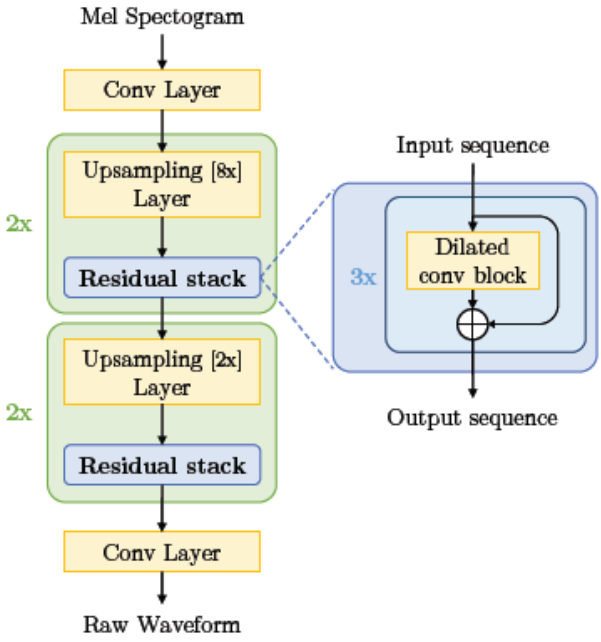
Αρχιτεκτονική Παραγωγικού Δικτύου (Generator's architecture)

Ο Generator είναι ένα πλήρως συνελικτικό δίκτυο (βλ. κεφάλαιο 4.2) το οποίο δέχεται στην είσοδο του ένα σπεκτρόγραμμα s κλίμακας Mel και παράγει μια κυματομορφή ήχου x. Επιπλέον, επειδή το σπεκτρόγραμμα έχει ανάλυση 256 επί την ελάχιστη χρονική ανάλυση, το δίκτυο αποτελείται από μια σειρά από κυλιόμενα συνελικτικά μέρη, ώστε να αυξηθεί ο ρυθμός δειγματοληψίας των δεδομένων εισόδου (upsampling). Προκειμένου να επιτευχθεί η αύξηση των δειγμάτων (upsampling) κάθε συνελικτικό block έχει μέγεθος πυρήνα διπλάσιο από τη μετατόπιση (stride).

Σε αυτό το σημείο, αξίζει να αναφέρουμε ότι, αντίθετα με τα "παραδοσιακά" Παραγωγικά Αντιπαραθετικά Δίκτυα, το MelGan δεν δέχεται στην είσοδο του ένα διάνυσμα θορύβου, καθώς η χρήση μιας τέτοιας εισόδου δεν βελτιώνει ιδιαίτερα την απόδοση του δικτύου. Η βασική εξήγηση αυτής της συμπεριφοράς είναι καθαρά διαισθητική και δόθηκε στο [27] από τους Mathew et al. Σύμφωνα με αυτή την εξήγηση λοιπόν, η προσθήκη θορύβου στην είσοδο ενός Παραγωγικού Νευρωνικού Δικτύου δεν είναι απαραίτητη όταν η εξωτερική πληροφορία ("conditioning information") είναι ιδιαίτερα σημαντική.

Στα Συνελικτικά Παραγωγικά Δίκτυα που εκπαιδεύονται για την παραγωγή εικόνων τίθεται μια "επαγωγική" μεροληψία ("inductive bias"), δηλαδή μια μεροληψία που προέρχεται από διαισθητικές υποθέσεις, σύμφωνα με το οποίο, κοντινά pixel πάνω στην εικόνα συσχετίζονται, καθώς υπάρχει μεγάλη επικάλυψη των δεκτικών τους πεδίων. Με βάση αυτή την ιδέα, ο Generator ενός MelGan δικτύου εφαρμόζει μια μεροληψία ότι υπάρχει μια συσχέτιση υψηλής κλίμακας μεταξύ διαφορετικών χρονικών βημάτων. Γι' αυτό τον λόγο μετά την εφαρμογή ενός συνελικτικού block ακολουθεί μια σειρά από residual blocks που πραγματοποιούν διασταλμένες συνελίξεις, δηλαδή blocks υπολειπόμενων συνδέσεων όπως αυτά που είδαμε στα Wavenets. Με αυτό τον τρόπο, σε κάθε διαδοχικό επίπεδο, ενεργοποιήσεις που έχουν μεγάλη χρονική απόσταση

έχουν και σημαντική επικάλυψη της εισόδου τους. Με βάση τα παραπάνω, εποπτικά, η αρχιτεκτονική ενός MelGAN Generator είναι η εξής:



Σχήμα 10: Αρχιτεκτονική Generator Δικτύου MelGan. Το upsampling πραγματοποιείται σε 4 στάδια (8x, 8x, 2x, 2x), ενώ κάθε υπολειπόμενο block διασταλμένων συνελίξεων αποτελείται από 3 επίπεδα, με διαστολή 1, 3, και 9 αντίστοιχα. Στα υπολειπόμενα block διασταλμένων συνελίξεων το μέγεθος του πυρήνα παραμένει σταθερό και ίσο με 3. Έτσι, στο τελικό επίπεδο το δεκτικό πεδίο του υπολειπόμενου block έχει μέγεθος 27 χρονικών βημάτων.

Μια ιδιαίτερα σημαντική παράμετρος για την επιτυχή εκπαίδευση του δικτύου είναι η προσεκτική επιλογή μιας τεχνικής κανονικοποίησης. Η συνηθέστερη τέτοια τεχνική σε Παραγωγικά Αντιπαραθετικά Δίκτυα, που καλούνται να αντιμετωπίσουν προβλήματα παραγωγής Εικόνας, είναι η κανονικοποίηση σε επίπεδο δείγματος (“instance normalization”). Ωστόσο, στην περίπτωση παραγωγής ηχητικού σήματος, αυτή η τεχνική κανονικοποίησης έχει ως αποτέλεσμα να χαθεί σημαντική τονική πληροφορία, που με την σειρά του έχει ως αποτέλεσμα την παραγωγή μεταλλικού ήχου.

Έτσι, έπειτα από δοκιμές, βρέθηκε ότι η πιο αποτελεσματική τεχνική κανονικοποίησης είναι η κανονικοποίηση βάρους (“weight normalization”), δηλαδή η κανονικοποίηση γίνεται μεταξύ των βαρών σε κάθε επίπεδο του δικτύου. Με την συγκεκριμένη τεχνική λοιπόν, απλά παραμετροποιούνται οι πίνακες βαρών, με αποτέλεσμα να διαχωριστεί το μέτρο του διανύσματος βαρών από την κατεύθυνση.

Αρχιτεκτονική Διαχωριστικού Δικτύου

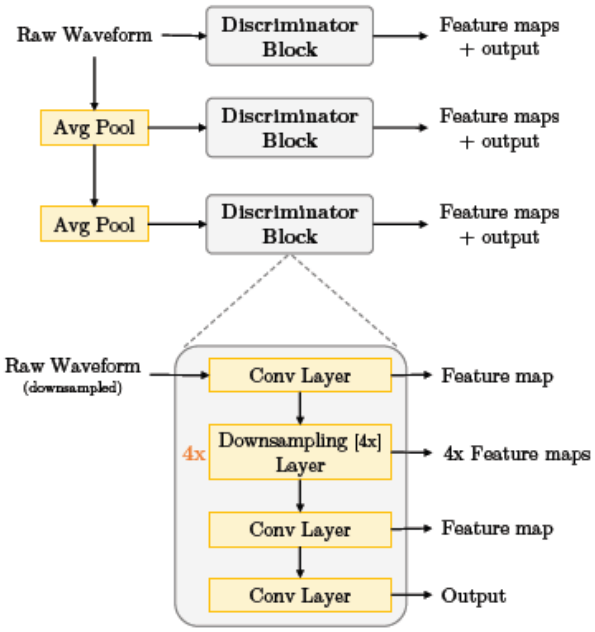
Η αρχιτεκτονική του Discriminator βασίστηκε στο Conditional GAN μοντέλο, το οποίο είχε αρχικά αναπτυχθεί για σύνθεση εικόνων υψηλής ευκρίνειας [27]. Η βασική ιδέα ενός τέτοιου μοντέλου είναι η εκπαίδευση τριών Διαχωριστικών Δικτύων (D1, D2, D3) τα οποία έχουν ακριβώς την ίδια δομή ως δίκτυα, αλλά "λειτουργούν" σε διαφορετική ηχητική κλίμακα. Πιο συγκεκριμένα, ο πρώτος Discriminator (D1) λειτουργεί στην κλίμακα του απευθείας ηχητικού σήματος, ενώ οι D2 και D3 δέχονται δεδομένα τα οποία έχουν γίνει downsampled στο μισό και στο ένα τέταρτο αντίστοιχα. Το downsampling πραγματοποιείται με την εφαρμογή ενός κυλιόμενου average pooling με μέγεθος πυρήνα (kernel size) ίσο με 4.

Η ιδέα αυτή πηγάζει απ' το γεγονός ότι τα ηχητικά σήματα παρουσιάζουν μια δομή σε διαφορετικές κλίμακες. Με αυτή την λογική λοιπόν, οι τρεις Discriminators θα μάθουν features τα οποία αντιστοιχούν σε διαφορετικές συχνοτικές αναλύσεις του σήματος ήχου. Πιο συγκεκριμένα, όπως είναι γνωστό από τον τομέα της επεξεργασίας σήματος, με την μείωση των δειγμάτων ενός σήματος χάνεται και η πληροφορία του σήματος που αντιστοιχεί σε υψηλές συχνότητες, καθώς η διαδικασία του downsampling αντιστοιχεί μια την εφαρμογή ενός low pass φίλτρου στο σήμα. Έτσι, ο Discriminator που έχει πρόσβαση στο downsampled σήμα, αγνοεί πλήρως σε πληροφορία υψηλών συχνοτήτων, με αποτέλεσμα να μαθαίνει τα features χαμηλών συχνοτήτων που είναι "υπεύθυνα" για τον διαχωρισμό σε αληθινά και "ψεύτικα" δείγματα.

Κάθε Discriminator βασίζεται σε αυτόν του PatchGan μοντέλου [28], το οποίο βέβαια έχει σχεδιαστεί για την μεταφορά εικόνας από ένα στυλ σε κάποιο άλλο (π.χ. ημέρα – νύχτα). Η βασική ιδέα του PatchGan μοντέλου είναι η χρήση ενός Discriminator που διαχωρίζει τις εικόνες σε αληθινές ή ψεύτικες, λαμβάνοντας υπόψιν μόνο ένα συγκεκριμένο μέρος (patch) της εικόνας, μεγέθους $N \times N$. Ταυτόχρονα, ο Discriminator είναι ένα συνελικτικό δίκτυο που διατρέχει την εικόνα, εξάγοντας τον μέσο όρο όλων των αποκρίσεων ώστε να αποφασίσει την προέλευση του κάθε δείγματος. Με αυτό τον τρόπο, ο Discriminator αποτελείται από λιγότερες παραμέτρους, εκπαιδεύεται ταχύτερα και μπορεί να εφαρμοστεί σε εικόνες τυχαίου μεγέθους.

Αντίστοιχα, οι Discriminators του MelGan αποτελούνται από μια σειρά από μετατοπισμένα συνελικτικά επίπεδα με μεγάλο μέγεθος πυρήνα. Με αυτή την έννοια ο Discriminator εφαρμόζουν κυλιόμενες συνελίξεις με επικαλυπτόμενα παράθυρα, καθώς το μέγεθος πυρήνα επιλέγεται αρκετά μεγάλο σε σχέση με την εκάστοτε μετατόπιση του παραθύρου. Έτσι, οι discriminators του μοντέλου μαθαίνουν να κατηγοριοποιούν την είσοδο που δέχονται, λαμβάνοντας υπόψιν μόνο μικρά μέρη της για κάθε παράθυρο. Η απώλεια του Discriminator υπολογίζεται πάνω στα επικαλυπτόμενα παράθυρα, με αποτέλεσμα το μοντέλο να μαθαίνει να διατηρεί τη συνοχή μεταξύ διαφορετικών patches. Με βάση τα παραπάνω, θα μπορούσαμε να θεωρήσουμε ότι ο Discriminator είναι ένα Μαρκοβιανό δίκτυο, καθώς μοντελοποιεί το σήμα εισόδου ως ένα τυχαίο Μαρκοβιανό πεδίο (Random Markov Field), υποθέτοντας ανεξαρτησία δειγμάτων που απέχουν περισσότερο από την διάμετρο του εφαρμοζόμενου παραθύρου.

Συνολικά, το διαχωριστικό δίκτυο με τους τρεις Discriminators παρουσιάζεται στο παρακάτω σχήμα:



Σχήμα 11: Διαχωριστικό Δίκτυο

Σε ότι αφορά την τεχνική κανονικοποίησης, οι Discriminators χρησιμοποιούν και αυτή με την σειρά τους την κανονικοποίηση βαρών, όπως αυτή εξηγήθηκε στο Παραγωγικό Δίκτυο. Επιπλέον, κάθε Discriminator block, αποτελείται από 4 μετατοπισμένες συνελίξεις με την μετατόπιση (stride) να είναι ίση με 4.

Εκπαίδευση Δικτύου

Για την εκπαίδευση του MelGan δικτύου, ύστερα από διαδοχικά πειράματα, ως βέλτιστη συνάρτηση κόστους ("Loss Function") επιλέχθηκε η Hinge Loss για GANs, όπως ορίστηκε από τους Lim et al [29]. Η απώλεια Hinge χρησιμοποιείται κυρίως σε ταξινομητές (classifiers) που πληρούν το κριτήριο μέγιστου περιθωρίου. Τα πιο χαρακτηριστικά μοντέλα που χρησιμοποιούν αυτή την συνάρτηση απώλειας είναι οι Μηχανές Διανυσμάτων Υποστήριξης ("Support Vector Machines" (SVMs)).

Εάν θεωρήσουμε ότι η έξοδος παίρνει τις τιμές $t = \pm 1$, τότε η Hinge απώλεια της πρόβλεψης y ορίζεται ως:

$$\ell(y) = max(0,1 - t \cdot y) \qquad (13)$$

Στα πλαίσια της εκπαίδευσης ενός Παραγωγικού Αντιπαραθετικού Δικτύου, η συνάρτηση κόστους Hinge ορίζεται ως εξής:

$$\mathcal{L}_D(\hat{G}, D) = \mathbb{E}_x\big[max(0, 1 - D_k(x))\big] + \mathbb{E}_{s,z}\left[max\left(0,1 + D_k\left(\hat{G}(s,z)\right)\right)\right] \qquad (14)$$

$$\mathcal{L}_G(G, \widehat{D_k}) = \mathbb{E}_{s,z}\left[\sum_{k\,=\,1,2,3} - \widehat{D_k}\big(G(s,z)\big)\right] \qquad (15)$$

,όπου x είναι η κυματομορφή ήχου, s η πληροφορία που προδιαθέτει το δίκτυο (conditioning information), όπως είναι το Mel-spectrogram και z το διάνυσμα θορύβου.

Επιπλέον, πέρα από το σήμα του Διαχωριστικού πεδίου, ο Generator εκπαιδεύεται και με βάση τον στόχο αντιστοίχισης των features, όπως προτάθηκε από τους Larsen et al. [30] Με αυτή την μέθοδο ελαχιστοποιείται η L1 απόσταση μεταξύ του feature map του Discriminator για τα πραγματικά και ψεύτικα δεδομένα. Ως L1 απόσταση μεταξύ δύο διανυσμάτων ορίζεται το άθροισμα των μέτρων τους. Διαισθητικά, μπορούμε να θεωρήσουμε ότι ο Discriminator επιχειρεί να μάθει μια μετρική ομοιότητας, ώστε να μπορεί να διαχωρίζει τα αληθινά από τα ψεύτικα δεδομένα. Η συνάρτηση κόστους της αντιστοίχισης των features ορίζεται ως:

$$\mathcal{L}_{FM}(G, D_k) = \mathbb{E}_{x,s\sim p_{data}}\left[\sum_{i=1}^{T}\frac{1}{N_i}\left\|D_k^{(i)}\big(G(s)\big) - D_k^{(i)}(x)\right\|_1\right] \qquad (16)$$

, όπου το $D_k^{(i)}$αναπαριστά το feature map του i^{οστού} επιπέδου, για το k block του Discriminator και το N_i τον αριθμό των "CNN μονάδων" για αυτό το επίπεδο. Η αντιστοίχιση των features πραγματοποιείται σε κάθε επίπεδο, για όλα τα blocks του Discriminator.

Τελικά, ο Generator εκπαιδεύεται με βάση τον παρακάτω στόχο:

$$\min_G\left(\mathbb{E}_{s,z}\left[\sum_{k\,=\,1,2,3} - \widehat{D_k}\big(G(s,z)\big)\right] + \lambda\sum_{k\,=\,1,2,3}\mathcal{L}_{FM}(G, D_k)\right) \qquad (17)$$

,όπου λ = 10

7 Πρακτικό μέρος – Υλοποίηση

Η βασική ιδέα της δικής μας υλοποίησης είναι ο συνδυασμός των δύο μοντέλων που αναλύθηκαν προηγουμένως, δηλαδή των “A Universal Music Translation Network” και “MelGAN” με σκοπό τη μεταφορά ενός μουσικού μέρους από ένα σύνολο οργάνων σε κάποιο άλλο. Σε αυτό το κεφάλαιο θα αναφερθούμε στην επιλογή των δεδομένων εκπαίδευσης, στην αρχιτεκτονική του συνολικού δικτύου, καθώς και στον καθορισμό της διαδικασίας εκπαίδευσης του δικτύου. Η υλοποίηση πραγματοποιήθηκε σε γλώσσα Python και πιο συγκεκριμένα με χρήση της βιβλιοθήκης PyTorch, η οποία ενδείκνυται για εκπαίδευση βαθιών νευρωνικών δικτύων.

7.1 Επιλογή και Προεπεξεργασία Δεδομένων Εκπαίδευσης

Για την εκπαίδευση του δικτύου χρησιμοποιήσαμε μέρος του συνόλου δεδομένων “MusicNet” το οποίο και αναλύθηκε εκτενέστερα στο κεφάλαιο 5.2. Πιο συγκεκριμένα, επιλέξαμε δεδομένα που προέρχονται μόνο από το domain “Beethoven Solo Piano”. Η διαδικασία που ακολουθήθηκε προκειμένου να εξάγουμε τα δεδομένα εκπαίδευσης είναι η ακόλουθη.

Αρχικά κατεβάσαμε το πλήρες σύνολο δεδομένων το οποίο παρέχεται στον παρακάτω σύνδεσμο: https://homes.cs.washington.edu/~thickstn/start.html. Πιο συγκεκριμένα, επιλέξαμε την raw μορφή των δεδομένων. Έτσι, συνολικά το σύνολο δεδομένων αποτελείται από 330 μονοφωνικά wav αρχεία. Κάθε .wav αρχείο αντιστοιχεί σε μια ολοκληρωμένη μουσική σύνθεση και έχει ρυθμό δειγματοληψίας 44.100 Hz, που αποτελεί και τον συνηθέστερο ρυθμό δειγματοληψίας των CD. Επιπλέον, σε κάθε wav αρχείο αντιστοιχεί και ένα αρχείο .csv το οποίο περιλαμβάνει metadata για το αντίστοιχο μουσικό έργο

Στην συνέχεια, μέσω του parse.py script χωρίσαμε τα δεδομένα σε domains. Όπως είναι κατανοητό η πληροφορία για το domain που ανήκει κάθε wav αρχείο περιλαμβάνεται στο εκάστοτε csv. Από την παραπάνω διαδικασία προέκυψαν τα παρακάτω domains

1. Bach Solo Piano
2. Bach Solo Cello
3. Beethoven Accompanied Violin
4. Beethoven Solo Piano
5. Beethoven String Quartet
6. Cambini Wind Quintet

Κατά την διαδικασία του parsing, το script παρέχει και την συνολική διάρκεια των δεδομένων εκπαίδευσης για κάθε domain. Τα αποτελέσματα αυτά παρουσιάζονται στον παρακάτω πίνακα :

Domain	**Διάρκεια (σε Seconds)**
Bach Solo Piano	8863
Bach Solo Cello	2965
Beethoven Accompanied Violin	4871
Beethoven Solo Piano	37903
Beethoven String Quartet	12853
Cambini Wind Quintet	2577

Πίνακας 1: Συνολική διάρκεια (σε δευτερόλεπτα) κάθε μουσικού domain

Από τα παραπάνω αποτελέσματα γίνεται εμφανής η διαφορά στα διαθέσιμα δεδομένα για κάθε μουσικό domain. Ο βασικό λόγος που επιλέχθηκε λοιπόν το domain “Beethoven Solo Piano” είναι επειδή διαθέτει την μεγαλύτερη εκπροσώπηση, τουλάχιστον σε ό,τι αφορά την συνολική διάρκεια.

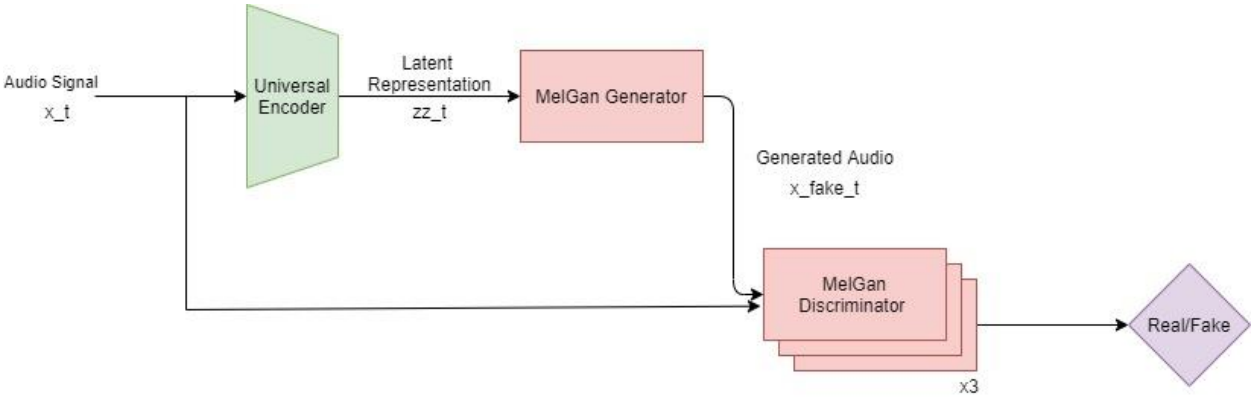
Ως επόμενο βήμα, χωρίσαμε το σύνολο δεδομένων “Beethoven Solo Piano” σε train set και test set. Πιο συγκεκριμένα, κρατήσαμε το 90% των .wav αρχείων στα δεδομένα εκπαίδευσης και το υπόλοιπο 10% ως test data. Έτσι προέκυψαν 83 αρχεία για εκπαίδευση και 9 για test.

Τέλος, για να δημιουργήσουμε το Dataset με το οποίο θα εκπαιδεύσουμε το δίκτυο μας, χωρίσαμε κάθε wav αρχείο σε μικρότερα αρχεία, διάρκειας 10 δευτερολέπτων. Επιπρόσθετα, πραγματοποιήσαμε επαναδειγματοληψία στα δεδομένα μας με συχνότητα 16 kHz ώστε να μειώσουμε τα συνολικό μήκος της εισόδου, οπότε και να αυξήσουμε την ταχύτητα του μοντέλου μας. Με τον χωρισμό των δεδομένων εισόδου σε μικρότερα μουσικά μέρη 10 δευτερολέπτων, παρατηρήσαμε ότι ορισμένα από τα αρχικά αλλά και τα περισσότερα από τα τελικά μουσικά μέρη αποτελούνται είτε από πλήρη ησυχία (πριν και μετά από τη μουσική σύνθεση), είτε από θόρυβο (προετοιμασία, παλαμάκια κ.α.). Γι’ αυτό τον λόγο χειροκίνητα αφαιρέσαμε τέτοια ανεπιθύμητα δεδομένα.

Με βάση τα παραπάνω καταλήξαμε με ένα σύνολο δεδομένων που αποτελείται από 3330 δεδομένα εκπαίδευσης διάρκειας 10 δευτερολέπτων και 384 test data. Παρατηρούμε λοιπόν, ότι βρισκόμαστε αρκετά κοντά στον αρχικό χωρισμό των διαθέσιμων δεδομένων, με ποσοστά 90%-10% σε train και test set αντίστοιχα.

7.2 Αρχιτεκτονική Δικτύου

Εποπτικά, το συνολικό μοντέλο που εκπαιδεύσαμε αποτελείται από τον κωδικοποιητή (Encoder) του “A Universal Music Translation Network” και από το GAN μοντέλο “MelGAN”, το οποίο δέχεται στην είσοδο του, αντί για το Mel – σπεκτρόγραμμα του σήματος ήχου, την “πληροφορία” που παράγει ο κωδικοποιητής. Συνολικά, το δίκτυο παρουσιάζεται στο παρακάτω σχήμα:

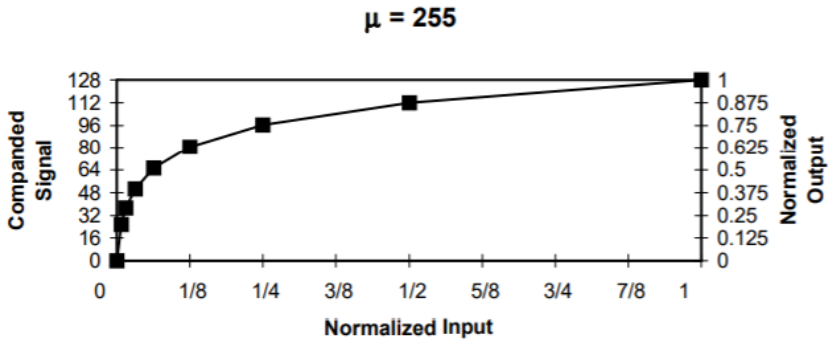


Σχήμα 1: Αρχιτεκτονική τελικού δικτύου (Universal MelGan)

7.2.1 Universal Encoder

Σε ότι αφορά τον κωδικοποιητή του “A Universal Music Translation Network”, θα αναφερόμαστε σε αυτόν για το συνολικό δίκτυο μας ως Universal Encoder. Ο Universal Encoder που χρησιμοποιήσαμε λοιπόν είναι ίδιος με αυτόν που ορίστηκε στο κεφάλαιο 6.1.2, δηλαδή αποτελείται από τις ίδιες παραμέτρους. Ο Encoder που χρησιμοποιήσαμε ήταν ήδη εκπαιδευμένος στο πλήρες σύνολο δεδομένων “MusicNet” και παρέχεται από τους δημιουργούς του μοντέλου. Με αυτή την έννοια, στο PyTorch αρχικά δημιουργήσαμε ένα δίκτυο ίδιας αρχιτεκτονικής με τον Universal Encoder, και στην συνέχεια, μέσω της συνάρτησης load_state_dict() φορτώσαμε τις παραμέτρους του pre-trained Universal Encoder.

Ο Universal Encoder δέχεται στην είσοδο του απευθείας το σήμα ήχου, το οποίο πρώτα έχει υποστεί μ-law companding, το οποίο αναλύθηκε εκτενέστερα στο κεφάλαιο 6.1.1, ώστε η είσοδος να μετατραπεί σε συγκεκριμένες στάθμες κβάντισης. Μάλιστα, ο Universal Encoder δεν δέχεται κανονικοποιημένες τις τιμές του μ-law compression, αλλά η είσοδος παραμένει στο εύρος [0, 128]. Παρακάτω παρουσιάζεται ένα ενδεικτικό σχήμα του μ-law companding:



Σχήμα 2: μ-law companding

Όπως ήδη αναφέραμε, ο Universal Encoder παράγει μια κρυμμένη ακολουθία που αποτελείται από 64 κανάλια, πραγματοποιώντας μείωση των δειγμάτων εισόδου με παράγοντα 800. Έτσι αν η είσοδος αποτελείται από ένα σήμα μήκους len(input), τότε ο Universal Encoder θα παράξει έναν tensor διαστάσεων [len(batch_size), 64, len(signal)], δηλαδή έναν τρισδιάστατο tensor, ο οποίος στην πρώτη διάσταση έχει μήκος όσο και το batch size που θα ορίσουμε κατά την εκπαίδευση.

Το PyTorch, αντί για numpy.arrays, δηλαδή πίνακες, χειρίζεται τη μορφή δεδομένων “Tensor”. Οι Tensors λοιπόν, μπορούν να θεωρηθούν παρόμοιοι με τους numpy.arrays, καθώς μπορούν να εφαρμοστούν όλες οι πράξεις πινάκων μεταξύ τους. Ωστόσο, η κύρια διαφορά τους είναι ότι οι Tensors μπορούν να τρέξουν είτε σε CPU ή σε GPU, διευκολύνοντας σε πολύ σημαντικό βαθμό την εκπαίδευση βαθιών νευρωνικών δικτύων, η οποία επιταχύνεται σημαντικά εάν πραγματοποιηθεί σε μια GPU.

7.2.2 MelGAN Δίκτυο

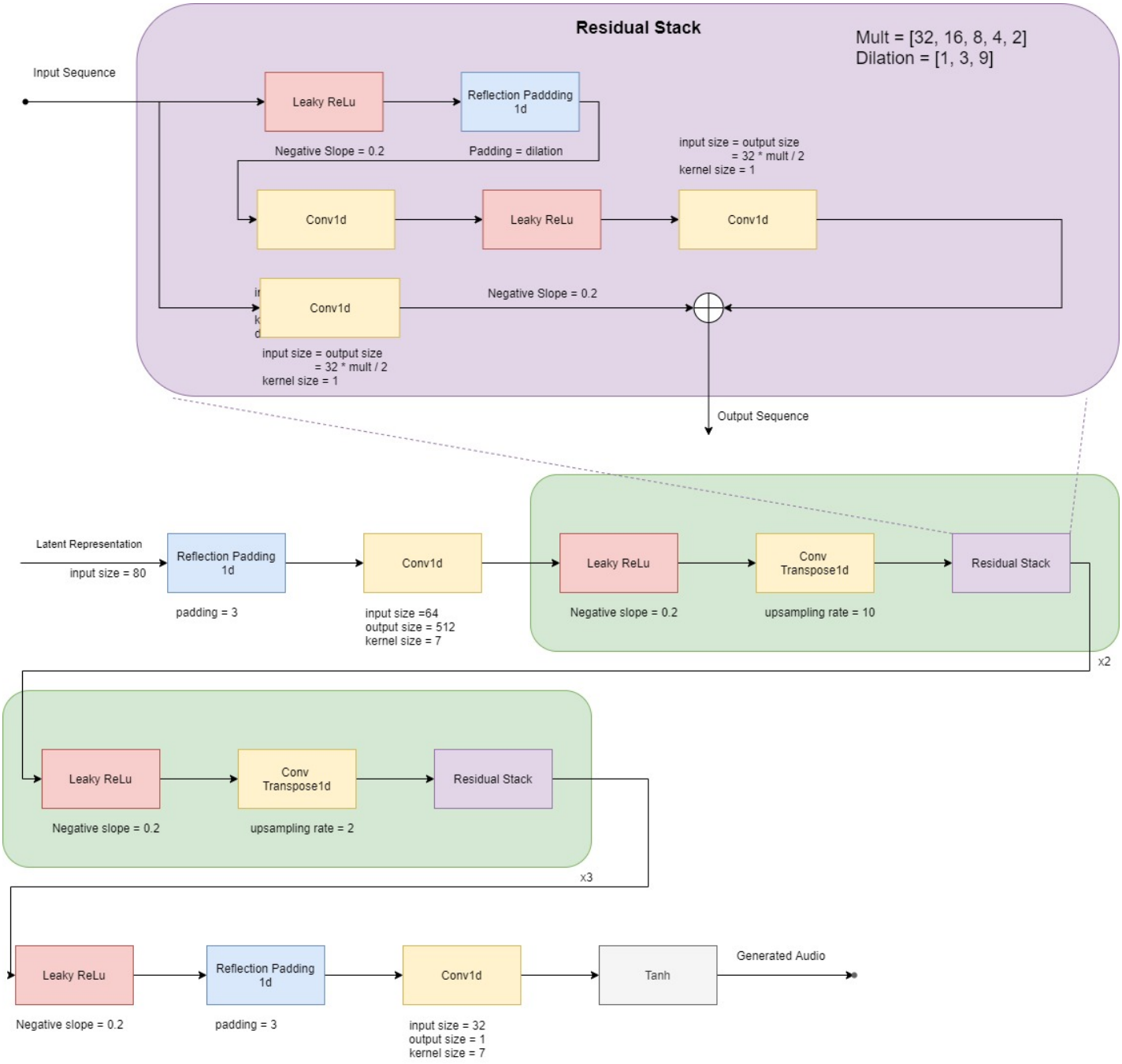
Σύμφωνα με τους δημιουργούς του MelGAN, το δίκτυο δέχεται στην είσοδο του ένα σπεκτρόγραμμα του αρχικού σήματος εισόδου. Το σπεκτρόγραμμα που χρησιμοποιούν για το αρχικό δίκτυο αποτελείται από 80 κανάλια και πραγματοποιεί downsampling στο πεδίο του χρόνου με παράγοντα 256. Θα πρέπει λοιπόν να προσαρμόσουμε κατάλληλα το δίκτυο μας, ώστε να δέχεται την είσοδο του την ενδιάμεση αναπαράσταση που παράγει ο Universal Encoder και αντίστοιχα, ο Generator να ανακατασκευάζει ένα σήμα ίδιου μήκους με το αρχικό.

Από την στιγμή που η νέα είσοδος που λαμβάνει ο MelGan generator αποτελείται από 80, και όχι 64 κανάλια, πρέπει να προσαρμόσουμε το πρώτο layer του δικτύου το οποίο δέχεται την είσοδο. Έτσι ορίζουμε στο πρώτο Convolutional Layer να έχει input size = 64, ενώ το output size του, το κρατάμε σταθερό και ίσο με 512 κανάλια, όπως ορίζουν οι δημιουργοί του αρχικού μοντέλου.

Όπως αναλύσαμε στο κεφάλαιο 6.2.2, ο αρχικός Generator του MelGan πραγματοποιεί upsampling σε 4 στάδια με παράγοντες [8, 8, 2, 2] και ένα residual block να ακολουθεί κάθε

upsampling layer. Στην δική μας περίπτωση, για να πραγματοποιηθεί upsampling με παράγοντα 800, όσο δηλαδή και το downsampling που εφαρμόστηκε στην είσοδο από τον Universal Encoder, δημιουργούμε 5 upsampling blocks με παράγοντες [10, 10, 2, 2, 2]

Το δίκτυο του Generator παρουσιάζεται αναλυτικά στο παρακάτω σχήμα:



Σχήμα 3: Αρχιτεκτονική Universal MelGan Generator

Σε ότι αφορά τον Discriminator του δικτύου μας, διατηρήσαμε ακριβώς την ίδια δομή με αυτή που προτείνουν οι δημιουργοί του αρχικού δικτύου. Έτσι, έχουμε 3 διαφορετικούς Discriminators που δέχονται το σήμα εισόδου σε διαφορετική κλίμακα. Πιο συγκεκριμένα για τον 2ο και 3ο Discriminator πραγματοποιείται ένα downsampling με παράγοντα 2 και 4 αντίστοιχα.

7.2.3 Παράμετροι Δικτύων

Με βάση τα παραπάνω, προκύπτει ο παρακάτω πίνακας σε ότι αφορά τις παραμέτρους με τις οποίες ορίζουμε κάθε δίκτυο

Δίκτυο	**Πλήθος Παραμέτρων**
Universal Encoder	1,982,528
Universal Generator	18,998,850
Universal Discriminators (συνολικά 3)	5,641,362
Συνολικές Παράμετροι	26,622,740
Συνολικές Εκπαιδεύσιμες Παράμετροι	24,640,212

Πίνακας 2: Παράμετροι Δικτύων

Γίνεται λοιπόν προφανές ότι η εκπαίδευση του μοντέλου πρέπει οπωσδήποτε να πραγματοποιηθεί σε GPU, ώστε να ολοκληρωθεί σε κάποιο εύλογο χρονικό διάστημα

7.3 Διαδικασία Εκπαίδευσης

Η βασική ιδέα για την εκπαίδευση του δικτύου είναι να μάθει πολύ καλά να ανακατασκευάζει τα δεδομένα που προέρχονται από αυτό το μουσικό Domain. Έτσι, όταν στην φάση του testing του δώσουμε στην είσοδο κάποιο μουσικό μέρος από διαφορετικό domain, αυτό θα επιχειρήσει να το μεταφέρει στο domain πάνω στο οποίο εκπαιδεύτηκε.

Αφού ορίσουμε τα δίκτυα απ' τα οποία αποτελείται το μοντέλο μας, πρέπει να ορίσουμε τους DataLoaders τόσο για το train set αλλά και για το test set. Επιπλέον πρέπει να ορίσουμε την κατάλληλη διαδικασία εκπαίδευσης.

Αρχικά ορίζουμε τους DataLoaders για τα δεδομένα μας. Η κλάση torch.utils.data.DataLoader του PyTorch ορίζει έναν iterator του αντίστοιχου Dataset. Σε κάθε εποχή λοιπόν, για κάθε αρχείο

εκπαίδευσης , το οποίο όπως αναφέραμε προηγουμένως αποτελείται από ένα μουσικό μέρος 10 δευτερολέπτων, επιλέγουμε τυχαία ένα κομμάτι 1 δευτερολέπτου. Για την εκπαίδευση επιλέξαμε σταθερό batch size = 16, ενώ ορίσαμε ως συνολικό πλήθος εποχών τις 3000. Με αυτό τον τρόπο, κάθε εποχή αποτελείται από 209 batches. Αντίστοιχα, τα test δεδομένα προκύπτουν από την τυχαία επιλογή 4 δευτερολέπτων από κάθε αρχείο.

Για τον Generator αλλά και για τους 3 discriminators χρησιμοποιήσαμε τον Adam optimizer, ο οποίος χρησιμοποιείται πολύ συχνά ως ενναλακτική επιλογή του Stochastic Gradient Descent, κυρίως σε προβλήματα επεξεργασίας εικόνας και επεξεργασίας φυσικής γλώσσας. Η βασική διαφορά του Adam Optimizer σε σχέση με το Stochastic Gradient Descent, είναι ότι το δεύτερο διατηρεί έναν κοινό ρυθμό εκπαίδευσης (Learning rate) για την ανανέωση όλων των βαρών του δικτύου, καθ' όλη την διάρκεια της εκπαίδευσης. Αντίθετα, ο Adam optimizer, προσδίδει έναν ξεχωριστό ρυθμό εκπαίδευσης σε κάθε παράμετρο του δικτύου, ο οποίος μάλιστα προσαρμόζεται κατάλληλα κατά την διάρκεια της εκπαίδευσης. Έτσι, υπό μια έννοια, κατά την διαδικασία εκπαίδευσης, εκπαιδεύεται τόσο το μοντέλο μας, όσο και ο optimizer.

Προκειμένου να ορίσουμε τον Adam Optimizer πρέπει να ορίσουμε τις παραμέτρους του. Ως αρχικό ρυθμό εκπαίδευσης (learning rate) επιλέξαμε την τιμή 0.0001. Επιπλέον, ως β1 το οποίο αντιστοιχεί στον συντελεστή κλίσης του ρυθμού εκπαίδευσης, επιλέξαμε την τιμή 0.5, και ως β2 (συντελεστής του τετραγώνου της κλίσης) επιλέχθηκε η τιμή 0.9.

Τέλος, η ανανέωση των βαρών του συνολικού μοντέλου πραγματοποιείται με backpropagation των συναρτήσεων απώλειας του MelGAN μοντέλου, όπως αυτές ορίστηκαν στο κεφάλαιο 6.2.2. Με αυτή την έννοια, έχουμε 3 ειδών απώλειες: την απώλεια του Generator, τις απώλειες των 3 Discriminators αλλά και την απώλεια που προκύπτει απ' το Feature Matching σε κάθε layer των Discriminators.

8 Αξιολόγηση Αποτελεσμάτων

Η εκπαίδευση του δικτύου πραγματοποιήθηκε στην Pro Version του Google Colab, η οποία παρέχει αυξημένη πρόσβαση σε GPU. Αρχικά, ορίσαμε 3000 εποχές, ενώ η GPU που κατά κύριο λόγω παρείχε το Google Colab (ανάλογα με την ζήτηση, είναι πολύ πιθανό να αλλάξει η GPU) ήταν η Tesla P100-PCIE-16GB. Η εκπαίδευση διήρκησε συνολικά 5 ημέρες.

Παρακάτω θα αναφερθούμε στα κριτήρια αξιολόγησης του μοντέλου που εκπαιδεύσαμε, αλλά και σε πιθανές ιδέες για μελλοντικές επεκτάσεις, οι οποίες βέβαια ξεπερνούν τον σκοπό αυτής της εργασίας.

8.1 Αξιολόγηση Μοντέλου

Η αξιολόγηση των Παραγωγικών Αντιπαραθετικών Δικτύων είναι ένα βασικό πρόβλημα τέτοιων μοντέλων, αλλά και ένας ερευνητικός τομέας που εξελίσσεται τα τελευταία χρόνια. Τα περισσότερα βαθιά νευρωνικά δίκτυα εκπαιδεύονται με βασικό στόχο να ελαχιστοποιήσουν την συνάρτηση απώλειας τους, μέχρις ότου προέλθει σύγκλιση σε μια τιμή. Ωστόσο, από την στιγμή που τα Παραγωγικά Αντιπαραθετικά Δίκτυα δεν έχουν μια συγκεκριμένη συνάρτηση απώλειας, αλλά επιχειρούν να εκπαιδεύσουν ταυτόχρονα τον Generator και τον Discriminator, η εφαρμογή μιας τέτοιας μεθόδου αξιολόγησης δεν είναι δυνατή.

Με αυτή την έννοια, οι τιμές των απωλειών/ συναρτήσεων κόστους, τόσο του Παραγωγικού όσο και του Αντιπαραθετικού δικτύου, δεν είναι ενδεικτικές για την πρόοδο της εκπαίδευσης και την ικανότητα του Generator να παράγει αξιόπιστα δεδομένα. Γι’ αυτόν τον λόγο, έχει αναπτυχθεί μια σειρά από ποιοτικές αλλά και ποσοτικές τεχνικές εκτίμησης της ποιότητας και ποικιλίας των παραγόμενων δεδομένων. Οι περισσότερες από αυτές τις τεχνικές εφαρμόζονται με ιδιαίτερη επιτυχία σε προβλήματα παραγωγής εικόνων. Ωστόσο, ορισμένες από αυτές μπορούν να υιοθετηθούν και σε προβλήματα παραγωγής ηχητικών σημάτων.

Βέβαια, θα πρέπει να έχουμε υπόψιν, ότι από την στιγμή που δεν υπάρχει κάποια καθολική μέθοδος αξιολόγησης των GAN μοντέλων, δεν υπάρχει και συμφωνημένος τρόπος σύγκρισης της αποτελεσματικότητας τους. Αυτό μάλιστα, αποτελεί ένα πρόβλημα που παραμένει ανοιχτό και ήδη από το 2018 γίνεται μεγάλη προσπάθεια ώστε να βρεθεί μια συγκεκριμένη μέθοδος αξιολόγησης αυτών των δικτύων.

Επιπλέον, βάσει των παραπάνω δεδομένων, δεν είναι καθόλου ξεκάθαρο πότε θα πρέπει να σταματήσει η εκπαίδευση ενός τέτοιου μοντέλου. Γι’ αυτόν τον λόγο μια συνήθης πρακτική είναι να σώζουμε τα μοντέλα ανά τακτά διαστήματα εποχών. Στην δική μας περίπτωση επιλέξαμε να σώζουμε το μοντέλο μας ανά 1000 επαναλήψεις του training loop, δηλαδή ανά 1000 batches, που πρακτικά ισοδυναμούν με 5 εποχές.

Στην συνέχεια, θα αναλύσουμε ορισμένες από τις συνηθέστερες τεχνικές αξιολόγησης των Παραγωγικών Αντιπαραθετικών Δικτύων και το αν ήταν δυνατόν να τις εντάξουμε ως μεθόδους αξιολόγησης στο δικό μας μοντέλο.

Inception Score

Μια από τις βασικότερες τεχνικές ποσοτικής αξιολόγησης του μοντέλου είναι η μέτρηση του Inception Score. Το Inception Score αποτελεί μια αντικειμενική μετρική που αξιολογεί την ποιότητα συνθετικών εικόνων. Η μετρική αυτή προτάθηκε από τον Tim Salimans το 2016 [31] και από τότε χρησιμοποιείται ευρέως για την αξιολόγηση τέτοιων μοντέλων.

Η τεχνική αυτή, εισάγει ένα ήδη εκπαιδευμένο βαθύ νευρωνικό δίκτυο το οποίο κατηγοριοποιεί τις παραγόμενες εικόνες. Πιο συγκεκριμένα το δίκτυο αυτό ονομάζεται Inception v3 model και από αυτό πήρε το όνομα του η μετρική. Με βάση την έξοδο του Inception δικτύου προκύπτει και το Inception Score. Είναι λοιπόν προφανές ότι για ένα πρόβλημα όπως το δικό μας, που αφορά την παραγωγή ηχητικών σημάτων, δεν είναι δυνατόν να εφαρμόσουμε αυτή την τεχνική.

Ποιοτική Αξιολόγηση Αποτελεσμάτων

Μια συνήθης τεχνική αξιολόγησης ενός GAN μοντέλου είναι η ποιοτική αξιολόγηση από ανθρώπους. Πιο συγκεκριμένα, ένα καθορισμένο δείγμα ανθρώπων, που συνήθως αποτελεί έναν συνδυασμό από ειδικούς του χώρου (στην δική μας περίπτωση μουσικούς), ερασιτέχνες αλλά και ανθρώπους που δεν έχουν κάποια σχέση με το πεδίο στο οποίο εντάσσεται το πρόβλημα. Μάλιστα, αυτή η τεχνική αποτελεί και την συνηθέστερη αλλά και πιο διαισθητικά σωστή μέθοδος αξιολόγησης ενός GAN μοντέλου.

Στην δική μας περίπτωση, απευθυνθήκαμε σε ένα μικρό δείγμα 5 ανθρώπων, ένας εξ ’αυτών επαγγελματίας μουσικός, δύο ερασιτέχνες και δύο που δεν είχαν κάποια επαφή με οποιοδήποτε μουσικό όργανο ή θεωρία της μουσικής. Προκειμένου να αξιολογηθεί το μοντέλο μας, επιλέξαμε 3 test δεδομένα και 2 δεδομένα εκπαίδευσης τα οποία ανήκαν στο domain που εκπαιδεύσαμε το δίκτυο μας. Ο σκοπός λοιπόν, ήταν να περάσουμε αυτά τα δεδομένα μέσα απ’ το μοντέλο και αυτό να επιχειρήσει να τα ανακατασκευάσει. Αυτή η διαδικασία γινόταν επαναληπτικά ανά 5 εποχές, ώστε να μπορέσουμε να αξιολογήσουμε ποιοτικά την εξέλιξη του μοντέλου μας.

Έτσι, κατά μέσο όρο ανά 100 εποχές διαθέταμε τα παραγόμενα δείγματα για αξιολόγηση, τόσο σε σύγκριση με τα αυθεντικά μουσικά μέρη, όσο και με τα παραγόμενα δείγματα προηγούμενων εποχών.

Συνολικά, παρατηρήσαμε, ότι τα παραγόμενα δείγματα ξεκίνησαν να παραπέμπουν στα αρχικά μετά από 200 – 250 εποχές, καθώς προηγουμένως η έξοδος ήταν πολύ θορυβώδης. Στην συνέχεια,

η ποιότητα των παραγόμενων δεδομένων βελτιωνόταν αισθητά ανά 100 εποχές, μέχρι περίπου και την 900ή εποχή. Από αυτό το σημείο, μέχρι και την 1300ή εποχή η βελτίωση γινόταν με πολύ μικρό ρυθμό, ενώ από εκείνο περίπου το σημείο μέχρι και την 1800η εποχή η ποιότητα των δειγμάτων παρέμεινε σταθερή και πολύ κοντά στην αρχική.

Βέβαια, σε σχέση με τα πρωτότυπα δεδομένα, αυτά που παρήγαγε το μοντέλο μας είχαν ένα πολύ μικρό επίπεδο θορύβου. Επιπλέον, στα test δεδομένα, ήταν πιθανό τα παραγόμενα δεδομένα να μην έχουν τις ίδιες δυναμικές με τα αρχικά. Εντούτοις, αυτά τα δύο προβλήματα τα συναντάμε και στα δίκτυα που βασίστηκε το μοντέλο μας, δηλαδή τα "A Universal Music Translation Network" και "MelGAN".

Σφάλμα Ανακατασκευής

Η μετρική που καθόρισε ποσοτικά την εξέλιξη και την πορεία εκμάθησης του μοντέλου μας ήταν το σφάλμα ανακατασκευής της ενδιάμεσης αναπαράστασης που προκύπτει απ' την έξοδο του Universal Encoder.

Τα περισσότερα Παραγωγικά Αντιπαραθετικά Δίκτυα που επιχειρούν να ανακατασκευάσουν αυθεντικά δεδομένα, δεχόμενα στην είσοδο τους μια διαφορετική αναπαράσταση τους, συνήθως χρησιμοποιούν ως ποσοτική μετρική αξιολόγησης το σφάλμα ανακατασκευής των αυθεντικών δεδομένων. Ωστόσο, εμείς ακολουθήσαμε την προτροπή των δημιουργών του MelGAN δικτύου και μετράμε το σφάλμα μεταξύ της κρυμμένης αναπαράστασης που προκύπτει από την έξοδο του Universal Encoder, μεταξύ των πραγματικών και παραγόμενων δειγμάτων.

Γι' αυτό τον λόγο, όποτε δημιουργούσαμε παραγόμενα δεδομένα για ποιοτική αξιολόγηση, μετρούσαμε και καταγράφαμε το σφάλμα ανακατασκευής που αντιστοιχούσε σε εκείνη την εποχή. Μάλιστα, αποθηκεύαμε και το δίκτυο που παράγει το μικρότερο δυνατό σφάλμα, καθ' όλη την διάρκεια της εκπαίδευσης.

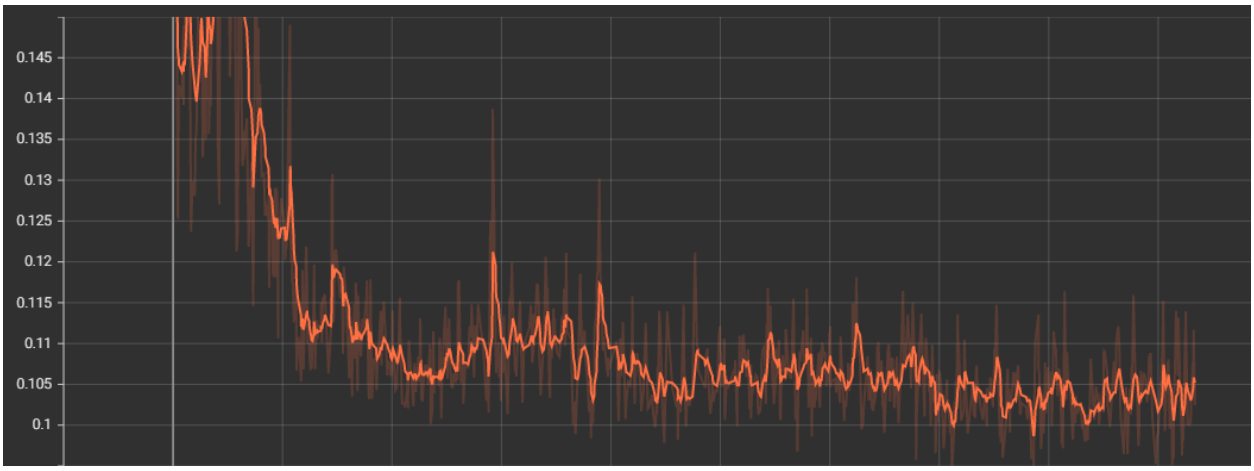
Σε αυτό το σημείο να αναφέρουμε ότι ως σφάλμα ανακατασκευής πήραμε την μέση τιμή της L1 απόστασης μεταξύ της παραγόμενης και πραγματικής ενδιάμεσης κρυμμένης αναπαράστασης. Ενδεικτικά, η L1 απόσταση ορίζεται ως εξής:

$$L1Loss\ Function\ = \sum_{i=1}^{n} |zz_{true} - zz_{Predicted}|$$

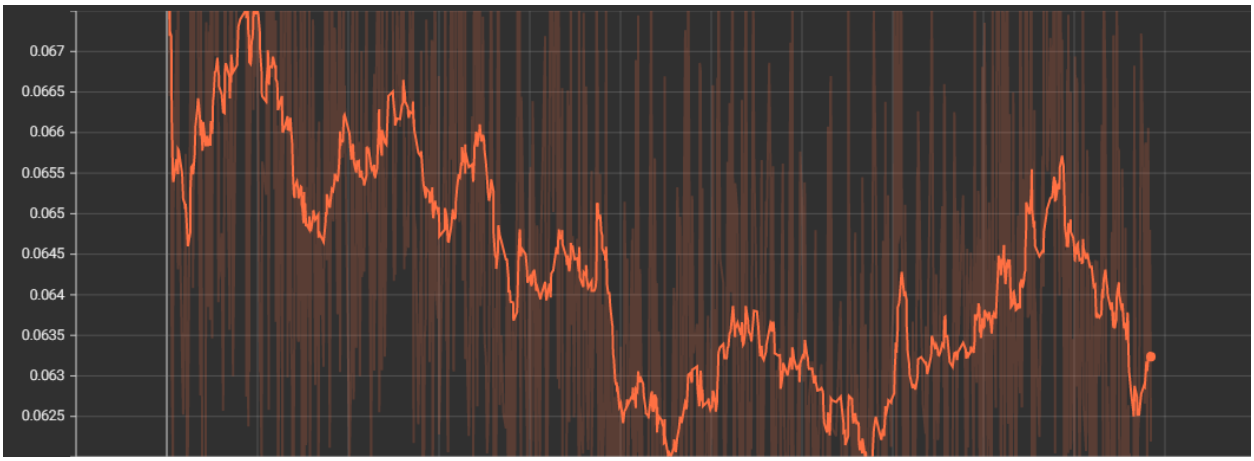
, όπου ως zz ορίζουμε την έξοδο του Universal Encoder.

Ενδεικτικά παρουσιάζουμε το κανονικοποιημένο σφάλμα ανακατασκευής για τις εποχές 1-50 αλλά και 1270 – 1800, όπου και θεωρήσαμε ότι πλέον το μοντέλο είχε συγκλίνει στο βέλτιστο

δυνατό. Από τα παρακάτω διαγράμματα γίνεται εμφανής αφενός η ταχεία εκπαίδευση του δικτύου στις πρώτες εποχές αλλά και η σύγκλιση της ποιότητας των αποτελεσμάτων κατά το τέλος της εκπαίδευσης.



Σχήμα 1: Κανονικοποιημένο Σφάλμα Ανακατασκευής για τις εποχές 1-50



Σχήμα 2: Κανονικοποιημένο Σφάλμα Ανακατασκευής για τις εποχές 1280-180

8.2 Τέλος Εκπαίδευσης - Αξιολόγηση στο πρόβλημα μεταφοράς μουσικής

Ενδεικτικά το μικρότερο σφάλμα ανακατασκευής μετρήθηκε στην εποχή 1370. Η εκπαίδευση του δικτύου συνεχίστηκε μέχρι την εποχή 1800. Αφού λοιπόν δεν παρατηρήσαμε κάποια ποιοτική βελτίωση των παραγόμενων δεδομένων, ούτε και μείωση του σφάλματος ανακατασκευής αποφασίσαμε να διακόψουμε την εκπαίδευση και να κρατήσουμε το δίκτυο με το μικρότερο σφάλμα ανακατασκευής μέχρι εκείνη την χρονική στιγμή.

Ακολούθως, παρείχαμε στο δίκτυο test δεδομένα από διαφορετικά μουσικά domains. Πιο συγκεκριμένα επιλέξαμε δείγματα 10 δευτερολέπτων από τους εξής domains:

1. Bach Solo Cello
2. Beethoven String Quartet
3. Cambini Wind Quintet
4. Classical Guitar
5. Hard Rock
6. Jazz

Δεδομένου ότι το MelGan δίκτυο είχε εκπαιδευτεί μόνο σε δεδομένα που ανήκουν στο μουσικό domain “Beethoven Solo Piano”, δίνοντας στον Generator την κρυμμένη αναπαράσταση ενός διαφορετικού μουσικού domain, περιμέναμε να πραγματοποιήσει μια μουσική μεταφορά του συγκεκριμένου μουσικού μέρους σε Piano.

Πράγματι, το μοντέλο ανταποκρίνεται αρκετά καλά σε αυτό το task, καθώς προσδίδει χαρακτηριστικά του παιξίματος ενός πιάνο αλλά και το ηχόχρωμα του στα παραγόμενα δεδομένα. Το μοντέλο παρουσιάζει πολύ αξιόπιστα και αληθοφανή δεδομένα που προέρχονται απ’ τα Domains Beethoven String Quartet και Cambini Wind Quintet. Αντίθετα, παρουσιάζεται υψηλός θόρυβος και ανακρίβειες στην μεταφορά μουσικών μερών του domain Bach Solo Cello.

Ωστόσο, το μεγαλύτερο ενδιαφέρον επικεντρώνεται στα 3 τελευταία μουσικά domains, πάνω στα οποία δεν έχει εκπαιδευτεί ούτε ο pre-trained Universal Encoder. Παρατηρούμε ότι σε αυτές τις περιπτώσεις η ποιότητα των αποτελεσμάτων ποικίλει σημαντικά. Πιο συγκεκριμένα, η μεταφορά από Hard Rock σε Solo Piano κρίθηκε ανεπιτυχής, καθώς στα παραγόμενα δεδομένα έχουν ενταχθεί πολύ υψηλά επίπεδα θορύβου. Αντίθετα, η μεταφορά από το Domain Classical Guitar, είναι ιδιαίτερα κοντά στα αυθεντικά δεδομένα. Αυτό πιθανότατα οφείλεται στο ότι το πιάνο και η κιθάρα είναι αρκετά κοντινά όργανα ως προς τη μουσική τους έκταση αλλά και καθώς και τα δύο είναι πολυφωνικά. Τέλος, ιδιαίτερα πετυχημένη παρουσιάζεται η μεταφορά από το Domain της Jazz, καθώς σε πολλές περιπτώσεις τα παραγόμενα αποτελέσματα παραπέμπουν στον τρόπο που ένας πραγματικός πιανίστας θα απέδιδε ένα Jazz κομμάτι που η βασική μελωδία παίζεται από πνευστά όργανα.

9 Μελλοντικές Επεκτάσεις

Παρατηρούμε ότι η υλοποίηση που πραγματοποιήσαμε παράγει υποσχόμενα αποτελέσματα, καθώς τα παραγόμενα δείγματα προσομοιάζουν σε μεγάλο βαθμό το πως θα έπαιζε ένας πιανίστας ένα μουσικό μέρος που στην πραγματικότητα εκτελείται από ένα διαφορετικό σύνολο οργάνων. Οι μελλοντικές επεκτάσεις του μοντέλου που προτάθηκε θα μπορούσαν να κινηθούν προς δύο κατευθύνσεις.

Αρχικά, μπορούμε να εκπαιδεύσουμε το δίκτυο μειώνοντας τις συνολικές εκπαιδεύσιμες παραμέτρους. Ενδεικτικά, μπορούμε να μειώσουμε το βάθος των συνελικτικών δικτύων του MelGAN Generator και να μελετήσουμε την ποιότητα των παραγόμενων δειγμάτων για διαφορετικά ποσοστά μείωσης. Αντίστοιχα, θα μπορούσαμε να μελετήσουμε την ποιότητα των παραγόμενων δειγμάτων εάν μειώσουμε το πλήθος των διαχωριστικών δικτύων (Discriminators).

Μια δεύτερη ιδέα θα ήταν να αντικαταστήσουμε τον pre-trained Universal Encoder που χρησιμοποιήσαμε με ένα νέο δίκτυο το οποίο θα πραγματοποιούσε κωδικοποίηση του σήματος ήχου. Η βασική διαφορά σε αυτή την περίπτωση είναι ότι το δίκτυο αυτό μπορεί να θεωρηθεί ως μια επέκταση του MelGAN Generator ο οποίος τελικά θα έχει μια μορφή ενός Autoencoder δικτύου. Έτσι η εκπαίδευση θα πραγματοποιηθεί end-to-end, δίχως να χρησιμοποιήσουμε κάποιο ήδη εκπαιδευμένο δίκτυο που θα έχει την μορφή ενός WaveNet μοντέλου.

10 Βιβλιογραφία

[1] https://study.com/academy/lesson/what-is-music-definition-terminology-characteristics.html

[2] UKEssays. (November 2018). Why Do the Same Notes Sound Different in Instruments?. Retrieved from https://www.ukessays.com/essays/physics/notes-sound-instruments-8936.php?vref=1

[3] http://www.nagyvaryviolins.com/tonequality.html

[4] https://www.britannica.com/art/arrangement

[5] Mincham, J. (2016) the Cantatas of Johan Sebastian Bach. http://www.jsbachcantatas.com/documents/chapter-85-bwv-29/

[6] "The art of Arranging", Shirley M. Rider

[7] "The Technique of Orchestration" Kent Wheeler Kennan, Prentice-Hall; 2nd edition (January 1, 1972)

[8] L. Wyse. 2017. Audio Spectrogram Representations for Processing with Convolutional Neural Networks. Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN. Anchorage, US. May, 2017. 1(1). pp 37-41

[9] Nistal, Javier & Lattner, Stefan & Richard, Gaël. (2020). Comparing Representations for Audio Synthesis Using Generative Adversarial Networks.

[10] Cheuk, Kin Wai & Agres, Kat & Herremans, Dorien. (2020). The impact of Audio input representations on neural network based music transcription.

[11] S. Dieleman, A. van den Oord, and K. Simonyan, “The challenge of realistic music generation: modelling raw audio at scale,” in NeurIPS, Montréal, Canada, Dec. 2018, pp. 8000–8010.

[12] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. CoRR abs/1609.03499, 2016.

[13] Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. Gansynth: Adversarial neural audio synthesis. arXiv preprint arXiv:1902.08710, 2019a

[14] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville. MelGAN: Generative adversarial networks for conditional waveform synthesis. arXiv preprint arXiv:1910.06711, 2019.

[15] Lawrence R. Rabiner, Ronald W. Schafer "Ψηφιακή Επεξεργασία Φωνής, Θεωρία και Εφαρμογές", εκδόσεις Π.Χ. Πασχαλίδης

[16] Κωνσταντίνος Διαμαντάρας "Τεχνητά Νευρωνικά Δίκτυα", εκδόσεις Κλειδάριθμος

[17] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., Simonyan, K.: Neural audio synthesis of musical notes with WaveNet autoencoders. In: ICML. (2017)

[18] Mor, N., Wolf, L., Polyak, A., and Taigman, Y. Autoencoder-based music translation. In InternationalConference on Learning Representations, 2019. URL https://openreview.net/forum?id=HJGkisCcKm

[19] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron C Courville, ´ "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in Advances in Neural Information Processing Systems, 2019, pp. 14881–14892.

[20] van den Oord, Aäron, Kalchbrenner, Nal, and Kavukcuoglu, Koray. Pixel recurrent neural networks. arXiv preprint arXiv:1601.06759, 2016a.

[21] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. CoRR, abs/1512.03385, 2015.

[22] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., Simonyan, K.: Neural audio synthesis of musical notes with WaveNet autoencoders. In: ICML. (2017)

[23] Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). In: International Conference on Learning Representations (ICLR). (2017)

[24] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron C Courville, ´ "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in Advances in Neural Information Processing Systems, 2019, pp. 14881–14892.

[25] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Advances in neural information processing systems, pp. 2672–2680, 2014.

[26] https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29

[27] Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8798–8807, 2018b.

[28] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.1125–1134, 2017.

[29] Lim, J. H. and Ye, J. C. Geometric gan. arXiv preprint arXiv:1705.02894, 2017.

[30] Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300, 2015.

[31] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In Advances in Neural Information Processing Systems, pp. 2234–2242, 2016.