ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΥΤΟΜΑΤΗ ΕΞΑΓΩΓΗ ΗΧΗΤΙΚΩΝ ΠΗΓΩΝ ΑΠΟ ΗΧΟΓΡΑΦΗΣΕΙΣ ΜΕ ΒΑΘΕΙΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Άγγελος Μπούσης

Επιβλέπων: Νικόλαος Μητιανούδης Αναπλ. Καθηγητής ΔΠΘ

Οκτώβριος 2020, Ξάνθη

Στους γονείς μου

Ευχαριστίες

Ευχαριστώ πολύ τους γονείς μου που στηρίζουν κάθε μου προσπάθεια με όλη τους την αγάπη. Θα ήθελα επιπλέον να ευχαριστήσω τους φίλους μου και "τα αδέρφια μου" που ήταν πάντα δίπλα μου, και για αυτό το υπέροχο τελευταίο καλοκαίρι στην Ξάνθη. Τέλος, ευχαριστώ τον κύριο Νικόλαο Μητιανούδη, για την εμπιστοσύνη που μου έδειξε, και την πολύτιμη καθοδήγησή του.

Περιεχόμενα

Σημειογραφία

α	Ένα βαθμωτό μέγεθος(ακέραιο ή πραγματικό)
$\boldsymbol{\alpha}$	Ένα διάνυσμα
$\mathbf{A}$	Ένας πίνακας
$\mathbf{I}$	Μοναδιαίος πίνακας με διαστάσεις που υποδηλώνονται από το εκάστοτε πρόβλημα
$diag(\boldsymbol{\alpha})$	Ένας τετραγωνικός, διαγώνιος πίνακας με διαγώνιες εισόδους που δίνονται από το $\boldsymbol{\alpha}$
$\mathbb{A}$	Ένα σύνολο
$\mathbb{N}$	Το σύνολο των φυσικών αριθμών
$\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$	Το σύνολο των ακεραίων αριθμών
$[\alpha : b] = \{\alpha, \alpha + 1, \dots, b\} \subset \mathbb{Z}$	Το σύνολο των ακεραίων από το α εώς το b για $\alpha, b \in \mathbb{Z}$
$\mathbb{R}$	Το σύνολο των πραγματικών αριθμών
$\mathbb{R}_{\geq 0} = \{\alpha \in \mathbb{R} \vert \alpha \geq 0\}$	Το σύνολο των μη αρνητικών πραγματικών αριθμών
$[\alpha, b]$	Το πραγματικό εύρος που περιέχει τα α και b
$(\alpha, b]$	Το πραγματικό εύρος που εξαιρεί το α αλλά περιέχει το b
$\mathbb{A} \backslash \mathbb{B}$	Το σύνολο που περιέχει τα στοιχεία του $\mathbb{A}$ που δεν υπάρχουν στο $\mathbb{B}$
$\mathbb{A} \subset \mathbb{B}$	Το σύνολο $\mathbb{A}$ αποτελεί υποσύνολο του $\mathbb{B}$
$\mathbb{C}$	Το σύνολο των μιγαδικών αριθμών
$j = \sqrt{-1}$	Η φανταστική μονάδα
$\overline{\exp(jz)} = \exp(-jz)$	Ο συζυγής μιγαδικός
$\vert\alpha\vert$	Απόλυτη τιμή ενός αριθμού $\alpha \in \mathbb{R}$(ή $\alpha \in \mathbb{C}$)
$\mathbb{R}^N$	Ο πραγματικός χώρος συντεταγμένων διάστασης $N \in \mathbb{N}$
$\mathbb{C}^N$	Ο μιγαδικός χώρος συντεταγμένων διάστασης $N \in \mathbb{N}$
α_i	Στοιχείο i ενός διανύσματος $\boldsymbol{\alpha}$, με την δεικτοδότηση να ξεκινά από το 1
$A_{i,j} = A(i, j)$	Στοιχείο i, j ενός πίνακα $\mathbf{A}$
$A_{i,j,k}$	Στοιχείο (i, j, k) ενός 3-D tensor $\mathbf{A}$

$\boldsymbol{\alpha}^T$	Το ανάστροφο διάνυσμα του διανύσματος $\boldsymbol{\alpha}$
$\mathbf{A}^T$	Ο ανάστροφος πίνακας του πίνακα $\mathbf{A}$
$\mathbf{A} \odot \mathbf{B}$	Γινόμενο στοιχείο προς στοιχείο(Hadamard) των $\mathbf{A}$ και $\mathbf{B}$
$\langle \boldsymbol{\alpha} \vert \mathbf{b} \rangle$	Εσωτερικό γινόμενο των διανυσμάτων $\boldsymbol{\alpha}$ και $\mathbf{b}$
$\frac{dy}{dx}$	Παράγωγος του y ως προς το x
$\frac{\partial y}{\partial x}$	Μερική παράγωγος του y ως προς το x
$\nabla_{\mathbf{x}} y$	Βαθμίδα του y ως προς το $\mathbf{x}$
$\int_{\mathbb{S}} f(\mathbf{x}) d\mathbf{x}$	Πεπερασμένο ολοκλήρωμα ως προς το $\mathbf{x}$ κατά το μήκος του συνόλου $\mathbb{S}$
$P(a)$	Μια πιθανοτική κατανομή κατά μήκος μιας διακριτής μεταβλητής
$p(a)$	Μια πιθανοτική κατανομή κατά μήκος μιας συνεχούς μεταβλητής ή μιας μεταβλητής της οποίας ο τύπος δεν έχει καθορισθεί
$a \sim P$	Τυχαία μεταβλητή η οποία έχει κατανομή P
$\mathbb{E}_{x \sim P}[f(x)]$ ή $\mathbb{E}f(x)$	Αναμενόμενη ή προσδοκώμενη τιμή της $f(x)$ ως προς την $P(x)$
$Var(f(x))$	Διασπορά της $f(x)$
$H(x)$	Εντροπία του Shannon μιας τυχαίας μεταβλητής x
$D_{KL}(P\|\|Q)$	Συνάρτηση απόκλισης Kullback-Leibler των P και Q
$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Γκαουσιανή κατανομή κατα μήκος του $\mathbf{x}$ με μέση τιμή $\boldsymbol{\mu}$ και συνδιασπορά $\boldsymbol{\Sigma}$
$f : \mathbb{A} \to \mathbb{B}$	Συνάρτηση f με πεδίο ορισμού $\mathbb{A}$ και εύρος τιμών $\mathbb{B}$
$f(\mathbf{x}; \boldsymbol{\theta})$	Μια συνάρτηση του $\mathbf{x}$ παραμετροποιημένη από το $\boldsymbol{\theta}$
$\log x$	Ο φυσικός ή νεπέριος λογάριθμος του x
$\log_2 x$	Ο λογάριθμος του x με βάση το 2
$\log_{10} x$	Ο λογάριθμος του x με βάση το 10
$\sigma(x)$	Η λογιστική σιγμοειδής συνάρτηση, $\sigma(x) = \frac{1}{1+\exp(-x)}$
$\text{sign}(x) = \begin{cases} -1, & \alpha\nu \quad x < 0, \\ 0, & \alpha\nu \quad x = 0, \\ 1, & \alpha\nu \quad x > 0. \end{cases}$	Η συνάρτηση προσήμου
$\lfloor x \rfloor = \max\{m \in \mathbb{Z}, x \in \mathbb{R} \mid m \leq x\}$	Η συνάρτηση δαπέδου του πραγματικού αριθμού x
$\|\|\mathbf{x}\|\|_2$	L^2 νόρμα του $\mathbf{x}$
x^+	Το θετικό μέρος του x, δηλ., $\max\{0, x\}$
p_{data}	Η κατανομή γέννησης δεδομένων

$\hat{p}_{data}$	Η εμπειρική κατανομή που καθορίζεται από το σύνολο εκπαίδευσης
$\mathbf{x}^{(i)}$	Η i-οστή είσοδος(ή δείγμα) από ένα σύνολο δεδομένων
$y^{(i)}$ ή $\mathbf{y}^{(i)}$	Ο i-οστός στόχος που σχετίζεται με το $\mathbf{x}^{(i)}$ για μάθηση με επιτήρηση

Περίληψη

Ο διαχωρισμός ηχητικών πηγών(Audio Source Separation) από μουσικές ηχογραφήσεις, αποτελεί σημαντική πρόκληση για το πεδίο της μουσικής εξόρυξης πληροφορίας(Music Information Retrieval). Στην παρούσα διπλωματική εργασία, εξετάζουμε μεθόδους για την εξαγωγή ηχητικών πηγών από ηχογραφήσεις, χρησιμοποιώντας συνελικτικά νευρωνικά δίκτυα(Convolutional Neural Networks ή CNNs). Στόχος μας είναι να βελτιώσουμε την απόδοση της ποιότητας του διαχωρισμού των φωνητικών(Vocals), του μπάσου(Bass) και της μουσικής συνοδείας(Accompaniment) από το συνολικό σήμα της μίξης(Mixture).

Προσεγγίζουμε το πρόβλημα με μια τεχνική που έχει φέρει εξαιρετικά αποτελέσματα στο πεδίο της όρασης υπολογιστών: Την ταξινόμηση εικονοστοιχείων(Pixel-Wise Classification). Χρησιμοποιούμε ως συνάρτηση κόστους την δυαδική διασταυρωμένη εντροπία(Binary Cross-Entropy), και προεκπαιδεύουμε το συνελικτικό νευρωνικό δίκτυο ως έναν autoencoder χρησιμοποιώντας φασματογραφήματα φωνητικών(ή μπάσου αντίστοιχα). Η τεχνική της ταξινόμησης εικονοστοιχείων, άμεσα εκτιμά την ταμπέλα της ηχητικής πηγής για κάθε χρονο-συχνοτικό(T-F) στοιχείο στην εικόνα του φασματογραφήματος και έτσι περιορίζονται συνήθεις εργασίες προεπεξεργασίας και μετεπεξεργασίας. Το προτεινόμενο δίκτυο εκπαιδεύεται με βάση την ιδανική δυαδική μάσκα(Ideal Binary Mask ή IBM) ως ταμπέλα στόχος εξόδου(Output Target Label). Η ιδανική δυαδική μάσκα ανιχνεύει την κυρίαρχη ηχητική πηγή κάθε χρονοσυχνοτικού στοιχείου στο φασματογράφημα(Spectrogram) του πλάτους ενός σήματος μίξης, θεωρώντας κάθε χρονο-συχνοτικό στοιχείο ως ένα εικονοστοιχείο με μια επισυναπτόμενη ταμπέλα(ανάλογα την ηχητική πηγή). Χρησιμοποιούμε την δυαδική διασταυρωμένη για να ελαχιστοποιήσουμε την μέση πιθανότητα σφάλματος μεταξύ του στόχου και της προβλεπόμενης ταμπέλας για κάθε εικονοστοιχείο. Προσεγγίζοντας το προβλήμα χρησιμοποιώντας την τεχνική της ταξινόμησης εικονοστοιχείων, εξαλείφουμε ένα από τα πιο κοινώς χρησιμοποιούμενα βήματα μετεπεξεργασίας: το φιλτράρισμα Wiener.

Παρουσιάζουμε μονοφωνικές και στερεοφωνικές υλοποιήσεις στα 22.05kHz. Επιπρόσθετα, παρουσιάζουμε και μια μονοφωνική υλοποίηση στα 44.1kHz. Απεικονίζουμε αναλυτικά, τις καμπύλες των μετρικών απόδοσης και απωλειών, για κάθε μία υλοποίηση.

Τέλος, παρουσιάζουμε τα συμπεράσματα από το πειραματικό μέρος, και προτείνουμε διάφορους τρόπους βελτίωσης.

Κεφάλαιο 1

Εισαγωγή

Τα τεχνητά νευρωνικά δίκτυα(Artificial Neural Networks) έχουν κερδίσει την προσοχή των επιστημόνων αρκετές φορές κατά το παρελθόν. Τα κύρια γεγονότα που συνέβαλαν σε αυτό είναι 1) η εφεύρεση του Perceptron Algorithm [2] το 1957, 2) η εφεύρεση του Backporpagation Algorithm [4] το 1986 και 3) η επιτυχία της βαθιάς μάθησης στην αναγνώριση φωνής [22] και στην κατηγοριοποίηση εικόνας [24] το 2012, οδηγώντας έτσι στην αναγέννηση της βαθιάς μάθησης(Deep Learning) περιλαμβάνοντας βαθιά πρόσο - τροφοδοτούμενα νευρωνικά δίκτυα(Feedforward Neural Networks), συνελικτικά νευρωνικά δίκτυα και LSTM(Long Short Term Memory). Όμως η βαθιά μάθηση, δεν οφείλεται μόνο στα προαναφερθέντα σημεία.

Μεταξύ του 1990 και 2010, η ταχύτητα των κεντρικών μονάδων επεξεργασίας(CPUs) των προσωπικών μας υπολογιστών, αυξήθηκε κατά περίπου με έναν παράγοντα του 5,000. Ως εκ τούτου, είναι πιθανόν σήμερα να τρέξουμε μικρά deep-learning μοντέλα στο προσωπικό μας laptop, πράγμα αδύνατο να επιτευχθεί 25 με 30 χρόνια πριν. Όμως, τυπικά deep-learning μοντέλα τα οποία χρησιμοποιούνται σε εφαρμογές διαχωρισμού ηχητικών πηγών, αναγνώρισης φωνής(Speech Recognition), όρασης υπολογιστών(Computer Vision) κ.λπ., απαιτούν πολύ μεγαλύτερη επεξεργαστική ισχύ από ότι ένα λάπτοπ, με βάση τα σημερινά δεδομένα, μπορεί να παράσχει. Κατά τη δεκαετία του 2000 και μετέπειτα, και λόγω της ολοένα και αυξανόμενης βιομηχανίας του gaming, εταιρίες όπως η NVIDIA και η AMD έχουν επενδύσει δισεκατομμύρια δολάρια, για την ανάπτυξη γρήγορων, μαζικών παράλληλων chips(Graphical Processing Units ή GPUs), οδηγώντας σε παιχνίδια με φωτορεαλιστικά γραφικά [34].

Η γενικότερη αυτή πρόοδος οφέλησε την επιστημονική κοινότητα όταν, το 2007, η NVIDIA ξεκίνησε την ανάπτυξη της `CUDA`[1] . Η CUDA είναι μια προγραμματιστική διεπαφή για τις GPUs της NVIDIA. Έτσι, ένας μικρός αριθμός από GPUs, ξεκίνησε να αντικαθιστά τα μεγάλα συμπλέγματα(Clusters) από CPUs, που χρησιμοποιούνταν σε πολλές εφαρμογές υψηλής παραλληλοποίησης. Τα βαθεία νευρωνικά δίκτυα επομένως, εποφελήθηκαν σημαντικά, αφού αποτελούν-

[1] **https://developer.nvidia.com/cuda-zone**.

ται κυρίως από μικρούς πολλαπλασιασμούς πινάκων και είναι επίσης υψηλά παραλληλοποιήσιμα.

Έτσι, λόγω των παραπάνω γεγονότων και του ολοένα και αυξανόμενου ενδιαφέροντος για την βαθιά μάθηση, έχει επιτραπεί η πρόοδος σε πρακτικές εφαρμογές σε πολλές περιοχές της ψηφιακής επεξεργασίας σήματος(Digital Signal Processing). Ενώ αρχικά δηλαδή η βαθιά μάθηση χρησιμοποιήθηκε για ψηφιακή επεξεργασία εικόνας (Image Processing) [24], μετέπειτα υιοθετήθηκε ευρέως στην επεξεργασία ομιλίας(Speech Processing), στην μουσική, στην επεξεργασία ήχου περιβάλλοντος(Enviromental Sound Processing), καθώς επίσης και σε έναν μεγάλο αριθμό πεδίων όπως την γονιδιωματική(Genomics), την κβαντική χημεία(Quantum Chemistry), την επεξεργασία φυσικής γλώσσας(Natural Language Processing) κ.α. Ως αποτέλεσμα, προηγούμενες μέθοδοι ψηφιακής επεξεργασίας του ήχου όπως τα Gaussian Mixture Models, Hidden Markov Models και μη-αρνητική παραγοντοποίηση πίνακα(Non-Negative Matrix Factorization), έχουν ξεπεραστεί σε απόδοση από τα μοντέλα βαθιάς μάθησης, σε εφαρμογές όπου είναι διαθέσιμα επαρκή δεδομένα.

Στην παρούσα διπλωματική εργασία, ασχολούμαστε με τον διαχωρισμό ηχητικών πηγών. Ενώ ο διαχωρισμός πηγών έχει εφαρμογή σε πολλά πεδία όπως είναι η ιατρική, η οικονομία, οι επικοινωνίες, η χημεία κ.α., αυτή η διπλωματική εργασία θα εστιάσει στον διαχωρισμό πηγών από ηχητικές ηχογραφήσεις. Ο διαχωρισμός ηχητικών πηγών αποδικνύεται χρήσιμος για πολλές εφαρμογές όπως είναι το beat tracking [26], η εκτίμηση της θεμελιώδους συχνότητας [21], το remixing [31] και το upmixing. Παρουσιάζεται στον αναγνώστη το πως επιτυγχάνεται ο διαχωρισμός των φωνητικών, του μπάσου και της μουσικής συνοδείας, από ηχητικές ηχογραφήσεις διαφόρων ειδών όπως rock, hip-hop, electro house κ.α. Χρησιμοποιούνται βαθεία νευρωνικά δίκτυα και συγκεκριμένα συνελικτικά νευρωνικά δίκτυα, επεκτείνοντας την δουλειά που έχει γίνει από άλλους ερευνητές στο πεδίο της επεξεργασίας εικόνας, όπου τέτοια δίκτυα, έχοντας διαθέσιμα επαρκή δεδομένα και κατάλληλη επεξεργαστική ισχύ, επιτυγχάνουν εξαιρετικά αποτελέσματα.

1.1 Οργάνωση του τόμου

Η διπλωματική εργασία είναι οργανωμένη σε έξι κεφάλαια.

Στο Κεφάλαιο 2 περιγράφεται η βασική θεωρία μηχανικής μάθησης, οι διάφορες μορφές αυτής, παραδοσιακά μοντέλα ταξινόμησης και βασικά στοιχεία βελτιστοποίησης και πιθανοτήτων.

Στο Κεφάλαιο 3 περιγράφεται η βαθιά μάθηση, τα συνελικτικά νευρωνικά δίκτυα, η συνάρτηση κόστους που χρησιμοποιείται από το μοντέλο μας, διάφοροι αλγόριθμοι βελτιστοποίησης και τεχνικές ομαλοποίησης.

Στο Κεφάλαιο 4 παρουσιάζονται βασικές μέθοδοι ψηφιακής επεξεργασίας σήματος του ήχου, όπως είναι ο υπολογισμός του βραχυχρόνιου μετασχηματισμού Fourier και η αναπαράσταση φασματογραφήματος.

Στο Κεφάλαιο 5 παρουσιάζεται η βασική θεωρία του διαχωρισμού ηχητικών πηγών, οι μετρικές αξιολόγησης και περιγράφονται περιληπτικά τεχνικές που

αποτελούν το state-of-the-art για την επίλυση του προβλήματος διαχωρισμού ηχητικών πηγών.

Στο Κεφάλαιο 6 γίνεται η παρουσίαση του πειραματικού μέρους. Αναφέρεται το λογισμικό του πειράσματος, το σύνολο δεδομένων(Dataset) που χρησιμοποιήσαμε [37] και έπειτα δίνεται η μεθοδολογία, τα προτεινόμενα μοντέλα και τέλος γίνεται η ανάλυση των αποτελεσμάτων.

Στο Κεφάλαιο 7 γίνεται συζήτηση και παρουσίαση των συμπερασμάτων καθώς και προτάσεις μας για μελλοντική έρευνα.

Κεφάλαιο 2

Βασική θεωρία μηχανικής μάθησης

2.1 Εισαγωγή στη μηχανική μάθηση

Η μηχανική μάθηση είναι η επιστήμη (και η τέχνη) του να προγραμματίζεις ηλεκτρονικούς υπολογιστές έτσι ώστε εκείνοι να μπορούν να *μαθαίνουν* από τα δεδομένα. Μία πιο γενική απόδοση του όρου είναι η εξής

> *Η μηχανική μάθηση είναι το πεδίο της έρευνας το οποίο δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να χρειάζεται να προγραμματίζονται ρητά*
>
> *Arthur Samuel, 1959*

Ένα σύστημα μηχανικής μάθησης, τροφοδοτείται με πολλά παραδείγματα σχετικά με μια διεργασία, και βρίσκει μια στατιστική συσχέτιση μεταξύ των παραδειγμάτων αυτών, η οποία τελικά επιτρέπει στο σύστημα να αναπτύξει κάποιους κανόνες για την αυτοματοποίηση της διεργασίας αυτής. Τα δεδομένα που τροφοδοτούν το μαθηματικό αυτό μοντέλο, αποτελούν τα δεδομένα εκπαίδευσης(Training Data).

Οι αλγόριθμοι μηχανικής μάθησης επομένως κάνουν προβλέψεις και παίρνουν αποφάσεις χρησιμοποιώντας τα δεδομένα εκπαίδευσης και ο κύριος στόχος είναι να μάθουν οι υπολογιστές να υλοποιούν αυτήν την διαδικασία αυτόματα, χωρίς την παρέμβαση του ανθρώπου. Χρησιμοποιούνται από ένα μεγάλο εύρος εφαρμογών όπως είναι η όραση υπολογιστών, τα chat bots, το φιλτράρισμα της ηλεκτρονικής αλληλογραφίας κ.α., εφαρμογές δηλαδή στις οποίες είναι ανέφικτο να χρησιμοποιηθούν συγκεκριμένες εντολές για την εκτέλεση της διεργασίας. Οι αλγόριθμοι μηχανικής μάθησης γενικά ανήκουν σε τέσσερις κατηγορίες, οι οποίες περιγράφονται στα επόμενα εδάφια [34].

2.1.1 Μάθηση με επιτήρηση(Supervised Learning)

Η μάθηση επιτυγχάνεται απεικονίζοντας τα δεδομένα εισόδου σε γνωστές ταμπέλες(Labels), δεδομένου ενός συνόλου με παραδείγματα. Ουσιαστικά πρόκειται για την παρατήρηση διαφόρων δεδομένων ενός τυχαίου διανύσματος $\mathbf{x}$ σε σχέση με μια τιμή ή διάνυσμα $\mathbf{y}$, και έπειτα την πρόβλεψη του $\mathbf{y}$ από το $\mathbf{x}$, συνήθως από την εκτίμηση της $p(\mathbf{y}|\mathbf{x})$. Ο όρος μάθηση με επιτήρηση προέρχεται από το γεγονός, ότι ο στόχος y, παρέχεται από κάποιον στο σύστημα μηχανικής μάθησης και του λέει τι πρέπει να κάνει. Γενικά, σχεδόν όλες οι εφαρμογές της βαθιάς μάθησης ανήκουν σε αυτήν την κατηγορία όπως λόγου χάριν η αναγνώριση ομιλίας και η ταξινόμηση εικόνας. Αν και κατα κύριο λόγο, η μάθηση με επιτήρηση είναι κυρίως ταξινόμηση(Classification) και παλινδρόμηση(Regression), αυτή, συναντάται και σε προβλήματα αναγνώρισης αντικειμένου(Object Detection), κατάτμησης εικόνας(Image Segmentation), κ.α.

2.1.2 Μάθηση χωρίς επιτήρηση(Unsupervised Learning)

Αυτό το είδος μάθησης, αποτελείται από την εύρεση ενδιαφέροντων μετασχηματισμών των δεδομένων εισόδου, χωρίς την βοήθεια ταμπελών, για την απεικόνιση δεδομένων(Data Visualization), την συμπίεση δεδομένων(Data Compression), την εξάλειψη του θορύβου στα δεδομένα(Data Denoising) ή και για να καταλάβουμε καλύτερα τους συσχετισμούς που παρουσιάζονται στα διαθέσιμα δεδομένα.

2.1.3 Ιδιο-επιτηρούμενη μάθηση(Self-supervised Learning)

Αποτελεί ουσιαστικά υποκατηγορία της μάθησης με επιτήρηση, αλλά είναι αρκετά διαφορετική και επομένως αξίζει να αναφερθεί ξεχωριστά. Η ιδιο-επιτηρούμενη μάθηση, είναι μάθηση με επιτήρηση αλλά χωρίς να έχουν ορισθεί οι ταμπέλες ρητά από τον προγραμματιστή. Οι ταμπέλες συνεχίζουν να εμπλέκονται στο πρόβλημα όπως και στη μάθηση με επιτήρηση, όμως εξάγονται από τα δεδομένα εισόδου, συνήθως χρησιμοποιώντας έναν heuristic αλγόριθμο. Λόγου χάριν, οι autoencoders, αποτελούν στοιχείο της ιδιο-επιτηρούμενης μάθησης.

2.1.4 Ενισχυτική μάθηση(Reinforcement Learning)

Επι του παρόντος, η ενισχυτική μάθηση είναι μια ερευνητική περιοχή η οποία δεν έχει επιτύχει σημαντικά αποτελέσματα, πέρα από τα παίγνια. Αναμένεται όμως στο μέλλον να καλύψει ένα μεγάλο εύρος εφαρμογών όπως αυτοκίνητα χωρίς οδηγό, ρομπότς κ.α.

2.2 Βελτιστοποίηση με χρήση βαθμίδας

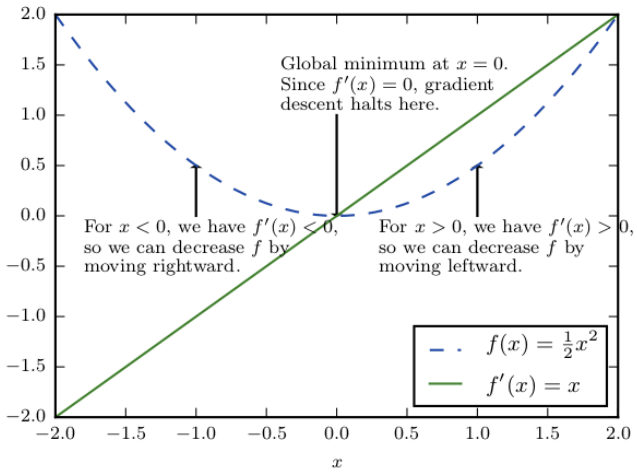
Οι περισσότεροι αλγόριθμοι μηχανικής και βαθιάς μάθησης, εμπεριέχουν την έννοια της βελτιστοποίησης [35]. Η βελτιστοποίηση αναφέρεται στην διαδικασία, είτε της

ελαχιστοποίησης, είτε της μεγιστοποίησης κάποιας συνάρτησης $f(\mathbf{x})$ προσδιορίζοντας το $\mathbf{x}$. Συνήθως επιδιώκουμε την ελαχιστοποίηση της $f(\mathbf{x})$. Η συνάρτηση που θέλουμε να ελαχιστοποιήσουμε ή μεγιστοποιήσουμε καλείται αντικειμενική συνάρτηση(Objective Function) ή κριτήριο. Όταν μιλάμε για ελαχιστοποίηση, τότε αυτή καλείται και συνάρτηση κόστους, συνάρτηση απωλειών ή συνάρτηση σφάλματος. Στα πλαίσια της διπλωματικής αυτής εργασίας, θα αναφερόμαστε συνήθως με τον όρο συνάρτηση κόστους.

Έστω ότι έχουμε μια συνάρτηση $y = f(x)$, όπου x και y είναι πραγματικοί αριθμοί. Η παράγωγος της συνάρτησης $f'(x)$ ή $\frac{dy}{dx}$ δίνει η κλίση της $f(x)$ στο σημείο x. Με άλλα λόγια, καθορίζει πως μια μικρή αλλαγή στην είσοδο επηρεάζει την έξοδο

$$f(x + \epsilon) \approx f(x) + \epsilon f'(x)$$

Η παράγωγος είναι επομένως χρήσιμη για την ελαχιστοποίηση μιας συνάρτησης επειδή μας λέει πως να αλλάξουμε το x, ώστε να επιτύχουμε μια μικρή βελτίωση στο y. Παραδείγματος χάριν, γνωρίζουμε ότι το $f(x - \epsilon\,\mathrm{sign}(f'(x)))$ είναι μικρότερο από το $f(x)$ για αρκετά μικρό ϵ. Επομένως μπορούμε να μειώσουμε το $f(x)$, κάνοντας μικρά βήματα στο x, με πρόσημο αντίθετο της παραγώγου. Αυτή είναι η πολύ γνωστή τεχνική Gradient Descent.



Σχήμα 2.1: Απεικόνιση του πως ο αλγόριθμος Gradient Descent, χρησιμοποιεί της παραγώγους μιας συνάρτησης για την εύρεση ελαχίστου. Πηγή: [35]

Συνήθως ελαχιστοποιούμε συναρτήσεις οι οποίες έχουν πολλαπλές εισόδους $f : \mathbb{R}^n \to \mathbb{R}$. Για να έχει νόημα η έννοια της ελαχιστοποίησης, η έξοδος πρέπει να είναι μοναδική και βαθμωτό μέγεθος.

Εδώ χρησιμοποιείται η έννοια των μερικών παραγώγων. Η μερική παράγωγος $\frac{\partial}{\partial x_i} f(\mathbf{x})$ μετράει πόσο η f αλλάζει, μόνο όσον αφορά την αύξηση της μεταβλητής x_i στο σημείο $\mathbf{x}$. Η βαθμίδα γενικοποιεί την έννοια της παραγώγου στην περίπτωση όπου η παράγωγος έχει ληφθεί ως προς το διάνυσμα. Η παράγωγος δηλαδή της

f, είναι το διάνυσμα που περιέχει όλες τις μερικές παραγώγους $\nabla_{\mathbf{x}} f(\mathbf{x})$.

Η κατευθυνόμενη παράγωγος στην κατεύθυνση $\mathbf{x}$(το μοναδιαίο διάνυσμα), είναι η καμπύλη της συνάρτησης f στην κατεύθυνση u. Με άλλα λόγια, η κατευθυνόμενη βαθμίδα είναι η παράγωγος της συνάρτησης $f(\mathbf{x} + \alpha\mathbf{u})$ ως προς το α, στο $\alpha = 0$. Χρησιμοποιώντας τον κανόνα της αλυσίδας, μπορούμε να δούμε ότι το $\frac{\partial}{\partial\alpha} f(\mathbf{x} + \alpha\mathbf{u})$, αντιστοιχεί στο $\mathbf{u}^T \nabla_x f(\mathbf{x})$, όταν $\alpha = 0$.

Για να ελαχιστοποιήσουμε την f, θα θέλαμε να βρούμε την κατεύθυνση για την οποία η f ελαχιστοποιείται με τον γρηγορότερο ρυθμό. Αυτό μπορεί να επιτευχθεί με την χρήση της κατευθυνόμενης παραγώγου

$$\min_{\mathbf{u}, \mathbf{u}^T\mathbf{u}=1} \mathbf{u}^T \nabla_{\mathbf{x}} f(\mathbf{x}) \tag{2.1}$$

$$= \min_{\mathbf{u}, \mathbf{u}^T\mathbf{u}=1} ||\mathbf{u}||_2 ||\nabla_{\mathbf{x}} f(\mathbf{x})||_2 \cos\theta \tag{2.2}$$

όπου θ είναι η γωνία μεταξύ του $\mathbf{u}$ και της βαθμίδας. Αντικαθιστώντας με $||\mathbf{u}||_2 = 1$ και αγνοώντας παράγοντες που δεν εξαρτώνται από το $\mathbf{u}$, αυτό απλοποιείται σε $\min_{\mathbf{u}} \cos\theta$. Αυτό ελαχιστοποιείται όταν το $\mathbf{u}$, δείχνει προς την αντίθετη κατεύθυνση από την βαθμίδα. Δηλαδή, η βαθμίδα δείχνει "προς την κορυφή του λόφου", και η αρνητική βαθμίδα "προς το χαμηλότερο σημείο του λόφου". Άρα μπορούμε να μειώσουμε την f, πηγαίνοντας προς την κατεύθυνση της αρνητικής βαθμίδας. Αυτό είναι γνωστό και ως η μέθοδος Steepest Descent. Η μέθοδος Steepest Descent προτείνει ένα νέο σημείο

$$\mathbf{x}' = \mathbf{x} - \epsilon \nabla_{\mathbf{x}} f(\mathbf{x}) \tag{2.3}$$

όπου ϵ είναι ο ρυθμός μάθησης(Learning Rate), ένα θετικό βαθμωτό μέγεθος που καθορίζει το μέγεθος του βήματος.

Ο Steepest Descent συγκλίνει όταν κάθε στοιχείο της βαθμίδας είναι μηδέν ή πολύ κοντά στο μηδέν.

2.3 Παραδοσιακά μοντέλα ταξινόμησης

2.3.1 Γραμμική παλινδρόμηση

Ένας από τους πιο απλούς αλγορίθμους μηχανικής μάθησης, είναι η γραμμική παλινδρόμηση(Linear Regression). Σκοπός είναι η δημιουργία ενός συστήματος, το οποίο μπορεί να λάβει ένα διάνυσμα $\mathbf{x} \in \mathbb{R}^n$ ως είσοδο, και να προβλέψει μια βαθμωτή τιμή $y \in \mathbb{R}$ ως έξοδο. Η έξοδος της γραμμικής παλινδρόμησης είναι γραμμική συνάρτηση της εισόδου. Έστω $\hat{y}$ η τιμή που προβλέπει το μοντέλο. Η έξοδος είναι

$$\hat{y} = \mathbf{w}^T \mathbf{x} \tag{2.4}$$

όπου $\mathbf{w} \in \mathbb{R}^n$ είναι ένα διάνυσμα από παραμέτρους. Μπορούμε να σκεφτούμε το $\mathbf{w}$ ως ένα σύνολο από βάρη, που καθορίζουν πως κάθε χαρακτηριστικό επηρεάζει την πρόβλεψη. Δηλαδή εάν ένα χαρακτηριστικό x_i λαμβάνει ένα θετικό βάρος w_i, τότε αυξάνοντας την τιμή του χαρακτηριστικού αυτού, αυξάνεται και η τιμή της πρόβλεψης $\hat{y}$. Αντίθετα, αν το χαρακτηριστικό λαμβάνει ένα αρνητικό βάρος, τότε

η αύξηση της τιμής του χαρακτηριστικού, μειώνει την τιμή της πρόβλεψης $\hat{y}$. Αν το βάρος ενός χαρακτηριστικού είναι μηδέν, τότε δεν έχει επίδραση στην πρόβλεψη.

Επόμενο βήμα, είναι ο καθορισμός μιας μετρικής απόδοσης P. Υποθέτουμε ότι έχουμε ένας πίνακα από m εισόδους δειγμάτων τα οποία δεν θα χρησιμοποιήσουμε για εκπαίδευση, αλλά για αξιολόγηση της απόδοσης του μοντέλου. Έχουμε επίσης ένα διάνυσμα με τις επιθυμητές τιμές της παλινδρόμησης y, για κάθε ένα από αυτά τα δείγματα. Επειδή το σύνολο δεδομένων, χρησιμοποιείται μόνο για αξιολόγηση, το ονομάζουμε σύνολο δοκιμής(Test Set). Αναφερόμαστε στον πίνακα εισόδου ως $\mathbf{X}^{(test)}$ και στο διάνυσμα της παλινδρόμησης ως $\mathbf{y}^{(test)}$.

Ένας τρόπος υπολογισμού της απόδοσης του μοντέλου, είναι ο υπολογισμός του μέσου τετραγωνικού σφάλματος(Mean Squared Error) του μοντέλου στο σύνολο δοκιμής. Εάν $\hat{\mathbf{y}}^{(test)}$ δίνει τις προβλέψεις του μοντέλου στο σύνολο δοκιμής, τότε το μέσο τετραγωνικό σφάλμα, δίνεται από

$$MSE_{test} = \frac{1}{m}\sum_{i}(\hat{\mathbf{y}}^{(test)} - \mathbf{y}^{(test)})_i^2 \tag{2.5}$$

Διαισθητικά, μπορούμε να δούμε ότι το σφάλμα γίνεται 0 όταν $\hat{\mathbf{y}}^{(test)} = \mathbf{y}^{(test)}$. Επιπρόσθετα γίνεται εύκολα αντιληπτό ότι το σφάλμα αυξάνεται όταν η δεύτερη νόρμα ή Ευκλίδεια απόσταση μεταξύ των προβλέψεων και των επιθυμητών εξόδων αυξάνεται

$$MSE_{test} = \frac{1}{m}||\hat{\mathbf{y}}^{(test)} - \mathbf{y}^{(test)}||_2^2 \tag{2.6}$$

Για την δημιουργία ενός αλγορίθμου μηχανικής μάθησης, χρειάζεται να σχεδιάσουμε έναν αλγόριθμο, ο οποίος θα βελτιώνει τα βάρη $\mathbf{w}$, με τρόπο που θα μειώνει το MSE_{test}, όταν ο αλγόριθμος αρχίζει να μαθαίνει από τις παρατηρήσεις στο σύνολο εκπαίδευσης $(\mathbf{X}^{(train)}, \mathbf{y}^{(train)})$. Μπορούμε απλώς να ελαχιστοποιήσουμε το MSE_{test}, λύνοντας για την βαθμίδα να ισούται με 0

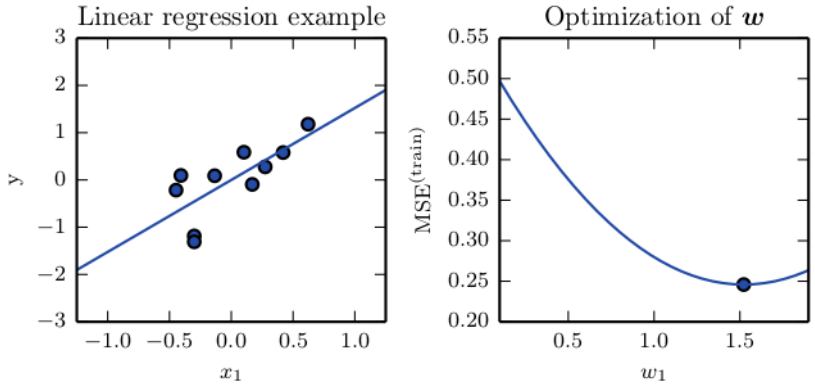
$$\nabla_w MSE_{train} = 0 \tag{2.7}$$

και συνεπώς, βρίσκουμε τα βέλτιστα βάρη $\mathbf{w}$.

Αξίζει να σημειωθεί ότι ο όρος γραμμική παλινδρόμηση, χρησιμοποιείται συχνά για ένα λίγο πιο διαφορετικό μοντέλο, το οποίο εισάγει τον όρο b

$$\hat{y} = \mathbf{w}^T\mathbf{x} + b \tag{2.8}$$

Αυτό ο όρος b, καλείται μεροληπτική παράμετρος του γραμμικού μετασχηματισμού, το οποίο απλώς σημαίνει ότι η ευθεία γραμμή που χωρίζει τα δεδομένα, δεν χρειάζεται να περνά από την αρχή των αξόνων.



Σχήμα 2.2: Ένα πρόβλημα της γραμμικής παλινδρόμησης, με ένα σύνολο εκπαίδευσης που αποτελείται από δέκα σημεία, καθένα από τα οποία αντιπροσωπεύει ένα χαρακτηριστικό. Για τον λόγο αυτό, το διάνυσμα βαρών $\mathbf{w}$ περιέχει μια μόνο παράμετρο για μάθηση, την w_1. (Αριστερά)Η γραμμική παλινδρόμηση μαθαίνει να ορίζει το w_1, έτσι ώστε η γραμμή $y = w_1 x$ να περνά όσο πιο κοντά γίνεται στα δεδομένα εκπαίδευσης. (Δεξιά)Η τιμή του w_1, για την οποία το μέσο τετραγωνικό σφάλμα ελαχιστοποιείται στο σύνολο εκπαίδευσης. Πηγή: [35]

2.3.2 Λογιστική παλινδρόμηση

Γενικά, η γραμμική παλινδρόμηση αντιστοιχεί στην παραμετρική οικογένεια κατανομών

$$p(y|\mathbf{x};\boldsymbol{\theta}) = \mathcal{N}(y;\boldsymbol{\theta}^T\mathbf{x},\mathbf{I}) \tag{2.9}$$

Μπορούμε να γενικοποιήσουμε την γραμμική παλινδρόμηση στο σενάριο ταξινόμησης, καθορίζοντας μια διαφορετική οικογένεια πιθανοτικών κατανομών. Αν έχουμε δύο κλάσεις, την κλάση 0 και την κλάση 1, τότε χρειάζεται μόνο να καθορίσουμε την πιθανότητα, σε μια από τις δύο κλασεις. Η πιθανότητα της κλάσης 1 καθορίζει την πιθανότητα της κλάσης 0, επειδή αυτές οι δύο τιμές πρέπει να αθροίζουν στο 1.

Η κανονική κατανομή στο εύρος των πραγματικών αριθμών, με την οποία εκφράσαμε την γραμμική παλινδρόμηση, είναι παραμετροποιημένη σε όρους μέσης τιμής. Οποιαδήποτε τιμή για αυτήν τη μέση τιμή είναι έγκυρη. Όταν όμως χρησιμοποιείται μια δυαδική μεταβλητή, η κατανομή που την εκφράζει είναι πιο περίπλοκη (βλ. εδάφιο 3.2.1), διότι η μέση τιμή πρέπει πάντα να είναι εντός του διαστήματος $(0, 1)$. Ένας τρόπος να επιλυθεί αυτό το πρόβλημα είναι να χρησιμοποιηθεί η λογιστική σιγμοειδής συνάρτηση(βλ. σχέση 3.17), ώστε να επιβληθεί στην έξοδο, να είναι εντός του εύρους $(0, 1)$, και ερμηνεύουμε αυτήν την τιμή ως την πιθανότητα

$$p(y = 1|\mathbf{x};\boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}^T\mathbf{x}) \tag{2.10}$$

Η προσέγγιση αυτή είναι γνωστή ως λογιστική παλινδρόμηση(Logistic Regression).

Σε αντίθεση με την γραμμική παλινδρόμηση, για τη λογιστική παλινδρόμηση, δεν υπάρχει κλειστή μορφή εξίσωσης για την εύρεση των βέλτιστων βαρών. Αντιθέτως πρέπει να ψάξουμε για τα βάρη αυτά, μεγιστοποιώντας την λογαριθμική πιθανοφάνεια, δηλαδή ελαχιστοποιώντας την αρνητική λογαριθμική πιθανοφάνεια χρησιμοποιώντας Gradient Descent(βλ. εδάφιο 2.7).

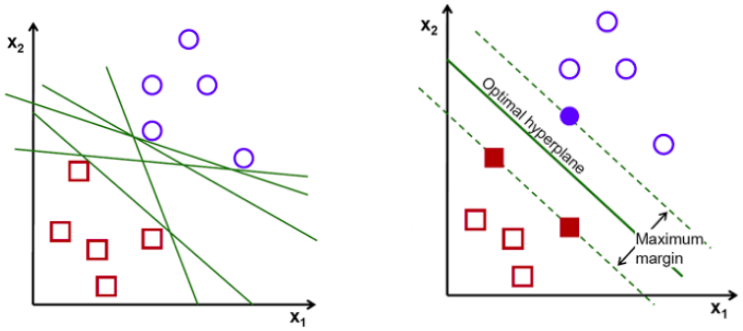
Γενικά, αυτή η στρατηγική μπορεί να εφαρμοσθεί σε οποιονδήποτε αλγόριθμο μηχανικής μάθησης με επιτήρηση, δηλαδή θεωρώντας μια παραμετρική οικογένεια δεσμευμένων πιθανοτικών κατανομών, για τα αντίστοιχα είδη των μεταβλητών εισόδου και εξόδου.

2.3.3 Support Vector Machines(SVM)

Τα Support Vector Machines [7], είναι ένα από τα διασημότερα μοντέλα για επίλυση προβλημάτων ταξινόμησης και παλινδρόμησης. Έστω το πρόβλημα της δυαδικής ταξινόμησης όπου χρησιμοποιούμε γραμμικά μοντέλα της μορφής

$$f(x) = \mathbf{w}^T \phi(\mathbf{x}) + b \tag{2.11}$$

όπου $\phi(\mathbf{x})$ δηλώνει έναν σταθερό μετασχηματισμό χαρακτηριστικών-χώρου, b η μεροληψία(Bias) και $\mathbf{w}$ το διάνυσμα των βαρών(Weights). Τα δεδομένα εκπαίδευσης αποτελούνται από N διανύσματα εισόδου $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$, με τις αντίστοιχες επιθυμητές τιμές εξόδου $\{y^{(1)}, \ldots, y^{(N)}\}$ όπου $y^{(i)} \in \{-1, 1\}$, και τα νέα σημεία $\mathbf{x}$ ταξινομούνται με βάση το πρόσημο του $f(\mathbf{x})$. Υποθέτουμε ότι τα δεδομένα μάθησης είναι γραμμικώς χωριζόμενα στο χώρο των χαρακτηριστικών, και έτσι, εκ της διευθέτησης του προβλήματος, υπάρχει τουλάχιστον μία επιλογή παραμέτρων $\mathbf{w}$ και b, ώστε η εξίσωση 2.11 να ικανοποιεί την συνθήκη $f(\mathbf{x}^{(i)}) > 0$ για σημεία με $y^{(i)} = +1$ και $f(\mathbf{x}^{(i)}) < 0$ για σημεία με $y^{(i)} = -1$, και έτσι να ισχύει για όλα τα σημεία μάθησης $y^{(i)} f(\mathbf{x}^{(i)}) > 0$. Φυσικά μπορεί να υπάρχουν πολλές λύσεις που να διαχωρίζουν ακριβώς τις κλάσεις. Η λύση που δίνουν τα Support Vector Machines, προσεγγίζουν το πρόβλημα από την πλευρά του περιθωρίου(Margin), το οποίο ορίζεται ως η μικρότερη απόσταση μεταξύ του συνόρου απόφασης και οποιουδήποτε από τα δεδομένα.



(a) Πιθανά υπερεπίπεδα διαχωρισμού. (b) Γραμμικός ταξινομητής με μέγιστο περιθώριο.

Σχήμα 2.3: Παράδειγμα δυαδικής ταξινόμησης δύο γραμμικώς χωριζόμενων κλάσεων. Πηγή: [53]

Γενικότερα, η κάθετη απόσταση ενός σημείου $\mathbf{x}$, από ένα υπερεπίπεδο $f(\mathbf{x}) = 0$, όπου το υπερεπίπεδο είναι στη μορφή (2.11), δίνεται από $|f(\mathbf{x})|/||\mathbf{w}||$. Επιπρόσθετα, μας ενδιαφέρουν λύσεις που ταξινομούν όλα τα σημεία σωστά, έτσι ώστε $y^{(i)}f(\mathbf{x}^{(i)}) > 0$ για κάθε i. Επομένως η απόσταση του σημείου $\mathbf{x}^{(i)}$ από την επιφάνεια απόφασης δίνεται από

$$\frac{y^{(i)}f(\mathbf{x}^{(i)})}{||\mathbf{w}||} = \frac{y^{(i)}(\mathbf{w}^T\phi(\mathbf{x}^{(i)}) + b)}{||\mathbf{w}||} \tag{2.12}$$

Επιθυμούμε να βελτιστοποιήσουμε τις παραμέτρους $\mathbf{w}$ και b, ώστε να μεγιστοποιήσουμε την απόσταση του περιθωρίου. Αυτό επιτυγχάνεται λύνοντας

$$\underset{\mathbf{w},b}{argmax}\left\{\frac{1}{||\mathbf{w}||}\min_i[y^{(i)}(\mathbf{w}^T\phi(\mathbf{x}^{(i)}) + b)]\right\} \tag{2.13}$$

2.4 Χωρητικότητα, υπερπροσαρμογή και υποπροσαρμογή

Είναι πολύ σημαντικό για τη μηχανική μάθηση, ο αλγόριθμός μας να μπορεί να δουλεύει καλά και σε νέες εισόδους που δεν έχει ξαναδεί. Η ικανότητα αυτή καλείται γενίκευση(Generalization). Συνήθως, όταν εκπαιδεύουμε ένα μοντέλο μηχανικής μάθησης, έχουμε πρόσβαση σε ένα σύνολο εκπαίδευσης. Μπορούμε να υπολογίσουμε κάποιο είδος σφάλματος στο σύνολο εκπαίδευσης, το οποίο καλείται σφάλμα εκπαίδευσης(Training Error), και μπορούμε να το μειώσουμε. Αυτό που διαχωρίζει τη μηχανική μάθηση από την βελτιστοποίηση(Optimization), είναι ότι θέλουμε και το σφάλμα γενίκευσης(Generalization Error) ή αλλιώς σφάλμα στο σύνολο δοκιμής(Test Error), να είναι επίσης χαμηλό.

Συνήθως υπολογίζουμε το σφάλμα γενίκευσης ενός μοντέλου μηχανικής μάθησης, μετρώντας την απόδοσή του σε ένα σύνολο δοκιμής από δείγματα που συλέχθηκαν ξεχωριστά από αυτά του συνόλου εκπαίδευσης.

Στο παράδειγμα της γραμμικής παλινδρόμησης(βλ. εδάφιο 2.3.1), η εκπαίδευση του μοντέλου έγινε ελαχιστοποιώντας το σφάλμα εκπαίδευσης

$$\frac{1}{m^{(train)}}||\mathbf{X}^{(train)}\mathbf{w} - \mathbf{y}^{(train)}||_2^2 \tag{2.14}$$

αλλά εμάς μας ενδιαφέρει και το σφάλμα στο σύνολο δοκιμής $\frac{1}{m^{(test)}}||\mathbf{X}^{(test)}\mathbf{w} - \mathbf{y}^{(test)}||_2^2$

Επομένως οι παράγοντες που καθορίζουν το πόσο καλά αποδίδει ένας αλγόριθμος μηχανικής μάθησης είναι η ικανότητά του να

1) Κάνει το σφάλμα εκπαίδευσης όσο το δυνατόν μικρότερο.

2) Κάνει τη διαφορά, μεταξύ του σφάλματος εκπαίδευσης και σφάλματος στο σύνολο δοκιμής, όσο το δυνατόν ελάχιστη.

Αυτοί οι δύο παράγοντες αντιστοιχούν σε δύο κεντρικές προκλήσεις για τη μηχανική μάθηση: Την υποπροσαρμογή(Underfitting) και την υπέρπροσαρμογή(Overfitting). Η υποπροσαρμογή συμβαίνει όταν το μοντέλο δεν είναι ικανό να επιτύχει μια ικανοποιητικά μικρή τιμή σφάλματος στο σύνολο εκπαιδεύσης. Όσον αφορά την υπέρπροσαρμογή, αυτή συμβαίνει όταν η διαφορά μεταξύ του σφάλματος εκπαίδευσης και σφάλματος στο σύνολο δοκιμής είναι αρκετά μεγάλη.

Μπορούμε να ελέγξουμε αν ένα μοντέλο έχει την τάση να κάνει υπερπροσαρμογή ή υποπροσαρμογή, αλλάζοντας την χωρητικότητά(Capacity) του. Ένας απλοϊκός ορισμός για την χωρητικότητα του μοντέλου, είναι η ικανότητά του να προσαρμόζεται σε ένα μεγάλο σύνολο από συναρτήσεις. Μοντέλα με χαμηλή χωρητικότητα, πιθανόν να δυσκολευθούν να προσαρμοστούν στα δεδομένα. Μοντέλα με υψηλή χωρητικότητα αντίθετα, πιθανόν να υπερπροσαρμόζονται, μαθαίνοντας πολύ καλά τις ιδιότητες του συνόλου εκπαίδευσης, το οποίο αποτελεί μειονέκτημα για την απόδοση στο σύνολο δοκιμής.

Ένας τρόπος να ελέγξουμε την χωρητικότητα του αλγορίθμου μάθησης, είναι η σωστή επιλογή του χώρου υποθέσεων(Hypothesis Space), δηλαδή ένα σύνολο συναρτήσεων που μπορούν να επιλεγούν ως λύση από τον αλγόριθμο. Παραδείγματος χάριν, ο αλγόριθμος γραμμικής παλινδρόμησης, έχει το σύνολο όλων των γραμμικών συναρτήσεων ως είσοδο και ως τον χώρο υποθέσεων του. Μπορούν επιπρόσθετα να συμπεριληφθούν πολυώνυμα, αντί μόνο για γραμμικές συναρτήσεις ως χώρος υποθέσεων.

Ένα πολυώνυμο πρώτου βαθμού δίνει το μοντέλο γραμμικής παλινδρόμησης, με την εκτίμηση

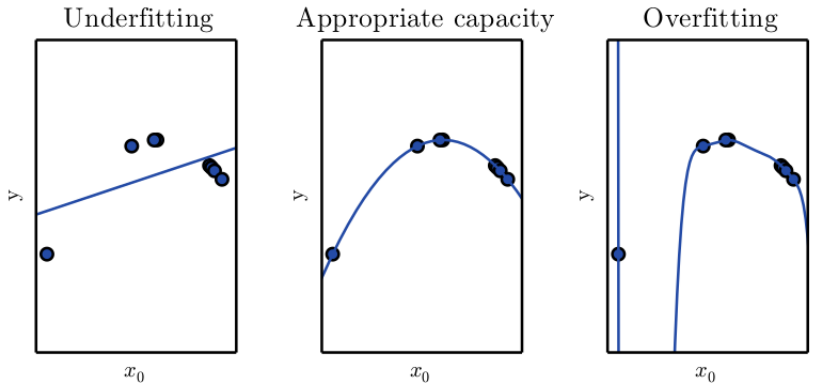
$$\hat{y} = b + wx \tag{2.15}$$

Μπορούμε να χρησιμοποιήσουμε επίσης ένα πολυώνυμο δευτέρου βαθμού

$$\hat{y} = b + w_1 x + w_2 x^2 \tag{2.16}$$

Επιπρόσθετα μπορούν να προστεθούν και άλλες δυνάμεις του x, παραδείγματος χάριν το παρακάτω πολυώνυμο ενάτου βαθμού

$$\hat{y} = b + \sum_{i=1}^{9} w_i x^i \tag{2.17}$$



Σχήμα 2.4: Παράδειγμα προσαρμογής τριών μοντέλων. Αριστερά, απεικονίζεται μια γραμμική συνάρτηση να προσαρμόζεται στα δεδομένα με σημαντικό υποπροσαρμογή. Φαίνεται δηλαδή ότι δεν μπορεί να εναρμονιστεί με την φυσική καμπύλωση που διέπει τα δεδομένα. Στο κέντρο, απεικονίζεται μια πολυωνυμική συνάρτηση δευτέρου βαθμού, η οποία προσαρμόζεται σωστά στα δεδομένα και γενικοποιεί σωστά σε σημεία που δεν έχει ξαναδεί. Δεξιά, απεικονίζεται ένα πολυώνυμο ενάτου βαθμού που υπερπροσαρμόζεται στα δεδομένα. Πηγή: [35]

Οι αλγόριθμοι μηχανικής μάθησης γενικά θα αποδώσουν καλύτερα όταν η χωρητικότητα είναι η κατάλληλη, για την πραγματική πολυπλοκότητα του προβλήματος που έχουν να επιλύσουν και για τα δεδομένα εκπαίδευσης που παρέχονται. Μοντέλα με μη επαρκή χωρητικότητα, δεν είναι ικανά να επιλύσουν πολύπλοκα προβλήματα. Από την άλλη πλευρά, μοντέλα με μεγάλη χωρητικότητα μπορούν να επιλύουν πιο πολύπλοκα προβλήματα, αλλά όταν η χωρητικότητα είναι μεγαλύτερη από την απαιτούμενη για την επίλυση του αντίστοιχου προβλήματος, τότε μπορεί να υπερπροσαρμόζονται.

2.5 Εκτιμητές, μεροληψία και διασπορά

Το πεδίο της στατιστικής, μας δίνει τη δυνατότητα, να επιλύσουμε προβλήματα μηχανικής μάθησης, όχι μόνο για το σύνολο εκπαίδευσης, αλλά και στο σύνολο δοκιμής, δηλαδή μας βοηθά στη γενίκευση του προβλήματος. Η γενίκευση, η υποπροσαρμογή και η υπερπροσαρμογή, μπορούν να περιγραφούν από τις έννοιες της εκτίμησης παραμέτρων, της μεροληψίας και της διασποράς.

2.5.1 Εκτίμηση σημείου

Η εκτίμηση σημείου είναι μια προσπάθεια για την "καλύτερη" πρόβλεψη ενός σημείου μιας ποσότητας που μας ενδιαφέρει, παραδείγματος χάριν, μιας απλής παραμέτρου, ενός διανύσματος παραμέτρων ή και μιας ολόκληρης συνάρτησης. Έστω $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ ένα σύνολο m ανεξάρτητων και όμοια κατανεμημένων

σημείων δεδομένων. Ο εκτιμητής σημείου είναι οποιαδήποτε συνάρτηση των δεδομένων

$$\hat{\boldsymbol{\theta}}_m = g(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}) \tag{2.18}$$

όπου $\boldsymbol{\theta}$ η παράμετρος και $\hat{\boldsymbol{\theta}}$ ο εκτιμητής αυτής. Ένας καλός εκτιμητής επομένως είναι μια συνάρτηση της οποίας η έξοδος, είναι κοντά στην πραγματική τιμή του $\boldsymbol{\theta}$.

2.5.2 Μεροληψία

Η μεροληψία ενός εκτιμητή ορίζεται ως

$$bias(\hat{\boldsymbol{\theta}}_m) = \mathbb{E}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\theta} \tag{2.19}$$

όπου $\mathbb{E}(\hat{\boldsymbol{\theta}}_m)$ η αναμενόμενη τιμή των δεδομένων και $\boldsymbol{\theta}$ η πραγματική τιμή του $\boldsymbol{\theta}$, που χρησιμοποιείται για να καθορίσει την κατανομή των δεδομένων. Ένας εκτιμητής $\hat{\boldsymbol{\theta}}_m$ είναι αμερόληπτος(Unbiased) αν $bias(\hat{\boldsymbol{\theta}}_m) = 0$, το οποίο υποδεικνύει ότι $\mathbb{E}(\hat{\boldsymbol{\theta}}_m) = \theta$. Ο εκτιμητής $\hat{\boldsymbol{\theta}}_m$ είναι ασυμπτωτικά αμερόληπτος άν $\lim_{m\to\infty} bias(\hat{\boldsymbol{\theta}}_m) = 0$, το οποίο υποδεικνύει ότι $\lim_{m\to\infty} \mathbb{E}(\hat{\boldsymbol{\theta}}_m) = \theta$

2.5.3 Διασπορά και ντε φάκτο σφάλμα

Μια άλλη ιδιότητα ενός εκτιμητή είναι το πόσο αναμένουμε να μεταβάλλεται ως συνάρτηση των δεδομένων. Η διασπορά ενός εκτιμητή είναι απλώς η διασπορά

$$Var(\hat{\theta}) \tag{2.20}$$

όπου η τυχαία μεταβλητή είναι το σύνολο εκπαίδευσης. Επιπλέον, η τετραγωνική ρίζα της διασποράς καλείται το ντε φάκτο σφάλμα(Standard Error) $SE(\hat{\theta})$.

Η διασπορά, ή το ντε φάκτο σφάλμα ενός εκτιμητή, παρέχει ένα μέτρο του πόσο αναμένουμε να μεταβάλλεται η εκτίμηση που κάναμε στα δεδομένα, καθώς επαναδειγματοληπτούμε το σύνολο από την υποκείμενη διαδικασία παραγωγής των δεδομένων. Θέλουμε δηλαδή αμερόληπτους εκτιμητές, όπως επίσης και σχετικά μικρή διασπορά.

Το ντε φάκτο σφάλμα της μέσης τιμής δίνεται από

$$SE(\hat{\mu}_m) = \sqrt{Var\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right]} = \frac{\sigma}{\sqrt{m}} \tag{2.21}$$

όπου σ^2 είναι η πραγματική διασπορά των δειγμάτων x^i. Το ντε φάκτο σφάλμα της μέσης τιμής, είναι πολύ σημαντικό για προβλήματα μηχανικής μάθησης. Συχνά, εκτιμούμε το σφάλμα γενίκευσης, απλώς υπολογίζοντας τον δειγματικό μέσο του σφάλματος στο σύνολο δοκιμής. Ο αριθμος των δεδομένων στο σύνολο δοκιμής καθορίζει την ακρίβεια της εκτίμησης. Από το κεντρικό οριακό θεώρημα, το οποίο μας λέει ότι η μέση τιμή θα κατανεμηθεί προσεγγιστικά με βάση την κανονική κατανομή, μπορούμε να χρησιμοποιήσουμε το ντε φάκτο σφάλμα για τον υπολογισμό της πιθανότητας ότι η η πραγματική αναμενόμενη τιμή πέφτει σε οποιοδήποτε επιλεγμένο διάστημα.

2.6 Εκτίμηση μέγιστης πιθανοφάνειας

Αντί να προσπαθούμε να μαντέψουμε μια συνάρτηση αν θα είναι καλός εκτιμητής ή όχι, πρώτα επιλέγοντάς την και έπειτα αναλύοντάς την με βάση τη μεροληψία και τη διασπορά, θα θέλαμε να έχουμε κάποια βασική αρχή με την οποία να μπορούμε να παράγουμε συναρτήσεις που να είναι και καλοί εκτιμητές.

Η πιο γνωστή αρχή είναι αυτή της μέγιστης πιθανοφάνειας(Maximum Likelihood Principle).

Έστω ένα σύνολο από m δείγματα $\mathbb{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$, τα οποία έχουν συλλεχθεί ανεξάρτητα από τη πραγματική, αλλά άγνωστη κατανομή δεδομένων $p_{data}(\mathbf{x})$.

Έστω $p_{model}(\mathbf{x}; \boldsymbol{\theta})$ μια παραμετρική οικογένεια από πιθανοτικές κατανομές σε κάποιον χώρο που ορίζεται από το $\boldsymbol{\theta}$. Δηλαδή η $p_{model}(\mathbf{x}; \boldsymbol{\theta})$ αντιστοιχεί κάθε $\mathbf{x}$ σε έναν πραγματικό αριθμό εκτιμώντας την πραγματική πιθανότητα $p_{data}(\mathbf{x})$.

Ο εκτιμητής μέγιστης πιθανοφάνειας για το $\boldsymbol{\theta}$ ορίζεται ως

$$\boldsymbol{\theta}_{ML} = \underset{\boldsymbol{\theta}}{argmax} p_{model}(\mathbb{X}; \boldsymbol{\theta}) \tag{2.22}$$

$$= \underset{\boldsymbol{\theta}}{argmax} \prod_{i=1}^{m} p_{model}(\mathbf{x}^{(i)}; \boldsymbol{\theta}) \tag{2.23}$$

Αυτό το γινόμενο πολλών πιθανοτήτων μπορεί να προκαλέσει διάφορα προβλήματα, παραδείγματος χάριν, αριθμητικό underflow. Επομένως, είναι βολικότερο να πάρουμε τον λογάριθμο της πιθανοφάνειας

$$\boldsymbol{\theta}_{ML} = \underset{\boldsymbol{\theta}}{argmax} \sum_{i=1}^{m} \log p_{model}(\mathbf{x}^{(i)}; \boldsymbol{\theta}) \tag{2.24}$$

Επειδή το $argmax$ δεν αλλάζει αν κάνουμε κάποιου είδους κανονικοποίηση στη συνάρτηση κόστους, μπορούμε να διαιρέσουμε με m ώστε να λάβουμε μια έκδοση της (2.24) η οποία να είναι εκφρασμένη ως η αναμενόμενη τιμή σε σχέση με την εμπειρική κατανομή $\hat{p}_{data}$ που καθορίζεται από το σύνολο εκπαίδευσης

$$\boldsymbol{\theta}_{ML} = \underset{\boldsymbol{\theta}}{argmax} \mathbb{E}_{\mathbf{x} \sim \hat{p}_{data}} \log p_{model}(\mathbf{x}; \boldsymbol{\theta}) \tag{2.25}$$

Ένας τρόπος να περιγράψουμε την εκτίμηση της μέγιστης πιθανοφάνειας, είναι να την δούμε ως μια αρχή που ελαχιστοποιεί την ανομοιότητα ανάμεσα στην εμπειρική κατανομή $\hat{p}_{data}$, που καθορίζεται από το σύνολο εκπαίδευσης και την κατανομή του μοντέλου. Ο βαθμός αυτής της ανομοιότητας μετράται από την απόκλιση Kullback-Leibler οποία δίνεται από

$$D_{KL}(\hat{p}_{data} || p_{model}) = \mathbb{E}_{\mathbf{x} \sim \hat{p}_{data}} [\log \hat{p}_{data}(\mathbf{x}) - \log p_{model}(\mathbf{x})] \tag{2.26}$$

Ο όρος στα αριστερά είναι μια συνάρτηση της διαδικασίας παραγωγής των δεδομένων και όχι του μοντέλου. Αυτό σημαίνει, ότι όταν εκπαιδεύουμε το μοντέλο να ελαχιστοποιεί την απόκλιση Kullback-Leibler, χρειάζεται μόνο να ελαχιστοποιήσουμε την ποσότητα

$$- \mathbb{E}_{\mathbf{x} \sim \hat{p}_{data}} [\log p_{model}(\mathbf{x})] \tag{2.27}$$

το οποίο προφανώς είναι το ίδιο με την μεγιστοποίηση της σχέσης (2.25).

Αυτή η ελαχιστοποίηση, είναι η γνωστή έννοια της ελαχιστοποίησης της διασταυρωμένης εντροπίας, μεταξύ των κατανομών που προαναφέρθηκαν. Γενικά, η απώλεια(ή κόστος) που αποτελείται από την αρνητική λογαριθμική πιθανοφάνεια, είναι η διασταυρωμένη εντροπία μεταξύ της εμπειρικής κατανομής που καθορίζεται από το σύνολο εκπαίδευσης και της πιθανοτικής κατανομής που καθορίζεται από το μοντέλο μας. Η διασταυρωμένη εντροπία είναι μια από τις πιο γνωστές συναρτήσεις κόστους, που χρησιμοποιούνται κατά κόρον για την εκπαίδευση νευρωνικών δικτύων και θα αναλυθεί περαιτέρω στο επόμενο κεφάλαιο.

2.7 Stochastic Gradient Descent

Ένας από τους πιο σημαντικούς αλγορίθμους για την βαθιά μάθηση είναι ο Stochastic Gradient Descent(SGD), ο οποίος αποτελεί μια επέκταση του αλγορίθμου Gradient Descent που αναλύθηκε στο εδάφιο (2.2).

Ένα πολύ συχνό πρόβλημα για την μηχανική μάθηση, είναι το ότι μεγάλα σύνολα εκπαίδευσης είναι απαραίτητα για σωστή γενίκευση, όμως τα μεγάλα σύνολα εκπαίδευσης, καταναλώνουν και πολλούς υπολογιστικούς πόρους.

Η συνάρτηση κόστους που χρησιμοποιείται από έναν αλγόριθμο μηχανικής μάθησης, συχνά αποσυντίθεται ως ένα άθροισμα στα δεδομένα εκπαίδευσης με κάποια συνάρτηση κόστους ανά δείγμα. Παραδείγματος χάριν, η αρνητική δεσμευμένη λογαριθμική πιθανοφάνεια ή αλλιώς διασταυρωμένη εντροπία, για το σύνολο εκπαίδευσης, μπορεί να γραφεί ως

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},y\sim\hat{p}_{data}} L(\mathbf{x}, y, \boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} L(\mathbf{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}) \tag{2.28}$$

όπου L είναι η απώλεια ανά-δείγμα $L(\mathbf{x}, y, \boldsymbol{\theta}) = -\log p(y|\mathbf{x};\boldsymbol{\theta})$.

Για αυτές τις συναρτήσεις κόστους, ο Gradient Descent απαιτεί τον υπολογισμό

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} L(\mathbf{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}) \tag{2.29}$$

Το υπολογιστικό κόστος αυτής της πράξης είναι $O(m)$. Όσο το σύνολο εκπαίδευσης μεγαλώνει με δισεκατομμύρια από δείγματα, ο χρόνος για ένα απλό βήμα της βαθμίδας γίνεται αρκετά μεγάλος.

Η μεγάλη καινοτομία του Stochastic Gradient Descent, είναι η θεώρηση της βαθμίδας ως μια αναμενόμενη ή προσδοκώμενη τιμή. Η αναμενόμενη αυτή τιμή μπορεί προσεγγιστικά να υπολογισθεί, χρησιμοποιώντας ένα μικρό σύνολο δειγμάτων. Συγκεκριμένα, σε κάθε βήμα του αλγορίθμου, μπορούμε να δειγματοληπτήσουμε ένα minibatch από δείγματα $\mathbb{B} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m')}\}$, τα οποία έχουν ληφθεί ομοιόμορφα από το σύνολο εκπαίδευσης. Το μέγεθος του minibatch m', τυπικά επιλέγεται, να είναι ένας σχετικά μικρός αριθμός δειγμάτων, συνήθως από ένα μέχρι μερικές εκατοντάδες. Το μέγεθος m' συνήθως διατηρείται σταθερό, όσο μεγαλώνει το μέγεθος m του συνόλου εκπαίδευσης. Μπορεί να προσαρμόσουμε

ένα σύνολο εκπαίδευσης με εκατομμύρια δείγματα, χρησιμοποιώντας σε κάθε βήμα παραμέτρους που έχουν υπολογιστεί μόνο σε εκατοντάδες δείγματα.

Η εκτίμηση της βαθμίδας διαμορφώνεται ως εξής

$$\mathbf{g} = \frac{1}{m'} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{m'} L(\mathbf{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}) \tag{2.30}$$

χρησιμοποιώντας δείγματα από το minibatch $\mathbb{B}$. Ο Stochastic Gradient Descent τότε, ακολουθεί την εκτίμηση της βαθμίδας καθοδικά

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \mathbf{g} \tag{2.31}$$

όπου ϵ είναι ο ρυθμός μάθησης.

Κεφάλαιο 3

Βαθιά μάθηση

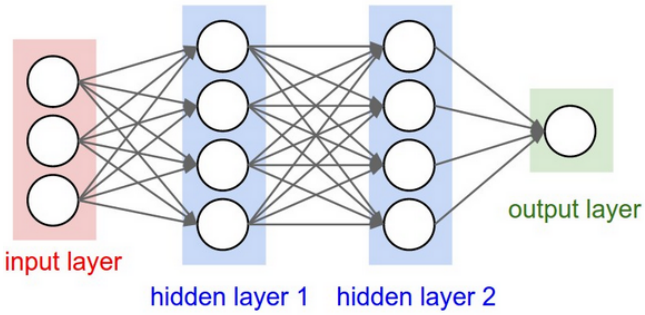
3.1 Βαθεία προσο-τροφοδοτούμενα δίκτυα

Τα βαθεία προσο-τροφοδοτούμενα δίκτυα ή προσο-τροφοδοτούμενα νευρωνικά δίκτυα ή αλλιώς multilayer perceptrons(MLPs), αποτελούν εξαιρετικής σημασίας μοντέλα βαθιάς μάθησης. Στόχος του προσο-τροφοδοτούμενου δικτύου είναι η προσέγγιση κάποιας συνάρτησης f^*. Παραδείγματος χάριν, ένας ταξινομητής $y = f^*(\mathbf{x})$, αντιστοιχεί το διάνυσμα εισόδου $\mathbf{x}$ σε μια κατηγορία y. Ένα προσο-τροφοδοτούμενο δίκτυο ορίζει την αντιστοίχιση $\mathbf{y} = f(\mathbf{x}; \theta)$, και μαθαίνει την τιμή των παραμέτρων θ, που οδηγούν στην καλύτερη προσεγγιστική συνάρτηση.

Αυτά τα μοντέλα ονομάζονται προσο-τροφοδοτούμενα επειδή η ροή της πληροφορίας ρέει μέσω της συνάρτησης με είσοδο $\mathbf{x}$, μέσω ενδιάμεσων υπολογισμών για τον ορισμό της f και τελικά καταλήγει στην έξοδο $\mathbf{y}$. Δεν υπάρχουν συνδέσεις ανατροφοδότησης στο δίκτυο. Όταν προσο-τροφοδοτούμενα δίκτυα επεκτείνονται ώστε να εμπεριέχουν και συνδέσεις ανατροφοδότησης, τότε αυτά καλούνται επανατροφοδοτούμενα νευρωνικά δίκτυα(Recurrent Neural Networks ή RNN).

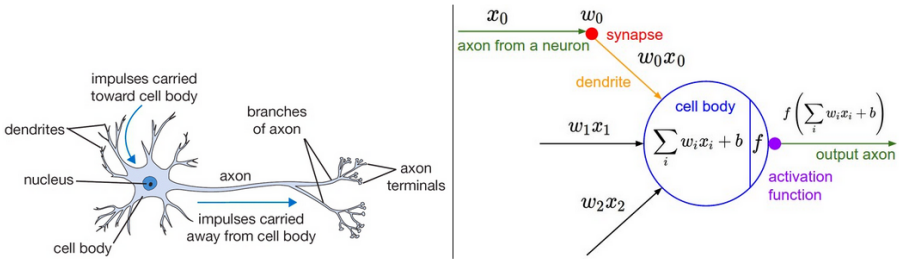
Η αναπαράσταση των προσο-τροφοδοτούμενων δικτύων, τυπικά γίνεται συνθέτοντας μαζί πολλές διαφορετικές συναρτήσεις. Το μοντέλο συσχετίζεται με έναν κατευθυνόμενο άκυκλο γράφο που περιγράφει πως αυτές οι συναρτήσεις συνθέτονται μαζί. Έστω, ότι έχουμε τρεις συναρτήσεις $f^{(1)}$, $f^{(2)}$, και $f^{(3)}$ συνδεδεμένες σε αλυσίδα, έτσι ώστε $f(\mathbf{x}) = f^{(3)}(f^{(2)}(f^{(1)}(\mathbf{x})))$. Στην περίπτωση αυτή, $f^{(1)}$ ονομάζεται το πρώτο επίπεδο(Layer) του δικτύου, $f^{(2)}$ ονομάζεται το δεύτερο επίπεδο, κ.ο.κ. Το συνολικό μήκος της αλυσίδας δίνει το βάθος του μοντέλου. Από εδώ προέρχεται ο όρος βαθιά μάθηση. Το τελικό επίπεδο είναι το επίπεδο εξόδου. Κατά την εκπαίδευση του δικτύου, προσπαθούμε η $f(\mathbf{x})$ να προσεγγίσει όσο το δυνατόν καλύτερα την $f^*(\mathbf{x})$. Τα δεδομένα εκπαίδευσης παρέχουν προσεγγιστικά παραδείγματα της $f^*(\mathbf{x})$ για διάφορα σημεία εκπαίδευσης. Κάθε παράδειγμα $\mathbf{x}$, συνοδεύεται από μια ταμπέλα $y \approx f^*(\mathbf{x})$. Τα παραδείγματα εκπαίδευσης καθορίζουν επακριβώς τι θα πρέπει το επίπεδο εξόδου να κάνει για κάθε σημείο $\mathbf{x}$. Πρέπει να παράξει μια τιμή η οποία είναι κοντά στο y. Η συμπεριφορά των άλλων επιπέδων, δεν καθορίζεται επακριβώς από τα δεδομένα εκπαίδευσης. Ο

αλγόριθμος εκμάθησης(Learning Algorithm) που χρησιμοποιείται πρέπει να αποφασίσει πως να χρησιμοποιήσει αυτά τα επίπεδα ώστε να παράξει το επιθυμητό αποτέλεσμα, δηλαδή μια προσέγγιση της f^*. Επειδή τα δεδομένα εκπαίδευσης δεν δείχνουν τις επιθυμητές εξόδους για κάθε ένα από αυτά τα επίπεδα, αυτά καλούνται κρυφά(Hidden Layers).



Σχήμα 3.1: Νευρωνικό δίκτυο αποτελούμενο από 3 επίπεδα, 2 κρυφά από 4 νευρώνες έκαστο και ένα επίπεδο εξόδου. Πηγή: [50]

Τέλος, τα δίκτυα αυτά καλούνται νευρωνικά, διότι η βασική ιδέα προέχρχεται από την νευροεπιστήμη(Neuroscience). Κάθε νευρώνας δέχεται σήματα εισόδου από τους δενδρίτες και παράγει σήματα εξόδου δια μέσου του μονού άξονα του. Ο άξονας σταδιακά, εμφανίζει διακλαδώσεις οι οποίες συνδέονται μέσω των συνάψεων με δενδρίτες άλλων νευρώνων. Το μαθηματικό μοντέλο που αντιστοιχεί στον βιολογικό νευρώνα, περιγράφεται ως εξής. Τα σήματα ταξιδεύουν δια μέσου των αξόνων (x_0), επιδρούν πολλαπλασιαστικά(w_0x_0) με τους δενδρίτες των άλλων νευρώνων, βασισμένα στη συναπτική δύναμη της αντίστοιχης σύναψης (w_0). Η βασική ιδέα, είναι ότι οι συναπτικές δυνάμεις(τα βάρη $\mathbf{w}$), έχουν την ικανότητα της μάθησης και ελέγχουν την ποσότητα και την κατεύθυνση της επιρροής τους(θετικά και αρνητικά βάρη) από τον έναν νευρώνα στον άλλον. Στο βασικό μοντέλο, οι δενδρίτες μεταφέρουν το σήμα στο κυτταρικό σώμα όπου όλα μαζί αθροίζονται. Αν το τελικό άθροισμα είναι πάνω από ένα συγκεκριμένο κατώφλι, τότε ο νευρώνας πυροδοτείται στέλνοντας έναν σπινθήρα δια μέσου του άξονά του. Στο μαθηματικό μοντέλο, θεωρούμε ότι δεν μας απασχολεί ο ακριβής χρόνος πυροδότησης του νευρώνα, παρά μόνο η συχνότητα της πυροδότησης. Μοντελοποιούμε την συχνότητα της πυροδότησης του νευρώνα με μια συνάρτηση ενεργοποίησης f [50].



Σχήμα 3.2: Απεικόνιση ενός βιολογικού νευρώνα(αριστερά) και του μαθηματικού του μοντέλου(δεξιά). Πηγή: [50]

3.2 Συνάρτηση κόστους

Μια σημαντική παράμετρος για τον σχεδιασμό ενός βαθέος νευρωνικού δικτύου είναι η επιλογή της συνάρτησης κόστους. Οι συναρτήσεις κόστους για τα βαθεία νευρωνικά δίκτυα είναι λίγο πολύ οι ίδιες όπως και για άλλα παραμετρικά μοντέλα, π.χ. τα γραμμικά.

Μία από τις σημαντικότερες έννοιες για την μηχανική μάθηση και αναγνώριση προτύπων(Pattern Recognition), είναι η έννοια της εντροπίας. Προερχόμενη από την θεωρία πληροφοριών(Information Theory), αποδικνύεται πολύ χρήσιμη ως συνάρτηση κόστους όταν χρησιμοποιούμε νευρωνικά δίκτυα σε συνδυασμό με τον Back-Propagation αλγόριθμο(βλ. εδάφιο 3.4) για την βελτιστοποίηση των βαρών. Θεωρούμε μια διακριτή μεταβλητή x και ελέγχουμε πόση πληροφορία λαμβάνεται κατά την παρατήρηση μια συγκεκριμένης τιμής της μεταβλητής αυτής. Η ποσότητα της πληροφορίας μπορεί να θεωρηθεί ως "ο βαθμός της εκπλήξεως" μαθαίνοντας την τιμή του x. Αν μαθαίναμε ότι ένα αρκετά απίθανο γεγονός συνέβη, τότε θα είχαμε λάβει αρκετά περισσότερη πληροφορία από ότι αν μαθαίναμε ότι το γεγονός αυτό ήταν αρκετά πιθανό και αν γνωρίζαμε ότι το γεγονός ήταν βέβαιο ότι θα συνέβαινε, τότε δεν θα είχαμε λάβει καθόλου πληροφορία. Το μέτρο της πληροφορίας εξαρτάται επομένως από την πιθανοτική κατανομή $p(x)$, και ψάχνουμε μια ποσότητα $h(x)$, η οποία είναι μονότονη συνάρτηση της πιθανότητας $p(x)$ και εκφράζει την πληροφορία που περιέχεται. Η μορφή της $h(\cdot)$ μπορεί να βρεθεί θεωρώντας ότι τα έχουμε δύο γεγονότα x και y ασυσχέτιστα μεταξύ τους. Τότε η πληροφορία που λαμβάνουμε από την παρατήρηση τους, θα είναι το άθροισμα των πληροφοριών τους ξεχωριστά, έτσι ώστε $h(x, y) = h(x) + h(y)$. Δύο ασυσχέτιστα γεγονότα θα είναι στατιστικά ανεξάρτητα και επομένως $p(x, y) = p(x)p(y)$. Προκύπτει ότι

$$h(x) = -\log_2 p(x) \tag{3.1}$$

όπου το αρνητικό πρόσημο δηλώνει ότι η πληροφορία θα είναι θετική ή μηδέν. Αν θεωρηθεί τώρα ότι ένας αποστολέας θέλει να μεταδώσει την τιμή μίας τυχαίας μεταβλητής σε έναν δέκτη, τότε η μέση ποσότητα πληροφορίας μετάδοσης δίνεται

από

$$H(x) = -\sum_{x} p(x) \log_2 p(x) \tag{3.2}$$

Αυτή η σημαντική ποσότητα, καλείται εντροπία της τυχαίας μεταβλητής x.

Στις περισσότερες περιπτώσεις το παραμετρικό μοντέλο καθορίζει μια κατανομή $p(\mathbf{y}|\mathbf{x};\theta)$ και απλά χρησιμοποιούμε την αρχή της μέγιστης πιθανοφάνειας. Τα περισσότερα μοντέρνα νευρωνικά δίκτυα, εκπαιδεύονται με βάση την μέγιστη πιθανοφάνεια. Αυτό σημαίνει ότι η συνάρτηση κόστους είναι απλώς η αρνητική λογαριθμική πιθανοφάνεια(Negative Log-Likelihood) ή αλλιώς η διασταυρωμένη εντροπία ανάμεσα στα δεδομένα εκπαίδευσης και στην κατανομή του μοντέλου. Αυτή η συνάρτηση κόστους δίνεται από

$$J(\theta) = -\mathbb{E}_{x,y\sim\hat{p}_{data}} \log p_{model}(\mathbf{y}|\mathbf{x}) \tag{3.3}$$

Η ακριβής μορφή της συνάρτησης κόστους μπορεί να διαφέρει από μοντέλο σε μοντέλο, αφού εξαρτάται από την ακριβή μορφή του $\log p_{model}$. Το δικό μας πρόβλημα, αφορά την περίπτωση της δυαδικής ταξινόμησης(αν υπάρχουν φωνητικά(μπάσο) ή όχι, βλ. εδάφιο 6.2), επομένως στο επόμενο εδάφιο θα αναλύσουμε την μορφή της συνάρτησης κόστους για την κατανομή Bernoulli. Σημειώνεται ότι για προβλήματα ταξινόμησης πολλών κλάσεων χρησιμοποιείται η κατανομή Multinoulli με Softmax μονάδες στην έξοδο του δικτύου.

3.2.1 Η κατανομή Bernoulli

Η κατανομή Bernoulli είναι μια κατανομή μιας δυαδικής τυχαίας μεταβλητής. Ελέγχεται από μια παράμετρο $\phi \in [0,1]$, η οποία δίνει την πιθανότητα στην τυχαία μεταβλητή να ισούται με το 1. Έχει τις εξής ιδιότητες

$$P(x=1) = \phi \tag{3.4}$$

$$p(x=0) = 1-\phi \tag{3.5}$$

$$P(x=x) = \phi^x(1-\phi)^{1-x} \tag{3.6}$$

$$\mathbb{E}_x[x] = \phi \tag{3.7}$$

$$Var_x(x) = \phi(1-\phi) \tag{3.8}$$

3.2.2 Σιγμοειδής μονάδες για Bernoulli κατανομές στην έξοδο του δικτύου

Η επιλογή της συνάρτησης κόστους, είναι στενά συνδεδεμένη με την επιλογή την μονάδας εξόδου στο δίκτυο. Τις περισσότερες φορές χρησιμοποιείται η διασταυρωμένη εντροπία μεταξύ της κατανομής των δεδομένων και της κατανομής του μοντέλου. Επομένως, η μορφή της μονάδας εξόδου, καθορίζει τη μορφή της συνάρτησης διασταυρωμένης εντροπίας.

Οποιοδήποτε είδος μονάδας νευρωνικού δικτύου που χρησιμοποιείται ως έξοδός του, μπορεί επίσης να χρησιμοποιηθεί και ως κρυφή μονάδα. Εδώ θα αναλυθεί η έξοδος, αλλά η βασική αρχή είναι η ίδια και για το εσωτερικό του δικτύου. Θα

θεωρήσουμε ότι το προσο-τροφοδοτούμενο δίκτυο, παρέχει ένα σετ από κρυφά χαρακτηριστικά που καθορίζονται από $\mathbf{h} = f(\mathbf{x}; \theta)$. Ο ρόλος του επιπέδου εξόδου του δικτύου, είναι να παρέχει κάποιον επιπλέον μετασχηματισμό στα χαρακτηριστικά ώστε να ολοκληρωθεί η διεργασία από το δίκτυο.

Πολλές διεργασίες, απαιτούν λοιπόν την πρόβλεψη της τιμής μιας δυαδικής μεταβλητής y. Προβλήματα ταξινόμησης με δύο κλάσεις ανήκουν σε αυτήν την κατηγορία.

Η προσέγγιση με την αρχή της μέγιστης πιθανοφάνειας, είναι να ορίσουμε μια κατανομή Bernoulli στο y, εξαρτώμενη από το $\mathbf{x}$.

Η κατανομή Bernoulli καθορίζεται από έναν αριθμό. Το νευρωνικό δίκτυο χρειάζεται να προβλέψει μόνο το $P(y = 1|\mathbf{x})$. Για να είναι ο αριθμός αυτός έγκυρος, πρέπει να βρίσκεται εντός του διαστήματος $[0, 1]$.

Μια σιγμοειδής(Sigmoid) μονάδα εξόδου ορίζεται ως

$$\hat{y} = \sigma(\mathbf{w}^T \mathbf{h} + b) \tag{3.9}$$

όπου σ είναι η λογιστική σιγμοειδής(Logistic Sigmoid) συνάρτηση

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \tag{3.10}$$

Η σιγμοειδής μονάδα εξόδου, μπορεί να θεωρηθεί ότι αποτελείται από δύο συνιστώσες. Αρχικά χρησιμοποιεί ένα γραμμικό επίπεδο για να υπολογίσει το $z = \mathbf{w}^T \mathbf{h} + b$. Εν συνεχεία, χρησιμοποιεί την σιγμοειδή συνάρτηση ενεργοποίησης(βλ. σχέση 3.17) για να μετατρέψει το z σε πιθανότητα.

Έστω τώρα μια μη κανονικοποιημένη πιθανοτική κατανομή $\widetilde{P}(y)$, της οποίας το άθροισμα των πιθανοτήτων δεν δίνει μονάδα. Διαιρώντας με κατάλληλη σταθερά μπορούμε να λάβουμε μια έγκυρη πιθανοτική κατανομή. Από την υπόθεση ότι οι μη κανονικοποιημένες λογαριθμικές πιθανότητες είναι γραμμικές στο y και στο z, μπορούμε να χρησιμοποιήσουμε την εκθετική συνάρτηση για να πάρουμε τις μη κανονικοποιημένες πιθανότητες. Έπειτα, κανονικοποιώντας προκύπτει μια κατανομή Bernoulli που ελέγχεται από ένας σιγμοειδή μετασχηματισμό του z

$$\log \widetilde{P}(y) = yz \tag{3.11}$$

$$\widetilde{P}(y) = \exp(yz) \tag{3.12}$$

$$P(y) = \frac{\exp(yz)}{\sum_{y'=0}^{1} \exp(y'z)} \tag{3.13}$$

$$P(y) = \sigma((2y - 1)z) \tag{3.14}$$

και επομένως η δυαδική διασταυρωμένη εντροπία ως συνάρτηση κόστους θα είναι

$$J(\theta) = -\log P(y|\mathbf{x}) \tag{3.15}$$

$$= -\log \sigma((2y - 1)z) \tag{3.16}$$

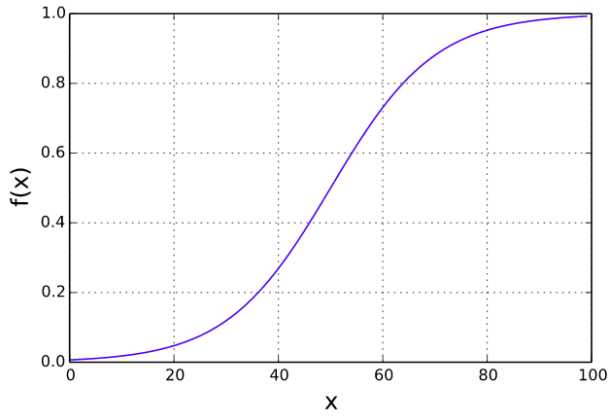
3.3 Συνάρτηση ενεργοποίησης

Οι συναρτήσεις ενεργοποίησης(Activation Functions), ουσιαστικά αποφασίζουν πότε ένας νευρώνας πρέπει να ενεργοποιηθεί και πότε όχι, μετασχηματίζοντας την έξοδό του, αναλόγως με τις ιδιότητες που διέπουν την εκάστοτε συνάρτηση ενεργοποίησης. Ένα νευρωνικό δίκτυο χρειάζεται να περιλαμβάνει μη γραμμικές τέτοιες συναρτήσεις, ώστε να δώσει ακριβή αποτελέσματα, εκτελώντας πολύπλοκες διεργασίες. Παρακάτω περιγράφονται οι πιο κοινές συναρτήσεις ενεργοποίησης:

- **Λογιστική σιγμοειδής(Logistic Sigmoid)**

  Η σιγμοειδής μη γραμμικότητα έχει την μαθηματική μορφή

$$f(x) = \sigma(x) = \frac{1}{1 + \exp(-x)} \tag{3.17}$$



Σχήμα 3.3: Η λογιστική σιγμοειδής συνάρτηση. Πηγή: [49]

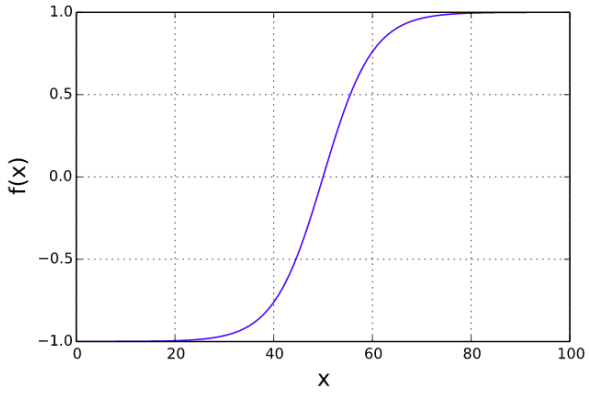
Όπως είδαμε και στο προηγούμενο εδάφιο κατά την ανάλυση της δυαδικής διασταυρωμένης εντροπίας, ως συνάρτησης κόστους, οι σιγμοειδής, ως μονάδες εξόδου, χρησιμοποιούνται για την πρόβλεψη μια δυαδικής μεταβλητής να ισούται με 1. Οι σιγμοειδής μονάδες φθάνουν σε κορεσμό στο μεγαλύτερο μέρος του πεδίου ορισμού τους. Όταν το x είναι πολύ θετικό, η σιγμοειδής μονάδα τείνει σε μια μεγάλη τιμή, ενώ όταν το x είναι πολύ αρνητικό, τείνει σε μια μικρή τιμή και είναι εξαρτάται σε πολύ μεγάλο βαθμό από την είσοδό της όταν το x είναι πολύ κοντά στο 0. Αυτός ο κορεσμός των σιγμοειδών μονάδων σε όλο το πεδίο ορισμού τους, μπορεί να κάνει την βελτιστοποίηση με χρήση βαθμίδας πολύ δύσκολη. Γι' αυτόν το λόγο, οι σιγμοειδής συναρτήσεις χρησιμοποιούνται σπάνια στις κρυφές μονάδες των προσο-τροφοδοτούμενων δικτύων. Χρησιμοποιούνται κυρίως ως μονάδες εξόδου, όπου η βελτιστοποίηση με χρήση βαθμίδας, μαζί με την κατάλληλη

συνάρτηση κόστους(δυαδική διασταυρωμένη εντροπία), αναιρούν αυτόν τον κορεσμό της λογιστικής σιγμοειδούς συνάρτησης στο επίπεδο εξόδου.

- **Υπερβολική εφαπτομένη(Hyperbolic Tangent)**

  Η συνάρτηση υπερβολικής εφαπτομένης, προσαρμόζει την είσοδο που δέχεται, στο σύνολο τιμών $[-1, 1]$. Σημαντικό πλεονέκτημα είναι ότι οι τιμές της υπερβολικής εφαπτομένης είναι κεντραρισμένες στο μηδέν, πράγμα το οποίο βοηθά στη διάδοση. Η μορφή της είναι

$$f(x) = \tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \tag{3.18}$$

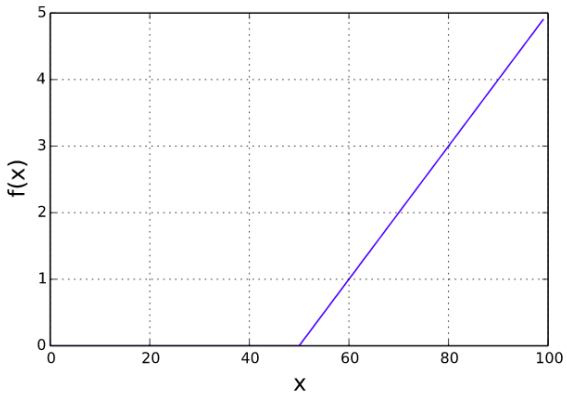


Σχήμα 3.4: Η συνάρτηση υπερβολικής εφαπτομένης. Πηγή: [49]

Αποτελεί και αυτή η συνάρτηση μια μορφή σιγμοειδούς, επομένως έχει και αυτή το ίδιο πρόβλημα κορεσμού και συνδέεται με την λογιστική σιγμοειδή ως εξής: $\tanh(x) = 2\sigma(2x) - 1$

- **Γραμμικώς ανορθωμένη μονάδα(Rectified Linear Unit ή ReLU)**
  Οι γραμμικώς ανορθωμένες μονάδες χρησιμοποιούν την συνάρτηση ενεργοποίησης

$$f(x) = x^+ = \max\{0, x\} \tag{3.19}$$

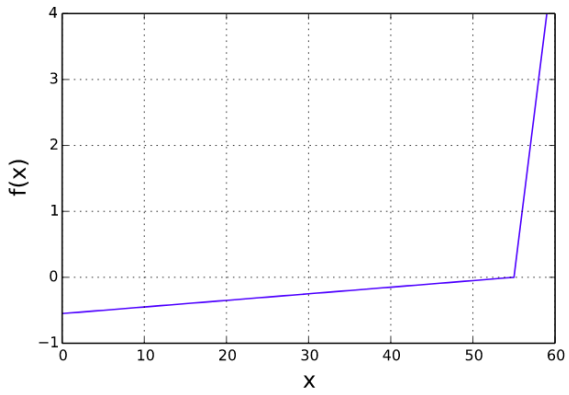


Σχήμα 3.5: Η συνάρτηση ReLU. Πηγή: [49]

Αυτές οι μονάδες είναι εύκολο να βελτιστοποιηθούν επειδή ομοιάζουν με τις γραμμικές μονάδες. Η μόνη διαφορά μεταξύ μιας γραμμικής μονάδας με μια ανορθωμένη γραμμική μονάδα είναι, ότι η ανορθωμένη είναι μηδενική στο μισό πεδίο ορισμού της. Αυτό οδηγεί τις παραγώγους να παραμένουν μεγάλες όταν η μονάδα είναι ενεργοποιημένη. Οι βαθμίδες δεν είναι μόνο μεγάλες, αλλά και σταθερές. Η δεύτερη παράγωγος είναι σχεδόν παντού 0, ενώ η παράγωγος της διαδικασίας ανόρθωσης είναι παντού ίση με 1, όταν η μονάδα είναι ενεργοποιημένη. Αυτό οδηγεί σε πολύ καλύτερη μάθηση και επίσης εξαλείφει το πρόβλημα του κορεσμού που παρουσιάζουν οι σιγμοειδής συναρτήσεις. Ένα μειονέκτημα, είναι ότι μια συνάρτηση ReLU, μπορεί να προκαλέσει "κόλλημα" στους νευρώνες και αυτοί να σταματήσουν να μαθαίνουν, και αυτό γιατί αν οι βαθμίδες που υπολογίζονται είναι αρνητικές, τότε η έξοδος θα είναι πάντοτε μηδενική.

- **Διαρρέουσα γραμμικώς ανορθωμένη μονάδα(Leaky ReLU)**
  Η Leaky ReLU σε αντίθεση με την απλή ReLU, δεν είναι μηδέν για $x < 0$, αλλά έχει μια μικρή αρνητική κλίση που καθορίζεται από την παράμετρο α, και η μορφή της είναι

$$f(x) = \begin{cases} x, & αν \quad x \geq 0, \\ \alpha x, & αλλιώς \end{cases} \tag{3.20}$$



Σχήμα 3.6: Η συνάρτηση Leaky ReLU. Πηγή: [49]

Άρα η Leaky ReLU διορθώνει το πρόβλημα με το "κόλλημα" στους νευρώνες που μπορεί να προκηθεί από την απλή ReLU.

3.4 Ο αλγόριθμος Back-Propagation

Όταν χρησιμοποιούμε ένα προσο-τροφοδοτούμενο νευρωνικό δίκτυο το οποίο δέχεται μια είσοδο $\mathbf{x}$ και παράγει μια έξοδο $\hat{\mathbf{y}}$, η πληροφορία ρέει προς τα μπροστά δια μέσου του δικτύου. Η είσοδος $\mathbf{x}$ παρέχει την αρχική πληροφορία η οποία στη συνέχεια διαδίδεται στις κρυφές μονάδες των επιπέδων του δικτύου και τελικά παράγει την έξοδο $\hat{\mathbf{y}}$. Αυτό καλείται εμπρόσθια διάδοση(Forward Propagation). Κατά τη διάρκεια της εκπαίδευσης του δικτύου, η εμπρόσθια διάδοση συνεχίζεται μέχρι να παραχθεί το βαθμωτό κόστος $J(\theta)$. Ο Back-Propagation αλγόριθμος [4] επιτρέπει στην πληροφορία από το κόστος, να αρχίσει να ρέει αντίστροφα, προς τα πίσω δηλαδή, διαμέσου του δικτύου, ώστε να υπολογιστεί η βαθμίδα.

Ουσιαστικά, ο Back-Propagation αλγόριθμος είναι η μέθοδος υπολογισμού της βαθμίδας, ενώ για να επιτευχθεί η μάθηση με την βαθμίδα αυτή, πρέπει να χρησιμοποιηθεί κάποιος άλλος αλγόριθμος, παραδείγματος χάριν, ο Stochastic Gradient Descent.

Για να ελαχιστοποιήσουμε λοιπόν τη συνάρτηση κόστους $J(\theta)$, πρέπει να υπολογίσουμε τη βαθμίδα. Επομένως με τον Back-Propagation μπορούν να υπολογιστούν οι παράγωγοι πολύπλοκων συναρτήσεων χρησιμοποιώντας τον κανόνα της αλυσίδας. Ο Back-Propagation, περιγράφεται συνήθως με την έννοια των υπολογιστικών γράφων(Computational Graphs). Ο αλγόριθμος, αποτελείται ουσιαστικά από πράξεις γινομένου Ιακωβιανού πινάκα και βαθμίδας για κάθε διεργασία στον γράφο.

3.5 Ομαλοποίηση

Ένα πολύ σημαντικό πρόβλημα για τη μηχανική μάθηση, και κατ' επέκταση και για την βαθιά μάθηση, είναι η δημιουργία αλγορίθμων που δεν θα αποδίδουν καλά μόνο στα δεδομένα εκπαίδευσης, αλλά και στα δεδομένα δοκιμής, δηλαδή σε νέες εισόδους. Υπάρχουν πολλές στρατηγικές που χρησιμοποιούνται για την μείωση του σφάλματος στο σύνολο δοκιμής, με κόστος πολλές φορές την αύξηση στο σφάλμα του συνόλου εκπαίδευσης. Αυτές οι στρατηγικές είναι γνωστές με τον όρο ομαλοποίηση(Regularization). Ο όρος ομαλοποίηση, μπορεί να εκφραστεί ως έννοια, ως οποιαδήποτε μετατροπή εφαρμόζουμε σε έναν αλγόριθμο μάθησης, ώστε να μειωθεί το σφάλμα γενίκευσης, αλλά όχι το σφάλμα εκπαίδευσης.

Πολλές προσεγγίσεις βασίζονται στον περιορισμό της χωρητικότητας των μοντέλων, όπως είναι τα νευρωνικά δίκτυα, η γραμμική ή λογιστική παλινδρόμηση, προσθέτοντας μια παράμετρο ποινής νόρμας $\Omega(\theta)$ στην συνάρτηση κόστους J. Ορίζουμε την ομαλοποιημένη συνάρτηση κόστους ως

$$\widetilde{J}(\theta; \mathbf{X}, \mathbf{y}) = J(\theta; \mathbf{X}, \mathbf{y}) + \alpha\Omega(\theta) \tag{3.21}$$

όπου $a \in [0, +\infty)$ είναι μια υπερπαράμετρος που καθορίζει το βάρος της παραμέτρου ποινής νόρμας Ω, σε σχέση με την συνάρτηση κόστους J. Θέτοντας α ίσο με μηδέν, δεν έχουμε ομαλοποίηση. Μεγαλύτερες τιμές του α αντιστοιχούν σε περισσότερη ομαλοποίηση.

3.5.1 Η παράμετρος ομαλοποίησης L^2

Μία από τις πιο κοινές παραμέτρους ποινής νόρμας είναι η L^2, η οποία είναι γνωστή και ως απόσβεση βάρους(Weight Decay). Η στρατηγική αυτή οδηγεί τα βάρη πιο κοντά στην αρχή των αξόνων, προσθέτοντας στην συνάρτηση κόστους έναν όρο

$$\Omega(\theta) = \frac{1}{2}||\mathbf{w}||_2^2 \tag{3.22}$$

Η L^2 ομαλοποίηση είναι επίσης γνωστή και ως ridge regression και Tikhonov regularization

3.5.2 Η παράμετρος ομαλοποίησης L^1

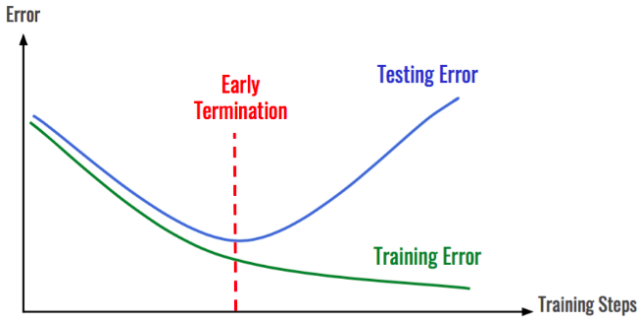
Ενώ η L^2 παράμετρος είναι η πιο κοινώς χρησιμοποιούμενη, υπάρχουν και άλλοι τρόποι για να εισαχθεί κάποια ποινή στη νόρμα. Μια επιλογή είναι η L^1 παράμετρος ομαλοποίησης για το μοντέλο παραμέτρων $\mathbf{w}$, η οποία ορίζεται ως

$$\Omega(\theta) = ||\mathbf{w}||_1 = \sum_i |w_i| \tag{3.23}$$

3.5.3 Πρόωρη παύση

Όταν η εκπαίδευση, γίνεται με μοντέλα αρκετά μεγάλα τα οποία έχουν επαρκή χωρητικότητα ώστε να υπερπροσαρμόζονται, παρατηρούμε ότι το σφάλμα εκπαίδευσης μειώνεται με σταθερό ρυθμό στο χρόνο, αλλά το σφάλμα στο σύνολο δοκιμής αρχίζει να έχει ξανά ανοδική πορεία.

Αυτό σημαίνει ότι μπορούμε να πάρουμε ένα μοντέλο με καλύτερο σφάλμα στο σύνολο δοκιμής, επιστρέφοντας τις ρυθμίσεις των παραμέτρων στο σημείο του χρόνου με το χαμηλότερο σφάλμα στο σύνολο δοκιμής. Κάθε φορά δηλαδή που το σφάλμα στο σύνολο δοκιμής βελτιώνεται, κρατάμε και ένα αντίγραφο των παραμέτρων του μοντέλου. Στο τέλος της εκπαίδευσης, επαναφέρουμε το καλύτερο μοντέλο.



Σχήμα 3.7: Πρόωρη παύση για περιορισμό της υπερπροσαρμογής. Πηγή: [51]

3.5.4 Dropout

Το Dropout [27], είναι μια μέθοδος αρκετά ισχυρή, αλλά και πολύ απλή υπολογιστικά, για την επίτευξη της ομαλοποίησης σε μια μεγάλη οικογένεια από μοντέλα. Για την περίπτωση των μοντέρνων νευρωνικών δικτύων, τα οποία βασίζονται σε μια σειρά από όμοιους μετασχηματισμούς και μη γραμμικότητες, η μέθοδος απλώς απενεργοποιεί κάποιες μονάδες του δικτύου, πολλαπλασιάζοντας την έξοδο των μονάδων αυτών με μηδέν.

Πιο συγκεκριμένα, για την εκπαίδευση με Dropout, χρησιμοποιούμε αλγόριθμο μάθησης που βασίζεται σε minibatches, ο οποίος κάνει μικρά βήματα, όπως παραδείγματος χάριν ο Stochastic Gradient Descent. Κάθε φορά που φορτώνουμε ένα παράδειγμα σε minibatch, χρησιμοποιούμε τυχαία μια διαφορετική δυαδική μάσκα, εφαρμόζοντάς την σε όλες τις εισόδους και τα κρυφές μονάδες του δικτύου. Η μάσκα για κάθε μονάδα δειγματοληπτείται ανεξάρτητα από κάθε άλλη μονάδα. Η πιθανότητα μια τιμή της μάσκας να είναι μονάδα, το οποίο προκαλεί σε μια μονάδα να περιέχεται στο δίκτυο, είναι μια υπερπαράμετρος καθορισμένη πριν ξεκινήσει η εκπαίδευση. Τυπικές τιμές για την πιθανότητα διατήρησης των μονάδων εισόδου είναι 0.8, ενώ για τις κρυφές μονάδες είναι 0.5. Έπειτα υλοποιείται η εμπρόσθια διάδοση, η Back-Propagation και η μάθηση συνεχίζεται ως συνήθως.

Αν υποθέσουμε δηλαδή ότι ένα διάνυσμα-μάσκα $\boldsymbol{\mu}$, καθορίζει ποιες μονάδες να περιέχονται, και $J(\boldsymbol{\theta}, \boldsymbol{\mu})$ καθορίζει το κόστος του μοντέλου, το οποίο καθορίζεται από τις παραμέτρους $\boldsymbol{\theta}$ και τη μάσκα $\boldsymbol{\mu}$. Τότε, η εκπαίδευση με Dropout, είναι απλώς η ελαχιστοποίηση του $\mathbb{E}_{\boldsymbol{\mu}} J(\boldsymbol{\theta}, \boldsymbol{\mu})$. Η προσδοκώμενη ή αναμενόμενη αυτή τιμή, περιέχει εκθετικά πολλούς όρους, αλλά μπορούμε να λάβουμε μια αμερόληπτη εκτίμηση της βαθμίδας, δειγματοληπτώντας τιμές του $\boldsymbol{\mu}$.

(a) Ένα τυπικό νευρωνικό δίκτυο (b) Μετά την εφαρμογή Dropout

Σχήμα 3.8: Απεικόνιση της μεθόδου Dropout. Πηγή: [27]

3.5.5 Επαύξηση του συνόλου δεδομένων

Ο καλύτερος τρόπος για να μπορέσει ένα μοντέλο μηχανικής μάθησης να γενικεύει καλύτερα, είναι να εκπαιδευτεί σε περισσότερα δεδομένα. Βεβαίως, στην πράξη, τα διαθέσιμα δεδομένα είναι περιορισμένα. Ένας τρόπος να ξεπεραστεί αυτό το πρόβλημα, είναι η δημιουργία "απομιμήσεων" των δεδομένων, και η προσθήκη τους στο σύνολο εκπαίδευσης.

Αυτή η προσέγγιση είναι ευκολότερη για προβλήματα ταξινόμησης. Ένας ταξινομητής, χρειάζεται να λάβει μια περίπλοκη, και υψηλών διαστάσεων είσοδο $\mathbf{x}$, και να εξάγει μια ταυτοποιήσιμη κατηγορία y. Αυτό σημαίνει ότι ο ταξινομητής πρέπει να είναι αμετάβλητος για ένα μεγάλο εύρος μετασχηματισμών. Μπορούμε να δημιουργήσουμε νέα ζεύγη $(\mathbf{x}, y)$, αρκετά εύκολα, απλώς μετασχηματίζοντας τις εισόδους $\mathbf{x}$ στο διαθέσιμο σύνολο εκπαίδευσης.

Η επαύξηση του συνόλου δεδομένων(Dataset Augmentation), έχει αποδειχθεί ιδιαίτερα αποδοτική για ένα συγκεκριμένο πρόβλημα ταξινόμησης: Την αναγνώριση αντικειμένων.

Αντιθέτως, για την περίπτωση της επεξεργασίας του ήχου, τα αποτελέσματα δεν είναι εξίσου αποδοτικά. Έχει γίνει η προσπάθεια από διάφορους ερευνητές τα προηγούμενα χρόνια για επαύξηση των μουσικών συνόλων δεδομένων, κυρίως με τις εξής τεχνικές [29], [38]:

- Ολίσθηση του pitch (Pitch Shifting)
- Χρονική παραμόρφωση (Time Stretching)
- Τυχαία εναλλαγή αριστερού και δεξιού καναλιού για κάθε όργανο
- Προσθήκη θορύβου στο υπόβαθρο, παραδείγματος χάριν συναυλίες με συνωστισμό, θόρυβο από το μετρό, από ομιλίες κ.λπ
- Συμπίεση δυναμικού εύρους (Dynamic Range Compression)

Μέχρι στιγμής, η συνολική βελτίωση από τέτοιου είδους επαύξηση των δεδομένων δεν θεωρείται σημαντική, $0.2dB$ στο SDR(βλ. σχέση 5.3) για τα φωνητικά χρησιμοποιώντας το DSD100 σύνολο δεδομένων [43].

3.5.6 Προσθήκη θορύβου στις ταμπέλες

Τα περισσότερα σύνολα δεδομένων, εμπεριέχουν έναν αριθμό λαθών στις ταμπέλες y. Είναι μεγάλο σφάλμα να μεγιστοποιήσουμε το $\log(y|\mathbf{x})$, όταν το y είναι λάθος. Ένας τρόπος για να το αποτρέψουμε αυτό, είναι να προσθέσουμε θόρυβο στις ταμπέλες. Λόγου χάριν, μπορούμε να υποθέσουμε ότι για μια μικρή σταθερά ϵ, η ταμπέλα y για το σύνολο εκπαίδευσης είναι σωστή με πιθανότητα $1-\epsilon$, ενώ σε αντίθετη περίπτωση οποιαδήποτε άλλη από τις πιθανές ταμπέλες μπορεί να είναι σωστή. Αυτή η τεχνική πιθανόν να είναι προτιμότερη για προβλήματα πολλών κλάσεων και όχι δυαδικής ταξινόμησης [45].

3.6 Βελτιστοποίηση

Αυτό το εδάφιο, εστιάζει σε μια συγκεκριμένη περίπτωση βελτιστοποίησης: Την εύρεση των παραμέτρων $\boldsymbol{\theta}$ ενός νευρωνικού δικτύου, ούτως ώστε να επιτευχθεί σημαντική ελάττωση της συνάρτησης κόστους $J(\boldsymbol{\theta})$, η οποία τυπικά περιλαμβάνει μια μετρική απόδοσης για το σύνολο εκπαίδευσης, καθώς επίσης και επιπρόσθετους όρους ομαλοποίησης.

Οι αλγόριθμοι βελτιστοποίησης που χρησιμοποιούνται για την εκπαίδευση βαθείων νευρωνικών δικτύων, διαφέρουν από τους παραδοσιακούς αλγορίθμους βελτιστοποίησης. Στη βαθιά μάθηση, συνήθως μας ενδιαφέρει μια μετρική απόδοσης P, η οποία ορίζεται ως προς το σύνολο δοκιμής και η οποία μπορεί να είναι αρκετά δύσκολο να ελεγχθεί. Δηλαδή δεν μπορούμε άμεσα να βελτιστοποιήσουμε την ποσότητα P, αλλά προσπαθούμε να ελλατώσουμε την συνάρτηση κόστους $J(\boldsymbol{\theta})$, ελπίζοντας πως έτσι θα βελτιωθεί και η ποσότητα P.

Τυπικά, η συνάρτηση κόστους μπορεί να γραφτεί ως η μέση τιμή στο σύνολο εκπαίδευσης

$$J(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x},y)\sim\hat{p}_{data}} L(f(\mathbf{x};\boldsymbol{\theta}), y) \tag{3.24}$$

όπου L είναι η συνάρτηση απωλειών ανά δείγμα, $f(\mathbf{x};\boldsymbol{\theta})$ είναι η πρόβλεψη της εξόδου όταν η είσοδος είναι $\mathbf{x}$, και $\hat{p}_{data}$ είναι η εμπειρική κατανομή. Για την περίπτωση της επιτηρούμενης μάθησης, το y είναι η ταμπέλα στόχος(Target Label).

Η παραπάνω εξίσωση (3.24), καθορίζει μια συνάρτηση κόστους, ως προς το σύνολο εκπαίδευσης. Συνήθως προτιμούμε να ελαχιστοποιούμε την συνάρτηση κόστους, για την οποία η αναμενόμενη ή μέση τιμή, λαμβάνεται ως πρός την κατανομή γέννησης δεδομένων p_{data}, αντί για το πεπερασμένο σύνολο εκπαίδευσης:

$$J^*(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x},y)\sim p_{data}} L(f(\mathbf{x};\boldsymbol{\theta}), y) \tag{3.25}$$

3.6.1 Ελαχιστοποίηση εμπειρικού ρίσκου

Η ποσότητα που δίνεται από την εξίσωση (3.25) είναι γνωστή ως ρίσκο. Σε αυτήν την περίπτωση η αναμενόμενη τιμή είναι παρμένη ως προς την κατανομή γέννησης δεδομένων p_{data}. Αν γνωρίζαμε την πραγματική κατανομή $p_{data}(\mathbf{x}, y)$, τότε η ελαχιστοποίηση του ρίσκου θα ήταν ένα πρόβλημα που επιλύεται με έναν αλγόριθμο βελτιστοποίησης. Όμως εμείς δεν γνωρίζουμε την $p_{data}(\mathbf{x}, y)$, αλλά

μόνο το ότι έχουμε ένα σύνολο εκπαίδευσης που αποτελείται από δείγματα, δηλαδή ένα πρόβλημα μηχανικής μάθησης.

Ο απλούστερος τρόπος να μετατρέψουμε ένα πρόβλημα μηχανικής μάθησης, σε ένα πρόβλημα βελτιστοποίησης είναι να ελαχιστοποιήσουμε το αναμενόμενο κόστος στο σύνολο εκπαίδευσης. Αυτό σημαίνει ότι αντικαθίσταται η πραγματική κατανομή $p(\mathbf{x}, y)$ με την εμπειρική κατανομή $\hat{p}(\mathbf{x}, y)$ που καθορίζεται από το σύνολο εκπαίδευσης. Ελαχιστοποιούμε το εμπειρικό ρίσκο

$$\mathbb{E}_{\mathbf{x},y\sim\hat{p}_{data}(\mathbf{x},y)}[L(f(\mathbf{x};\boldsymbol{\theta}),y)] = \frac{1}{m}\sum_{i=1}^{m} L(f(\mathbf{x}^{(i)};\boldsymbol{\theta}),y^{(i)}) \tag{3.26}$$

όπου m είναι ο αριθμός των δειγμάτων εκπαίδευσης.

Η διαδικασία αυτή καλείται ελαχιστοποίηση εμπειρικού ρίσκου. Σε προβλήματα βαθιάς μάθησης χρησιμοποιείται σπάνια διότι ως μέθοδος είναι επιρρεπής στην υπερπροσαρμογή.

3.6.2 Batch και minibatch αλγόριθμοι

Μια σημαντική διαφορά που διαχωρίζει, τους αλγορίθμους μηχανικής μάθησης, από τους απλούς αλγορίθμους βελτιστοποίησης, είναι το ότι η συνάρτηση κόστους συνήθως εκφράζεται ως ένα άθροισμα στα δεδομένα εκπαίδευσης. Δηλαδή, αλγόριθμοι βελτιστοποίησης για προβλήματα μηχανικής μάθησης, υπολογίζουν σε κάθε βήμα τις παραμέτρους, ως την αναμενόμενη τιμή της συνάρτησης κόστους, της οποίας η εκτίμηση γίνεται χρησιμοποιώντας μόνο ένα υποσύνολο των όρων της πλήρους συνάρτησης κόστους.

Παραδείγματος χάριν, προβλήματα εκτίμησης μέγιστης πιθανοφάνειας, από λογαριθμική σκοπιά, αποσυντίθενται σε ένα άθροισμα των δειγμάτων

$$\boldsymbol{\theta}_{ML} = \underset{\boldsymbol{\theta}}{argmax}\sum_{i=1}^{m}\log p_{model}(\mathbf{x}^{(i)},y^{(i)};\boldsymbol{\theta}) \tag{3.27}$$

Η μεγιστοποίηση του παραπάνω αθροίσματος, είναι ταυτόσιμη με την ελαχιστοποίηση την αναμενόμενης τιμής της εμπειρικής κατανομής που καθορίζεται από το σύνολο εκπαίδευσης

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},y\sim\hat{p}_{data}}\log p_{model}(\mathbf{x},y;\boldsymbol{\theta}) \tag{3.28}$$

Αλγόριθμοι βελτιστοποίησης που χρησιμοποιούν ολόκληρο το σύνολο εκπαίδευσης, καλούνται batch ή ντετερμινιστικοί μέθοδοι βαθμίδας(Deterministic Gradient Methods), διότι επεξεργάζονται όλα τα δεδομένα εκπαίδευσης ταυτόχρονα σε ένα μεγάλο batch. Πρέπει να σημειωθεί ότι αυτή η ορολογία, μπορεί να έχει αμφίσημη σημασία, επειδή η λέξη "batch", χρησιμοποιείται επίσης για την περιγραφή του minibatch που χρησιμοποιείται από τον Minibatch Stochastic Gradient Descent. Τυπικά ο όρος "Batch Gradient Descent", υποδηλώνει ότι χρησιμοποιούμε το πλήρες σύνολο εκπαίδευσης, ενώ ο όρος "batch" δεν υποδηλώνει το ίδιο. Παραδείγματος χάριν, είναι κοινή η χρήση του όρου "batch size", για να περιγράψουμε το μέγεθος του minibatch.

Οι αλγόριθμοι βελτιστοποίησης που χρησιμοποιούν ένα μόνο δείγμα τη φορά, καλούνται μερικές φορές στοχαστικοί ή και online μέθοδοι. Ο όρος "online" χρησιμοποιείται συνήθως όταν τα δείγματα λαμβάνονται από μια συνεχόμενη ροή δειγμάτων αντί για ένα προκαθορισμένου μεγέθους σύνολο εκπαίδευσης στο οποίο γίνονται αρκετά περάσματα.

Οι περισσότεροι αλγόριθμοι που χρησιμοποιούνται για βαθιά μάθηση, βρίσκονται κάπου στο ενδιάμεσο, δηλαδή χρησιμοποιούν περισσότερα του ενός δείγματος, αλλά όχι όμως το σύνολο των δειγμάτων του συνόλου εκπαίδευσης. Αυτές οι μέθοδοι καλούνται minibatch ή στοχαστικές minibatch, ή απλώς στοχαστικές. Το μέγεθος του minibatch, καθορίζεται κυρίως από τους εξής παράγοντες:

- Μεγαλύτερα batches παρέχουν πιο ακριβή εκτίμηση της βαθμίδας, αλλά δεσμεύουν περισσότερη RAM.
- Πολυπύρηνες αρχιτεκτονικές συνήθως δεν μπορούν να χρησιμοποιήσουν την πλήρη επεξεργαστική ισχύ τους για υπερβολικά μικρά μεγέθη των batches. Αυτό οδηγεί στην επιλογή ενός απόλυτου ελάχιστου μεγέθους batch.
- Αν όλα τα δείγματα του batch πρόκειται να επεξεργαστούν παράλληλα, όπως δηλαδή είναι και η τυπική περίπτωση επεξεργασίας, τότε, το μέγεθος της μνήμης μεγαλώνει ανάλογα με το μέγεθος του batch. Επομένως ανάλογα με το διαθέσιμο hardware, υπάρχει ένας περιοριστικός παράγοντας στο μέγεθος του batch.
- Κάποια είδη hardware, μπορούν να αποδόσουν καλύτερα με συγκεκριμένα μεγέθη πινάκων. Ειδικά όταν χρησιμοποιούνται GPUs, είναι κοινό το μέγεθος του batch, να είναι δύναμη του 2, και να οδηγεί σε καλύτερους χρόνους εκτέλεσης.
- Τα μικρά batches, μπορούν να έχουν κάποια επίδραση ομαλοποίησης στο δίκτυο, πιθανόν επειδή εισάγουν θόρυβο στη διαδικασία της μάθησης. Το σφάλμα γενίκευσης, είναι συχνά βέλτιστο για batch μεγέθους 1. Όμως η εκπαίδευση με τόσο μικρό μέγεθος batch, πιθανόν να απαιτεί και μικρό ρυθμό μάθησης για την διατήρηση της σταθερότητας, λόγω της μεγάλης μεταβλητότητας στον υπολογισμό της βαθμίδας. Ο τελικός χρόνος εκτέλεσης επομένως θα είναι αρκετά μεγάλος, λόγω του ότι χρειάζονται πολύ περισσότερα βήματα.

Είναι πάρα πολύ σημαντικό, η επιλογή των minibatches να γίνεται με τυχαιότητα. Ο υπολογισμός ενός αμερόληπτου εκτιμητή της αναμενόμενης βαθμίδας από ένα σύνολο δειγμάτων, απαιτεί ότι τα δείγματα αυτά είναι στατιστικώς ανεξάρτητα. Επιπρόσθετα, θέλουμε οι διαδοχικές εκτιμήσεις βαθμίδων να είναι και αυτές στατιστικώς ανεξάρτητες, ώστε και τα διαδοχικά minibatches των δειγμάτων, να είναι επίσης στατιστικώς ανεξάρτητα.

3.6.3 Συχνά προβλήματα βελτιστοποίησης

Τοπικά ελάχιστα

Ένα από τα πιο σημαντικά χαρακτηριστικά ενός προβλήματος βελτιστοποίησης κυρτών συναρτήσεων, είναι ότι μπορεί να αναχθεί σε πρόβλημα εύρεσης τοπικού ελαχίστου. Οποιοδήποτε τοπικό ελάχιστο, εγγυάται να είναι και ολικό ελάχιστο. Όταν βελτιστοποιούμε μια κυρτή συνάρτηση, γνωρίζουμε ότι έχουμε φτάσει σε μια ικανοποιητική λύση αν βρούμε ένα κρίσιμο σημείο οποιουδήποτε είδους.

Για μη κυρτές συναρτήσεις όμως, όπως είναι τα νευρωνικά δίκτυα, είναι πιθανό να έχουμε πολλά τοπικά ελάχιστα. Ειδικά για τα βαθεία νευρωνικά δίκτυα, έχουμε έναν υπερβολικά μεγάλο αριθμό τοπικών ελαχίστων, αν και αυτό δεν αποτελεί απαραίτητα πρόβλημα.

Τα νευρωνικά δίκτυα, έχουν πολλαπλά τοποικά ελάχιστα, λόγω του προβλήματος της ταυτοποιησιμότητας του μοντέλου(Model Identifiability Problem). Ένα μοντέλο είναι ταυτοποιήσιμο αν ένα ικανοποιητικά μεγάλο σύνολο εκπαίδευσης μπορεί να αποκλείσει όλες τις παραμέτρους του μοντέλου εκτός από μία. Παραδείγματος χάριν, στα νευρωνικά δίκτυα, υπάρχει η μη ταυτοποιησιμότητα, γνωστή ως συμμετρία βαρών χώρου(Weight Space Symmetry), όπου έχουμε m επίπεδα με n μονάδες το καθένα, ώστε συνολικά να υπάρχουν $n!^m$ πιθανοί τρόποι διάταξης των κρυφών μονάδων.

Εκτός της συμμετρίας βαρών χώρου, υπάρχουν και άλλες περιπτώσεις μη ταυτοποιησιμότητας για τα νευρωνικά δίκτυα. Λόγου χάριν, για οποιοδήποτε δίκτυο που χρησιμοποιεί γραμμικώς ανορθωμένες μονάδες ή του οποίου η έξοδος είναι η μέγιστη των εισόδων(Maxout Network), μπορούμε να πολλαπλασιάσουμε τα βάρη εισόδου με έναν παράγοντα α, αν επιπλέον πολλαπλασιάσουμε τα βάρη εξόδου με έναν παράγοντα $1/\alpha$. Αυτό σημαίνει ότι αν η συνάρτηση κόστους, δεν περιλαμβάνει όρους, όπως είναι η L^2 παράμετρος ποινής νόρμας, η οποία εξαρτάται άμεσα από τα βάρη και όχι από τις εξόδους του μοντέλου, τότε οποιοδήποτε τοπικό ελάχιστο δικτύου που χρησιμοποιεί γραμμικώς ανορθωμένες μονάδες, ή δικτύου του οποίου η έξοδος είναι η μέγιστη των εισόδων [25], βρίσκεται πάνω σε μια παραβολή $(m \times n)$-διαστάσεων με τα αντίστοιχα τοπικά ελάχιστα.

Αυτά τα προβλήματα ταυτοποιησιμότητας, σημαίνουν ότι η συνάρτηση κόστους του νευρωνικού δικτύου μπορεί να έχει έναν υπερβολικά μεγάλο αριθμό τοπικών ελαχίστων. Όμως, όλα αυτά τα τοπικά ελάχιστα που προκύπτουν από την μη ταυτοποιησιμότητα, έχουν ισοδύναμες τιμές στη συνάρτηση κόστους που σημαίνει ότι δεν αποτελούν πρόβλημα της μορφής της μη κυρτότητας.

Αποδεικνύεται δηλαδή ότι τα τοπικά ελάχιστα δεν αποτελούν σημαντικό πρόβλημα, παρα μόνο όταν έχουν μεγάλο κόστος σε σχέση με το ολικό ελάχιστο. Γενικά το πρόβλημα αυτό παραμένει ένα ανοικτό θέμα για την ερευνητική κοινότητα, αλλά το πιο πιθανό είναι ότι τα περισσότερα τοπικά ελάχιστα έχουν μια μικρή τιμή που προκύπτει από τη συνάρτηση κόστους, και επομένως δεν είναι τόσο σημαντικό να βρεθεί ένα πραγματικό ολικό ελάχιστο. Αντίθετα αρκεί να βρεθεί ένα σημείο στον χώρο των παραμέτρων το οποίο να έχει χαμηλό αλλά όχι ελάχιστο κόστος.

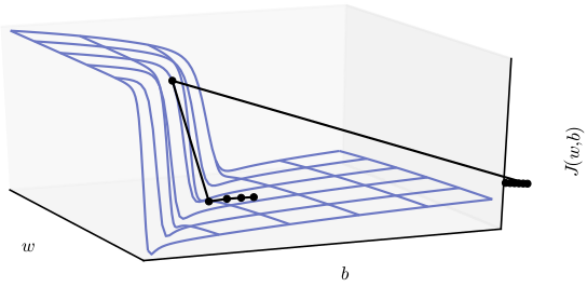
Εκτοξευόμενες και εξαφανιζόμενες βαθμίδες

Μια άλλη δυσκολία που έχουν να αντιμετωπίσουν οι αλγόριθμοι βελτιστοποίησης νευρωνικών δικτύων, προκύπτει όταν ο υπολογιστικός γράφος γίνεται υπερβολικά βαθύς. Τα προσο-τροφοδοτούμενα με πολλά επίπεδα έχουν τέτοιους γράφους.

Λόγου χάριν, υποθέτουμε ότι ένας υπολογιστικός γράφος, περιέχει ένα μονοπάτι, το οποίο αποτελείται από επαναλαμβανόμενες επαναλήψεις πολλαπλασιασμών με έναν πίνακα $\mathbf{W}$. Μετά από t βήματα, αυτό είναι ισοδύναμο με το να πολλαπλασιάσουμε με $\mathbf{W}^t$. Υποθέτουμε ότι ο $\mathbf{W}$ έχει ιδιο-παραγοντοποίηση $\mathbf{W} = \mathbf{V}diag(\boldsymbol{\lambda})\mathbf{V}^{-1}$. Επομένως προκύπτει

$$\mathbf{W}^t = (\mathbf{V}diag(\boldsymbol{\lambda})\mathbf{V}^{-1})^t = \mathbf{V}diag(\boldsymbol{\lambda})^t\mathbf{V}^{-1} \tag{3.29}$$

Οποιεσδήποτε ιδιοτιμές λ_i οι οποίες δεν είναι κοντά σε μια απόλυτη τιμή του 1, είτε θα εκτοξευθούν αν είναι μεγαλύτερες του 1 σε πλάτος, είτε θα εξαφανιστούν αν είναι μικρότερες του 1 σε πλάτος. Αυτό είναι το πρόβλημα των εκτοξευόμενων και εξαφανιζόμενων βαθμίδων(Exploding and Vanishing Gradient Problem), και αναφέρεται στο γεγονός, ότι οι βαθμίδες σε έναν τέτοιο γράφο, κλιμακώνονται ανάλογα το $diag(\boldsymbol{\lambda})^t$.



Σχήμα 3.9: Η συνάρτηση κόστους για υψηλώς μη γραμμικά βαθεία νευρωνικά δίκτυα, συχνά εμπεριέχει έντονες μη γραμμικότητες στον χώρο των παραμέτρων, οι οποίες οφείλονται στον πολλαπλασιασμό πολλών παραμέτρων. Αυτές οι μη γραμμικότητες, οδηγούν σε πολύ μεγάλες παραγώγους σε κάποια σημεία του χώρου. Όταν οι παράμετροι είναι κοντά σε μιά απότομη περιοχή, σε μια πλαγιά δηλαδή, όπως απεικονίζεται στο σχήμα, τότε ένα βήμα του Gradient Descent, μπορεί να λειτουργήσει ως καταπέλτης και να στείλει τις παραμέτρους πολύ μακριά, χάνοντας έτσι το μεγαλύτερο κομμάτι της βελτιστοποίησης που είχε επιτευχθεί μέχρι στιγμής.

3.6.4 Βασικοί αλγόριθμοι

Έχουμε ήδη αναφέρει τους βασικότερους αλγορίθμους βελτιστοποίησης για μηχανική μάθηση. Τον Gradient Descent ο οποίος ακολουθεί την βαθμίδα καθοδικά, καθόλη τη διάρκεια της εκπαίδευσης, αλλά και μια βελτιωμένη έκδοσή του, τον Stochastic Gradient Descent που χρησιμοποιεί αμερόληπτα επιλεγμένα minibatches.

Ενώ ο Stochastic Gradient Descent παραμένει μια αρκετά δημοφιλής στρατηγική βελτιστοποίησης, η μάθηση μπορεί καποιες φορές να είναι αργή. Μια άλλη μέθοδος, που βελτιώνει την ταχύτητα της μάθησης είναι η μέθοδος της **ορμής**(Momentum). Η μέθοδος αυτή βοηθά κυρίως στην αντιμετώπιση της υψηλής κυρτότητας, καθώς και μικρών αλλά σταθερών βαθμίδων ή και θορυβικών βαθμίδων.

Μια παραλλαγή της ορμής, είναι η **ορμή Nesterov**. Η ορμή Nesterov, μπορεί να ερμηνευθεί και ως η μέθοδος που προσπαθεί να προσθέσει έναν διορθωτικό παράγοντα στην μέθοδο της ορμής. Για την περίπτωση του Stochastic Gradient Descent όμως, η ορμή Nesterov δεν βελτιώνει τον ρυθμό της σύγκλισης.

3.6.5 Αλγόριθμοι με προσαρμοζόμενους ρυθμούς μάθησης

Οι ερευνητές που ασχολούνται με νευρωνικά δίκτυα, έχουν συνειδητοποιήσει εδώ και καιρό, ότι ο ρυθμός μάθησης είναι μια από τις πιο δύσκολες στον καθορισμό, υπερπαραμέτρους, διότι επηρεάζει σημαντικά την απόδοση του μοντέλου. Γενικά, το κόστος είναι συχνά αρκετά ευαίσθητο, σε κάποιες κατευθύνσεις του χώρου παραμέτρων σε αντίθεση με κάποιες άλλες. Η μέθοδος της ορμής που αναφέρθηκε στο προηγούμενο εδάφιο, μπορεί μετριάσει κατά κάποιον τρόπο αυτά τα προβλήματα, αλλά το κάνει προσθέτοντας μια επιπλέον υπερπαράμετρο. Διερωτόμαστε επομένως αν υπάρχει άλλος τρόπος. Θεωρώντας ότι οι κατευθύνσεις της ευαισθησίας που αναφέρθηκαν, είναι κατα κάποιον τρόπο τοποθετημένες σε άξονες, θα μπορούσαμε να χρησιμοποιήσουμε ξεχωριστούς ρυθμούς μάθησης για κάθε παράμετρο και αυτόματα να προσαρμόζουμε αυτούς τους ρυθμούς μάθησης κατά τη διάρκεια της εκπαίδευσης του δικτύου.

Χαρακτηριστικοί αλγόριθμοι που ανήκουν σε αυτήν την κατηγορία είναι ο **AdaGrad**, ο **RMSProp** και ο διασημότερος που είναι ο **Adam**, ο οποίος χρησιμοποιείται στην συγκεκριμένη διπλωματική εργασία και θα αναλυθεί στο επόμενο εδάφιο.

3.6.6 Ο αλγόριθμος Adam

Ο αλγόριθμος Adam, είναι ένας αλγόριθμος βελτιστοποίησης με προσαρμοζόμενο ρυθμό μάθησης, και το όνομά του προέρχεται από την φράση "Adaptive Moments". Θεωρείται πιθανόν, ως μια παραλαγή του συνδυασμού του RMSProp και της ορμής με μερικές διαφοροποιήσεις.

Αλγόριθμος 1: Ο αλγόριθμος Adam

Απαίτησε: Μέγεθος βήματος ϵ (Προτεινόμενο: 0.001)
Απαίτησε: Εκθετικούς ρυθμούς απόσβεσης για τους εκτιμητές ορμής, ρ_1 και ρ_2 στο διάστημα $[0, 1)$(Προτεινόμενα: 0.9 και 0.999 αντίστοιχα)
Απαίτησε: Αρχικές παραμέτρους θ
Αρχικοποίησε την πρώτη και δεύτερη μεταβλητή της ορμής $\mathbf{s} = \mathbf{0}$, $\mathbf{r} = \mathbf{0}$;
Αρχικοποίησε τον χρόνο του βήματος $t = 0$;
Όσο *δεν ισχύει το κριτήριο τερματισμού* **Κάνε**
  Δειγματολήπτησε ένα minibatch από m δείγματα του συνόλου εκπαίδευσης $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ με τις αντίστοιχες επιθυμητές εξόδους να είναι $\mathbf{y}^{(i)}$;
  Υπολόγισε την βαθμίδα: $\mathbf{g} \leftarrow \frac{1}{m} \nabla_{\boldsymbol{\theta}} \sum_i L(f(\mathbf{x}^{(i)}; \boldsymbol{\theta}), \mathbf{y}^{(i)})$;
  $t \leftarrow t + 1$;
  Ανανέωσε την μεροληπτική πρώτη εκτίμηση της ορμής: $\mathbf{s} \leftarrow \rho_1 \mathbf{s} + (1 - \rho_1)\mathbf{g}$;
  Ανανέωσε την μεροληπτική δεύτερη εκτίμηση της ορμής: $\rho \leftarrow \rho_2 \mathbf{r} + (1 - \rho_2)\mathbf{g} \odot \mathbf{g}$;
  Διόρθωσε την μεροληψία της πρώτης ορμής: $\hat{\mathbf{s}} \leftarrow \frac{\mathbf{s}}{1-\rho_2^t}$;
  Υπολόγισε την ανανέωση: $\boldsymbol{\Delta\theta} = -\epsilon \frac{\hat{\mathbf{s}}}{\sqrt{\hat{\mathbf{r}}}+\delta}$ (Οι πράξεις εφαρμόζονται στοιχείο προς στοιχείο);
  Εφάρμοσε την ανανέωση: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{\Delta\theta}$;
Τέλος

3.6.7 Batch Normalization

Πολλές τεχνικές βελτιστοποίησης, δεν είναι ακριβώς αλγόριθμοι, αλλά μπορούν να χαρακτηριστούν ως πρότυπα τα οποία ειδικεύονται στο να εξάγουν αλγορίθμους ή υπορουτίνες οι οποίες μπορούν να ενταχθούν σε πολλούς άλλους διαφορετικούς αλγορίθμους.

Η τεχνική Batch Normalization είναι μια από τις σημαντικότερες και πιο πρόσφατες καινοτομίες, για την βελτιστοποίηση βαθέων νευρωνικών δικτύων, και στην πραγματικότητα δεν αποτελεί καν αλγόριθμο βελτιστοποίησης. Αντίθετα, είναι μια μέθοδος προσαρμοζόμενης επαναπαραμετροποίησης.

Τα πολύ βαθιά μοντέλα περιέχουν την σύνθεση πολλών συναρτήσεων και επιπέδων. Η βαθμίδα μας λέει πως να ανανεώσουμε κάθε παράμετρο, υπό την προϋπόθεση ότι τα άλλα επίπεδα δεν αλλάζουν. Στην πράξη, αναβαθμίζουμε όλα τα επίπεδα ταυτόχρονα. Όταν κάνουμε την αναβάθμιση, μη αναμενόμενα αποτελέσματα μπορούν να συμβούν επειδή πολλές συναρτήσεις που είναι συντεθιμένες μεταξύ τους, αλλάζουν ταυτόχρονα, χρησιμοποιώντας ανανεώσεις που υπολογίστηκαν υπό την προϋπόθεση ότι οι άλλες συναρτήσεις παρέμειναν σταθερές.

Το Batch Normalization παρέχει έναν απλό τρόπο για την επαναπαραμετροποίηση οποιουδήποτε σχεδόν βαθέος δικτύου, μειώνει σημαντικά το πρόβλημα των συντονισμένων ανανεώσεων κατά μήκος των διάφορων επιπέδων

και μπορεί να εφαρμοσθεί σε οποιαδήποτε είσοδο ή κρυφό επίπεδο σε ένα δίκτυο. Έστω $\mathbf{H}$ αποτελεί ένα minibatch από ενεργοποιήσεις του επιπέδου προς κανονικοποίηση, διατεταγμένο ως πίνακας, με τις ενεργοποιήσεις για κάθε δείγμα να απεικονίζονται σε μια γραμμή του πίνακα. Για την κανονικοποίηση του $\mathbf{H}$, το αντικαθιστούμε με

$$\mathbf{H}' = \frac{\mathbf{H} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{3.30}$$

όπου $\boldsymbol{\mu}$ είναι ένα διάνυσμα που περιέχει τη μέση τιμή κάθε μονάδας και $\boldsymbol{\sigma}$ είναι ένα διάνυσμα που περιέχει την τυπική απόκλιση κάθε μονάδας. Ουσιαστικά το διάνυσμα $\boldsymbol{\mu}$ και το διάνυσμα $\boldsymbol{\sigma}$, εφαρμόζονται σε κάθε γραμμή του πίνακα $\mathbf{H}$ και οι πράξεις γίνονται στοιχείο προς στοιχείο. Το υπόλοιπο του δικτύου λειτουργεί στο $\mathbf{H}'$, ακριβώς με τον ίδιο τρόπο που το κανονικό δίκτυο το έπρατε στο $\mathbf{H}$.

Κατά τον χρόνο εκπαίδευσης ισχύει

$$\boldsymbol{\mu} = \frac{1}{m} \sum_i \mathbf{H}_{i,:} \tag{3.31}$$

και

$$\boldsymbol{\sigma} = \sqrt{\delta + \frac{1}{m} \sum_i (\mathbf{H} - \boldsymbol{\mu})_i^2} \tag{3.32}$$

όπου δ είναι μια μικρή θετική τιμή(π.χ. 10^{-8}), η οποία εφαρμόζεται προς αποφυγή της περίπτωσης όπου η βαθμίδα δεν ορίζεται για $\sqrt{z}$ στο $z = 0$. Εν συνεχεία, κάνουμε back-propagation κατά μήκος αυτών των διεργασιών, για τον υπολογισμό της μέσης τιμής και της τυπικής απόκλισης και για την εφαρμογή τους στην κανονικοποίηση του $\mathbf{H}$. Αυτό σημαίνει ότι η βαθμίδα δεν θα υλοποιήσει ποτέ μια διεργασία η οποία δρα έτσι ώστε να αυξάνονται η τυπική απόκλιση και η μέση τιμή του h_i. Οι διεργασίες κανονικοποίησης που προαναφέρθηκαν, αποσβένουν την επίδραση μιας τέτοιας ενέργειας και μηδενίζουν τη συνιστώσα της ενέργειας αυτής στη βαθμίδα. Αυτή ήταν και η μεγάλη καινοτομία της μεθόδου Batch Normalization.

Τέλος, για τον χρόνο δοκιμής, τα $\boldsymbol{\mu}$ και $\boldsymbol{\sigma}$, μπορούν να αντικατασταθούν από μέσους όρους που συλλέχθησαν κατά τον χρόνο εκπαίδευσης. Αυτό επιτρέπει στο μοντέλο να μπορεί να αξιολογηθεί σε ένα μόνο δείγμα, χωρίς να χρειάζεται να χρησιμοποιηθούν οι ορισμοί των μ και σ, οι οποίες εξαρτώνται από ολόκληρο το minibatch.

3.7 Συνελικτικά νευρωνικά δίκτυα

Τα συνελικτικά νευρωνικά δίκτυα(Convolutional Neural Networks ή CNNs) έχουν παίξει σημαντικό ρόλο στην ιστορία της βαθιάς μάθησης. Είναι από τα πρώτα βαθεία μοντέλα που απέδωσαν τόσο καλά, και επιπλέον ήταν από τα πρώτα που χρησιμοποιήθηκαν για την επίλυση σημαντικών προβλημάτων και ανάπτυξη εμπορικών εφαρμογών, και παραμένουν η κύρια επιλογή για ανάπτυξη εμπορικών εφαρμογών βαθιάς μάθησης στις μέρες μας.

Επιπρόσθετα, χρησιμοποιήθηκαν για να κερδιθούν αρκετοί διαγωνισμοί. Η έντονη πρόσφατη ενασχόληση με συνελικτικά νευρωνικά δίκτυα, ξεκίνησε το 2012, όταν ο Krizhevsky κ.α., κέρδισαν τον διαγωνισμό αναγνώρισης αντικειμένων ImageNet.

Τα συνελικτικά νευρωνικά δίκτυα, αποτελούν ένα ειδικευμένο είδος νευρωνικών δικτύων για την επεξεργασία δεδομένων τα οποία έχουν τοπολογία πλέγματος. Λόγου χάριν, δεδομένα χρονοσειρών(Time-Series Data) μπορούν να ερμηνευθούν ως ένα μονοδιάστατο(1-D) πλέγμα, λαμβάνοντας δεδομένα ανά τακτικά χρονικά διαστήματα, και επιπλέον δεδομένα εικόνων, μπορούν να ερμηνευθούν ως δισδιάστατα(2-D) πλέγματα από εικονοστοιχεία(Pixels). Τα συνελικτικά νευρωνικά δίκτυα, έχουν επιτύχει εξαιρετικά αποτελέσματα σε πολλές πρακτικές εφαρμογές. Το όνομά τους, υποδηλώνει ότι το δίκτυο υλοποιεί την μαθηματική πράξη της συνέλιξης η οποία θα αναλυθεί στο επόμενο εδάφιο.

3.7.1 Η διαδικασία της συνέλιξης

Έστω μια συνάρτηση $x(t)$, όπου $x, t \in \mathbb{R}$ και έστω μια συνάρτηση βάρους $w(\alpha)$, όπου $w, \alpha \in \mathbb{R}$. Η μαθηματική πράξη της συνέλιξης ορίζεται ως

$$s(t) = \int x(\alpha) w(t - \alpha) d\alpha \tag{3.33}$$

Τυπικά η πράξη της συνέλιξης συμβολίζεται με αστερίσκο.

$$s(t) = (x * w)(t) \tag{3.34}$$

Για την ορολογία των συνελικτικών δικτύων, ο πρώτος όρος(η συνάρτηση x) της συνέλιξης συχνά καλείται ως **είσοδος**, και ο δεύτερος όρος(η συνάρτηση w) ως το **kernel**. Η έξοδος συχνά καλείται ως **χάρτης χαρακτηριστικών**(Feature Map).

Στην πράξη, όταν δουλεύουμε με δεδομένα σε έναν υπολογιστή, η μεταβλητή t είναι διακριτοποιημένη και επομένως μπορεί να πάρει μονο ακέραιες τιμές. Αν υποθέσουμε ότι τα x και w ορίζονται μόνο για τον ακέραιο t, τότε μπορούμε να ορίσουμε την διακριτή συνέλιξη

$$s(t) = (x * w)(t) = \sum_{\alpha=-\infty}^{\infty} x(\alpha) w(t - \alpha) \tag{3.35}$$

Σε εφαρμογές μηχανικής μάθησης, συνήθως η είσοδος είναι ένας πολυδιάστατος πίνακας δεδομένων και το kernel είναι συνήθως ένας πολυδιάστατος πίνακας από παραμέτρους οι οποίες προσαρμόζονται από τον αλγόριθμο μάθησης. Θα αναφερόμαστε σε αυτούς τους πολυδιάστατους πίνακες ως tensors. Επειδή κάθε στοιχείο της εισόδου και του kernel, πρέπει να αποθηκευθεί εξωτερικά και ξεχωριστά, συνήθως υποθέτουμε ότι αυτές οι συναρτήσεις είναι μηδέν οπουδήποτε εκτός από το πεπερασμένο σύνολο των σημείων για τα οποία αποθηκεύουμε τις τιμές. Στην πράξη αυτό σημαίνει ότι μπορούμε να υλοποιήσουμε το άπειρο άθροισμα ως άθροισμα σε έναν πεπερασμένο αριθμό από στοιχεία του πίνακα.

Συχνά, χρησιμοποιούμε συνελίξεις σε περισσότερους από έναν άξονες τη φορά. Λόγου χάριν, αν χρησιμοποιήσουμε μια δισδιάστατη εικόνα I ως είσοδο, πιθανόν να θέλουμε να χρησιμοποιήσουμε ένα kernel δύο διαστάσεων K

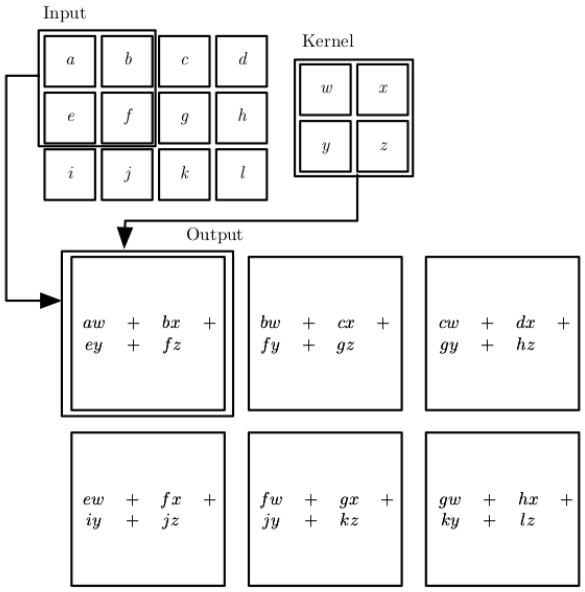
$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \tag{3.36}$$

Η συνέλιξη είναι αντιμεταθετική και επομένως γράφουμε ισοδυνάμως

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \tag{3.37}$$

Η αντιμεταθετική ιδιότητα της συνέλιξης, προκύπτει από το "αναποδογύρισμα" του kernel σε σχέση με την είσοδο. Πολλές βιβλιοθήκες νευρωνικών δικτύων υλοποιούν μια παρόμοια συνάρτηση η οποία καλείται δια-συσχέτιση(cross-correlation), η οποία είναι η ίδια όπως η συνέλιξη αλλά χωρίς "αναποδογύρισμα" του kernel

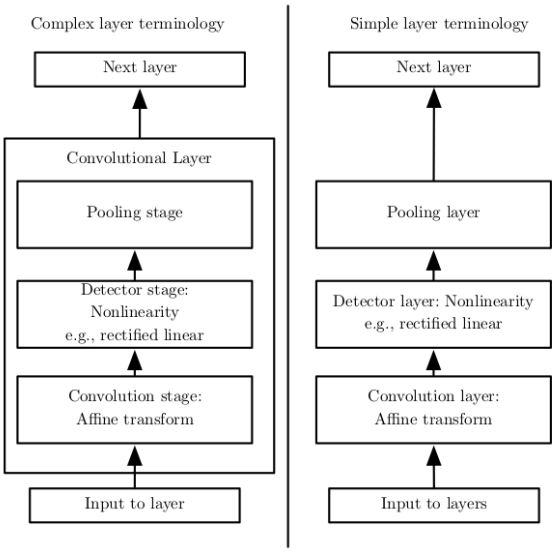
$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n)K(m,n) \tag{3.38}$$



Σχήμα 3.10: Παράδειγμα δισδιάστατης συνέλιξης χωρίς "αναποδογύρισμα" του kernel. Οι έξοδοι περιορίζονται μόνο στις θέσεις όπου το kernel "πέφτει" ακριβώς εντός της εικόνας, το οποίο καλείται και "έγκυρη" συνέλιξη κάποιες φορές.

3.7.2 Ομαδοποίηση

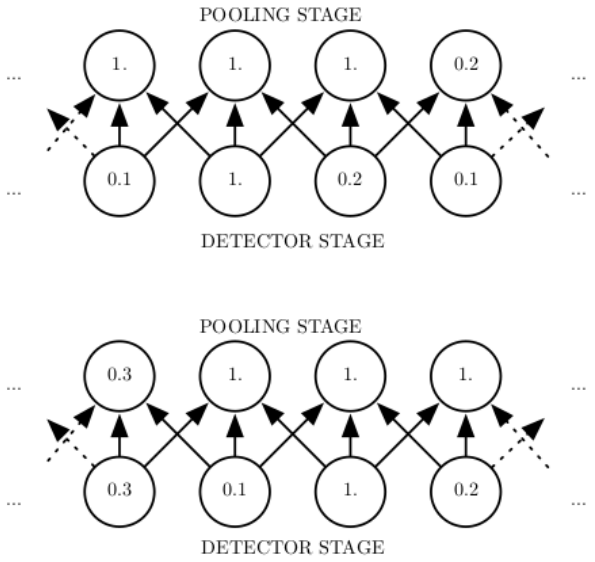
Ένα τυπικό επίπεδο ενός συνελικτικού δικτύου αποτελείται από τρία στάδια. Στο πρώτο στάδιο, το επίπεδο εφαρμόζει διάφορες παράλληλες συνελίξεις για την παραγωγή ενός συνόλου γραμμικών συναρτήσεων. Στο δεύτερο στάδιο, κάθε γραμμική ενεργοποίηση περνά δια μέσου μιας μη γραμμικής συνάρτησης ενεργοποίησης, όπως είναι οι γραμμικώς ανορθωμένες συναρτήσεις(βλ. σχέση 3.19). Αυτό το στάδιο, καλείται συχνά και **στάδιο ανίχνευσης**(detector stage). Στο τρίτο στάδιο, χρησιμοποιούμε μια **συνάρτηση ομαδοποίησης**(pooling function) για να διαμορφώσουμε την έξοδο του επιπέδου.



Σχήμα 3.11: Απεικόνιση των συνιστωσών ενός τυπικού συνελικτικού νευρωνικού δικτύου. Μπορούν να περιγραφούν με δύο ορολογίες. (Αριστερά)Το συνελικτικό δίκτυο μπορεί να ερμηνευθεί ως ένας μικρός αριθμός από συσχετιζόμενα πολύπλοκα επίπεδα, με κάθε ένα από αυτά να έχει πολλά "στάδια". Σε αυτήν την περίπτωση, η αντιστοίχιση μεταξύ των kernel tensors και των επιπέδων του δικτύου είναι ένα-προς-ένα. (Δεξιά)Το συνελικτικό δίκτυο μπορεί να ερμηνευθεί ως ένας μεγάλος αριθμός απλών επιπέδων. Κάθε βήμα της επεξεργασίας, θεωρείται από μόνο του ένα επίπεδο. Αυτό σημαίνει ότι δεν έχει κάθε επίπεδο παραμέτρους.

Μια συνάρτηση ομαδοποίησης, αντικαθιστά την έξοδο του δικτύου σε μια συγκεκριμένη τοποθεσία, με έναν στατιστικό μέσο όρο των γειτονικών εξόδων. Παραδείγματος χάριν, η διαδικασία **μεγίστης ομαδοποίησης**(max pooling) δηλώνει την μέγιστη έξοδο σε μια τετραγωνική γειτονιά. Άλλες διάσημες συναρτήσεις ομαδοποίησης περιέχουν τον μέσο όρο μιας τετραγωνικής γειτονιάς, την L^2 νόρμα μιας τετραγωνικής γειτονιάς, ή έναν σταθμισμένο μέσο όρο βασιζόμενο στην απόσταση από το κεντρικό εικονοστοιχείο.

Σε κάθε περίπτωση, η ομαδοποίηση βοηθά να κάνουμε την αναπαράσταση σχεδόν αμετάβλητη σε μικρές αλλαγές της εισόδου. Η αμεταβλητότητα για τοπικές μικρές αλλαγές μπορεί να είναι μια ιδιαίτερα χρήσιμη ιδιότητα αν μας ενδιαφέρει περισσότερο αν κάποιο χαρακτηριστικό είναι παρόν, παρά το που ακριβώς βρίσκεται. Λόγου χάριν, όταν ψάχνουμε να βρούμε αν μια εικόνα περιέχει ένα πρόσωπο, τότε δεν χρειάζεται να ξέρουμε ακριβώς την τοποθεσία των ματιών με ακρίβεια εικονοστοιχείου, αλλά αρκεί να γνωρίζουμε ότι υπάρχει ένα μάτι στο αριστερό μέρος του προσώπου και ένα στο δεξί.

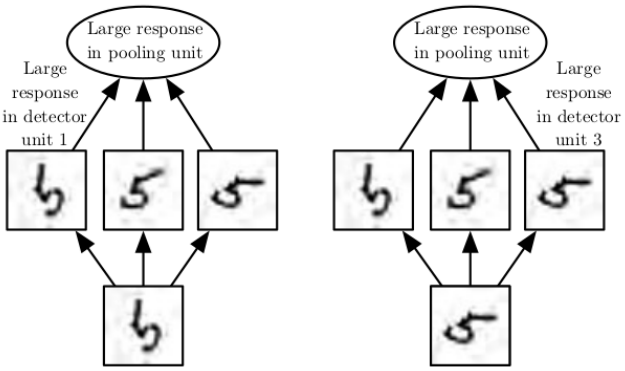


Σχήμα 3.12: Η διαδικασία της μεγίστης ομαδοποίησης εισάγει αμεταβλητότητα. (Πάνω)Μια όψη ενδιάμεσης εξόδου ενός συνελικτικού επιπέδου. Η κάτω γραμμή απεικονίζει εξόδους της μη γραμμικότητας. Η πάνω γραμμή δείχνει εξόδους της μεγίστης ομαδοποίησης, με βήμα απόφασης(Stride) του ενός εικονοστοιχείου μεταξύ των περιοχών ομαδοποίησης και μια περιοχή ομαδοποίησης πλάτους τριών εικονοστοιχείων. (Κάτω)Μια όψη του ίδιου δικτύου, αφότου η είσοδος έχει ολισθήσει δεξιά κατά ένα εικονοστοιχείο. Κάθε τιμή στην κάτω γραμμή έχει αλλάξει, αλλά μονο οι μισές τιμές στην πάνω γραμμή έχουν αλλάξει, επειδή οι μονάδες μεγίστης ομαδοποίησης, είναι ευαίσθητες μόνο ως προς την μέγιστη τιμή της γειτονιάς και όχι ως προς την ακριβή της τοποθεσία.

Η χρήση της ομαδοποίησης μπορεί να θεωρηθεί και ως η προσθήκη ενός εκ των προτέρων περιορισμού, ούτως ώστε η συνάρτηση που το επίπεδο μαθαίνει να είναι αμετάβλητη σε μικρές εισόδους. Όταν αυτή η υπόθεση είναι σωστή, τότε η στατιστική απόδοση του μοντέλου μπορεί να βελτιωθεί δραματικά.

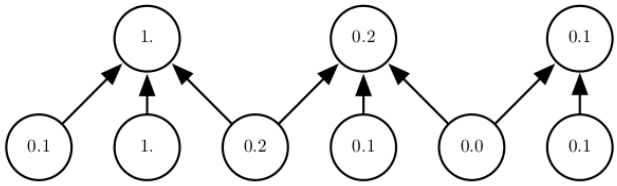
Η ομαδοποίηση σε χωρικές περιοχές προκαλεί αμεταβλητότητα σε μικρές αλλαγές, αλλά αν υλοποιήσουμε την ομαδοποίηση σε εξόδους ξεχωριστά παραμετροποιημένων συνελιξεων, τότε τα χαρακτηριστικά μπορούν να μάθουν σε

ποιους μετασχηματισμούς να γίνουν ανεξάρτητα(βλ. σχήμα 3.13).



Σχήμα 3.13: Παράδειγμα της μάθησης της αμεταβλητότητας. Μια μονάδα ομαδοποίησης η οποία ομαδοποιεί πολλαπλά χαρακτηριστικά τα οποία μαθαίνονται με ξεχωριστές παραμέτρους, μπορεί να μάθει να είναι αμετάβλητη σε μετασχηματισμούς της εισόδου. Εδώ απεικονίζεται πως ένα σύνολο από τρία φίλτρα και το πως μια μονάδα μεγίστης ομαδοποίησης μπορεί να μάθει να είναι αμετάβλητη στην περιστροφή. Και τα τρία φίλτρα προτίθενται να αναγνωρίσουν ένα χειρόγραφο 5. Κάθε φίλτρο προσπαθεί να ταιριάξει έναν ελαφρά διαφορετικό προσανατολισμό του 5. Όταν το 5 εμφανίζεται στην είσοδο, το αντίστοιχο φίλτρο θα το ταιριάξει και θα προκαλέσει μια μεγάλη ενεργοποιήση σε μια μονάδα ανίχνευσης. Τότε, η μονάδα μεγίστης ομαδοποίησης, έχει μια μεγάλη ενεργοποίηση, ανεξάρτητα από ποια μονάδα ανίχνευσης ενεργοποιήθηκε. Εδώ απεικονίζεται το πως το δίκτυο επεξεργάζεται δύο διαφορετικές εισόδους, οδηγώντας στην ενεργοποίηση δύο διαφορετικών μονάδων ανίχνευσης. Η επίδραση την μονάδας ομαδοποίησης είναι σχεδόν η ίδια και για τις δύο περιπτώσεις. Την αρχή αυτή εκμεταλλεύονται τα δίκτυα μεγίστης εξόδου και άλλα συνελικτικά δίκτυα.

Επειδή η ομαδοποίηση συνοψίζει τις τιμές σε μια μεγάλη γειτονιά, είναι πιθανό να χρησιμοποιηθούν λιγότερες μονάδες ομαδοποίησης σε σχέση με τις μονάδες ανίχνευσης, δηλώνοντας συνοπτικές στατιστικές για περιοχές οι οποίες χωρίζονται ανά k εικονοστοιχεία αντί για 1. Αυτό βελτιώνει την υπολογιστική απόδοση του δικτύου επειδή το επόμενο επίπεδο έχει k φορές λιγότερες εισόδους να επεξεργαστεί. Όταν ο αριθμός των παραμέτρων στο επόμενο επίπεδο, είναι συνάρτηση του μεγέθους της εισόδου του(π.χ. ένα πλήρως συνεκτικό επίπεδο), τότε αυτή η μείωση στο μέγεθος της εισόδου μπορεί να οδηγήσει σε βελτιωμένη στατιστική απόδοση και μειωμένες απαιτήσεις σε μνήμη.



Σχήμα 3.14: Εδώ, γίνεται χρήση της μεγίστης ομαδοποίησης με μέγεθος 3 και 1 βήμα απόφασης μεταξύ ομαδοποιήσεων του 2. Αυτό μειώνει το μέγεθος της αναπαράστασης με έναν παράγοντα του 2, το οποίο μειώνει το υπολογιστικό και στατιστικό κόστος για το επόμενο επίπεδο. Μπορεί να παρατηρηθεί ότι η πιο δεξιά περιοχή υποδειγματοληψίας έχει μικρότερο μέγεθος αλλά πρέπει να περιέχεται αν δεν θέλουμε να αγνοήσουμε κάποιες από τις μονάδες ανίχνευσης.

3.7.3 Παραλλαγές της βασικής πράξης της συνέλιξης

Όταν αναφερόμαστε στη συνέλιξη για την περίπτωση των νευρωνικών δικτύων, συνήθως δεν αναφερόμαστε ακριβώς στην κανονική διακριτή πράξη της συνέλιξης όπως είναι γνωστή στη μαθηματική βιβλιογραφία. Οι συναρτήσεις συνέλιξης που χρησιμοποιούνται στην πράξη διαφέρουν ελαφρώς.

Πρώτον, όταν αναφερόμαστε στη συνέλιξη στην περίπτωση των νευρωνικών δικτύων, συνήθως εννοούμε μια λειτουργία η οποία αποτελείται από πολλές παράλληλες εφαρμογές της συνέλιξης, επειδή η συνέλιξη με ένα απλό μονό kernel μπορεί να εξάγει μόνο ένα είδος χαρακτηριστικού, μολονότι μπορεί να το κάνει σε πολλές χωρικές τοποθεσίες. Αυτό που συνήθως θέλουμε, είναι κάθε επίπεδο του δικτύου να εξάγει πολλά είδη χαρακτηριστικών σε πολλές τοποθεσίες.

Επιπρόσθετα, η είσοδος συνήθως δεν είναι απλώς ένα πλέγμα με πραγματικές τιμές. Αντιθέτως, είναι ένα πλέγμα από διανυσματικές παρατηρήσεις. Λόγου χάριν, μια έγχρωμη εικόνα έχει κόκκινη, πράσινη και μπλε ένταση για κάθε εικονοστοιχείο. Σε ένα συνελικτικό δίκτυο με πολλά επίπεδα, η είσοδος στο δεύτερο επίπεδο είναι η έξοδος του πρώτου επιπέδου, το οποίο συνήθως περιέχει την έξοδο πολλών διαφορετικών συνελίξεων σε κάθε θέση. Επομένως όταν δουλεύουμε με εικόνες, συνήθως ερμηνεύουμε την είσοδο και έξοδο της συνέλιξης ως 3-D tensors, με έναν δείκτη σε διαφορετικά κανάλια και δύο δείκτες στις χωρικές συντεταγμένες του κάθε καναλιού. Οι υλοποιήσεις σε λογισμικό(βλ. εδάφια 6.1.1 και 6.1.2), βασίζονται σε λειτουργία batch, και έτσι, στην πραγματικότητα χρησιμοποιούν 4-D tensors, με τον τέταρτο άξονα να αφορά τα διαφορετικά δείγματα του batch, αλλά θα παραλείψουμε τον batch άξονα στην παρακάτω περιγραφή για λόγους απλότητας.

Επειδή τα συνελικτικά δίκτυα συνήθως χρησιμοποιούν πολυκαναλική συνέλιξη, οι γραμμικές λειτουργίες στις οποίες βασίζονται δεν εγγυώνται την αντιμεταθετικότητα, ακόμη και όταν χρησιμοποιείται το "αναποδογύρισμα" του kernel. Αυτές οι πολυκαναλικές λειτουργίες είναι αντιμεταθετικές, μόνο στην περίπτωση όπου ο αριθμός των καναλιών της εξόδου, είναι ίδιος με τον αριθμό των καναλιών της εισόδου, για κάθε λειτουργία.

Αν υποθέσουμε ότι έχουμε ένα 4-D kernel tensor $\mathbf{K}$ με το στοιχείο $K_{i,j,k,l}$ να δίνει την ισχύ της σύνδεσης μεταξύ μιας μονάδας στο κανάλι i της εξόδου και μιας μονάδας στο κανάλι j της εισόδου, με k γραμμές και l στήλες μεταξύ της μονάδας εξόδου και της μονάδας εισόδου. Υποθέτουμε ότι η είσοδος αποτελείται από παρατηρήσιμα δεδομένα $\mathbf{V}$ με το στοιχείο $V_{i,j,k}$ να δίνει την τιμή της μονάδας εισόδου για το κανάλι i, με γραμμή j και στήλη k. Υποθέτουμε ότι η έξοδος είναι $\mathbf{Z}$ και έχει ίδια μορφή με το $\mathbf{V}$. Αν το $\mathbf{Z}$ παράγεται από τη συνέλιξη του $\mathbf{K}$ με το $\mathbf{V}$ χωρίς "αναποδογύρισμα" του $\mathbf{K}$, τότε

$$Z_{i,j,k} = \sum_{l,m,n} V_{l,j+m-1,k+n-1} K_{i,l,m,n} \tag{3.39}$$

όπου το άθροισμα δια των l, m και n είναι για όλες τις τιμές για τις οποίες οι λειτουργίες δεικτοποίησης του tensor εντός του αθροίσματος είναι έγκυρες. Λόγω σημειογραφίας γραμμικής άλγεβρας, υπάρχει το -1 στην προηγούμενη σχέση, επειδή θεωρούμε ως πρώτη θέση του πίνακα το 1. Για γλώσσες όπως είναι η Python η οποία είναι zero-based, η παραπάνω έκφραση γίνεται απλούστερη.

Όπως προαναφέρθηκε στο προηγούμενο εδάφιο, λόγω υπολογιστικού κόστους ίσως χρειαστεί να αγνοήσουμε κάποιες θέσεις του kernel(με το κόστος τη μείωσης της ακρίβειας της εξαγωγής χαρακτηριστικών). Μπορεί κανείς να το συλλογιστεί αυτό ως υποδειγματοληψία της εξόδου της πλήρους συνάρτησης συνέλιξης. Δηλαδή, αν θέλουμε να δειγματοληπτήσουμε μόνο κάθε s εικονοστοιχεία σε κάθε κατεύθυνση της εξόδου, τότε μπορούμε να καθορίσουμε μια υποδειγματοληφθείσα συνάρτηση συνέλιξης c

$$Z_{i,j,k} = c(\mathbf{K}, \mathbf{V}, s)_{i,j,k} = \sum_{l,m,n} [V_{l,(j-1)\times s+m,(k-1)\times s+n} K_{i,l,m,n}] \tag{3.40}$$

Αναφερόμαστε στο s ως **βήμα απόφασης**(Stride) της υποδειγματοληφθείσας συνάρτησης. Είναι επίσης πιθανό να καθορισθούν ξεχωριστά βήματα απόφασης για κάθε κατεύθυνση της κίνησης.

Ένα σημαντικό χαρακτηριστικό για κάθε υλοποίηση συνελικτικού δικτύου, είναι η προσθήκη μηδενικών(Zero Padding) στην είσοδο $\mathbf{V}$ ώστε να την κάνουμε μεγαλύτερη. Χωρίς αυτό το χαρακτηριστικό, το μέγεθος της αναπαράστασης μειώνεται κατά ένα εικονοστοιχείο λιγότερο από το μέγεθος του kernel για κάθε επίπεδο. Η προσθήκη μηδενικών στην είσοδο, μας επιτρέπει να ελέγξουμε το μέγεθος του kernel και το μέγεθος της εξόδου ανεξάρτητα. Υπάρχουν τρεις διαφορετικές περιπτώσεις προσθήκης μηδενικών:

- **Έγκυρη συνέλιξη(Valid Convolution)**
  Σε αυτήν την περίπτωση, δεν χρησιμοποιείται κάθολου η προσθήκη μηδενικών, και το kernel συνέλιξης επιτρέπεται να επισκεφθεί μόνο στις θέσεις όπου ολόκληρο το kernel περιέχεται εντός της εικόνας. Δηλαδή όλα εικονοστοιχεία στην έξοδο αποτελούν μια συνάρτηση του ίδιου αριθμού εικονοστοιχείων της εισόδου. Όμως το μέγεθος της εξόδου μειώνεται σε κάθε επίπεδο. Αν λόγου χάριν η εικόνα της εισόδου έχει πλάτος m και το kernel πλάτος k, τότε η έξοδος θα είναι $m - k + 1$. Η μείωση του μεγέθους της εξόδου μπορεί να είναι αρκετά μεγάλη για μεγάλα kernels.

- **Ταυτόσιμη συνέλιξη(Same Convolution)**
  Σε αυτή την περίπτωση, προστίθενται μηδενικά τόσο, ώστε να διατηρηθεί το μέγεθος της εξόδου ίσο με το μέγεθος της εισόδου. Το δίκτυο μπορεί να περιέχει όσα συνελικτικά επίπεδα γίνεται, ανάλογα με το διαθέσιμο hardware, αφού η πράξη της συνέλιξης δεν διαφοροποιεί τις αρχιτεκτονικές πιθανότητες για το επόμενο επίπεδο. Τα εικονοστοιχεία εισόδου κοντά στο σύνορο όμως, επηρεάζουν λιγότερα εικονοστοιχεία εξόδου από ότι τα εικονοστοιχεία κοντά στο κέντρο. Αυτό μπορεί να κάνει τα εικονοστοιχεία του συνόρου να μην αναπαρίστανται πλήρως στο μοντέλο.

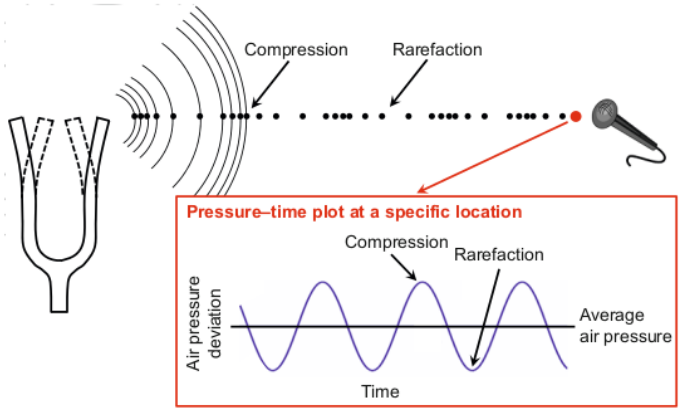
- **Πλήρης συνέλιξη(Full Convolution)**
  Η παραπάνω περίπτωση της μη πλήρης αναπαράστασης των εικονοστοιχείων στο σύνορο, οδηγεί στην ακραία περίπτωση της πλήρης συνέλιξης, όπου προστίθενται αρκετά μηδενικά ώστε κάθε εικονοστοιχείο να είναι επισκέψιμο k φορές σε κάθε κατεύθυνση, οδηγώντας σε μια εικόνα εξόδου πλάτους $m + k - 1$. Στην περίπτωση αυτή, τα εικονοστοιχεία της εξόδου κοντά στο σύνορο αποτελούν μια συνάρτηση από λιγότερα εικονοστοιχεία από ότι τα εικονοστοιχεία εξόδου κοντά στο κέντρο. Αυτό μπορεί να κάνει δύσκολη τη μάθηση ενός απλού kernel το οποίο αποδίδει καλά σε όλες τις θέσεις του συνελικτικού χάρτη χαρακτηριστικών. Συνήθως η βέλτιστη ποσότητα προσθήκης μηδενικών(όσον αφορά την απόδοση ταξινόμησης στο σύνολο δοκιμής), εμπίπτει κάπου ανάμεσα στις περιπτώσεις της έγκυρης και ταυτόσιμης συνέλιξης.

Κεφάλαιο 4

Βασικές μέθοδοι ψηφιακής επεξεργασίας σήματος του ήχου

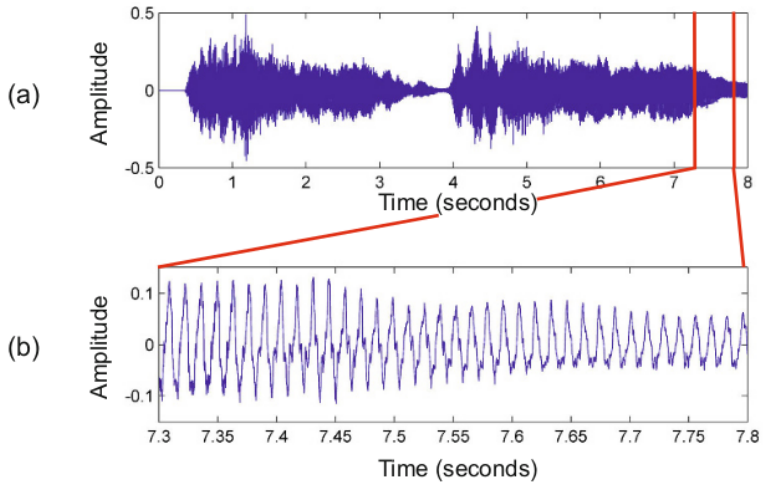
4.1 Κύματα και κυματομορφές

Ο ήχος παράγεται από ένα αντικείμενο που δονείται όπως οι φωνητικές χορδές ενός τραγουδιστή, οι χορδές μιας κιθάρας και το διάφραγμα ενός μεταλλικού τυμπάνου. Αυτές οι δονήσεις προκαλούν μετατοπίσεις και ταλαντώσεις στα μόρια του αέρα, οδηγώντας σε τοπικές περιοχές συμπίεσης και αραίωσης. Η εναλλασσόμενη πίεση ταξιδεύει μέσω του αέρα ως κύμα, από την πηγή της εώς τον αποδέκτη. Μπορεί τότε να ληφθεί ως ήχος από έναν άνθρωπο ή να μετατραπεί σε ηλεκτρικό σήμα από ένα μικρόφωνο [30].



Σχήμα 4.1: Διαπασών που δονείται. Η ταλάντωση της πίεσης διαδίδεται ως διαμήκες κύμα μέσω του αέρα. Η κυματομορφή δείχνει την απόκλιση της πίεσης του αέρα, από την μέση πίεση ως προς το χρόνο, σε ένα συγκεκριμένο σημείο του χώρου(στο μικρόφωνο). Πηγή: [30]

Γραφικά, η αλλαγή στην πίεση του αέρα σε μια συγκεκριμένη τοποθεσία, μπορεί να αναπαρασταθεί από ένα διάγραμμα πίεσης-χρόνου, ή αλλιώς από την κυματομορφή του ήχου.



Σχήμα 4.2: (a) Κυματομορφή των πρώτων 8 δευτερολέπτων από μια ηχογράφηση της πέμπτης συμφωνίας του Μπετόβεν. (b) Μεγέθυνση της περιοχής μεταξύ των 7.3 και 7.8 δευτερολέπτων. Πηγή: [30]

Αν τα σημεία υψηλής και χαμηλής πίεσης αέρα επαναλαμβάνονται εναλλασσό-

μενα και τακτικά, τότε η κυματομορφή καλείται περιοδική. Η περίοδος του κύματος καθορίζεται ως ο χρόνος ολοκλήρωσης ενός κύκλου. Η συχνότητα, μετράται σε Hertz(Hz) και είναι αντίστροφη της περιόδου. Όσο υψηλότερη είναι η συχνότητα ενός ημιτονοειδούς κύματος, τόσο υψηλότερα ακούγεται. Η ανθρώπινη ακοή έχει εύρος συχνοτήτων μεταξύ περίπου 20Hz και 20kHz. Το ημίτονο μπορεί να θεωρηθεί ως το πρωτότυπο της ακουστικής μιας μουσικής νότας. Μερικές φορές ο ήχος του ημιτονοειδούς κύματος καλείται αρμονικός ήχος ή τόνος.

4.2 Η βασική ιδέα της ανάλυσης Fourier

Η βασική ιδέα της ανάλυσης Fourier είναι να συγκρίνουμε το σήμα με ημιτονοειδής κυματομορφές διαφόρων συχνοτήτων $\omega \in \mathbb{R}$ (μετρημένες σε Hz). Κάθε τέτοιο ημίτονο ή τόνος μπορεί να θεωρηθεί ως μια πρωτότυπη ταλάντωση. Ως αποτέλεσμα, λαμβάνουμε για κάθε παράμετρο συχνότητας $\omega \in \mathbb{R}$ έναν συντελεστή πλάτους $d_\omega \in \mathbb{R}_{\geq 0}$, μαζί με έναν συντελεστή φάσης $\phi_\omega \in \mathbb{R}$. Στην περίπτωση που ο συντελεστής d_ω είναι μεγάλος, υπάρχει υψηλή ομοιότητα μεταξύ του σήματος και της ημιτόνου συχνότητας ω, και το σήμα περιέχει περιοδική ταλάντωση σε αυτήν την συχνότητα. Αν το d_ω είναι μικρό, το σήμα δεν περιέχει περιοδική συνιστώσα σε αυτήν την συχνότητα.

Ορίζουμε την συνημιτονοειδή συνάρτηση $\cos_{\omega,\phi} : \mathbb{R} \to \mathbb{R}$ ως εξής

$$\cos_{\omega,\phi}(t) = \sqrt{2}\cos(2\pi(\omega t - \phi)) \tag{4.1}$$

Για μια συγκεκριμένη συχνότητα $\omega \in \mathbb{R}$, ορίζουμε τους συντελεστές πλάτους και φάσης αντίστοιχα

$$d_\omega = \max_{\phi \in [0,1)} \left(\int_{t \in \mathbb{R}} f(t) \cos_{\omega,\phi}(t) dt \right) \tag{4.2}$$

$$\phi_\omega = \underset{\phi \in [0,1)}{argmax} \left(\int_{t \in \mathbb{R}} f(t) \cos_{\omega,\phi(t)} \, dt \right) \tag{4.3}$$

Ορίζουμε επιπρόσθετα τον μιγαδικό συντελεστή

$$c_\omega = \frac{d_\omega}{\sqrt{2}} \exp(2\pi j(-\phi_\omega)) \tag{4.4}$$

Έστω το πραγματικό σήμα $f : \mathbb{R} \to \mathbb{R}$. Για κάθε συχνότητα $\omega \in \mathbb{R}$, λαμβάνουμε έναν μιγαδικό συντελεστή $c_\omega \in \mathbb{C}$. Αυτή η συλλογή από συντελεστές κωδικοποιείται από μια μιγαδική συνάρτηση $\hat{f} : \mathbb{R} \to \mathbb{C}$, η οποία αντιστοιχεί σε κάθε συχνοτική παράμετρο τον συντελεστή c_ω

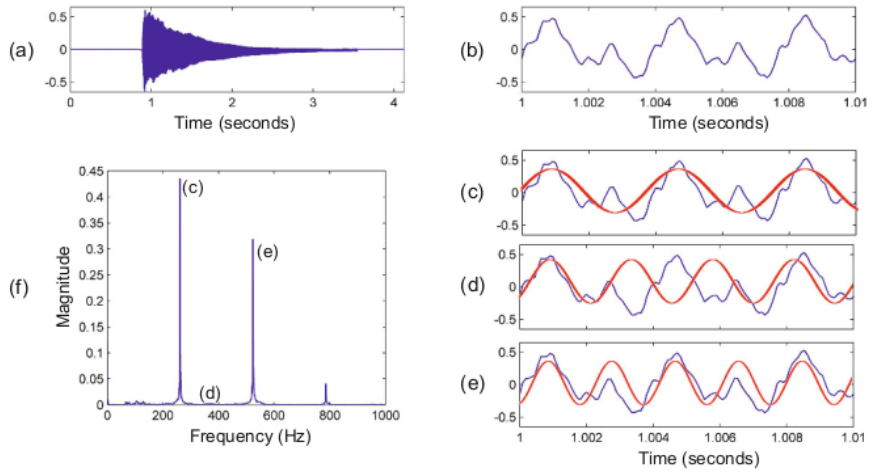
$$\hat{f}(\omega) = c_\omega \tag{4.5}$$

Η συνάρτηση $\hat{f}$ αποτελεί τον μετασχηματισμό Fourier του πραγματικού σήματος f, και οι τιμές $\hat{f}(\omega) = c_\omega$ καλούνται συντελεστές Fourier. Ο μετασχηματισμός

Fourier μπορεί να υπολογιστεί από τις εξής σχέσεις

$$\hat{f}(\omega) = \int_{t\in\mathbb{R}} f(t)\exp(-2\pi j\omega t)dt \tag{4.6}$$

$$= \int_{t\in\mathbb{R}} f(t)\cos(-2\pi\omega t)dt + j\int_{t\in\mathbb{R}} f(t)\sin(-2\pi\omega t)dt \tag{4.7}$$



Σχήμα 4.3: (a) Κυματομορφή μιας νότας $C4(261.6\text{Hz})$ σε πιάνο. (b) Εστίαση σε τμήμα των 10ms που ξεκινά σε χρόνο $t = 1\text{sec}$. (c-e) Σύγκριση της κυματομορφής με διάφορα ημίτονα διαφορετικών συχνοτήτων ω. (f) Συντελεστές πλάτους d_ω ως προς την συχνότητα ω. Πηγή: [30]

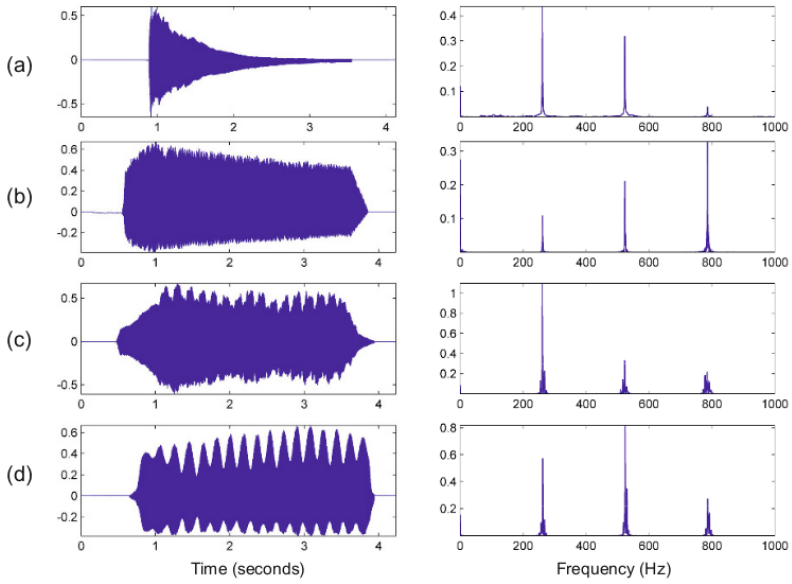
Ο μετασχηματισμός Fourier επομένως, διαχωρίζει το σήμα σε συνιστώσες συχνότητας. Μια πολύ σημαντική ιδιότητα του είναι ότι το πραγματικό σήμα, μπορεί να ανακατασκευαστεί από τους συντελεστές d_ω και ϕ_ω. Πρόκειται ουσιαστικά, για την υπέρθεση όλων των ημιτονοειδών κυματομορφών, όλων των πιθανών συχνοτήτων, με τους αντίστοιχους συντελεστές d_ω και ϕ_ω. Αυτή η σταθμισμένη υπέρθεση καλείται και αναπαράσταση Fourier του πρωτότυπου σήματος. Η ανακατασκευή του σήματος δίνεται από τις εξής σχέσεις

$$f(t) = \int_{\omega\in\mathbb{R}\geq 0} d_\omega\sqrt{2}\cos(2\pi(\omega t - \phi_\omega))d\omega \tag{4.8}$$

$$= \int_{\omega\in\mathbb{R}} c_\omega\exp(2\pi j\omega t)d\omega \tag{4.9}$$

Το πρωτότυπο σήμα και ο μετασχηματισμός Fourier αυτού, περιέχουν την ίδια ποσότητα πληροφορίας, αν και η αναπαράστασή της γίνεται με διαφορετικούς

τρόπους. Το σήμα αναπαριστά την πληροφορία ως προς τον χρόνο, ενώ ο μετασχηματισμός Fourier ως προς την συχνότητα.



Σχήμα 4.4: Κυματομορφές και πλάτος του μετασχηματισμού Fourier μιας νότας $C4(261.6\text{Hz})$ από τέσσερα διαφορετικά όργανα. (a) Πιάνο. (b) Τρομπέτα. (c)Βιολί. (d) Φλάουτο. Πηγή: [30]

4.3 Δειγματοληψία και κβαντισμός

Τα αναλογικά σήματα έχουν ένα συνεχόμενο εύρος τιμών σε χρόνο και σε πλάτος, το οποίο γενικώς, οδηγεί σε άπειρο αριθμό τιμών. Ως γνωστόν, ένας υπολογιστής μπορεί να αποθηκεύει και να επεξεργάζεται έναν πεπερασμένο αριθμό από τιμές, επομένως η κυματομορφή πρέπει να μετατραπεί σε κάποια διακριτή αναπαράσταση(Discrete Representation). Η διαδικασία αυτή καλείται ψηφιοποίηση(Digitization). Η πιο κοινή προσέγγιση για την ψηφιοποίηση ηχητικών σημάτων, αποτελείται από την δειγματοληψία και τον κβαντισμό.

Στην επεξεργασία σήματος, ο όρος δειγματοληψία, παραπέμπει στην διαδικασία μετατροπής του συνεχούς χρόνου σήματος(Continuous-Time[CT] Signal) σε διακριτού χρόνου σήμα(Discrete-Time[DT] Signal), το οποίο καθορίζεται από ένα διακριτό υποσύνολο $\mathbb{I}$ του άξονα του χρόνου. Έτσι ένα DT-σήμα ορίζεται ως συνάρτηση $\mathbf{x} : \mathbb{I} \to \mathbb{R}$, όπου το πεδίο $\mathbb{I}$ αντιστοιχεί σε σημεία του χρόνου. Το DT-σήμα μπορεί να επεκταθεί από το πεδίο $\mathbb{I}$, στο πεδίο $\mathbb{Z}$, θέτοντας όλες τις τιμές μηδέν για σημεία $\mathbb{Z}\backslash\mathbb{I}$, και άρα θεωρούμε $\mathbb{I} = \mathbb{Z}$. Η πιο κοινή διαδικασία δειγματοληψίας είναι ο μετασχηματισμός ενός CT-σήματος $f : \mathbb{R} \to \mathbb{R}$ σε ένα DT-σήμα $\mathbf{x} : \mathbb{Z} \to \mathbb{R}$, και είναι γνωστή ως ισαπέχουσα δειγματοληψία(Equidistant

Sampling). Για έναν πραγματικό αριθμό $T > 0$, το DT-σήμα $\mathbf{x}$ λαμβάνεται θέτοντας

$$x(n) = f(n \cdot T) \tag{4.10}$$

για $n \in \mathbb{Z}$. Η τιμή $x(n)$ καλείται δείγμα και έχει ληφθεί σε χρόνο $t = n \cdot T$ του πρωτότυπου αναλογικού σήματος f. Το αντίστροφο της περιόδου δειγματοληψίας T, $F_s = 1/T$ καλείται ρυθμός δειγματοληψίας(σε Hz).

Τυπικοί ρυθμοί δειγματοληψίας είναι 8kHz για τηλεφωνία, 32kHz για ψηφιακό ραδιόφωνο, 44.1kHz για ηχογραφήσεις CD, και 48kHz μέχρι 96kHz για επαγγελματικές ηχογραφήσεις σε στούντιο. Γενικώς, η δειγματοληψία εισάγει κάποια απώλεια, λόγω του ότι πληροφορία χάνεται κατά την διαδικασία αυτή. Για να ανακατασκευαστεί πλήρως το πρωτότυπο αναλογικό σήμα f πρέπει να ισχύει το θεώρημα δειγματοληψίας του Nyquist [20]

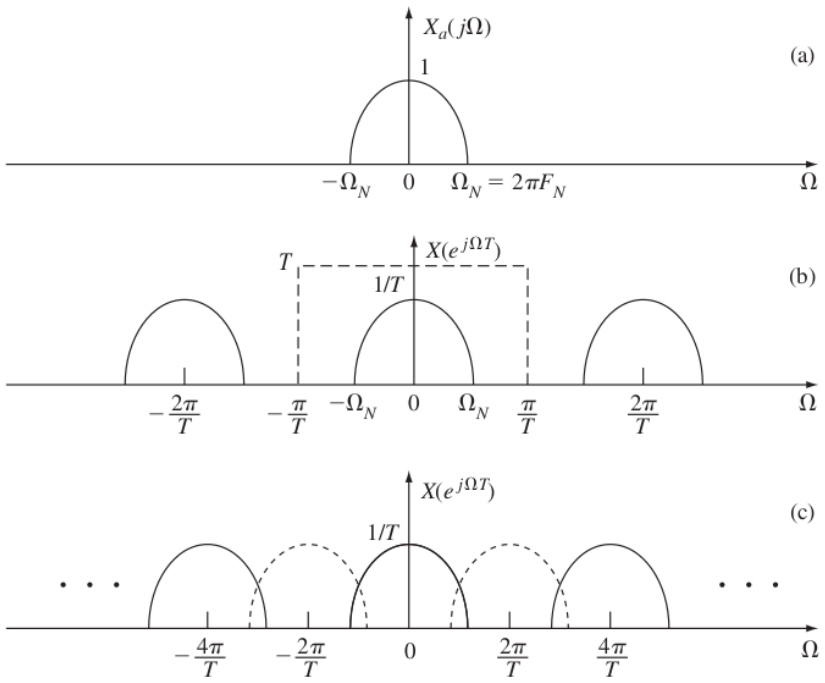
> *Θεώρημα δειγματοληψίας:* Αν ένα σήμα $f(t)$ έχει ένα περιορισμένου εύρους ζώνης(Bandlimited) μετασχηματισμό Fourier $X_a(j\Omega)$, τέτοιο ώστε $X_a(j\Omega) = 0$ για $\Omega \geq 2\pi F_N$, τότε η $f(t)$ μπορεί μοναδικά να ανακατασκευαστεί, από ισαπέχοντα δείγματα $f(n \cdot T)$, $-\infty < n < \infty$, αν ο ρυθμός δειγματοληψίας $F_s = 1/T$, ικανοποιεί την συνθήκη $F_s \geq 2F_N$

όπου

i) Ω η αναλογική συχνότητα

ii) $\omega = \Omega T$ η κανονικοποιημένη συχνότητα

iii) F_N η συχνότητα Nyquist



Σχήμα 4.5: Αναπαράσταση της δειγματοληψίας στο πεδίο της συχνότητας. (a) Μετασχηματισμός Fourier περιορισμένου εύρους ζώνης του σήματος f. (b) DTFT(βλ. εδάφιο (4.4)) των δειγμάτων $x(n) = f(n \cdot T)$ όταν $Fs = 1/T > 2F_N$(περίπτωση υπερδειγματοληψίας). (c) Φαινόμενο aliasing, DTFT των δειγμάτων $x(n) = f(n \cdot T)$ όταν $Fs = 1/T < 2F_N$(υποδειγματοληψία). Πηγή: [20]

Τα παραπάνω αποτελούν ένα πρώτο βήμα για την μετατροπή του σήματος από αναλογικό σε ψηφιακό. Εν συνεχεία, πρέπει να αντικατασταθούν οι συνεχείς τιμές των πιθανών πλατών από ένα διακριτό εύρος πιθανών τιμών($\mathbb{G} \subset \mathbb{R}$). Αυτή η διαδικασία είναι γνωστή ως κβαντισμός. Η κβάντιση μπορεί να μοντελοποιηθεί ως μια συνάρτηση που ονομάζεται κβαντιστής, $Q : \mathbb{R} \to \mathbb{G}$, και αντιστοιχεί σε κάθε τιμή πλάτους $\alpha \in \mathbb{R}$, μια τιμή $Q(\alpha) \in \mathbb{G}$. Πολλοί κβαντιστές απλώς, περικόπτουν και στρογγυλοποιούν την αναλογική τιμή σε κάποια μονάδα ακριβείας. Παραδείγματος χάριν, ένας τυπικός ομοιόμορφος κβαντιστής, με βήμα κβάντισης ίσο με κάποια τιμή Δ, μπορεί να ορισθεί ως

$$Q(a) = \text{sign}(a) \cdot \Delta \cdot \left\lfloor \frac{|\alpha|}{\Delta} + \frac{1}{2} \right\rfloor \tag{4.11}$$

4.4 Μετασχηματισμός Fourier για σήματα διακριτού χρόνου(DTFT)

Ορίζουμε την ενέργεια $E(\mathbf{x})$ ενός σήματος $\mathbf{x} \in \mathbb{C}^{\mathbb{Z}}$ ως

$$E(\mathbf{x}) = \sum_{n\in\mathbb{Z}} |x(n)|^2 \tag{4.12}$$

Ο χώρος $\ell^2(\mathbb{Z}) \subset \mathbb{C}^{\mathbb{Z}}$, ορίζεται ως το σύνολο όλων των σημάτων που έχουν πεπερασμένη ενέργεια:

$$\ell^2(\mathbb{Z}) = \{\mathbf{x} : \mathbb{Z} \to \mathbb{C} | E(\mathbf{x}) < \infty\} \tag{4.13}$$

Επιπλέον ορίζουμε την εκθετική συνάρτηση

$$\exp_\omega : \mathbb{R} \to \mathbb{C},\ \exp_\omega(t) = \exp(2\pi j\omega t) \tag{4.14}$$

Έστω $\mathbf{x} \in \ell^2(\mathbb{Z})$ είναι ένα αυθαίρετο DT-σήμα πεπερασμένης ενέργειας. Τότε η Fourier αναπαράσταση του $\mathbf{x}$ είναι

$$x(n) = \int_{\omega\in[0,1)} c_\omega \exp_\omega(n) d\omega = \int_{\omega\in[0,1)} c_\omega \exp(2\pi j\omega n) d\omega \tag{4.15}$$

για $n \in \mathbb{Z}$. Επιπρόσθετα, οι συντελεστές c_ω δίνονται από την εξαρτώμενη από την συχνότητα συνάρτηση $\hat{x} : [0,1) \to \mathbb{C}$

$$c_\omega = \hat{x}(\omega) = \sum_{n\in\mathbb{Z}} x(n)\overline{\exp_\omega(n)} = \sum_{n\in\mathbb{Z}} \exp(-2\pi j\omega n) \tag{4.16}$$

το οποίο καλείται μετασχηματισμός Fourier του $\mathbf{x}$.

4.5 Διακριτός μετασχηματισμός Fourier(DFT)

Ο υπολογισμός του μετασχηματισμού Fourier των σημάτων, περιλαμβάνει τον υπολογισμό ολοκληρωμάτων ή άπειρων αθροισμάτων τα οποία, γενικώς, είναι υπολογιστικά ανέφικτα. Επιπλέον, ο μετασχηματισμός Fourier υπολογίζεται μόνο για πεπερασμένο αριθμό συχνοτήτων. Επομένως, η επιλογή των άπειρων αθροισμάτων και των συντελεστών Fourier, γίνεται έτσι ώστε ληφθεί ένας γραμμικός μετασχηματισμός, γνωστός και ως διακριτός μετασχηματισμός Fourier. Έστω ένα DT-σήμα $\mathbf{x} \in \ell^2(\mathbb{Z})$. Υποθέτουμε ότι η ενέργεια του $\mathbf{x}$ είναι συγκεντρωμένη στο διάστημα $[0 : N-1]$, δηλαδή $x(n) \approx 0$ για $n \in \mathbb{Z}\backslash[0 : N-1]$. Τότε λαμβάνουμε από την (4.16)

$$\hat{x}(\omega) = \sum_{n\in\mathbb{Z}} x(n)\overline{\exp_\omega(n)} \approx \sum_{n=0}^{N-1} x(n)\overline{\exp_\omega(n)} \tag{4.17}$$

για μια παράμετρο συχνότητας ω. Για το $\hat{x}$ μόνο οι συχνότητες $\omega \in [0,1)$ πρέπει να ληφθούν υπόψη. Στην πράξη, ο υπολογισμός του μετασχηματισμού Fourier, γίνεται μόνο για πεπερασμένο υποσύνολο συχνοτήτων. Συγκεκριμένα, για έναν αριθμό $K \in \mathbb{N}$, θεωρούμε τις συχνότητες $\omega = k/K$ για $k \in [0 : K-1]$, το οποίο αντιστοιχεί σε $1/K$ δειγματοληψία του χώρου συχνοτήτων $[0,1)$. Αν και ο αριθμός των N σημείων στο χρόνο και K των συχνοτήτων δεν σχετίζεται καθόλου, είναι βολικό να θεωρήσουμε $N = K$. Επιπλέον, έστω $\mathbf{x} \in \ell^2(\mathbb{Z})$, ένα σήμα που είναι μηδενικό έξω από το διάστημα $[0 : N-1]$, έτσι ώστε να επέρχεται ισότητα με την (4.17). Τέτοια DT-σήματα καλούνται επίσης και πεπερασμένου μήκους σήματα, όπου N το μήκος του σήματος. Κάθε τέτοιο σήμα $\mathbf{x}$, μπορεί να ταυτοποιηθεί ως ένα διάνυσμα $\mathbf{x} = (x(0), x(1), ..., x(N-1))^T \in \mathbb{C}^N$, όπου θεωρούμε $\mathbb{C}^N \subset \ell^2(\mathbb{Z})$. Δεν χρειάζονται όλες οι συχνότητες $\omega \in [0,1)$ για να χαρακτηρισθεί ένα σήμα μήκους N. Μόνο οι συχνότητες k/N για $k \in [0 : N-1]$ είναι αρκετές για την αναπαράσταση τέτοιων σημάτων. Επομένως, καθορίζουμε ένα διάνυσμα $\mathbf{u_k} \in \mathbb{C}^N$ για κάθε $k \in [0 : N-1]$, θέτοντας

$$u_k(n) = \exp_{k/N}(n) = \exp(2\pi kn/N) \tag{4.18}$$

για $n \in [0 : N-1]$. Δηλαδή, το διάνυσμα $\mathbf{u_k}$ αποτελείται από τα πρώτα N δείγματα της εκθετικής συνάρτησης $\exp_{k/N}$. Τότε η (4.17) μπορεί να εκφραστεί ως

$$\hat{x}(k/N) = \sum_{n=0}^{N-1} x(n)\overline{\exp_{k/N}(n)} = \mathbf{x}^T\overline{\mathbf{u_k}} = \langle \mathbf{x}|\mathbf{u_k}\rangle \tag{4.19}$$

Από την (4.19), οι συντελεστές Fourier, ενός σήματος $\mathbf{x}$ πεπερασμένου μήκους N, δίνονται από

$$X(k) = \langle \mathbf{x}|\mathbf{u_k}\rangle = \sum_{n=0}^{N-1} x(n)\exp(-\pi jkn/N) \tag{4.20}$$

για $k \in [0 : N-1]$. Έστω $\mathbf{X} = (X(0), X(1), ..., X(N-1))^T \in \mathbb{C}^N$ το διάνυσμα των συντελεστών Fourier. Εξ' ορισμού, ο διακριτός μετασχηματισμός Fourier, είναι η αντιστοίχιση $\mathbb{C}^N \to \mathbb{C}^N$, που αντιστοιχεί το διάνυσμα εισόδου $\mathbf{x}$ στο διάνυσμα εξόδου $\mathbf{X}$. Από την (4.20) προκύπτει ότι αυτή η αντιστοίχιση είναι γραμμική, και μπορεί να περιγραφεί από τον $(N \times N)$ πίνακα $\mathbf{DFT_N}$, ο οποίος δίνεται από

$$DFT_N(n,k) = \exp(-2\pi jkn/N) \tag{4.21}$$

Μπορεί να παρατηρηθεί, ότι υπάρχουν πολλοί συσχετισμοί μεταξύ των αριθμών $\exp(2\pi kn/N)$ για $k, n \in [0, N-1)$. Χρησιμοποιώντας τα εξής

i) $\rho = \exp(2\pi j/N)$

ii) $\rho^{kn} = \exp(2\pi jkn/N)$

iii) $\omega = \bar{\rho} = \exp(-2\pi j/N)$

λαμβάνουμε $DFT_N(n,k) = \omega^{kn}$. Αυτό δίνει τον γνωστό πίνακα

$$DFT_N = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \omega & \omega^2 & \dots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \dots & \omega^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \dots & \omega^{(N-1)(N-1)} \end{pmatrix} \tag{4.22}$$

Προφανώς ο $\mathbf{DFT_N}$ είναι συμμετρικός πίνακας. Οι στήλες του δίνονται από $\overline{\mathbf{u_k}}$ και οι γραμμές του από $\overline{\mathbf{u_k}}^{\mathbf{T}}$. Άρα, σε τελική ανάλυση, ο μετασχηματισμός Fourier $\hat{x}$ ενός DT-σήματος $\mathbf{x}$ πεπερασμένου μήκους N, μπορεί να υπολογιστεί για συχνότητες $\omega = k/N$, $k \in [0 : N-1]$ από το γινόμενο πίνακα-διανύσματος $\mathbf{X} = DFT_N \cdot \mathbf{x}$

4.6 Γρήγορος μετασχηματισμός Fourier(FFT)

Ο υπολογισμός του γινομένου πίνακα-διανύσματος $\mathbf{X} = DFT_N \cdot \mathbf{x}$, απαιτεί $O(N^2)$ πολλαπλασιασμούς και προσθέσεις, οι οποίοι είναι πάρα πολλοι για τις περισσότερες εφαρμογές, πράγμα που καθιστά τον υπολογισμό του DFT υπολογιστικά πολύπλοκο ή και ανέφικτο. Εδώ έρχεται να δώσει λύση ο γρήγορος μετασχηματισμός Fourier. Η βασική ιδέα βασίζεται σε μια παραγοντοποίηση του DFT πίνακα σε ένα γινόμενο από $O(logN)$ αραιούς πίνακες, κάθε ένας από τους οποίους υπολογίζεται με $O(N)$ πράξεις. Επομένως ο αλγόριθμος FFT απαιτεί μόνο $O(NlogN)$ πολλαπλασιασμούς και προσθέσεις. Η αρχική εφεύρεση του έγινε από τον Gauss, αλλά ανακαλύφθηκε εκ νέου από τους Cooley και Tukey το 1965 [3] και πλέον χρησιμοποιείται από δισεκατομμύρια τηλεπικοινωνιακές και άλλες συσκευές.

4.7 Βραχυχρόνιος μετασχηματισμός Fourier(STFT)

Η ενέργεια $E(f)$ ενός μετρούμενου σήματος $f \in \mathbb{C}^{\mathbb{R}}$ ορίζεται ως

$$E(f) = \int_{t \in \mathbb{R}} |f(t)|^2 dt \tag{4.23}$$

και ο χώρος $L^2(\mathbb{R}) \subset \mathbb{C}^{\mathbb{R}}$, ορίζεται ως το σύνολο όλων των σημάτων πεπερασμένης ενέργειας

$$L^2(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{C} | f \ \mu\epsilon\tau\rho o\acute{\upsilon}\mu\epsilon\nu o \ \kappa\alpha\iota \ E(f) < \infty\} \tag{4.24}$$

Ο μετασχηματισμός Fourier $\hat{f}$ ενός σήματος $f \in L^2(\mathbb{R})$, περιγράφει το συχνοτικό περιεχόμενο του σήματος. Συγκρίνοντας το σήμα με μια περιοδική εκθετική συνάρτηση $t \to \exp(2\pi j\omega t)$, έχει ως αποτέλεσμα έναν συντελεστή $\hat{f}(\omega)$ που εξάγει την συνολική ένταση των ταλαντώσεων, για ω Hz, του σήματος. Όμως, λόγω της φυσικής μη τοπικότητας της συνάρτησης ανάλυσης, η συχνοτική πληροφορία πάντοτε εξάγεται ως μέσος όρος σε ολόκληρο το πεδίο του χρόνου. Έτσι,

ξαφνικές αλλαγές και τοπικές διακυμάνσεις του σήματος, όπως η αρχή και το τέλος γεγονότων, δεν μπορούν να ανιχνευθούν καλά από τον μετασχηματισμό Fourier. Τοπικά φαινόμενα του σήματος γίνονται ολικά φαινόμενα με τον μετασχηματισμό Fourier. Αντίθετα, μικρές αλλαγές στη φάση του μετασχηματισμού Fourier, μπορεί να έχουν σημαντικές επιδράσεις στο πεδίο του χρόνου. Ο Dennis Gabor, παρουσίασε το 1946, τον βραχυχρόνιο μετασχηματισμό Fourier(Short-Time Fourier Transform). Αυτός ο μετασχημάτισμός, αποτελεί έναν συμβιβασμό, μεταξύ μιας αναπαράστασης χρόνου και συχνότητας, καθορίζοντας την ημιτονοειδή συχνότητα και το περιεχόμενο της φάσης σε τοπικές περιοχές του σήματος, καθώς αυτό μεταβάλλεται με τον χρόνο. Επομένως, ο STFT δεν λέει μόνο ποιες συχνότητες περιέχονται στο σήμα, αλλά και σε ποιά σημεία του χρόνου, δηλαδή σε ποιο διάστημα του χρόνου αυτές εμφανίζονται.

Δοθέντος ενός σήματος f, θέλουμε να βρούμε έναν μετασχηματισμό που να εξάγει το συχνοτικό περιεχόμενο του f, σε μια γειτονιά κάθε σημείου σε χρόνο t. Η βασική ιδέα είναι να θεωρήσουμε μόνο ένα μικρό τμήμα του σήματος γύρω από ένα σημείο t, όπου η επιρροή ενός σημείου εντός του τμήματος, μειώνεται με την αύξηση της απόστασης από το t. Μαθηματικά αυτό μοντελοποιείται, πολλαπλασιάζοντας με μια συνάρτηση παραθύρου, η οποία μπορεί να θεωρηθεί ως μια συνάρτηση βάρους, η οποία λειτουργεί τοπικά γύρω από το t. Αντί να χρησιμποιηθεί μια διαφορετική συνάρτηση παραθύρου σε κάθε σημείο t, χρησιμοποιείται μία συνάρτηση παραθύρου που λειτουργεί τοπικά γύρω από το σημείο $t = 0$. Αυτή η συνάρτηση ολισθαίνει στο χρόνο. Αν $f \in L^2(\mathbb{R})$ είναι ένα σήμα και $g : \mathbb{R} \to \mathbb{R}$ μία τέτοια συνάρτηση παραθύρου, τότε η συνάρτηση $f_{g,t}$ γύρω από το σημείο t, ορίζεται ως

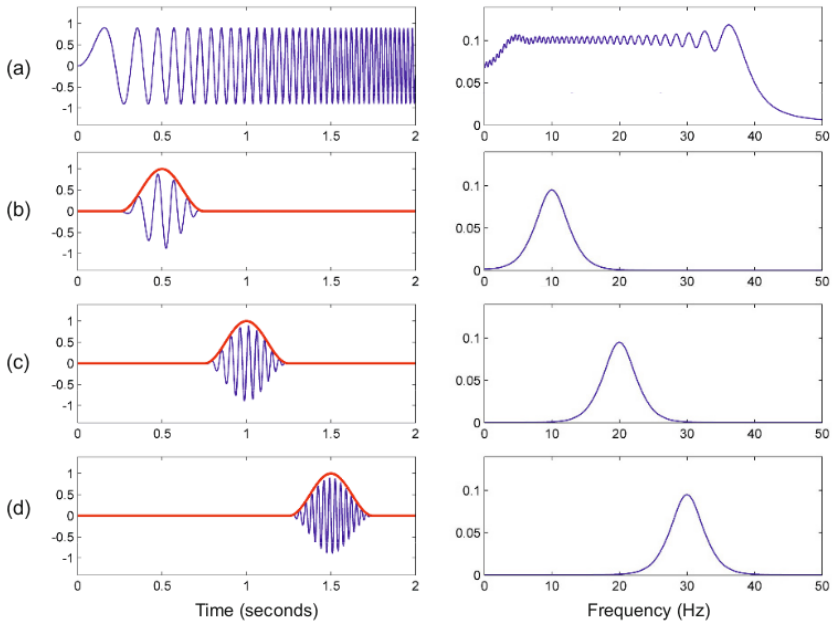
$$f_{g,t}(u) = f(u)g(u - t) \tag{4.25}$$

Γενικώς χρησιμοποιούνται και μιγαδικές συναρτήσεις παραθύρων $g : \mathbb{R} \to \mathbb{C}$ και απαιτείται $g \in L^2(\mathbb{R})$ και $||g||_2 \neq 0$. Επεκτείνοντας την σχέση (4.25), η $f_{g,t}$ ορίζεται ως

$$f_{g,t}(u) = f(u)\bar{g}(u - t) \tag{4.26}$$

Δοθέντος ενός σήματος $f \in L^2(\mathbb{R})$ καθώς επίσης και μιας συνάρτησης παραθύρου $g \in L^2(\mathbb{R})$, ο συνεχούς χρόνου STFT είναι μια συνάρτηση $\tilde{f}_g : \mathbb{R} \times \mathbb{R} \to \mathbb{C}$ και ορίζεται ως

$$\tilde{f}_g(t, \omega) = \widehat{f_{g,t}}(\omega) = \int_{u \in \mathbb{R}} f(u)\bar{g}(u - t) \exp(-2\pi j\omega u)du \tag{4.27}$$



Σχήμα 4.6: Ένα chirp σήμα και οι εκδόσεις του με συνάρτηση παραθύρου, καθώς επίσης και τα πλάτη των μετασχηματισμών Fourier. (a) Πρωτότυπο σήμα. (b) Παράθυρο κεντραρισμένο σε χρόνο $t = 0.5$s. (c) Παράθυρο κεντραρισμένο σε χρόνο $t = 1.0$s. (d) Παράθυρο κεντραρισμένο σε χρόνο $t = 1.5$s. Πηγή: [30]

4.8 Ο ρόλος της συνάρτησης παραθύρου

Η συνάρτηση παραθύρου g, παίζει σημαντικό ρόλο από την πλευρά της ψηφιακής επεξεργασίας σήματος. Συνήθως, η συνάρτηση παραθύρου επιλέγεται ώστε να είναι μηδενική εκτός κάποιου επιλεγμένου τμήματος, ούτως ώστε όταν ένα σήμα πολλαπλασιάζεται με την συνάρτηση παραθύρου, το γινόμενο να είναι επίσης μηδενικό εκτός του τμήματος. Ο φαινομενικά πιο απλός τρόπος να λάβουμε μια τοπική οπτική του σήματος f, είναι να το αφήσουμε ως έχει εντός του επιθυμητού τμήματος και να θέσουμε όλες τις τιμές μηδέν, που είναι εκτός του τμήματος. Μια τέτοια διαδικασία υλοποιείται χρησιμοποιώντας ένα ορθογωνικό παράθυρο(Rectangular Window)

$$f(t) = \begin{cases} 1, & a\nu - 0.5 \leq t \leq 0.5, \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{4.28}$$

Όμως αυτή η συνάρτηση παραθύρου έχει πολλά μειονεκτήματα, αφού γενικά οδηγεί σε ασυνέχειες στα σύνορα του τμήματος στο σήμα $f_{g,t}$. Τέτοιες απότομες αλλαγές οδηγούν σε θόρυβο εξαιτίας των παρεμβολών που απλώνονται σε όλο το φάσμα των συχνοτήτων. Αυτές οι θορυβικές συχνοτικές συνιστώσες προέρχονται από τις ιδιότητες του ορθογωνικού παραθύρου και όχι από το πρωτότυπο σήμα. Ο

λόγος είναι, ότι ο μετασχηματισμός Fourier του ορθογωνικού παραθύρου είναι

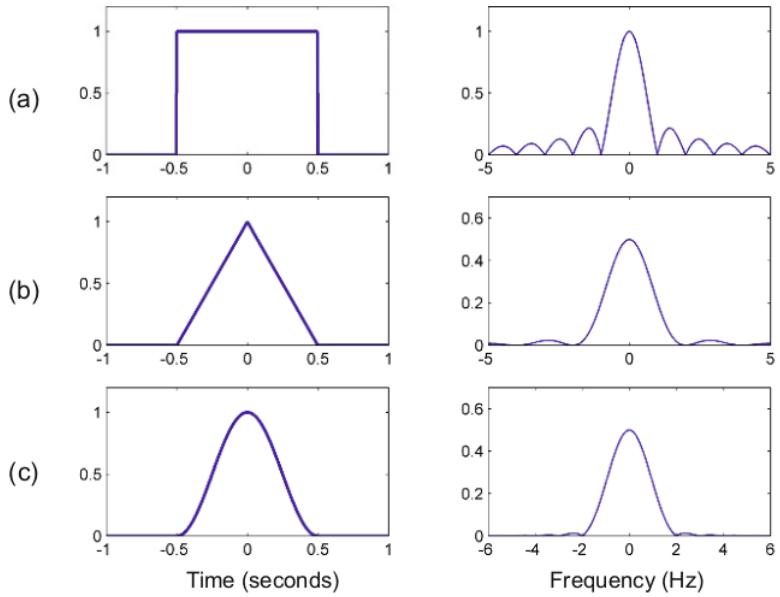
$$\mathrm{sinc}(t) = \begin{cases} \dfrac{\sin \pi t}{\pi t}, & \alpha\nu \quad t \neq 0, \\ 1, & \alpha\nu \quad t = 0. \end{cases} \tag{4.29}$$

και επομένως προκύπτει, ότι θορυβικοί κυματισμοί απλώνονται σε όλο το συχνοτικό φάσμα. Για την ελάττωση των επιπτώσεων στα σύνορα του επιλεγόμενου τμήματος, συνήθως επιλέγονται παράθυρα που είναι μη αρνητικά εντός του τμήματος, ενώ εκτός πέφτουν συνεχόμενα προς το μηδέν. Ένα τέτοιο παράδειγμα είναι το τριγωνικό παράθυρο(Triangular Window), το οποίο οδηγεί σε αρκετά μικρότερες θορυβικές κυματώσεις.

Ένα παράθυρο που χρησιμοποιείται αρκετά συχνά στην επεξεργασία σήματος και χρησιμοποιείται και στην συγκεκριμένη διπλωματική εργασία, είναι το παράθυρο Hann ή Hanning, το οποίο έχει πάρει το όνομά του από τον Julius von Hann. Το παράθυρο Hann g, είναι ένα παράθυρο υψωμένου συνημιτόνου(Raised Cosine) και ορίζεται ως

$$g(u) = \begin{cases} (1 + \cos(\pi u))/2, & \alpha\nu \quad -0.5 \leq u \leq 0.5, \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{4.30}$$

Είναι σχεδιασμένο ώστε στο σύνορο, να πέφτει απαλά προς το μηδέν, και έτσι οι θορυβικές κυματώσεις του μετασχηματισμού Fourier του παραθυρομένου σήματος, αποσβένονται σε μεγάλο βαθμό.



Σχήμα 4.7: Συναρτήσεις παραθύρου και μετασχηματισμοί τους κατά Fourier. (a) Ορθογωνικό παράθυρο. (b) Τριγωνικό παράθυρο. (c)Παράθυρο Hann. Πηγή: [30]

4.9 Διακριτή μορφή του βραχυχρόνιου μετασχηματισμού Fourier(Discrete STFT)

Στην πράξη, τα σήματα που χρησιμοποιούνται προέρχονται από δειγματοληψία και ο υπολογισμός του STFT γίνεται σε πεπερασμένο πλέγμα χρόνου-συχνότητας. Για λόγους απόδοσης, υλοποιούνται DFTs οι οποίοι υπολογίζονται από τον αλγόριθμο FFT.

Έστω $\mathbf{x}$ ένα DT-σήμα, που έχει ληφθεί από ένα CT-σήμα f από δειγματοληψία. Επιπρόσθετα, έστω $\mathbf{w}$ η δειγματοληπτική έκδοση μιας αναλογικής συνάρτησης παραθύρου g. Στην διακριτή περίπτωση, το παράθυρο μπορεί να ολισθήσει , υπό την έννοια των δειγμάτων. Λόγω προβλημάτων απόδοσης, συνήθως το παράθυρο υλοποιείται ώστε να ολισθαίνει με ακόμα μεγαλύτερα βήματα, τα οποία καθορίζεται από την παράμετρο hop size $H \in \mathbb{N}$(δοσμένη σε δείγματα).

Ορίζουμε την διακριτή έκδοση του STFT $\widetilde{\mathbf{x}^{\mathbf{w}}}$ ενός DT-σήματος $\mathbf{x}$ σε σχέση

με την συνάρτηση παραθύρου $\mathbf{w}$ ως

$$\widetilde{x^w}(m,\omega) = \sum_{n\in\mathbb{Z}} x(n)\overline{w}(n-mH)\exp(-2\pi j\omega(n-mH)) \tag{4.31}$$

$$= \sum_{n\in\mathbb{Z}} x(n+mH)\overline{w}(n)\exp(-2\pi j\omega n) \tag{4.32}$$

για $m \in \mathbb{Z}$ και $\omega \in [0,1)$. Αν η δειγματοληπτική έκδοση της συνάρτησης παραθύρου $\mathbf{w}$ είναι ένα πεπερασμένο σήμα, τότε το άθροισμα στην (4.31), γίνεται πεπερασμένο και έτσι μπορούμε να εφαρμόσουμε τον DFT ώστε να υπολογίσουμε τον διακριτό STFT για συγκεκριμένες συχνότητες.

Στην αναλογική περίπτωση, κάναμε την υπόθεση ότι η συνάρτηση παραθύρου ήταν κεντραρισμένη σε χρόνο μηδέν. Για την απλοποίηση των σχέσεων στη διακριτή περίπτωση, υποθέτουμε ότι η συνάρτηση παραθύρου επενεργεί μόνο στη θετική πλευρά του άξονα του χρόνου, κεντραρισμένη στο ήμισυ του μήκους του παραθύρου. Η περίπτωση μηδενικού κεντραρίσματος μπορεί εύκολα να ανακτηθεί, ξαναολισθαίνοντας το πρωτότυπο σήμα κατά το μισό του μήκους του παραθύρου.

Θεωρούμε ότι τα μη μηδενικά δείγματα του διακριτου παραθύρου $\mathbf{w}$ είναι $w(n)$ για $n \in [0 : N-1]$. Για κάθε πλαίσιο(Frame) με δείκτη $m \in \mathbb{Z}$, ορίζουμε το διάνυσμα $\mathbf{x_m} = (x_m(0), ..., x_m(N-1))^T \in \mathbb{C}^{\mathbb{N}}$ με

$$x_m(n) = x(n+mH)\overline{w}(n) \tag{4.33}$$

για $n \in [0 : N-1]$ και υπολογίζουμε το διάνυσμα $\mathbf{X_m} = (X_m(0), ..., X_m(N-1))^T \in \mathbb{C}^N$, χρησιμοποιώντας DFT μεγέθους N

$$\mathbf{X_m} = DFT_N \cdot \mathbf{x_m} \tag{4.34}$$

Έτσι λαμβάνουμε

$$\begin{aligned}\widetilde{x^w}(m,k/N) &= \sum_{n=0}^{N-1} x(n+mH)\overline{w}(n)\exp(-2\pi jkn/N)\\ &= \sum_{n=0}^{N-1} x_m(n)\exp(-2\pi jkn/N)\\ &= X_m(k)\end{aligned} \tag{4.35}$$

για $k \in [0 : N-1]$. Επομένως για κάθε χρονικό πλαίσιο $m \in \mathbb{Z}$, μπορεί να υπολογισθεί ο διακριτός STFT σε συχνότητες $\omega = k/N$ για $k \in [0 : N-1]$ χρησιμοποιώντας έναν DFT_N. Στην περίπτωση που το N είναι δύναμη του 2, αυτό γίνεται πολύ αποδοτικά χρησιμοποιώντας τον αλγόριθμο FFT.

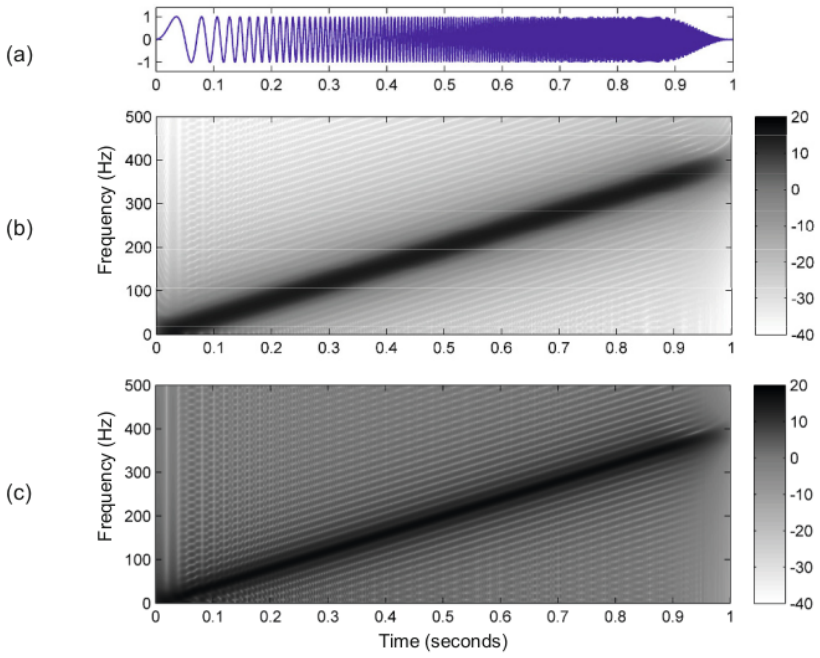
4.10 Αναπαράσταση φασματογραφήματος

Ο STFT ενός σήματος f, για κάθε σημείο στον χρόνο t και συχνότητα ω, εξάγει έναν μιγαδικό αριθμό $\widetilde{f}_g(t,\omega)$. Η πληροφορία αυτή συχνά οπτικοποιείται από το

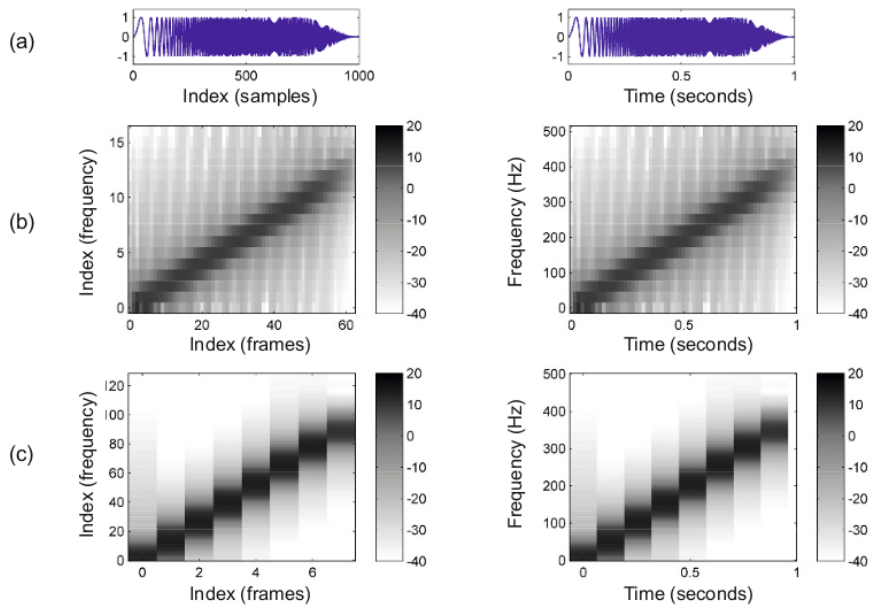
φασματογράφημα, το οποίο είναι μια δισδιάστατη αναπαράσταση του τετραγώνου του πλάτους

$$Spec(t,\omega) = |\widetilde{f}_g(t,\omega)|^2 = |\widetilde{f^g}(t,\omega)|^2 \tag{4.36}$$

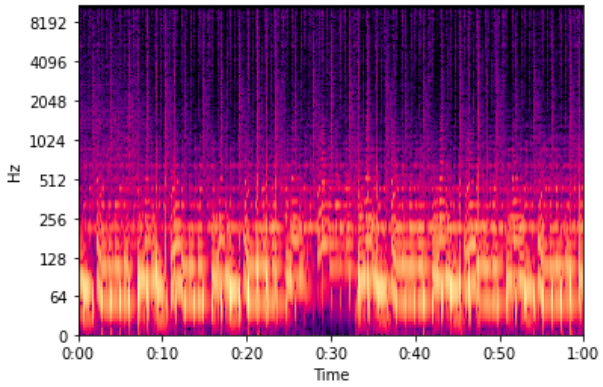
Όταν δημιουργούμε μια εικόνα ενός φασματογραφήματος, ο οριζόντιος άξονας αναπαριστά τον χρόνο, ο κάθετος άξονας την συχνότητα, και η διάσταση που δείχνει η τιμή του φασματογραφήματος για συγκεκριμένη συχνότητα και για συγκεκριμένο χρόνο, αναπαριστάται από την ένταση ή το χρώμα της εικόνας. Γενικά υπάρχουν πολλοί τρόποι να οπτικοποιήσουμε ένα φασματογράφημα. Για να δώσουμε έμφαση σε μουσικές συσχετίσεις ή συσχετίσεις τόνων, ο άξονας της συχνότητας συνήθως απεικονίζεται σε λογαριθμική κλίμακα, το οποίο οδηγεί σε μια λογαριθμική-συχνοτική αναπαράσταση(Log-Frequency Representation). Ο λογαριθμικός άξονας της συχνότητας βοηθάει στην κατανόηση, αφού η ανθρώπινη αντίληψη του pitch είναι εκ φύσεως λογαριθμική. Στην περίπτωση των ηχητικών σημάτων, χρησιμοποιείται συνήθως κλίμακα decibel για το πλάτος.



Σχήμα 4.8: Φασματογράφημα ενός chirp σήματος $f(t) = \sin(400\pi t^2)$ για $t \in [0,1]$ κάνοντας χρήση δύο διαφορετικών τύπων παραθύρου. Η κλίμακα του πλάτους είναι λογαριθμική. (a) Σήμα. (b) Φασματογράφημα με παράθυρο Hann μεγέθους 62.5ms. (c) Φασματογράφημα με ορθογωνικό παράθυρο μεγέθους 62.5ms. Πηγή: [30]



Σχήμα 4.9: Φασματογράφημα με χρήση διακριτού STFT. Αριστερά φαίνεται η διακριτή μορφή και δεξιά η φυσική τους αναπαράσταση. (a) Σήμα με $1/T = 1000$Hz. (b) Φασματογράφημα με $N = 32$ και $H = 16$. (c) Φασματογράφημα με $N = 256$ και $H = 128$. Πηγή: [30]

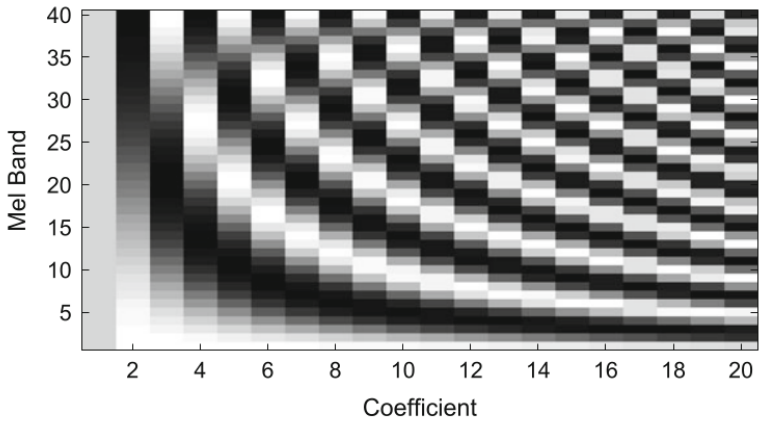


Σχήμα 4.10: Φασματογράφημα του μουσικού κομματιού "Rebekah Del Rio - Llorando (David August Edit Boiler Room Berlin)". Χρησιμοποιήθηκε η βιβλιοθήκη Librosa(βλ. εδάφιο 6.1.3) και STFT με μήκος $N = 4096$, hop size $H = 256$, παράθυρο Hann με μέγεθος $W = 1024$.

4.11 Mel Frequency Cepstral Coefficients

Οι Mel Frequency Cepstral Coefficients(MFCCs), προέρχονται από το πεδίο της επεξεργασίας ομιλίας, αλλά χρησιμοποιούνται επίσης και για την μοντελοποίηση της χροιάς στη μουσική. Το χαρακτηριστικό MFCC υπολογίζεται στο πεδίο της συχνότητας, παραγόμενο από το φασματογράφημα του σήματος [32].

Δοθέντων των πλατών σε ένα εύρος συχνοτήτων για κάθε πλαίσιο, η κλίμακα της συχνότητας σε Hertz πρώτα μετατρέπεται σε μια κλίμακα Mel. Από την αναπαράσταση Mel των πλατών ενός δοθέντος πλαισίου, λαμβάνεται ακολούθως ο λογάριθμος. Οι τιμές των λογαριθμικών πλατών που προκύπτουν στο εύρος της μπάντας Mel, τροφοδοτούνται σε έναν διακριτό μετασχηματισμό συνημιτόνου(Discrete Cosine Transform), ο οποίος αποτελεί μια μορφή του διακριτού μετασχηματισμού Fourier, και χρησιμοποιεί μόνο τις πραγματικές τιμές και όχι τις μιγαδικές. Ως αποτέλεσμα, εξάγεται ένα φάσμα που έχει υπολογιστεί στο εύρος των συχνοτήτων Mel. Το εξαγόμενο φάσμα αναπαριστά τους MFCCs για το εκάστοτε πλαίσιο. Αν η ίδια διαδικασία, υλοποιηθεί για όλα τα πλαίσια του μουσικού κομματιού, το αποτέλεσμα είναι προσωρινά διατεταγμένα διανύσματα από MFCCs, και ομοιάζει με φασματογράφημα που εξάγεται από STFT.



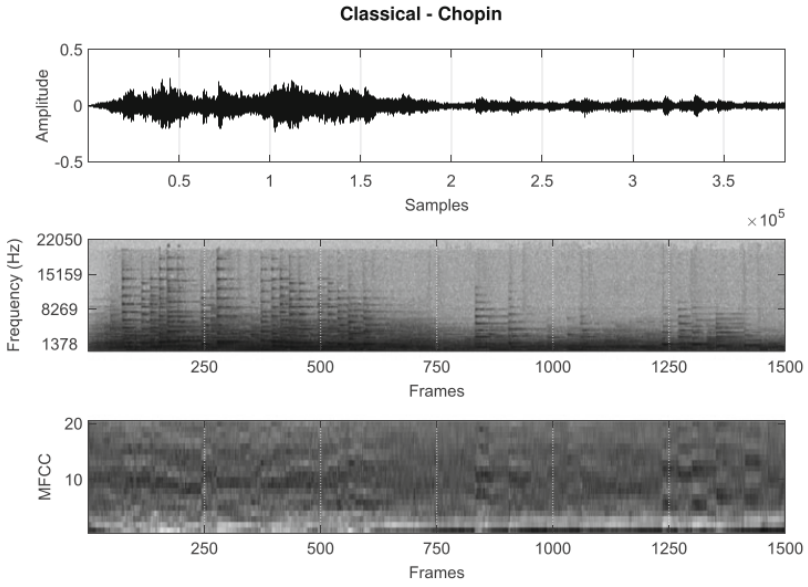
Σχήμα 4.11: Μπάντες συχνοτήτων των Mel Frequency Cepstral Coefficients. Πηγή: [32]

Ένα MFCC διάνυσμα περιγράφει τις περιοδικότητες στις τιμές των πλατών στο σύνολο του εύρους των συχνοτήτων του δοθέντος πλαισίου. Στην ψηφιακή επεξεργασία ήχου, συνήθως για κάθε πλαίσιο υπολογίζονται 13 με 25 MFCCs. Το πρώτο MFCC αντιστοιχεί στην μέση ενέργεια του σήματος για το εκάστοτε πλαίσιο. Συχνά παραλείπεται κατά την μοντελοποίηση χαρακτηριστικών για προβλήματα εξόρυξης μουσικής πληροφορίας. Η αύξηση των MFCC, οδηγεί σε υψηλότερες περιοδικότητες όπως φαίνεται και στο σχήμα (4.11).

Ο υπολογισμός των MFCC διανυσμάτων ή άλλων χαρακτηριστικών σε επίπεδο

πλαισίου, για ένα μουσικό κομμάτι, τυπικά έχει ως αποτέλεσμα δεκάδες χιλιάδες ξεχωριστά διανύσματα χαρακτηριστικών. Παραδείγματος χάριν, θεωρούμε ένα τραγούδι τριών λεπτών το οποίο έχει υποστεί δειγματοληψία στα 44,100Hz. Εφαρμόζοντας σε αυτό έναν εξαγωγέα χαρακτηριστικών σε επίπεδο πλαισίου, με μέγεθος πλαισίου 512 δείγματα και hop size 50%, παράγονται περισσότερα από 20,000 διανύσματα χαρακτηριστικών που το περιγράφουν.

$$N_{fu} = 1.5 \cdot (\frac{44,100 \quad samples/s}{512 \quad samples/frame}) \cdot 180\text{s} = 23,256 frames \tag{4.37}$$



Σχήμα 4.12: Πλάτος, φασματογράφημα και Mel Frequency Cepstral Coefficients ενός κλασικού κομματιού από τον Chopin. Πηγή: [32]

Κεφάλαιο 5

Διαχωρισμός ηχητικών πηγών

5.1 Εισαγωγή

Ο διαχωρισμός ηχητικών πηγών αποτελεί βασικό πρόβλημα για το πεδίο της ψηφιακής επεξεργασίας του ήχου, με εφαρμογές στην ομιλία, τη μουσική και τους ήχους του περιβάλλοντος. Η έρευνα σε αυτό το πεδίο έχει επιφέρει τεχνολογικές καινοτομίες, όπως την μετάβαση από την καλωδιακή τηλεφωνία, στην κινητή και σε hands free τηλέφωνα, την σταδιακή αντικατάσταση του στέρεο από 3-D ήχο και την εμφάνιση συνδεδεμένων συσκευών αποτελούμενες από ένα ή περισσότερα μικρόφωνα, οι οποίες μπορούν να εκτελούν διεργασίες ψηφιακής επεξεργασίας του ήχου οι οποίες προηγουμένως θεωρούνταν απίθανο να εκτελεσθούν [42].

Τα σήματα ομιλίας στον πραγματικό κόσμο, συχνά εμπλουτίζονται με πληροφορία από παρεμβαλόμενους ομιλητές, περιβαλλοντικό θόρυβο ή και από φαινόμενα αντήχησης. Τέτοια φαινόμενα χειροτερεύουν την ποιότητα της ομιλίας και μπορεί παραδείγματος χάριν να προκαλέσουν προβλήματα στην απόδοση συστημάτων αυτόματης αναγνώρισης ομιλίας(ASR). Ο διαχωρισμός ηχητικών πηγών επομένως είναι απαραίτητος σε τέτοιου είδους καταστάσεις. Περαιτέρω παραδείγματα αποτελούν η επικοινωνία με χρήση κινητών τηλεφώνων και συστημάτων handsfree, τα οποία απαιτούν το διαχωρισμό της φωνής του κοντινού ομιλητή σε σχέση με τους παρεμβαλόμενους ομιλητές και τους περιβαλλοντικούς θορύβους, πριν από την μετάδοση στον απομακρυσμένο δέκτη. Επιπρόσθετα συστήματα τηλεδιασκέψεων αντιμετωπίζουν το ίδιο πρόβλημα, με εξαίρεση το ότι στην περίπτωση αυτή, αρκετοί ομιλητές μπορούν να θεωρηθούν ότι συμμετέχουν και επομένως το πρόβλημα προσδιορίζεται διαφορετικά. Ο διαχωρισμός ηχητικών πηγών επιπλέον, αποτελεί σημαντικό βήμα προεπεξεργασίας για καινοτόμα συστήματα αυτόματης αναγνώρισης ομιλίας απομακρυσμένου μικροφώνου(Distantmicrophone ASR), όπως είναι διαθέσιμα σήμερα σε προσωπικούς βοηθούς, συστήματα πλοήγησης, τηλεοράσεις, παιχνιδομηχανές, συσκευές ιατρικής υπαγόρευσης ή και συστήματα καταγραφής των πρακτικών σε συνεδριάσεις. Σε

τελική ανάλυση, ο διαχωρισμός ηχητικών πηγών αποτελεί βασικό εργαλείο για την κατασκευή ρομπότς με ανθρώπινα χαρακτηριστικά, για συσκευές υποβοήθησης ακοής, αλλά και για συστήματα επιτήρησης με ικανότητες "υπερακοής", τα οποία μπορεί να ξεπερνούν τις ικανότητες της ανθρώπινης ακοής.

Ένα επιπλέον σημαντικό πεδίο διαχωρισμού ηχητικών πηγών με το οποίο ασχολούμαστε και στην συγκεκριμένη διπλωματική εργασία, είναι η μουσική. Οι μουσικές ηχογραφήσεις, τυπικά περιλαμβάνουν αρκετά όργανα τα οποία παίζονται ταυτόχρονα ζωντανά ή τα οποία έχουν υποστεί διαδικασία μίξης σε στούντιο ηχογράφησης. Επιπρόσθετα, soundtracks ταινιών περιέχουν επικαλυπτόμενη ομιλία με μουσική και ηχητικά εφέ. Ο διαχωρισμός ηχητικών πηγών έχει χρησιμοποιηθεί με επιτυχία για το upmixing μονοφωνικών και στερεοφωνικών ηχογραφήσεων σε 3-D ηχητικές μορφές ή και για εφαρμογές remixing. Βρίσκεται επιπλέον σε κωδικοποιητές ήχου οι οποίοι κωδικοποιούν μια δοθείσα ηχητική ηχογράφηση ως το άθροισμα διαφόρων ηχητικών πηγών τα οποία μπορούν έπειτα να επεξεργαστούν εύκολα. Μια επιπλέον χρησιμότητα είναι για συστήματα εξόρυξης μουσικής πληροφορίας, παραδείγματος χάριν την εξαγωγή της μελωδίας ή των στίχων ενός τραγουδιού από την διαχωρισμένη πηγή των φωνητικών.

5.2 Βασική θεωρία

5.2.1 Μονοκαναλικό έναντι Πολυκαναλικού

Υποθέτουμε ότι το παρατηρούμενο σήμα έχει I κανάλια τα οποία δεικτοποιούνται ως $i \in [1 : I]$. Με τον όρο κανάλι εννοούμε της έξοδο ενός μικροφώνου, στην περίπτωση όπου το παρατηρούμενο σήμα έχει ηχογραφηθεί από ένα ή περισσότερα μικρόφωνα, ή την είσοδο ενός ηχείου στην περίπτωση όπου πρόκειται να παιχτεί από ένα ή περισσότερα ηχεία. Ένα σήμα με $I = 1$ κανάλια καλείται μονοκαναλικό(Singlechannel) και αναπαρίσταται από ένα βαθμωτό μέγεθος $x(t)$, ενώ ένα σήμα με $I > 1$ κανάλια, καλείται πολυκαναλικό(Multichannel) και αναπαρίσταται από ένα διάνυσμα $\mathbf{x}(\mathrm{t})$, διαστάσεων $I \times 1$.

Η παρακάτω ανάλυση θα γίνει χρησιμοποιώντας πολυκαναλική σημειογραφία, αλλά ισχύει και για την μονοκαναλική περίπτωση.

5.2.2 Σημειακές έναντι διασκορπιστικών πηγών

Υποθέτουμε ότι υπάρχουν K ηχητικές πηγές οι οποίες δεικτοποιούνται ως $k \in [1 : K]$. Η λέξη "πηγή" αναφέρεται σε δύο διαφορετικά σενάρια. Μια σημειακή πηγή(Point Source) όπως λόγου χάριν ένας άνθρώπινος ομιλητής, ένα πουλί ή ένα ηχείο θεωρείται ότι παράγει ήχο από ένα μονό σημείο στον χώρο. Μπορεί να αναπαρασταθεί από ένα μονοκαναλικό σήμα. Αντίθετα, μια διασκορπιστική πηγή(Diffuse Source) όπως ένα αυτοκίνητο, ένα πιάνο ή η βροχή, συνεχώς παράγει ήχο από μια ολόκληρη περιοχή του χώρου. Οι ήχοι οι οποίοι παράγονται από διαφορετικά σημεία αυτής της περιοχής, είναι διαφορετικά αλλά όχι πάντοτε ανεξάρτητα μεταξύ τους. Επομένως, μια διασκορπιστική πηγή μπορεί να θεωρηθεί ως μια άπειρη συλλογή από σημειακές πηγές. Η εκτίμηση των ατομικών σημειακών

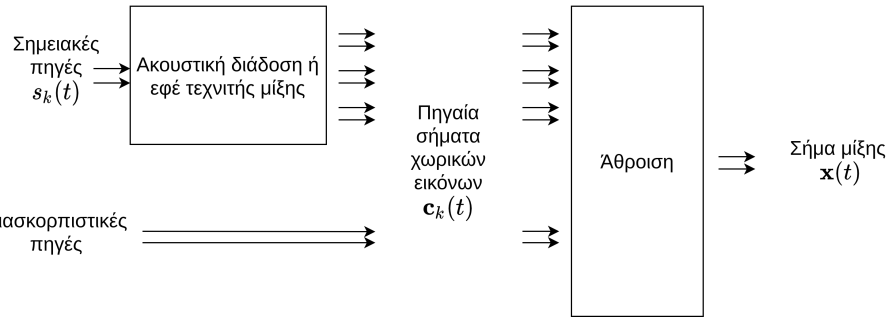
πηγών σε αυτήν την συλλογή θεωρείται αδιάφορη για τον διαχωρισμό ηχητικών πηγών και επομένως μια διασκορπιστική πηγή τυπικά αναπαρίσταται από το αντίστοιχο ηχογραφημένο σήμα από το μικρόφωνο ή τα μικρόφωνα και επεξεργάζεται ως ενιαίο.

5.2.3 Η διαδικασία της μίξης

Η διαδικασία της μίξης(Mixing Process) η οποία οδηγεί στο παρατηρούμενο σήμα, μπορεί γενικά να εκφραστεί σε δύο βήματα. Αρχικά, κάθε μονοκαναλική σημειακή πηγή $s_k(t)$ μετασχηματίζεται σε ένα $I \times 1$ πηγαίο σήμα χωρικής εικόνας $\mathbf{c_k}(t) = [c_{1k}(t), \ldots, c_{Ik}(t)]$, πιθανόν με μια ειδικευμένη μη γραμμική διαδικασία. Αυτή η διαδικασία η διαδικασία μπορεί να περιγράψει την ακουστική διάδοση από την σημειακή πηγή στο μικρόφωνο ή στα μικρόφωνα, περιλαμβάνοντας φαινόμενα αντίχησης είτε κάποια τεχνιτά εφέ μίξης. Οι διασκορπιστικές πηγές αντιθέτως, αναπαρίστανται άμεσα από τις $I \times 1$ χωρικές εικόνες $\mathbf{c}_k(t)$. Έπειτα, οι χωρικές εικόνες όλων των πηγών αθροίζονται για την εξαγωγή του $I \times 1$ παρατηρούμενου σήματος $\mathbf{x}(t) = [x_1(t), \ldots, x_I(t)]^T$ το οποίο καλείται το σήμα της μίξης(Mixture):

$$\mathbf{x}(t) = \sum_{k=1}^{K} \mathbf{c}_k(t) \tag{5.1}$$

Αυτό το άθροισμα υφίσταται εξαιτίας της υπέρθεσης των ηχητικών πηγών στην περίπτωση μιας ηχογράφησης με μικρόφωνο, είτε εξαιτίας της εξωτερικής άθροισης στην περίπτωση της τεχνιτής μίξης. Αυτό υποδηλώνει, ότι η χωρική εικόνα κάθε πηγής, αναπαριστά την συνεισφορά της πηγής στο τελικό σήμα της μίξης.



Σχήμα 5.1: Γενική διαδικασία μίξης για την περίπτωση $K = 4$ ηχητικών πηγών, εκ των οποίων οι τρεις πηγές σημειακές και μία διασκορπιστική, και $I = 2$ κανάλια. Πηγή: [42]

Σε αυτήν την ανάλυση, θεωρούμε ότι οι πηγές, είτε είναι αγνές, είτε παρεμβαλλόμενες, είτε είναι θόρυβος, αντιμετωπίζονται με τον ίδιο τρόπο. Όλα αυτά τα σήματα μπορεί να είναι είτε σημειακές είτε διασκορπιστικές πηγές. Η επιλογή των αγνών πηγών εξαρτάται από την εκάστοτε περίπτωση. Επιπλέον, η διάκριση μεταξύ παρεμβαλλόμενων πηγών και θορύβου, μπορεί να υφίσταται ή μπορεί και όχι, ανάλογα την περίπτωση.

5.2.4 Η τυπολογία των σεναρίων

Οι ερευνητές του διαχωρισμού ηχητικών πηγών, έχουν εφεύρει την εξής ορολογία για τον χαρακτηρισμό μια διαδικασία μίξης. Δοθέντος ενός σήματος προερχόμενο από μίξη, αυτό θεωρείται ως

- **Γραμμικό(Linear)** αν η διαδικασία της μίξης είναι γραμμική, και **μη γραμμικό(Non-linear)** σε διαφορετική περίπτωση.
- **Χρονικά αμετάβλητο(Time invariant)** αν η διαδικασία της μίξης είναι σταθερή κατά τη διάρκεια του χρόνου, και **χρονικά μεταβλητό(Time varying)** σε διαφορετική περίπτωση.
- **Στιγμιαίο(Instantaneous)** αν η διαδικασία της μίξης απλώς κλιμακώνει κάθε πηγαίο σήμα κατά έναν διαφορετικό παράγοντα σε κάθε κανάλι, **ανηχοϊκό(Anechoic)** αν επιπρόσθετα εφαρμόζει μια διαφορετική καθυστέρηση σε κάθε πηγή και σε κάθε κανάλι, και **συνελικτικό(Convolutive)** στην πιο γενική περίπτωση όπου το τελικό σήμα προκύπτει από την άθροιση πολλαπλών κλιμακούμενων και καθυστερημένων εκδόσεων των πηγών.
- **Υπερκαθορισμένο(Overdetermined)** αν δεν υπάρχει καμία διασκορπιστική πηγή και επιπρόσθετα, ο αριθμός των σημειακών πηγών είναι αυστηρά μικρότερος από τον αριθμό των καναλιών, **καθορισμένο(Determined)** αν δεν υπάρχει καμία διασκορπιστική πηγή και ο αριθμός των σημειακών πηγών είναι ίσος με τον αριθμό των καναλιών, και **υποκαθορισμένο(Underdetermined)** σε οποιαδήποτε άλλη περίπτωση.

Αυτή η κατηγοριοποίηση έχει περιορισμένη χρησιμότητα στην περίπτωση του ήχου. Γενικά όλες οι ηχητικές μίξεις είναι γραμμικές ή μπορούν να θεωρηθούν γραμμικές και επίσης είναι συνελικτικές. Η διάκριση μεταξύ υπερκαθορισμένου και υποκαθορισμένου σήματος μίξης, προκύπτει από το γεγονός ότι ένα σήμα προερχόμενο από μίξη, το οποίο είναι καθορισμένο ή υπερκαθορισμένο γραμμικο και χρονικά αμετάβλητο, μπορεί να διαχωριστεί πλήρως αντιστρέφοντας το σύστημα της μίξης, χρησιμοποιώντας έναν γραμμικό και χρονικά αμετάβλητο αντιστροφέα. Στην πράξη όμως, η πλειοψηφία των ηχητικών μίξεων περιέχουν τουλάχιστον μια διασκορπιστική πηγή(π.χ. θόρυβος του περιβάλλοντος) ή περισσότερες σημειακές πηγές από τα διαθέσιμα κανάλια. Επομένως, σε γενικές γραμμές, τα συστήματα διαχωρισμού ηχητικών πηγών, αντιμετωπίζουν ηχητικά σήματα μίξης τα οποία είναι υποκαθορισμένα, γραμμικά(χρονικά μεταβαλλόμενα ή αμετάβλητα) και συνελικτικά.

Πρόσφατα, έχει προταθεί μια εναλλακτική κατηγοριοποίηση βασισμένη στην ποσότητα της εκ των προτέρων διαθέσιμης πληροφορίας που έχουμε για το σήμα της μίξης που πρόκειται να επεξεργασθούμε. Το πρόβλημα διαχωρισμού καθορίζεται ως

- **Τυφλό(Blind)** όταν απολύτως καθόλου πληροφορία δεν δίνεται για τα πηγαία σήματα, την διαδικασία της μίξης ή την προοριζόμενη εφαρμογή.

- **Ασθενώς καθοδηγούμενο ή ημί-τυφλο(Weakly guided or semi-blind)** όταν γενική πληροφορία είναι διαθέσιμη για το γενικό πλαίσιο χρήσης, παραδείγματος χάριν την φύση των πηγών(ομιλία, μουσική, περιβαλλοντικοί ήχοι), τις θέσεις των μικροφώνων, το σενάριο της εγγραφής(εσωτερικός χώρος, εξωτερικός χώρος, επαγγελματικό στούντιο), και την προοριζόμενη εφαρμογή(υποβοήθηση ακοής, αναγνώριση ομιλίας κ.α.).

- **Ισχυρώς καθοδηγούμενο(Strongly guided)** όταν συγκεκριμένη πληροφορία είναι διαθέσιμη για το σήμα προς επεξεργασία, παραδείγματος χάριν η χωρική τοποθεσία των πηγών, το μοτίβο της δραστηριότητάς τους, η ταυτότητα των μεγαφώνων ή οι μουσικές παρτιτούρες.

- **Πληροφορημένο(Informed)** όταν πληροφορία υψηλής ακριβείας για τις πηγές και τη διαδικασία της μίξης, κωδικοποιείται και μεταδίδεται μαζί με τον ήχο.

Τέλος, το πρόβλημα του διαχωρισμού μπορεί να κατηγοριοποιηθεί, βασιζόμενο στην σειρά με την οποία τα δείγματα του σήματος της μίξης επεξεργάζονται. Καλείται **online** όταν το σήμα της μίξης λαμβάνεται σε πραγματικό χρόνο από μικρά μπλοκ από μερικές δεκάδες ή εκατοντάδες δείγματα και κάθε μπλοκ πρέπει να επεξεργασθεί με βάση τα προηγούμενα μπλόκ μόνο ή και με βάση λίγα μελλοντικά μπλοκ, αποδεχόμενοι όμως μια μικρή καθυστέρηση. Αντιθέτως, το πρόβλημα του διαχωρισμού καλείται **offline** ή **batch**, όταν η ηχογράφηση έχει ολοκληρωθεί και έχει υποστεί επεξεργασία ολόκληρη, χρησιμοποιώντας παρελθόντα και μελλοντικά δείγματα για την εκτίμηση ενός δοθέντος δείγματος των πηγών.

5.2.5 Αξιολόγηση του διαχωρισμού ηχητικών πηγών

Ο διαχωρισμός ηχητικών πηγών για σενάρια του πραγματικού κόσμου, είναι αδύνατον να είναι τέλειος. Για κάθε ηχητική πηγή, ο εκτιμητής της ή το πηγαίο σήμα χωρικής εικόνας, μπορεί να διαφέρει από το πραγματικό επιθυμητό σήμα με διάφορους τρόπους, μεταξύ των οποίων

- Παραμόρφωση(Distortion) του επιθυμητού σήματος, λόγου χάριν χαμηλοπερατό φιλτράρισμα, διακύμανση της εντάσεως στη διάρκεια του χρόνου κ.α.

- Παρεμβολές και θόρυβος από άλλες πηγές.

- "Μουσικός θόρυβος", λόγου χάριν απομονωμένοι ήχοι και σε συχνότητα και σε χρόνο, όμοιοι με αυτούς που παράγονται από κωδικοποίηση ήχου με απώλεια πληροφορίας σε πολύ χαμηλό bitrate.

Η αξιολόγηση αυτών των παραμορφώσεων είναι απαραίτητη ώστε να υπάρχει ένα μέτρο σύγκρισης των διαφορετικών αλγορίθμων και επιπλέον, βοηθά στην κατανόηση του πως μπορούμε να βελτιστοποιήσουμε τους αλγορίθμους αυτούς.

Ιδανικά, η αξιολόγηση πρέπει να βασίζεται στην απόδοση της υπο δοκιμής μεθόδου διαχωρισμού ηχητικών πηγών για την εκάστοτε εφαρμογή. Πράγματι,

η σημαντικότητα των διαφόρων τύπων παραμόρφωσης, εξαρτάται από την συγκεκριμένη εφαρμογή. Παραδείγματος χάριν, μια μικρή ποσότητα παραμόρφωσης του εξαγόμενου σήματος, θεωρείται αποδεκτή όταν ακούμε τα διαχωρισμένα σήματα, αλλά μπορεί να οδηγήσει σε μεγάλη μείωση στην απόδοση ενός συστήματος αναγνώρισης ομιλίας. Τα παράσιτα συνήθως μειώνονται κατά μεγάλο βαθμό όταν τα διαχωρισμένα σήματα, επαναμιξάρονται μαζί με έναν διαφορετικό τρόπο, ενώ πρέπει να αποφευχθούν με κάθε κόστος σε συστήματα υποβοήθησης ακοής. Γενικά, υπάρχουν κάποιες ντε φάκτο μετρικές απόδοσης διαθέσιμες ανάλογα με την εκάστοτε εφαρμογή.

Οι πιο ευρέως αποδεκτές μετρήσεις για την απόδοση των μοντέλων διαχωρισμού πηγών, αναπτύχθηκαν από τον Vincent κ.α. για τον τυφλό διαχωρισμό πηγών αρχικά [10], και ξαναχρησιμοποιήθηκαν για τον διαγωνισμό SiSEC [41].

Έστω η ηχητική πηγή s_k και $\hat{s}_k$ ο εκτιμητής αυτής. Χρησιμοποιώντας τις τεχνικές που αναπτύχθηκαν από τον Vincent, ο εκτιμητής της πηγής αναλύεται σε τέσσερις συνιστώσες

- s_{target} η εξαγόμενη ηχητική πηγή από τον αλγόριθμό μας
- e_{interf} ο όρος σφάλματος παρεμβολής
- e_{noise} ο όρος σφάλματος θορύβου
- e_{artif} ο όρος σφάλματος παρασίτων

Από την αποσύνθεση του εκτιμητή $\hat{s}_k$ ως

$$\hat{s}_k = s_{target} + e_{interf} + e_{noise} + e_{artif} \tag{5.2}$$

καθορίζονται τα εξής αριθμητικά κριτήρια απόδοσης υπολογίζοντας λόγους ενέργειας εκφρασμένους σε decibels(dB).

- Λόγος πηγής προς παραμόρφωση(Source to Distortion Ratio)

$$SDR = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2} \tag{5.3}$$

- Λόγος πηγής προς παρεμβολή(Source to Interference Ratio)

$$SIR = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2} \tag{5.4}$$

- Λόγος πηγής προς θόρυβο(Source to Noise Ratio)

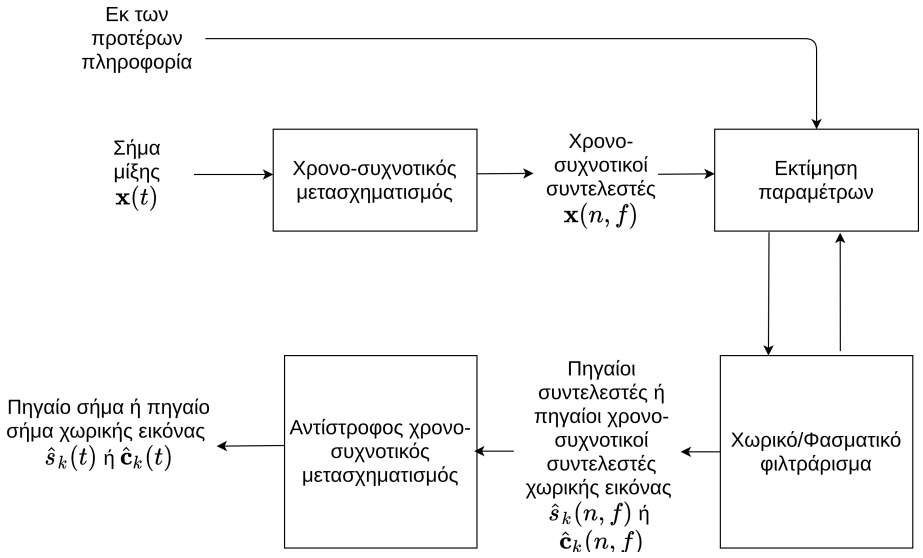
$$SNR = 10\log_{10}\frac{||s_{target} + e_{interf}||^2}{||e_{noise}||^2} \tag{5.5}$$

- Λόγος πηγής προς παράσιτα(Source to Artifacts Ratio)

$$SAR = 10\log_{10}\frac{||s_{target} + e_{interf} + e_{noise}||^2}{||e_{artif}||^2} \tag{5.6}$$

5.2.6 Γενικό σχήμα επεξεργασίας

Πολλές διαφορετικές προσεγγίσεις έχουν προταθεί για τον διαχωρισμό ηχητικών πηγών. Η πλειοψηφία αυτών ακολουθεί το παρακάτω σχήμα (5.2), το οποίο έχει εφαρμογή από κοινού και σε σενάρια μονοκαναλικά και πολυκαναλικά. Το σήμα προερχόμενο από μίξη, στο πεδίο του χρόνου $\mathbf{x}(t)$, μετασχηματίζεται και αναπαρίσταται στο πεδίο χρόνου-συχνότητας. Ένα μοντέλο από μιγαδικούς χρονο-συχνοτικούς συντελεστές του σήματος μίξης $\mathbf{x}(n,f)$ και των πηγών $s_k(n,f)$(και αντιστοίχως οι πηγαίες χωρικές εικόνες $\mathbf{c}_k(n,f)$) δημιουργείται αρχικά. Η επιλογή του μοντέλου γίνεται με βάση την απριόρι γνώση για το σενάριο. Οι παράμετροι του μοντέλου εκτιμούνται από το $\mathbf{x}(n,f)$ ή από διαχωρισμένα δεδομένα εκπαίδευσης, σύμφωνα με ένα προκαθορισμένο κριτήριο. Επιπρόσθετη συγκεκριμένη απριόρι πληροφορία, δύναται να χρησιμοποιηθεί για την βελτίωση της εκτιμήσεως των παραμέτρων, όταν αυτή είναι διαθέσιμη. Δοθέντων αυτών των παραμέτρων, λαμβάνεται μια χρονικά μεταβαλλόμενη μονή έξοδος(ή αντίστοιχα πολλαπλή έξοδος) ενός φίλτρου μιγαδικών τιμών, και εφαρμόζεται στο σήμα προερχόμενο από μίξη $\mathbf{x}(n,f)$, ούτως ώστε να λάβουμε μια εκτίμηση για τους μιγαδικούς χρονο-συχνοτικούς συντελεστές των πηγών $\hat{s}_k(n,f)$(αντίστοιχα των πηγαίων χωρικών εικόνων $\hat{\mathbf{c}}_k(n,f)$. Τελικά, ο χρονο-συχνοτικός μετασχηματισμός αντιστρέφεται, εξάγοντας τους εκτιμητές των πηγών στο πεδίο του χρόνου $\hat{s}_k(t)$(αντίστοιχα τους εκτιμητές πηγαίων χωρικών εικόνων $\hat{\mathbf{c}}_k(t)$).



Σχήμα 5.2: Γενικό σχήμα επεξεργασίας για μονοκαναλικό και πολυκαναλικό διαχωρισμό ηχητικών πηγών. Πηγή: [42]

5.3 Ιστορικές τάσεις και state-of-the-art μέθοδοι

Η πιο δημοφιλής μέθοδος προεπεξεργασία για το πεδίο του διαχωρισμού ηχητικών πηγών, περιλαμβάνει τον μετασχηματισμό του ηχητικού σήματος από το πεδίο του χρόνου σε φασματογράφημα [6], [15], [33], [46]. Γνωρίζοντας ότι η τιμή κάθε ενός χρονο-συχνοτικού(T-F) στοιχείου στο φασματογράφημα του πλάτους X, είναι μη αρνητική, η μέχρι τώρα έρευνα στον τυφλό διαχωρισμό πηγών(Blind Source Separation), τυπικά εφαρμόζει τεχνικές όπως την ανεξάρτητη ανάλυση υποχώρων(Independent Subspace Analysis) [6] και την μη αρνητική παραγοντοποίηση πίνακα(Non-Negative Matrix Factorization) [8]. Η πρότερη τεχνική(ISA), αποτελεί μια παραλαγή της ανεξάρτητης ανάλυσης συνιστωσών(Independent Component Analysis) [9], η οποία έχει προηγουμένως χρησιμοποιηθεί για την επίλυση του προβλήματος cocktail party [1]. Η ανεξάρτητη ανάλυση συνιστωσών, βασίζεται πάνω στην υπόθεση ότι ο αριθμός των παρατηρούμενων σημάτων προερχόμενων από μίξη, είναι ίσος ή μεγαλύτερος από τις εξόδους. Η παραλαγή ISA όμως, χαλαρώνει αυτήν την συνθήκη χρησιμοποιώντας το μη-αρνητικό φασματογράφημα $\mathbf{X}$. Η δεύτερη τεχνική της μη-αρνητικής παραγοντοποίησης πίνακα, χρησιμοποιείται συχνά για τυφλό διαχωρισμό πηγών(BSS), και ουσιαστικά αποσυνθέτει το φασματογράφημα $\mathbf{X}$ σε δύο μη-αρνητικούς πίνακες $\mathbf{L}$ και $\mathbf{R}$. Το γινόμενο των δύο αυτών πινάκων προσεγγίζει το φασματογράφημα $\mathbf{X}$, έτσι ώστε $\mathbf{LR} \approx \mathbf{X}$, με το $\mathbf{D}$ να είναι η διαφορά $\mathbf{D} = \mathbf{X} - \mathbf{LR}$. Ο πίνακας $\mathbf{D}$, θεωρείται εν συνεχεία ότι έχει τα χαρακτηριστικά της χροιάς των φωνητικών.

Η μη-αρνητική παραγοντοποίηση πίνακα αποτελούσε την πιο ευρέως διαδεδομένη τεχνική για διαχωρισμό ηχητικών πηγών κατά τη δεκαετία του 2000 [14], [13], [11]. Η βασική διαφορά μεταξύ των διαφόρων παραλαγών που βασίζονται στην μη-αρνητική παραγοντοποίηση πίνακα, είναι το πως σχεδιάζεται η συνάρτηση κόστους. Μια τυπική σχεδίαση θα μπορούσε να είναι το $min\ ||\mathbf{X} - \mathbf{LR}||^2$ ή $min\ D_{KL}(\mathbf{X}||\mathbf{LR})$, όπου D_{KL} είναι η συνάρτηση απόκλισης Kullback-Leibler. Η δημοτικότητα της μη-αρνητικής παραγοντοποίησης πίνακα, οφείλεται μερικώς στο γεγονός, ότι οι δύο πίνακες($\mathbf{L}$ και $\mathbf{R}$) μπορούν εύκολα να θεωρηθούν ως ένα σύνολο από διαφορετικούς τύπους μουσικών οργάνων(ή διαφορετικών κομματιών στη μουσική), στα οποία αναφερόμαστε ως $\mathbf{M}$. Για την κατανόηση αυτής της θεώρησης, υποθέτουμε αρχικά ότι οι στήλες του $\mathbf{L}$ είναι συναρτήσεις βάσης συχνότητας/τόνου l_i, και οι γραμμές του $\mathbf{R}$ είναι συναρτήσεις βάσης χρόνου r_i, όπου i είναι ένα εκ των μουσικών οργάνων(ή κομματιών) της μουσικής. Οι παραγοντοποιημένοι πίνακες($\mathbf{L}$ και $\mathbf{R}$), μπορούν να αποσυντεθούν ως το άθροισμα του γινομένου(Outer Product) των συναρτήσεων βάσης $\mathbf{LR} = \sum_{i \in M} l_i \times r_i$. Επομένως, η συνάρτηση βάσης συχνότητας l_i, μπορεί να θεωρηθεί ως χροιά του οργάνου i. Το αντίστοιχο σύνολο των συναρτήσεων βάσης χρόνου r_i, υποδικνύει πως ο ήχος του οργάνου i εξελίσσεται στη διάρκεια της μουσικής. Επιπροσθέτως, το $\mathbf{M}$ μερικές φορές διαιρείται σε δύο μέρη, θέτοντας περιορισμούς για το σύνολο των αρμονικών ή pitched οργάνων(π.χ. πιάνο), $h \in \mathbf{M}$, και για το σύνολο των συγκρουστικών οργάνων(π.χ. ντραμς), $p \in \mathbf{M}$.

Μια άλλη τεχνική που έχει εφαρμοσθεί για την επίτευξη του διαχωρισμού

ηχητικών πηγών [17], είναι η εύρωστη κύρια ανάλυση συνιστωσών(Robust Principal Component Analysis ή rPCA). Χρησιμοποιεί έναν επαυξημένο Lagrange πολλαπλασιαστή για τον επακριβή διαχωρισμό του φασματογραφήματος $\mathbf{X}$, σε έναν χαμηλόβαθμο πίνακα και αραιό πίνακα, $\mathbf{X} = \sum_{i \in \mathbf{M}} l_i \times r_i - \mathbf{D}$, και χρησιμοποιήθηκε ευρέως μέχρι το 2012 [23]. Ο προκύπτων πίνακας $\mathbf{LR}$ είναι μια χαμηλόβαθμη προσέγγιση του $\mathbf{X}$. Η χρήση της μεθόδου rPCA στον διαχωρισμό πηγών, οφείλεται πρώτον στο γεγονός ότι η συνάρτηση βάσης του $\mathbf{LR}$ προσεγγίζει το φασματογράφημα της μουσικής συνοδευτικής συνιστώσας του σήματος μίξης, και δεύτερον ο $\mathbf{D}$ είναι ένας αραιός πίνακας ο οποίος προσεγγίζει πολύ καλά το φασματογράφημα των διαχωρισμένων φωνητικών. Για την καλύτερη κατανόηση των παραπάνω, ας σημειωθεί ότι $\mathbf{X} \approx \mathbf{LR}$ και $\mathbf{X} \approx \sum_{i \in \mathbf{M}} l_i \times r_i$. Αν ο αριθμός των μουσικών οργάνων $|\mathbf{M}|$, είναι ο βαθμός του πίνακα $\mathbf{X}$, τότε το $\mathbf{LR}$ είναι μια χαμηλόβαθμη προσέγγιση του $\mathbf{X}$. Σε τελική ανάλυση, αφού τα φωνητικά υφίστανται στο ενδιάμεσο των αρμονικών οργάνων και των συγκρουστικών οργάνων, θεωρείται ότι αναπαρίστανται από τον $\mathbf{D}$.

Τα τελευταία χρόνια, λόγω της πολύ σημαντικής επιρροής που είχε η έρευνα του Krizhevsky και άλλων [24], στην κατηγοριοποίηση εικόνων, η χρήση μεθόδων βαθιάς μάθησης έχει κερδίσει την προσοχή των επιστημόνων. Τα περισσότερα συστήματα διαχωρισμού ηχητικών πηγών, που βασίζονται στην βαθιά μάθηση [40] [38], [33], [44], εκπαιδεύονται ώστε η είσοδος του δικτύου(δηλ., το πλάτος του φασματογραφήματος του σήματος μίξης), να ταιριάζει με τις επιθυμητές ταμπέλες(δηλ., το πραγματικό πλάτος του φασματογραφήματος της εκάστοτε επιθυμητής ηχητικής πηγής). Δοθέντων επαρκών δεδομένων εκπαίδευσης, τα νευρωνικά δίκτυα, τυπικά έχουν την ικανότητα να εκτιμούν καλές προσεγγίσεις οποιασδήποτε συνεχούς συνάρτησης [5], στην περίπτωσή μας, εκτιμάται το πλάτος του φασματογραφήματος, για κάθε ηχητική πηγή. Αυτά τα πλάτη των φασματογραφημάτων όμως, δεν αποτελούν ακόμη καλές αναπαραστάσεις των διαφορετικών πηγών. Τουναντίον στη γενική διαίσθηση που έχουμε για το πρόβλημα, αυτά τα συστήματα απαιτούν ένα βήμα μετεπεξεργασίας, ένα φιλτράρισμα Wiener, στο οποίο μια χαλαρή μάσκα(Soft Mask) υπολογίζεται για τα εκτιμώμενα πλάτη των φασματογραφημάτων, για κάθε επιθυμητή ηχητική πηγή εξόδου. Εν συνεχεία, αυτές οι μάσκες πολλαπλασιάζονται με το πλάτος του φασματογραφήματος του σήματος μίξης, για την επαναδημιουργία του κάθε εκτιμώμενου σήματος. Η χρήση αυτών των χαλαρών μασκών, τυπικά δίνει μια καλύτερη ποιότητα διαχωρισμού από το να χρησιμοποιούσαμε κατευθείαν την έξοδο του δικτύου για την σύνθεση του τελικού σήματος [38]. Αυτό υποδικνύει, ότι θα πρέπει να παραλείψουμε την διαδικασία μετεπεξεργασίας του φιλτραρίσματος Wiener, και να σχεδιάσουμε ένα δίκτυο το οποίο θα μαθαίνει άμεσα μια χαλαρή μάσκα.

Οι τελευταίες καινοτομίες στο πεδίο της όρασης υπολογιστών [28], έχουν βελτιστοποιήσει σε μεγάλο βαθμό τις τεχνικές κατηγοριοποίησης εικόνων, μεταβαίνοντας από το επίπεδο κατηγοριοποίησης ολόκληρης της εικόνας, στο επίπεδο κατηγοριοποίησης εικονοστοιχείων. Η κατηγοριοποίηση εικονοστοιχείων(Pixel-Wise Classification), στοχεύει στην κατηγοριοποίηση καθενός εικονοστοιχείου σε μια εικόνα. Επομένως, το πρόβλημα της κατηγοριοποίησης κάθε χρονο-συχνοτικού στοιχείου ενός φασματογραφήματος σε φωνητική ή μη-φωνητική συνιστώσα, μπορεί να θεωρηθεί πρόβλημα κατηγοριοποίησης

εικονοστοιχείων. Για προβλήματα κατηγοριοποίησης εικόνων, αρχιτεκτονικές βασιζόμενες σε συνελικτικά νευρωνικά δίκτυα, έχει αποδειχθεί ότι αποδίδουν εξαιρετικά [24].

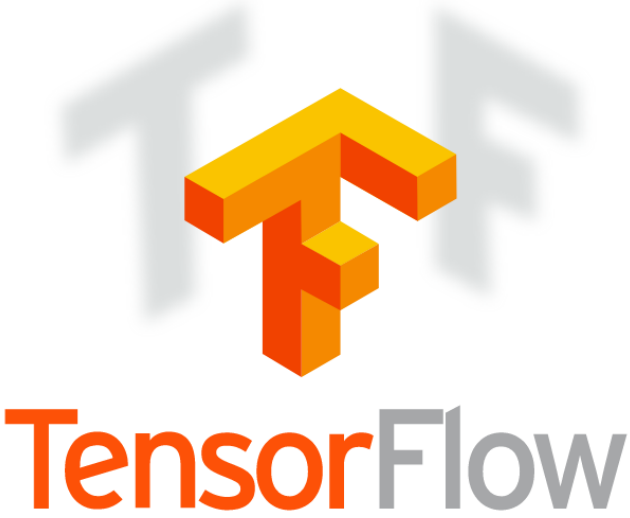
Άλλες state-of-the-art εναλλακτικές στη χρήση συνελικτικών νευρωνικών δικτύων, περιλαμβάνουν τη χρήση επανατροφοδοτούμενων νευρωνικών δικτύων [39] και αμφίδρομων δικτύων Long Short Term Memory(BLSTM) [44].

Κεφάλαιο 6

Πειραματικό μέρος

6.1 Λογισμικό του πειράματος

6.1.1 Η βιβλιοθήκη TensorFlow

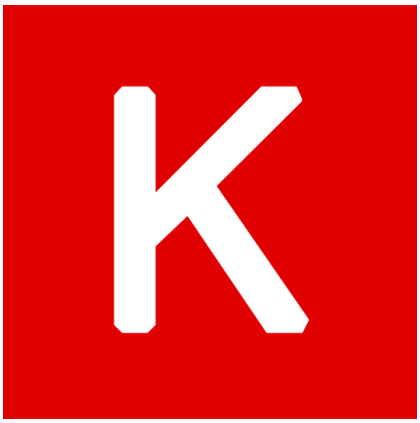


Η `TensorFlow`[1], είναι μια ανοιχτού κώδικα συμβολική μαθηματική βιβλιοθήκη λογισμικού και χρησιμοποιείται για εφαρμογές μηχανικής και βαθιάς μάθησης. Αναπτύχθηκε από την ομάδα Google Brain της Google για εσωτερική χρήση στην εταιρεία αρχικά. Η TensorFlow μπορεί να τρέξει σε πολλαπλές CPUs και GPUs, και είναι διαθέσιμη για συστήματα 64-bit Linux, macOS, Windows, καθώς και κινητές υπολογιστικές πλατφόρμες όπως Android και iOS. Η έκδοση 1.0.0 έγινε

[1] **https://www.tensorflow.org/**.

διαθέσιμη στις 11 Φεβρουαρίου, 2017, ενώ η 2.0.0 έγινε επισήμως διαθέσιμη τον Σεπτέμβρη του 2019. Για την υλοποίηση της συγκεκριμένης διπλωματικής εργασίας χρησιμοποιήθηκε η TensorFlow 2.0.

6.1.2 Η διεπαφή προγραμματισμού εφαρμογών Keras



Το `Keras`[2] είναι μια ανοιχτού κώδικα βιβλιοθήκη νευρωνικών δικτύων γραμμένη σε γλώσσα Python. Είναι ικανό να τρέχει πάνω στην πλατφόρμα του TensorFlow(και άλλων όπως Theano κ.λπ.) και έχει σχεδιαστεί ώστε να επιτρέπει τον εύκολο και γρήγορο πειραματισμό με βαθιά νευρωνικά δίκτυα, και εστιάζει στην φιλικότητα προς το χρήστη και στην επεκτασιμότητα. Κύριος δημιουργός και συντηρητής, είναι ο Francois Chollet. Το 2017, η υπεύθυνη ομάδα της Google για την TensorFlow, αποφάσισε να υποστηρίξει το Keras και να το συμπεριλάβει στη βιβλιοθήκη της.

6.1.3 Η βιβλιοθήκη Librosa

[2] **https://keras.io/**.

Το πεδίο της μουσικής εξόρυξης πληροφορίας, αν και σχετικά καινούριο, αναπτύσσεται ταχύτατα, και αυτό οφείλεται μερικώς, σε διάσημες μουσικές υπηρεσίες όπως iTunes, Shazam και Spotify. Μέχρι το 2015, χρησιμοποιούνταν κυρίως γλώσσες όπως MATLAB και C++ από τους ερεύνητες. Τα τελευταία χρόνια όμως χρησιμοποιείται κατά κόρον η γλώσσα Python για εφαρμογές βαθιάς μάθησης. Λόγω της μεγάλης επιτυχίας που είχε η βαθιά μάθηση σε πεδία σχετικά με την ψηφιακή επεξεργασία σήματος του ήχου, έγινε απαραίτητη η δημιουργία μια βιβλιοθήκης σε γλώσσα Python για αποτελεσματική και εύκολη ψηφιακή επεξεργασία του ήχου.

Η `Librosa`[3], έγινε παρουσιάστηκε για πρώτη φορά το 2015 και αποτελεί μια βιβλιοθήκη της Python, για ανάλυση του ήχου και της μουσικής. Προσφέρει ένα εύρος από εργαλεία όπως είναι η εξαγωγή φασματογραφήματος, βραχυχρόνιοι μετασχηματισμοί Fourier, αλλαγή ρυθμού δειγματοληψίας, κατασκευή τραπεζών φίλτρων, υπολογισμός διαφόρων συναρτήσεων παραθύρων, εξαγωγή των MFCCs και πολλά άλλα.

6.1.4 Η βιβλιοθήκη NumPy

Η `NumPy`[4], είναι μια βιβλιοθήκη ανοιχτού κώδικα της Python, για την υποστήριξη μεγάλων και πολυδιάστατων πινάκων, και προσφέρει μια μεγάλη συλλογή από μαθηματικές συναρτήσεις πολύ υψηλού επιπέδου, οι οποίες εφαρμόζονται στους πίνακες αυτούς. Πρόγονος του ήταν το Numeric, ενώ το 2005, ο Travis Oliphant δημιούργησε την σημερινή μορφή του NumPy εμπεριέχοντας πληροφορίες από το Numarray στο Numeric, προσθέτοντας παράλληλα πολλές επεκτάσεις.

6.1.5 Η βιβλιοθήκη h5py

Η βιβλιοθήκη `h5py`[5], αποτελεί μια διεπαφή της Python για την δυαδική μορφή δεδομένων HDF5. Επιτρέπει την αποθήκευση μεγάλων ποσοτήτων αριθμητικών δεδομένων και συνεργάζεται πολύ εύκολα με την βιβλιοθήκη NumPy κάνοντας την επεξεργασία των δεδομένων πιο εύκολη. Λόγου χάριν, μπορούμε να επεξεργαστούμε σύνολα πολλών Terabytes αποθηκευμένα στον δίσκο, σαν να ήταν πραγματικοί NumPy πίνακες. Δίνεται η δυνατότητα χιλιάδες σύνολα δεδομένων

[3] **https://librosa.org/**.
[4] **https://numpy.org/**.
[5] **https://h5py.org/**.

να αποθηκευτούν σε ένα μονό αρχείο, να κατηγοριοποιηθούν και να τους προστεθούν ετικέτες.

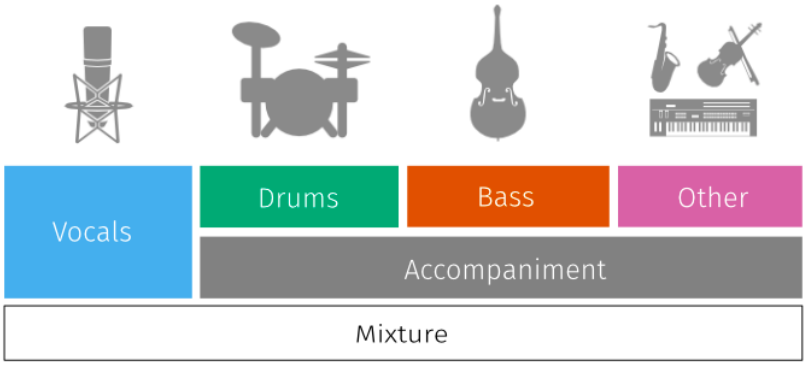
6.1.6 Google Colaboratory

Η διπλωματική εργασία, υλοποιήθηκε σε γλώσσα Python, χρησιμοποιώντας τις προαναφερθείσες βιβλιοθήκες και διεπαφές. Για τις υλοποιήσεις, χρησιμοποιήθηκε η πλατφόρμα `Google Colaboratory`[6]. Το Google Colaboratory(ή Colab) είναι μια δωρεάν υπηρεσία, η οποία βασίζεται στο Jupyter Notebook και δουλεύει στο cloud, όπου η μόνη απαίτηση είναι ένας web browser. Η εκτέλεση του κώδικα, γίνεται σε μια εικονική μηχανή η οποία φιλοξενείται στους διακομιστές της Google.

Όλα τα μοντέλα που θα παρουσιάσουμε σε επόμενο εδάφιο, εκπαιδεύτηκαν στις εικονικές μηχανές του Colab, οι οποίες έχουν τα εξής χαρακτηριστικά hardware:

- Intel(R) Xeon(R) CPU @ 2.00GHz
- NVIDIA Tesla K80 GPU
- 25.51GB RAM
- 68.40GB χώρο στον δίσκο

6.2 Το MUSDB18 σύνολο δεδομένων



Το MUSDB18 σύνολο δεδομένων [37] αποτελείται από 150 μουσικά κομμάτια, διάρκειας περίπου 10 ωρών. Αποτελείται από δύο φακέλους. Ο ένας αφορά το σύνολο εκπαίδευσης και αποτελείται από 100 μουσικά κομμάτια(περίπου 6.5 ώρες), και ο άλλος αφορά το σύνολο δοκιμής, και αποτελείται από 50 μουσικά κομμάτια(περίπου 3.5 ώρες). Οι μέθοδοι μάθησης με επιτήρηση, θα πρέπει να χρησιμοποιήσουν το σύνολο εκπαίδευσης για εκπαίδευση, ενώ θα πρέπει να χρησιμοποιηθούν και τα δύο σύνολα για την αξιολόγηση των μεθόδων. Αποτελεί το

[6] **https://colab.research.google.com/**.

επίσημο σύνολο δεδομένων για τον διεθνή διαγωνισμό SiSEC 2018 [41], για την αξιολόγηση αλγορίθμων διαχωρισμού πηγών. Όλα τα μουσικά κομμάτια:

- είναι πλήρη σε διάρκεια, στερεοφωνικά, έχουν υποστεί μίξη με χρήση επαγγελματικών workstations ψηφιακής επεξεργασίας του ήχου(DAWs), και επομένως, είναι αντιπροσωπευτικά για σενάρια πραγματικού κόσμου.
- είναι κωδικοποιημένα στα 44.1kHz.
- είναι χωρισμένα σε τέσσερις προκαθορισμένες κατηγορίες σε μορφή STEM. Ένα STEM[7] αρχείο είναι ένα αρχείο ήχου το οποίο περιέχει ένα μουσικό κομμάτι χωρισμένο σε τέσσερα μουσικά στοιχεία: Το φωνητικό stem, το μπάσο stem, το ντράμς stem και το αρμονικό stem.
- ανήκουν σε διάφορες κατηγορίες όπως jazz, electro, metal κ.α.

Τα δεδομένα για το MUSDB18 σύνολο δεδομένων έχουν συντεθεί από αρκετές διαφορετικές πηγές:

- 100 μουσικά κομμάτια έχουν παρθεί από το DSD100 σύνολο δεδομένων.
- 46 μουσικά κομμάτια έχουν παρθεί από το MedleyDB σύνολο δεδομένων με άδεια Creative Commons (BY-NC-SA 4.0).
- 2 κομμάτια προσφέρθηκαν από την Native Instruments.
- 2 κομμάτια είναι από την καναδική ροκ μπάντα "The Easton Ellises" .

6.3 Μεθοδολογία και προτεινόμενα μοντέλα

Στο εδάφιο αυτό παρουσιάζουμε μια μέθοδο βασιζόμενη σε συνελικτικά νευρωνικά δίκτυα, με εκτίμηση χαλαρής μάσκας [40]. Αρχικά, περιγράφουμε πως το πρωτότυπο σήμα μίξης μετασχηματίζεται σε ένα σύνολο από κομμάτια φασματογραγημάτων, τα οποία χρησιμοποιούνται ως είσοδος του προτεινόμενου μοντέλου. Εν συνεχεία περιγράφουμε την αρχιτεκτονική του μοντέλου μαζί με τον ορισμό της ιδανικής δυαδικής μάσκας και της συνάρτησης κόστους που χρησιμοποιείται. Επιπλέον, συζητάμε διάφορα προβλήματα σχετικά με την υλοποίηση και τον σχεδιασμό του συνελικτικού νευρωνικού δικτύου και τέλος, περιγράφουμε πως η έξοδος του νευρωνικού δικτύου μετασχηματίζεται σε τρία ξεχωριστά σήματα, τα φωνητικά, το μπάσο και τη μουσική συνοδεία. Δίνουμε παραδείγματα μονοκαναλικών και στερεοφωνικών μοντέλων.

Αξίζει να σημειωθεί το εξής [52]. Έχει προηγουμένως αναφερθεί ότι η ανθρώπινη ακοή έχει εύρος 20Hz-20kHz(βλ. εδάφιο 4.1). Γενικά αυτό είναι το εύρος που προσπαθούμε να δουλέψουμε όταν σχεδιάζουμε διάφορα συστήματα και ασχολούμαστε με την ψηφιακή επεξεργασία σήματος του ήχου, αλλά στην πραγματικότητα είναι δύσκολα να ακούσουμε τα όρια του παραπάνω εύρους

[7] **https://www.stems-music.com/stems-faq/**.

συχνοτήτων. Συνήθως για τους πολύ νέους το πάνω όριο είναι περίπου στα 15-17kHz, 13-15kHz για τους μεσήλικες, και ακόμα χαμηλότερα για τους πιο ηλικιωμένους. Επιπρόσθετα η ομιλία έχει ένα ανώτατο όριο στα 10kHz. Από την άλλη πλευρά, οτιδήποτε κάτω από τα 160Hz ακούγεται ανεπαίσθητα. Επομένως, με δειγματοληψία στα 22.05kHz και συνεπώς μια συχνότητα αποκοπής στα 11.025kHz, δεν χάνουμε αρκετή πληροφορία ακόμη και για την περίπτωση της μουσικής. Τα περισσότερα state-of-the-art συστήματα [40], [44] κάνουνε υποδειγματοληψία από τα 44.1kHz στα 22.05kHz ή και ακόμη χαμηλότερα [48] στα 16kHz. Εμείς στη συνέχεια θα παρουσιάσουμε και μια έκδοση στα 22.05kHz και μια στα 44.1kHz.

6.3.1 Βασική μεθοδολογία

Στάδιο προεπεξεργασίας

Κατά το στάδιο της προεπεξεργασίας, δημιουργείται η είσοδος του συνελικτικού νευρωνικού δικτύου. Αρχικά, επιλέγονται αμερόληπτα κομμάτια των 30 δευτερολέπτων. Έπειτα, εξαιρούμε τα κομμάτια τα οπόια είναι μικρότερα από 30 δευτερόλεπτα. Εφαρμόζουμε τον βραχυχρόνιο μετασχηματισμό Fourier(STFT) στα κομμάτια αυτά $\mathbf{x}$ ώστε να λάβουμε το φασματογράφημα του πλάτους $\mathbf{X}$, και το φασματογράφημα της φάσης $\mathbf{pX}$. Για κάθε βήμα του γρήγορου μετασχηματισμού Fourier(FFT), χρησιμοποιούμε την συνάρτηση παραθύρου Hanning, με μέγεθος παραθύρου W τα 46.44ms, hop size 11.61ms και έναν $4\times$ παράγοντα προσθήκης μηδενικών(Zero Padding). Θέτοντας το ρυθμό δειγματοληψίας f_s στα 22.05kHz, κάθε βήμα του FFT έχει μέγεθος $N = 4096$, $W = 1024$ και $H = 256$. Αυτή η επιλογή των ρυθμίσεων στον STFT, έχει επιλεγεί με βάση τον αλγόριθμο Sinusoidal Partials Tracking [36].

Ο Sinusoidal Partials Tracking αλγόριθμος συνδέει τα φασματικά peaks σε ένα σύνολο από κομμάτια. Κάθε κομμάτι μοντελοποιεί ένα χρονικά μεταβαλλόμενο ημίτονο. Τα κομμάτια αυτά καλούνται partials όταν αντιπροσωπεύουν το ντετερμινιστικό μέρος του ηχητικού σήματος. Στην συγκεκριμένη μελέτη, το μέσο μήκος ενός φωνητικού partial, έχει βρεθεί να είναι περίπου 9 συνεχόμενα πλαίσια(Frames) και ένας $4\times$ παράγοντας προσθήκης μηδενικών, βελτιώνει την ποιότητα του διαχωρισμού στην ιδανική περίπτωση. Επομένως, μπορούμε να υποθέσουμε ότι αυτές οι ρυθμίσεις θα επιτρέψουν την χρήση αρκετών χρονικών και φασματικών στοιχείων για την σωστή εκπαίδευση του συνελικτικού νευρωνικού δικτύου.

Η είσοδος του προτεινόμενου συνελικτικού δικτύου αποτελείται από ένα στιγμιότυπο εικόνας του φασματογραφήματος του πλάτους $\mathbf{X}$, με διαστάσεις (9×2049), το οποίο είναι ένα κομμάτι φασματογραφήματος $(9 \times 256 \times 1,000)/22,050 = 104.49$ms και 11.025kHz. Κάθε φασματογράφημα επομένως είναι μήκους 9 πλαισίων και χρησιμοποιείται hop size των 8 πλαισίων που αντιστοιχεί σε χρόνο 92.88ms. Ως εκ τούτου, υπάρχει επικάλυψη του ενός πλαισίου. Χρησιμοποιώντας τις προαναφερθέντες ρυθμίσεις του δικτύου, κάθε μουσικό κομμάτι των 30 δευτερολέπτων αντιστοιχεί σε $30 \times 1000/92.88 = 323$ τεμάχια εισόδου. Για το σύνολο εκπαίδευσης υπάρχουν 100 μουσικά σήματα, και αφού αποκλείσουμε αυτά

που είναι κάτω των 30 δευτερολέπτων, απομένουν 94. Επομένως τα συνολικά τεμάχια εισόδου για εκπαίδευση είναι $323 \times 94 = 30,362$. Για να βελτιώσουμε περαιτέρω την εκπαίδευση με ομαλοποίηση, ανακατεύουμε κάθε τεμάχιο εισόδου με τυχαιότητα.

Αρχιτεκτονική δικτύου με ιδανική δυαδική μάσκα και συνάρτηση κόστους διασταυρωμένης εντροπίας

Ο παρακάτω πίνακας (6.1) δείχνει την αρχιτεκτονική του δικτύου του προτεινόμενου συνελικτικού νευρωνικού δικτύου μαζί με τις ρυθμίσεις και τον αντίστοιχο αριθμό των παραμέτρων και χαρακτηριστικών εκπαίδευσης. Η εκπαίδευση του δικτύου έγινε σε μουσική, με ταμπέλες άσσων και μηδενικών, όπου δηλώνεται αν υπάρχουν φωνητικά(με άσσο) ή όχι(με μηδενικό). Ο προκύπτων χάρτης των χαρακτηριστικών(Saliency Map), δημιουργήθηκε με καθοδηγούμενη οπισθοδιάδοση του συνελικτικού νευρωνικού δικτύου, και απεικονίζει τα φωνητικά σε επίπεδο χρονοσυχνοτικών στοιχείων.

Επίπεδο(Τύπος)	Διαστάσεις εξόδου	Αριθμός παραμέτρων
Είσοδος (Επίπεδο εισόδου)	[(None, 18441)]	0
reshape(Reshape)	(None, 2049, 9, 1)	0
conv2d(Conv2D)	(None, 2049, 9, 32)	1184
conv2d_1(Conv2D)	(None, 2049, 9, 16)	18448
max_pooling2d(MaxPooling2D)	(None, 2049, 9, 16)	0
conv2d_2(Conv2D)	(None, 2049, 9, 64)	36928
conv2d_3(Conv2D)	(None, 2049, 9, 32)	73760
max_pooling2d(MaxPooling2D)	(None, 2049, 9, 32)	0
dropout(Dropout)	(None, 2049, 9, 32)	0
flatten(Flatten)	(None, 65568)	0
dense(Dense)	(None, 2048)	134285312
dropout_1(Dropout)	(None, 2048)	0
dense_1(Dense)	(None, 512)	1049088
dense_2(Dense)	(None, 18441)	9460233
Συνολικοί παράμετροι: 144,924,953		

Πίνακας 6.1: Αρχιτεκτονική του προτεινόμενου μοντέλου μαζί με τις ρυθμίσεις και τον αντίστοιχο αριθμό των παραμέτρων και χαρακτηριστικών προς εκπαίδευση.

Στη συγκεκριμένη πειραματική μελέτη, χρησιμοποιούμε την ιδανική δυαδική μάσκα ως ταμπέλα στόχων αντί της χρήσης ασθενών ταμπελών(Weak Labels), δηλαδή ταμπελών που υποδεικνύουν αν υπάρχουν φωνητικά σε ένα χρονοσυχνοτικό ή όχι. Η ιδανική δυαδική μάσκα μπορεί επισήμως να ορισθεί ως εξής. Αν θεωρήσουμε τον $F \times T$ πίνακα $\mathbf{X}$, ο οποίος υποδηλώνει το φασματογράφημα του πλάτους, όπου F είναι ο αριθμός των συχνοτικών στοιχείων $F = (\lfloor \frac{N}{2} \rfloor + 1)$, όπου N το μέγεθος του FFT και T ο αριθμός των πλαισίων. Δοθέντος του φασματογραφήματος του πλάτους των φωνητικών $\mathbf{X_V}$ και του φασματογραφήματος

του πλάτους του σήματος μίξης $\mathbf{X}$, η ιδανική δυαδική μάσκα των φωνητικών $\mathbf{B}$, διαστάσεων $F \times T$, υπολογίζεται ως

$$B(n,t) = \begin{cases} 1, & a\nu \quad X_V(n,t) > X(n,t), \\ 0, & a\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{6.1}$$

όπου $t \in [1 : T]$ είναι ο δείκτης στοιχείου χρόνου και $n \in [1 : F]$ ο δείκτης στοιχείου συχνότητας. Η ιδανική δυαδική μάσκα της μουσικής συνοδείας σημειώνεται ως $\overline{\mathbf{B}} = |1 - \mathbf{B}|$

Ο προκύπτων πίνακας $\mathbf{B}$ αποτελεί τις ταμπέλες στόχων του νευρωνικού δικτύου. Μαζί με τις προβλέψεις του δικτύου, $Y(n,t)$, οι οποίες διαμορφώνονται από την σιγμοειδή έξοδο του τελικού επιπέδου, μπορούμε να υπολογίσουμε την δυαδική διασταυρωμένη(βλ. σχέση 3.15) για κάθε χρονο-συχνοτικό στοιχείο

$$J(n,t) = -\log \sigma((2B(n,t) - 1)Y(n,t)) \tag{6.2}$$

Το δίκτυο εκπαιδεύεται, ούτως ώστε να ελαχιστοποιηθεί η διασταυρωμένη εντροπία. Επιπρόσθετα, για να βελτιώσουμε την απόδοση του δικτύου, τα βάρη αρχικοποιήθηκαν αρχικά με το Xavier's initializer [16]. Για την περαιτέρω βελτίωση των αρχικών αυτών βαρών, εκπαιδεύσαμε το συνελικτικό νευρωνικό δίκτυο ως autoencoder, χρησιμοποιώντας τμήματα φασματογραφήματος μόνο από τα φωνητικά για 250 εποχές. Αυτά τα αρχικά βάρη μας επέτρεψαν να εκπαιδεύσουμε το δίκτυο πιο αποτελεσματικά.

Επειδή εφαρμόσαμε έναν $4\times$ παράγοντα προσθήκης μηδενικών στο πεδίο της συχνότητας κατά τη διάρκεια του υπολογισμού του STFT, θέσαμε τις διαστάσεις του συνελικτικού φίλτρου ίσες με (3×12), όπου το 3 αναπαριστά τον χρόνικό στοιχείο και το 12 το συχνοτικό στοιχείο. Η χρονική διάσταση στο επίπεδο ομαδοποίησης δεν μειώθηκε καθώς αυτό μπορεί να προσθέσει παράσιτα στο σήμα. Η συχνοτική διάσταση όμως στο επίπεδο μεγίστης ομαδοποίησης, μειώθηκε. Αυτή η διαδικασία είναι χονδρικά ανάλογη με τον υπολογισμό των MFCCs(βλ. εδάφιο 4.11), το οποίο έχει εμπειρικά αποδείξει ότι παρέχει χρήσιμα χαρακτηριστικά για εφαρμογές ταξινόμησης μουσικής [12], [19], [18]. Επιπρόσθετα, χρησιμοποιούνται η τεχνική Dropout για ομαλοποίηση με πιθανότητα 0.5 και συναρτήσεις ενεργοποίησης γραμμικών ανορθωμένων μονάδων για τα ενδιάμεσα επίπεδα.

Εκπαίδευση

Για την εκπαίδευση του δικτύου, χρησιμοποιήσαμε μέγεθος batch ίσο με 256, το οποίο αποτελεί δύναμη του 2, σύμφωνα με τις υποδείξεις του εδαφίου (3.6.2). Επιπρόσθετα χρησιμοποιήσαμε τα εξής `Keras Callbacks`[8]

- **tf.keras.callbacks.ModelCheckpoint**
  Η δωρεάν έκδοση του Colab που χρησιμοποιήσαμε, είναι συχνά ασταθής και θέτει συχνούς περιορισμούς στην διάρκεια της εκπαίδευσης. Επομένως

[8] **https://keras.io/api/callbacks/**.

συνδέσαμε το `Google Drive`[9] με το Google Colaboratory, και χρησιμοποιώντας το συγκεκριμένο Keras Callback, αποθηκεύαμε στο Google Drive με το τέλος κάθε εποχής, το καλύτερο μοντέλο με βάση την απόδοση της δυαδικής διασταυρωμένης εντροπίας στο σύνολο δοκιμής.

- **tf.keras.callbacks.CSVLogger**
  Για την εξαγωγή των τιμών σε ένα υπολογιστικό φύλλο επέκτασης .csv και την εύκολη απεικόνιση τους.

- **tf.keras.callbacks.EarlyStopping**
  Αυτό το Keras Callback, υλοποιεί την περίπτωση ομαλοποίησης που περιγράφηκε στο εδάφιο (3.5.3). Εφαρμόζουμε δηλαδή πρόωρη παύση στην εκπαίδευση, όταν το μοντέλο μας αρχίζει και κάνει overfit. Η μετρική με βάση την οποία υλοποιήσαμε την πρόωρη παύση είναι η δυαδική διασταυρωμένη εντροπία στο σύνολο δοκιμής, και βάλαμε ένα όριο αναμονής των 100 εποχών. Αν δηλαδή για 100 εποχές δεν υπάρξει βελτίωση, επιστρέφεται το μοντέλο με τα καλύτερα βάρη μέχρι εκείνο το σημείο.

- **tf.keras.callbacks.TensorBoard**
  Για απεικόνιση καμπυλών και χρήσιμων στοιχείων για την εκπαίδευση.

Χρησιμοποιούμε επιπρόσθετα, τον αλγόριθμο βελτιστοποίησης ADAM(βλ. εδάφιο 3.6.6) με τις προκαθορισμένες ρυθμίσεις του Keras, και με ρυθμό μάθησης $1e-4$.

Μετεπεξεργασία

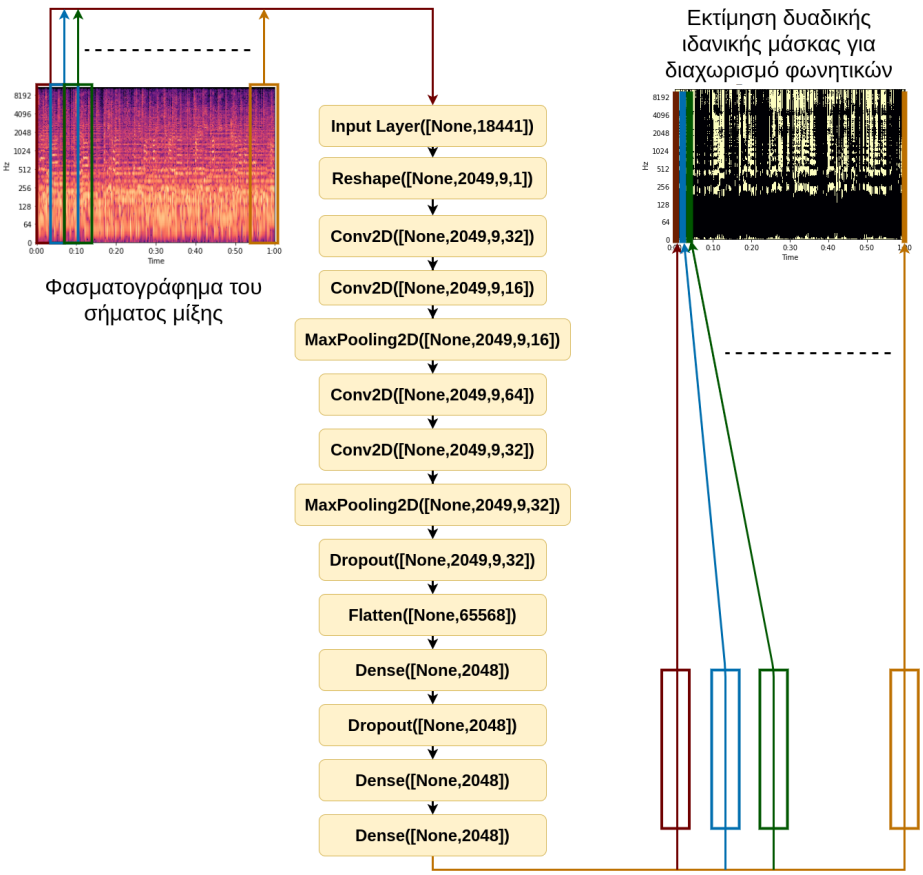
Στο στάδιο αυτό γίνεται η μετατροπή της εκτιμώμενης από το δίκτυο χαλαρής μάσκας, σε δύο ηχητικά σήματα, τα φωνητικά και τη μουσική συνοδεία. Για να το επιτύχουμε αυτό, η έξοδος του συνελικτικού νευρωνικού δικτύου μετασχηματίζεται από $(1 \times 18,441)$ σε $(9 \times 2,049)$ για να επανακατασκευάσουμε τα 9 πλαίσια. Η εκτιμώμενη έξοδος του δικτύου, πριν την προεπεξεργασία, θεωρείται ότι η χαλαρή μάσκα του εκτιμώμενου φασματογραφήματος των φωνητικών, το οποίο σημαίνει ότι κάθε χρονο-συχνοτικό στοιχείο βρίσκεται εντός του εύρους $[0, 1]$. Αυτή η υπόθεση δικαιολογείται από το γεγονός ότι η ιδανική δυαδική μάσκα, επιλέχθηκε ως η ταμπέλα στόχος κατά τη διάρκεια της εκπαίδευσης και επομένως χρησιμοποιήθηκε για τον υπολογισμό της διασταυρωμένης εντροπίας με την σιγμοειδή συνάρτηση ενεργοποίησης στην έξοδο του δικτύου. Κάθε τιμή κάθε χρονο-συχνοτικού στοιχείου στην χαλαρή μάσκα μπορεί να ερμηνευθεί ως η πιθανότητα ϵ, ότι το χρονο-συχνοτικό αυτό στοιχείο ανήκει στα φωνητικά. Για να βελτιώσουμε περαιτέρω την ποιότητα του διαχωρισμού, θέτουμε το ϵ ίσο με το μηδέν, όταν $\epsilon < \theta$, και ίσο με τη μονάδα σε αντίθετη περίπτωση, για $\theta = 0.15$.

Η αρχιτεκτονική του νευρωνικού δικτύου που περιγράφεται παραπάνω, δέχεται 9 ηχητικά πλαίσια ως είσοδο. Για να εκτιμήσουμε μια μονή χαλαρή μάσκα $\mathbf{M_V}$ για τον διαχωρισμό των φωνητικών από την ηχητική ηχογράφηση, ακολουθούμε τα εξής βήματα. Πρώτον, επικαλυπτόμενα κομμάτια φασματογραφήματος(καθένα

[9] **https://drive.google.com/**.

με μήκος 9 πλαισίων), τροφοδοτούνται στο δίκτυο με hop size του ενός πλαισίου. Τα μεσαία πλαίσια κάθε εκτιμώμενης χαλαρής μάσκας συγχωνεύονται για την δημιουργία της $\mathbf{M_V}$. Η χαλαρή μάσκα $\mathbf{M_S}$ για την λήψη της μουσικής συνοδείας από μια ηχητική ηχογράφηση, μπορεί να υπολογιστεί από το $1 - \mathbf{M_V}$.

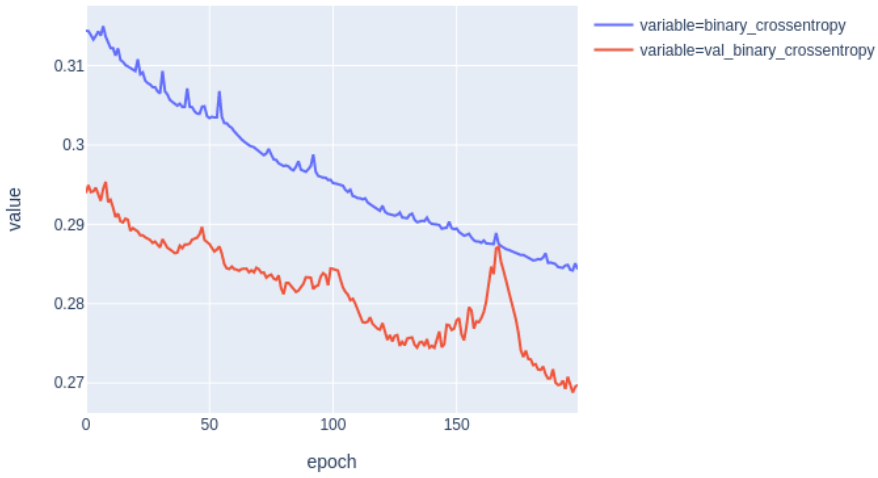
Τελικά, τα απομονωμένα φωνητικά λαμβάνονται υπολογίζοντας τον αντίστροφο βραχυχρόνιο μετασχηματισμό Fourier(Inverse STFT), του πολλαπλασιασμού στοιχείο προς στοιχείο μεταξύ του εκτιμώμενου πλάτους του φασματογραφήματος των φωνητικών $\mathbf{M_V}$, του πλάτους του φασματογραφήματος του σήματος μίξης $\mathbf{X}$ και της φάσης του φασματογραφήματος του σήματος μίξης $\mathbf{pX}$. Ομοίως, μπορούμε να λάβουμε το σήμα της απομονωμένης μουσικής συνοδείας, υπολογίζοντας τον αντίστροφο βραχυχρόνιο μετασχηματισμό Fourier, του πολλαπλασιασμού στοιχείο προς στοιχείο μεταξύ του πλάτους του φασματογραφήματος της μουσικής συνοδείας $\mathbf{M_S}$, και του πλάτους του φασματογραφήματος του σήματος μίξης $\mathbf{X}$, χρησιμοποιώντας τη φάση του φασματογραφήματος του σήματος μίξης $\mathbf{pX}$.



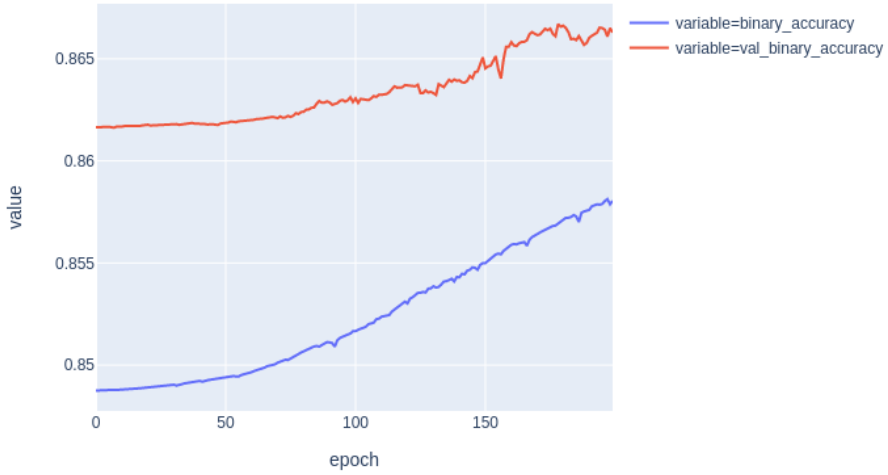
Σχήμα 6.1: Απεικόνιση της αρχιτεκτονικής για την εκτίμηση μια δυαδικής ιδανικής μάσκας. Εκτιμάται πρώτα η χαλαρή μάσκα και έπειτα με όριο $\theta = 0.15$, εξάγεται η ιδανική δυαδική μάσκα. Τα φασματογραφήματα πάρθηκαν από το μουσικό κομμάτι "REM - Losing My Religion".

6.3.2 Μονοφωνική υλοποίηση εξαγωγής φωνητικών στα 22.05kHz

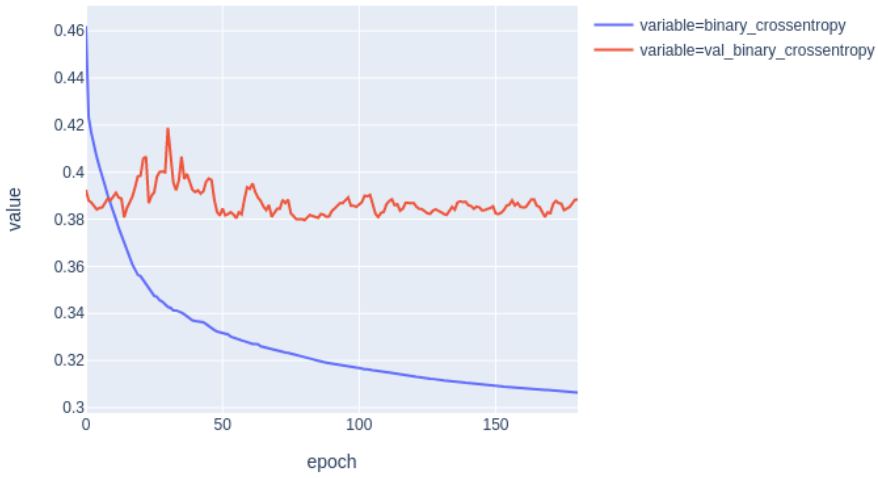
Χρησιμοποιώντας την μεθοδολογία που περιγράφηκε στο εδάφιο (6.3.1), παρουσιάζουμε τα αποτελέσματα από την εκπαίδευση του δικτύου.



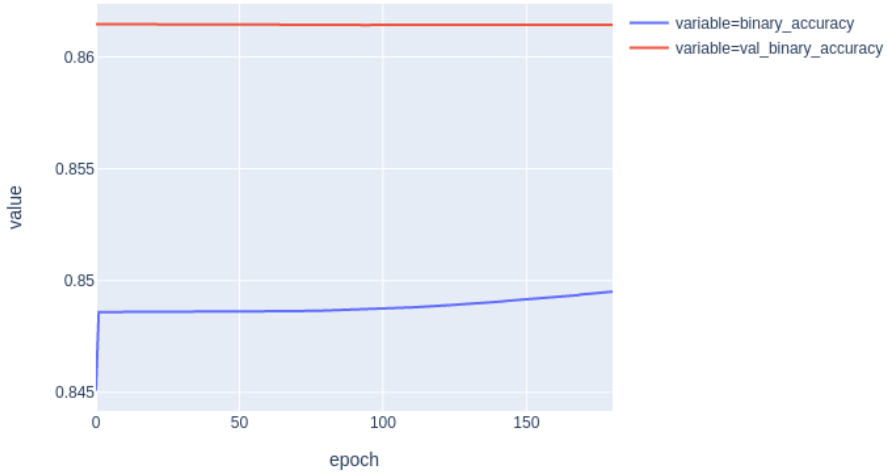
Σχήμα 6.2: Απεικόνιση της εξέλιξης της συνάρτησης κόστους για την εκπαίδευση του autoencoder. Η χαμηλότερη τιμή της συνάρτησης κόστους ήταν 0.26879.



Σχήμα 6.3: Απεικόνιση της εξέλιξης της ακρίβειας για την εκπαίδευση του autoencoder. Η υψηλότερη τιμή της ακρίβειας στο σύνολο δοκιμής ήταν 0.86670.



Σχήμα 6.4: Απεικόνιση της εξέλιξης της συνάρτησης κόστους για την εκπαίδευση του δικτύου. Η χαμηλότερη τιμή της συνάρτησης κόστους ήταν 0.37965.

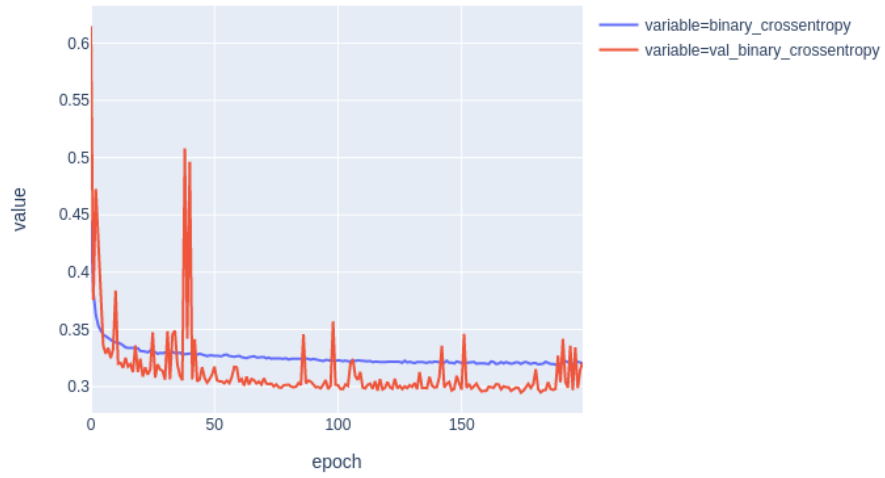


Σχήμα 6.5: Απεικόνιση της εξέλιξης της ακρίβειας για την εκπαίδευση του δικτύου. Η υψηλότερη τιμή της ακρίβειας στο σύνολο δοκιμής ήταν 0.86146.

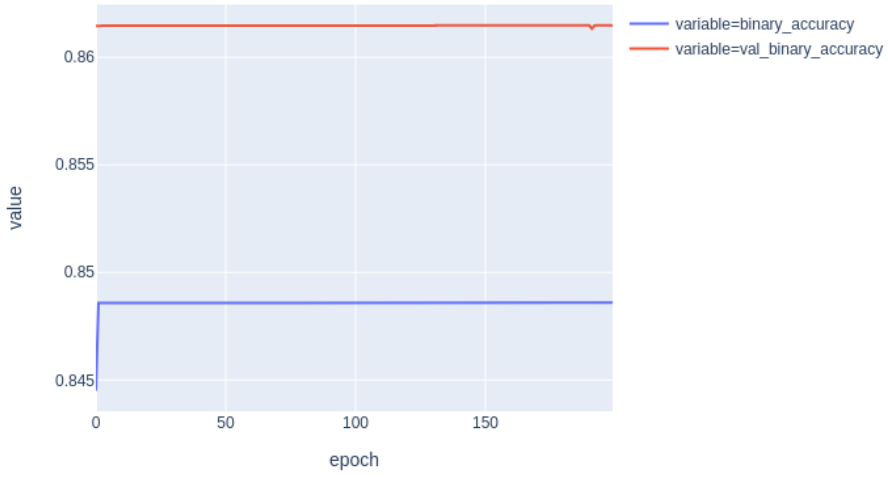
Μια άλλη τεχνική που χρησιμοποιείται συχνά, για την ταχύτερη σύγκλιση του μοντέλου, είναι το Batch Normalization(βλ. εδάφιο 3.6.7). Δοκιμάζοντας την παραπάνω τεχνική στο δίκτυο μας, δεν παρατηρήσαμε κάποια βελτίωση στον χρόνο εκτέλεσης, αλλά τουναντίον, είχαμε και χειρότερη ποιότητα διαχωρισμού πηγών. Στον πίνακα (6.2), παρουσιάζουμε το μοντέλο με χρήση Batch Normalization και στα σχήματα (6.6) και (6.8) παρουσιάζουμε τα αποτελέσματα.

Επίπεδο(Τύπος)	Διαστάσεις εξόδου	Αριθμός παραμέτρων
Είσοδος (Επίπεδο εισόδου)	[(None, 18441)]	0
reshape(Reshape)	(None, 2049, 9, 1)	0
conv2d(Conv2D)	(None, 2049, 9, 32)	1184
batch_normalization(BatchNormalization)	(None, 2049, 9, 32)	128
conv2d_1(Conv2D)	(None, 2049, 9, 16)	18448
batch_normalization_1(BatchNormalization)	(None, 2049, 9, 32)	64
max_pooling2d(MaxPooling2D)	(None, 2049, 9, 16)	0
conv2d_2(Conv2D)	(None, 2049, 9, 64)	36928
batch_normalization_2(BatchNormalization)	(None, 2049, 9, 64)	256
conv2d_3(Conv2D)	(None, 2049, 9, 32)	73760
batch_normalization_3(BatchNormalization)	(None, 2049, 9, 32)	128
max_pooling2d_1(MaxPooling2D)	(None, 2049, 9, 32)	0
dropout(Dropout)	(None, 2049, 9, 32)	0
flatten(Flatten)	(None, 65568)	0
dense(Dense)	(None, 2048)	134285312
dropout_1(Dropout)	(None, 2048)	0
dense_1(Dense)	(None, 512)	1049088
dense_2(Dense)	(None, 18441)	9460233
Συνολικοί παράμετροι: 144,925,529		

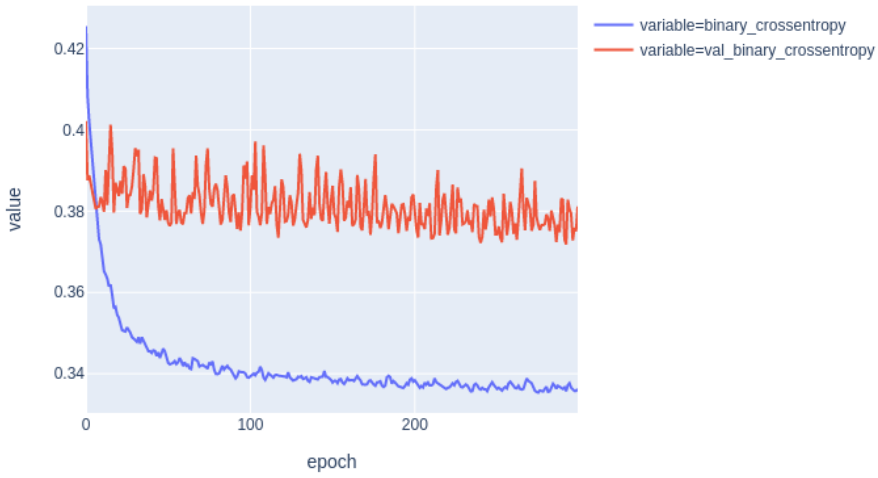
Πίνακας 6.2: Αλλαγή του προτεινόμενου μοντέλου με χρήση Batch Normalization και L2 ποινής νόρμας.



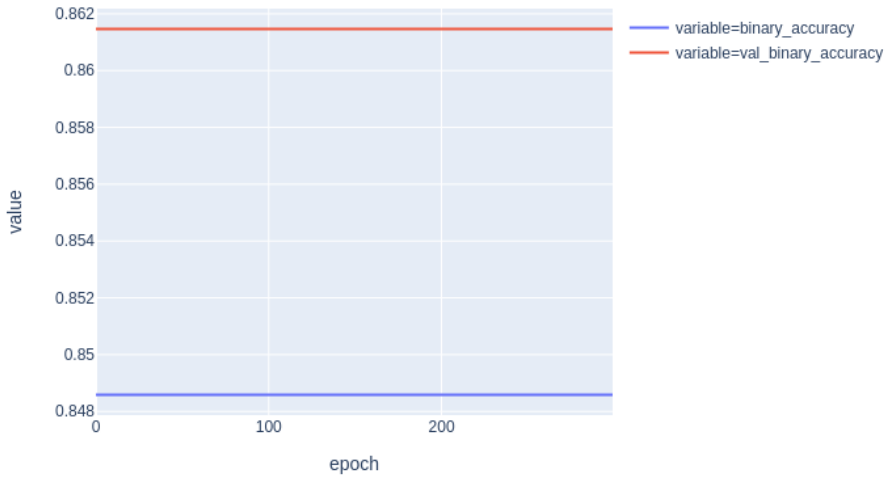
Σχήμα 6.6: Απεικόνιση της εξέλιξης της συνάρτησης κόστους για την εκπαίδευση του autoencoder του δικτύου του πίνακα (6.2). Η χαμηλότερη τιμή της συνάρτησης κόστους στο σύνολο δοκιμής ήταν 0.29487.



Σχήμα 6.7: Απεικόνιση της εξέλιξης της ακρίβειας για την εκπαίδευση του δικτύου. Η υψηλότερη τιμή της ακρίβειας στο σύνολο δοκιμής ήταν 0.86150.



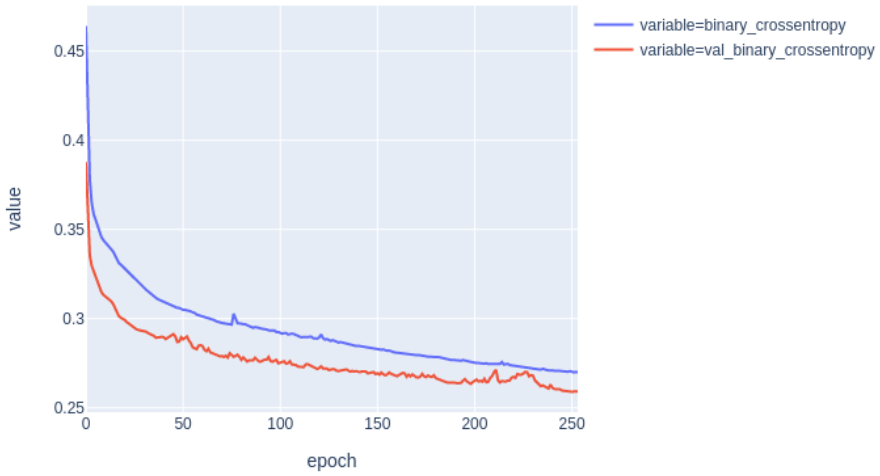
Σχήμα 6.8: Απεικόνιση της εξέλιξης της συνάρτησης κόστους για την εκπαίδευση του δικτύου του πίνακα (6.2). Η χαμηλότερη τιμή της συνάρτησης κόστους στο σύνολο δοκιμής ήταν 0.37181.



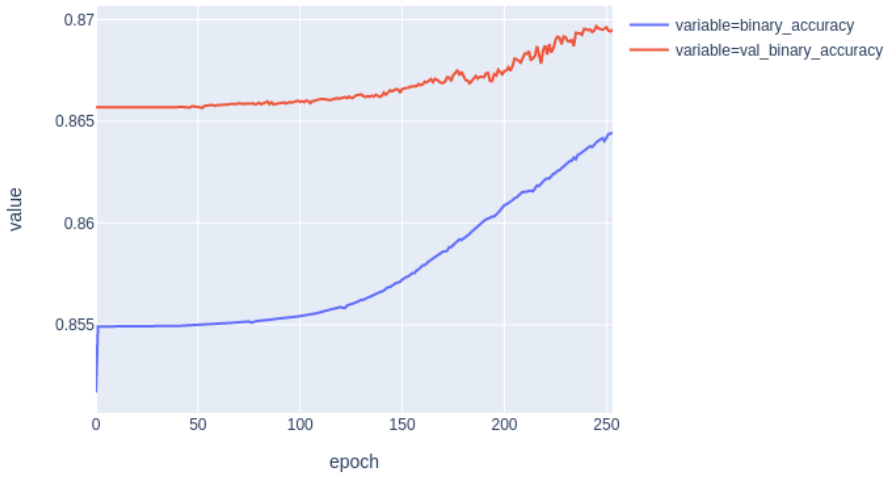
Σχήμα 6.9: Απεικόνιση της εξέλιξης της ακρίβειας για την εκπαίδευση του δικτύου. Η υψηλότερη τιμή της ακρίβειας στο σύνολο δοκιμής ήταν 0.86146.

6.3.3 Στερεοφωνικές υλοποιήσεις για εξαγωγή φωνητικών στα 22.05kHz

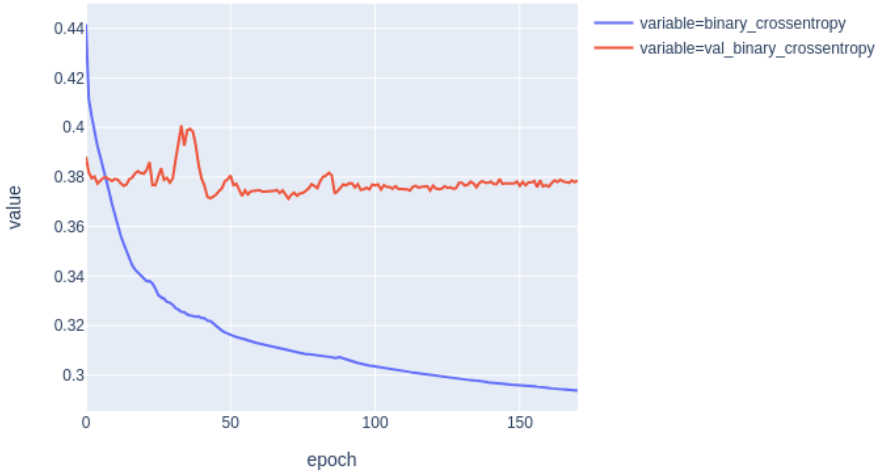
- **Υλοποίηση με ανακάτεμα των δύο καναλιών**
  Στην συγκεκριμένη υλοποίηση, χρησιμοποιήσαμε εναλλάξ πληροφορία από το αριστερό και δεξιό κανάλι. Για να κρατήσουμε τα μεγέθη των πινάκων ίδια όπως στην βασική μεθοδολογία του εδαφίου (6.3.1), λάβαμε τις περιττές θέσεις από το αριστερό κανάλι και τις άρτιες θέσεις από το δεξιό κανάλι. Το δίκτυο που χρησιμοποιήσαμε είναι αυτό του πίνακα (6.1). Οι ρυθμίσεις του STFT είναι όπως στην βασική μεθοδολογία.



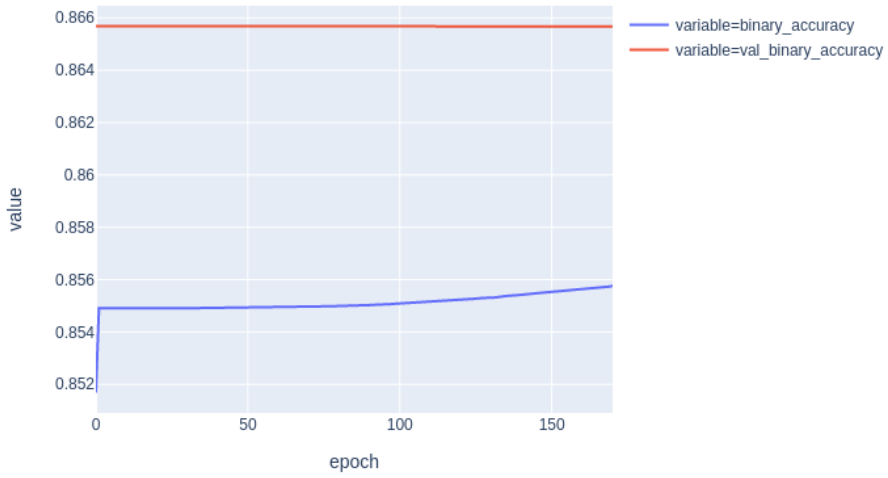
Σχήμα 6.10: Απεικόνιση της εξέλιξης της συνάρτησης κόστους για την εκπαίδευση του autoencoder. Η χαμηλότερη τιμή της συνάρτησης κόστους στο σύνολο δοκιμής ήταν 0.25898.



Σχήμα 6.11: Απεικόνιση της εξέλιξης της ακρίβειας για την εκπαίδευση του autoencoder. Η υψηλότερη τιμή της ακρίβειας στο σύνολο δοκιμής ήταν 0.86968.



Σχήμα 6.12: Απεικόνιση της εξέλιξης της συνάρτησης κόστους για την εκπαίδευση του δικτύου. Η χαμηλότερη τιμή της συνάρτησης κόστους στο σύνολο δοκιμής ήταν 0.37126.

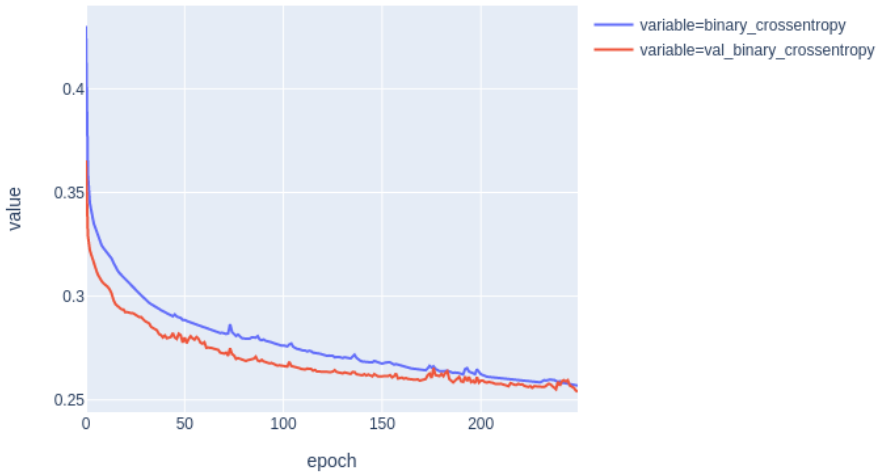


Σχήμα 6.13: Απεικόνιση της εξέλιξης της ακρίβειας για την εκπαίδευση του δικτύου. Η υψηλότερη τιμή της ακρίβειας στο σύνολο δοκιμής ήταν 0.86569.

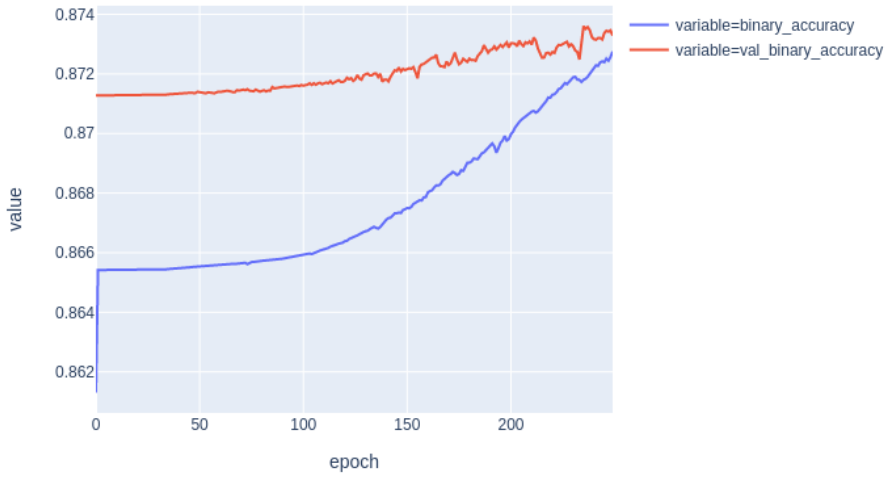
- **Υλοποίηση με χρήση πλήρους πληροφορίας και από τα δύο κανάλια**
  Για την συγκεκριμένη υλοποίηση, η βασική μεθοδολογία διαφοροποιείται ως προς το μέγεθος των πινάκων και ως προς την αρχιτεκτονική του δικτύου. Πιο συγκεκριμένα, χρησιμοποιούμε την πλήρη πληροφορία και από τα δύο κανάλια, δηλαδή το δίκτυο πλέον δέχεται 2 φασματογραφήματα ως είσοδο. Τις ταμπέλες τις δημιουργούμε ομοίως με την υλοποίηση της βασικής μεθοδολογίας του εδαφίου (6.3.1), δηλαδή τις εξάγουμε ως μέσο όρο των δύο καναλιών και στην έξοδο έχουμε την εκτίμηση μιας μονής χαλαρής μάσκας, την οποία στη συνέχεια, με κατάλληλη παράμετρο ορίου θ, την μετατρέπουμε σε δυαδική μάσκα. Οι ρυθμίσεις του STFT είναι όπως στη βασική μεθοδολογία.

Επίπεδο(Τύπος)	Διαστάσεις εξόδου	Αριθμός παραμέτρων
Είσοδος (Επίπεδο εισόδου)	[(None, 18441,2)]	0
reshape(Reshape)	(None, 2049, 9, 2)	0
conv2d(Conv2D)	(None, 2049, 9, 32)	2336
conv2d_1(Conv2D)	(None, 2049, 9, 16)	18448
max_pooling2d(MaxPooling2D)	(None, 2049, 9, 16)	0
conv2d_2(Conv2D)	(None, 2049, 9, 64)	36928
conv2d_3(Conv2D)	(None, 2049, 9, 32)	73760
max_pooling2d_1(MaxPooling2D)	(None, 2049, 9, 32)	0
dropout(Dropout)	(None, 2049, 9, 32)	0
flatten(Flatten)	(None, 65568)	0
dense(Dense)	(None, 2048)	134285312
dropout_1(Dropout)	(None, 2048)	0
dense_1(Dense)	(None, 512)	1049088
dense_2(Dense)	(None, 18441)	9460233
Συνολικοί παράμετροι: 144,926,105		

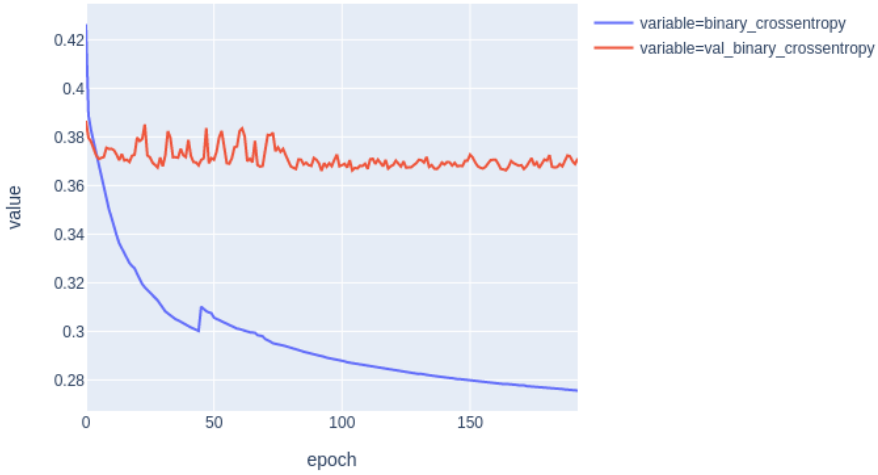
Πίνακας 6.3: Αρχιτεκτονική του προτεινόμενου στερεοφωνικού μοντέλου χρησιμοποιώντας πλήρη πληροφορία και από τα δύο κανάλια. Απεικονίζονται οι ρυθμίσεις και ο αντίστοιχος αριθμός των παραμέτρων και χαρακτηριστικών προς εκπαίδευση.



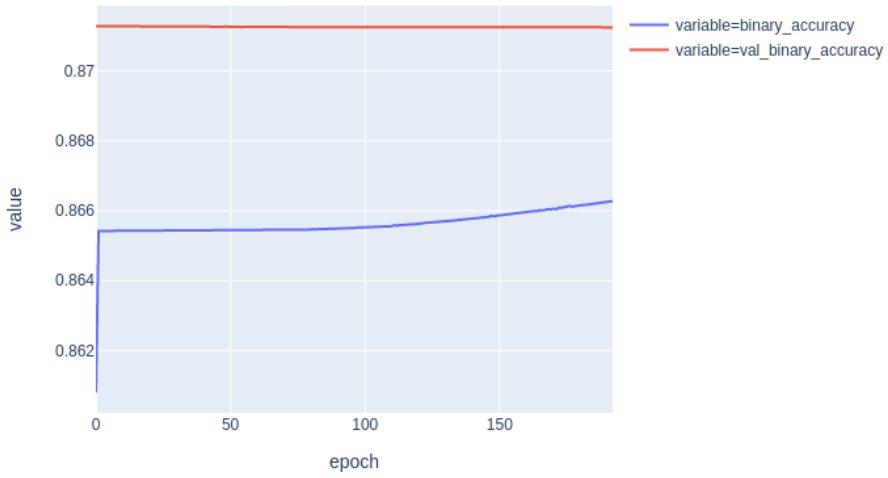
Σχήμα 6.14: Απεικόνιση της εξέλιξης της συνάρτησης κόστους για την εκπαίδευση του autoencoder. Η χαμηλότερη τιμή της συνάρτησης κόστους στο σύνολο δοκιμής ήταν 0.25394.



Σχήμα 6.15: Απεικόνιση της εξέλιξης της ακρίβειας για την εκπαίδευση του autoencoder. Η υψηλότερη τιμή της ακρίβειας στο σύνολο δοκιμής ήταν 0.87360.



Σχήμα 6.16: Απεικόνιση της εξέλιξης της συνάρτησης κόστους για την εκπαίδευση του δικτύου. Η χαμηλότερη τιμή της συνάρτησης κόστους στο σύνολο δοκιμής ήταν 0.36604.



Σχήμα 6.17: Απεικόνιση της εξέλιξης της ακρίβειας για την εκπαίδευση του δικτύου. Η υψηλότερη τιμή της ακρίβειας στο σύνολο δοκιμής ήταν 0.87127.

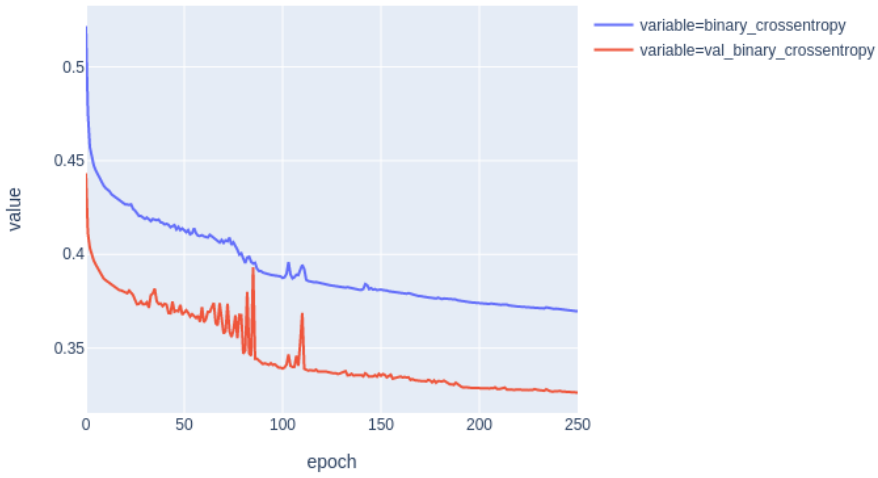
6.3.4 Μονοφωνική υλοποιήση εξαγωγής φωνητικών υψηλής ποιότητας στα 44.1kHz

Για την συγκεκριμένη υλοποίηση, ακολουθούμε ξανά την μεθοδολογία περιγράφηκε στο εδάφιο (6.3.1) κάνοντας κάποιες αλλαγές στις ρυθμίσεις του βραχυχρόνιου μετσχηματισμού Fourier. Συγκεκριμένα, με μήκος FFT $N = 4096$ και δειγματοληψία στα 44.1kHz, για να κρατήσουμε το μέγεθος του παραθύρου ίδιο στα 46.44ms, πρέπει να επιλέξουμε διακριτό μέγεθος παραθύρου

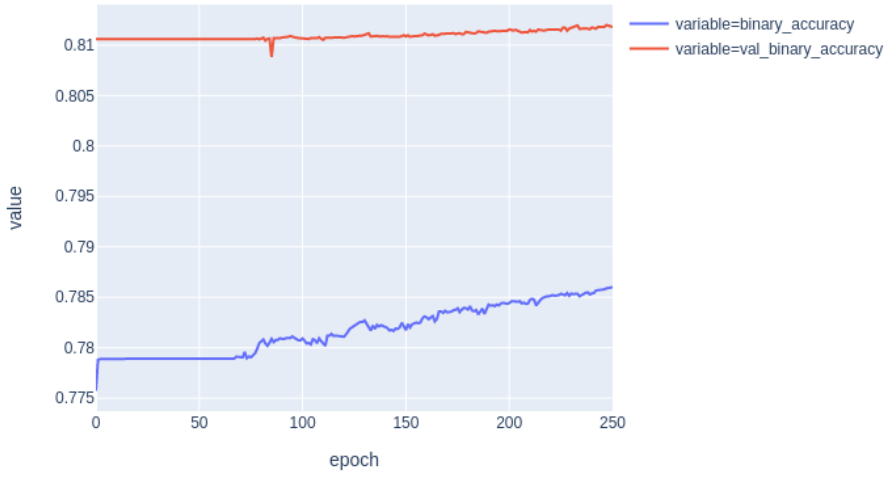
$$W = 46.44\text{ms} \cdot 44.1\text{kHz} = \lfloor 2048.004 \rfloor = 2048$$

Ομοίως το hop size για να διατηρηθεί στα 11.61ms επιλέγεται ως εξής

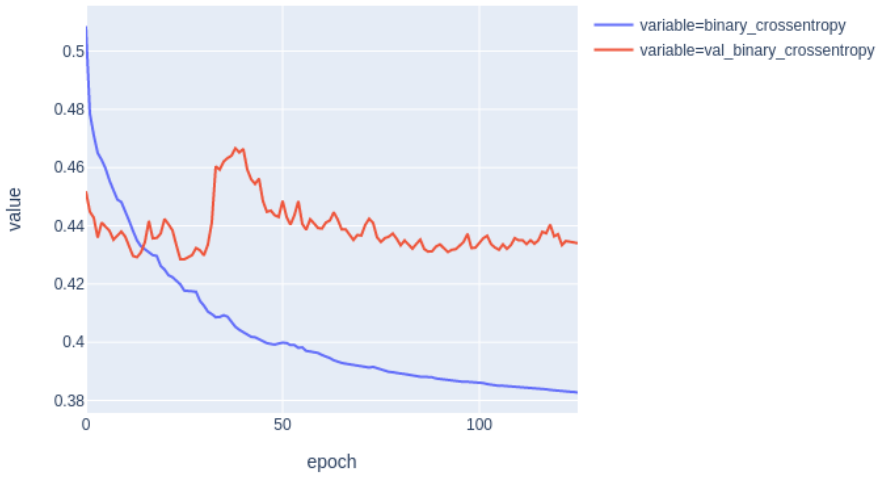
$$H = 46.44\text{ms} \cdot 44.1\text{kHz} = \lfloor 512.001 \rfloor = 512$$



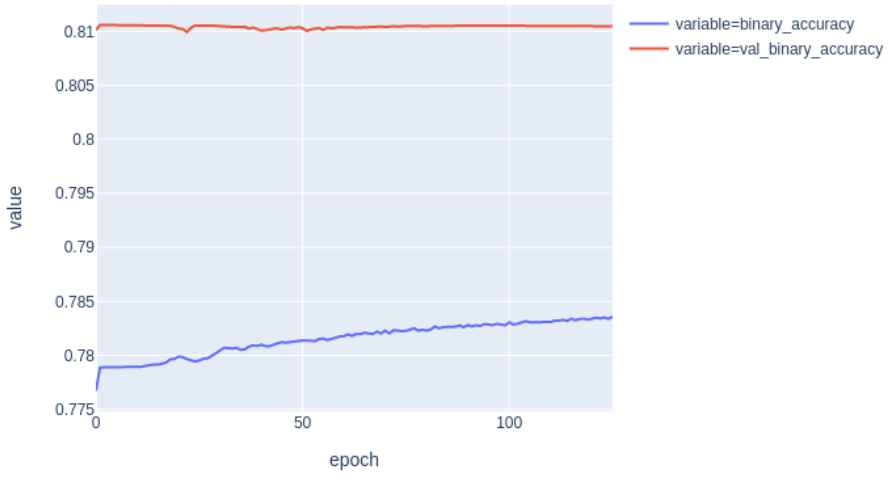
Σχήμα 6.18: Απεικόνιση της εξέλιξης της συνάρτησης κόστους για την εκπαίδευση του autoencoder. Η χαμηλότερη τιμή της συνάρτησης κόστους στο σύνολο δοκιμής ήταν 0.33372.



Σχήμα 6.19: Απεικόνιση της εξέλιξης της ακρίβειας για την εκπαίδευση του autoencoder. Η υψηλότερη τιμή της ακρίβειας στο σύνολο δοκιμής ήταν 0.81200.



Σχήμα 6.20: Απεικόνιση της εξέλιξης της συνάρτησης κόστους για την εκπαίδευση του δικτύου. Η χαμηλότερη τιμή της συνάρτησης κόστους στο σύνολο δοκιμής ήταν 0.42847.

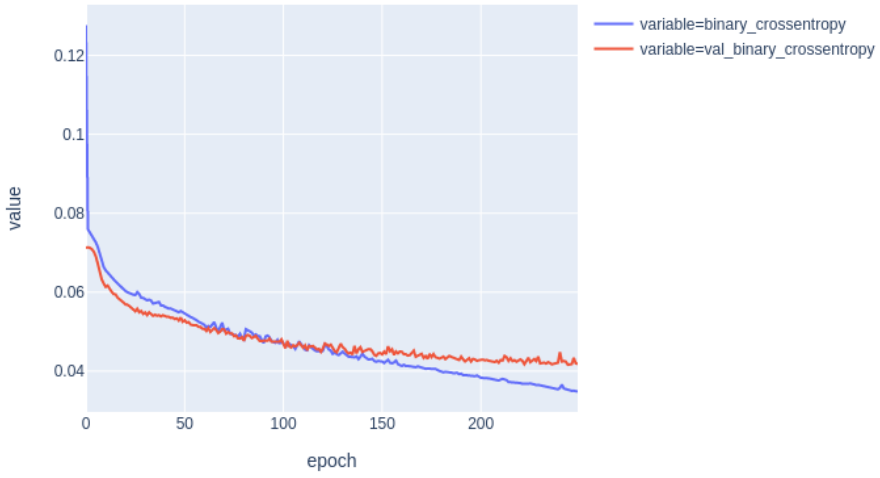


Σχήμα 6.21: Απεικόνιση της εξέλιξης της ακρίβειας για την εκπαίδευση του δικτύου. Η υψηλότερη τιμή της ακρίβειας στο σύνολο δοκιμής ήταν 0.81059.

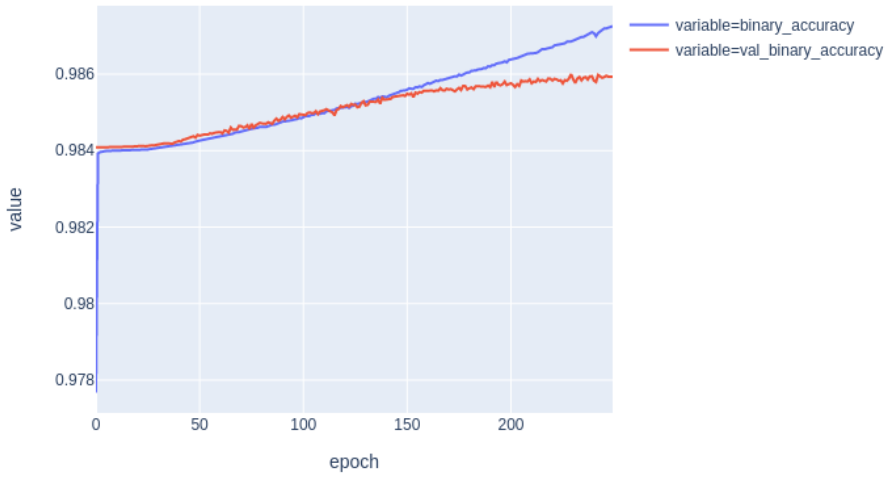
6.3.5 Μονοφωνική υλοποίηση εξαγωγής μπάσου στα 22.05kHz

Στην συγκεκριμένη υλοποίηση, ακολουθήσαμε την βασική μεθοδολογία του εδαφίου (6.3.1), με μόνη διαφοροποίηση τον τρόπο δημιουργίας της δυαδικής ιδανικής μάσκας. Συγκεκριμένα, δοθέντος του φασματογραφήματος του πλάτους του σήματος μίξης $\mathbf{X}$, και του φασματογραφήματος του πλάτους του μπάσου $\mathbf{X_B}$, δημιουργούμε την ιδανική δυαδική μάσκα ως εξής:

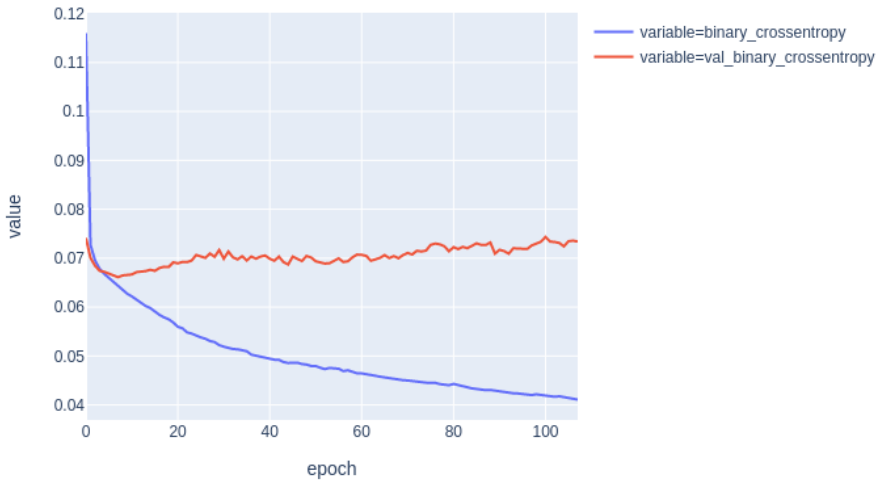
$$B(n,t) = \begin{cases} 1, & \alpha\nu \quad X_B(n,t) > X(n,t), \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{6.3}$$



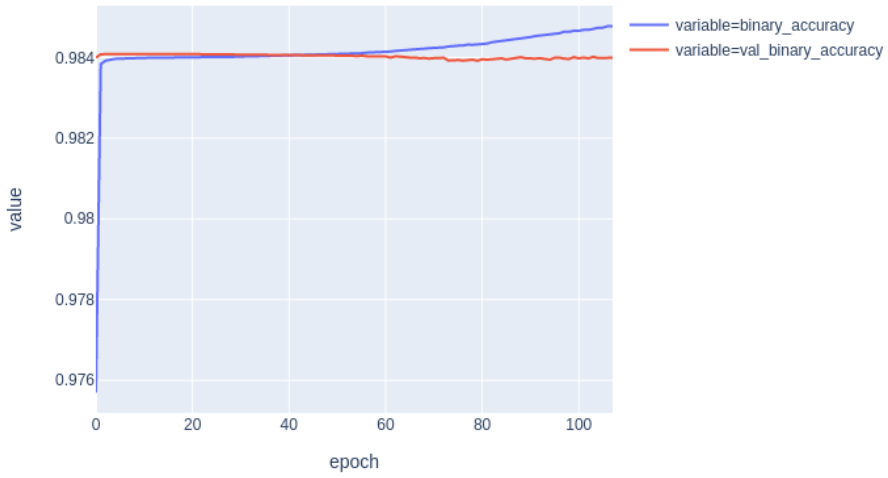
Σχήμα 6.22: Απεικόνιση της εξέλιξης της συνάρτησης κόστους για την εκπαίδευση του autoencoder. Η χαμηλότερη τιμή της συνάρτησης κόστους στο σύνολο δοκιμής ήταν 0.04160.



Σχήμα 6.23: Απεικόνιση της εξέλιξης της ακρίβειας για την εκπαίδευση του autoencoder. Η υψηλότερη τιμή της ακρίβειας στο σύνολο δοκιμής ήταν 0.98600.



Σχήμα 6.24: Απεικόνιση της εξέλιξης της συνάρτησης κόστους για την εκπαίδευση του δικτύου. Η χαμηλότερη τιμή της συνάρτησης κόστους στο σύνολο δοκιμής ήταν 0.06611.



Σχήμα 6.25: Απεικόνιση της εξέλιξης της ακρίβειας για την εκπαίδευση του δικτύου. Η υψηλότερη τιμή της ακρίβειας στο σύνολο δοκιμής ήταν 0.98408.

6.4 Ανάλυση των αποτελεσμάτων

Όλες μας οι υλοποιήσεις βασίστηκαν στη βασική μεθοδολογία του εδαφίου (6.3.1), με μερικές τροποποιήσεις ανάλογα την περίπτωση. Καταλήγουμε στα εξής συμπεράσματα

- Τα καλύτερα αποτελέσματα για εξαγωγή φωνητικών, επιτυγχάνονται από την μονοφωνική υλοποίηση στα 22.05kHz(βλ. εδάφιο 6.3.2) και από την στερεοφωνική υλοποίηση στα 22.05kHz, με χρήση πλήρους πληροφορίας και από τα δύο κανάλια (βλ. εδάφιο 6.3.3).
- Η χρήση Batch Normalization στην περίπτωση μας οδήγησε σε σημαντικό jitter στην καμπύλη κόστους του συνόλου δοκιμής. Τα αποτελέσματα ηχητικά δεν ήταν καλα.
- Η μονοφωνική υλοποίηση υψηλής ποιότητας στα 44.1kHz(βλ. εδάφιο 6.3.4), οδήγησε σε σημαντική υπερπροσαρμογή σε σύγκριση με την αντίστοιχη υλοποίηση στα 22.05kHz.

Τέλος, ο τρόπος με τον οποίο δημιουργείται η ιδανική δυαδική μάσκα παίζει σημαντικό ρόλο στην απόδοση του δικτύου και στην ποιότητα του διαχωρισμού. Δοκιμάσαμε τρεις διαφορετικούς τρόπους δημιουργίας της ιδανικής δυαδικής μάσκας.

- Πρώτος τρόπος

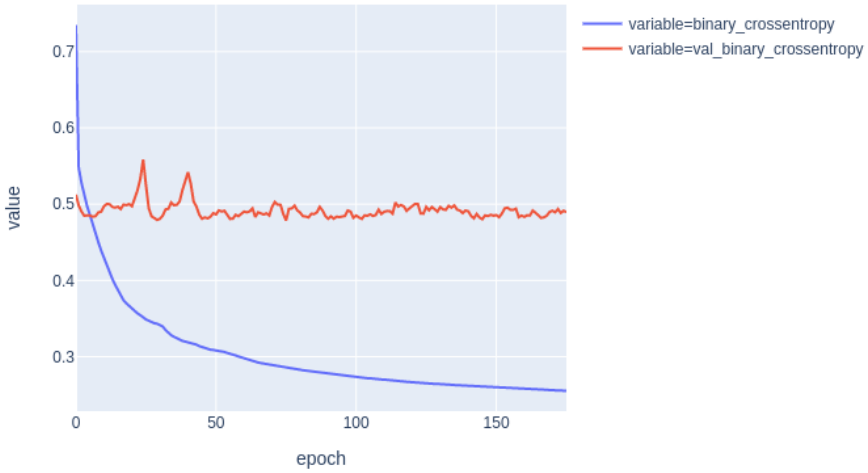
$$B(n,t) = \begin{cases} 1, & \alpha\nu \quad \dfrac{X_V(n,t)}{X(n,t)} > 0.15, \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{6.4}$$

Στην περίπτωση αυτή, είχαμε καλά αποτελέσματα μόνο στην περίπτωση της μονοφωνικής υλοποίησης εξαγωγής φωνητικών και μπάσου στα 22.05kHz. Στις υπόλοιπες υλοποιήσεις είχαμε σημαντική υπερπροσαρμογή.

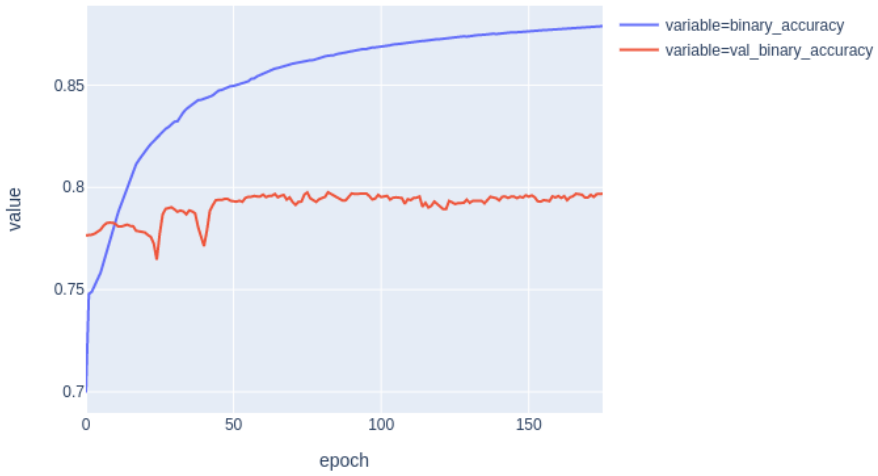
- Δεύτερος τρόπος

$$B(n,t) = \begin{cases} 1, & \alpha\nu \quad X_V(n,t) > X_S(n,t), \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{6.5}$$

όπου $\mathbf{X_S}$ το φασματογράφημα του πλάτους της μουσικής συνοδείας. Στην περίπτωση αυτή είχαμε σημαντική υπερπροσαρμογή σε όλες τις υλοποιήσεις και τα ηχητικά αποτελέσματα δεν ήταν καλά. Παρακάτω παρουσιάζουμε τα αποτελέσματα από της μονοφωνικής υλοποίησης εξαγωγής φωνητικών στα 22.05kHz.



Σχήμα 6.26: Απεικόνιση της εξέλιξης της συνάρτησης κόστους για την εκπαίδευση του δικτύου της μονοφωνικής υλοποίησης εξαγωγής φωνητικών στα 22.05kHz. Η χαμηλότερη τιμή της συνάρτησης κόστους στο σύνολο δοκιμής ήταν 0.47861.



Σχήμα 6.27: Απεικόνιση της εξέλιξης της ακρίβειας για την εκπαίδευση του δικτύου της μονοφωνικής υλοποίησης εξαγωγής φωνητικών στα 22.05kHz. Η υψηλότερη τιμή της ακρίβειας στο σύνολο δοκιμής ήταν 0.79762.

- Τρίτος τρόπος

$$B(n,t) = \begin{cases} 1, & \alpha\nu \quad X_V(n,t) > X(n,t), \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{6.6}$$

Είναι ο τρόπος της βασικής μεθοδολογίας με την οποία παρουσιάσαμε τα αποτελέσματα στα προηγούμενα εδάφια. Τα καλύτερα αποτελέσματα τα λάβαμε με τον τρίτο τρόπο.

Κεφάλαιο 7

Κατευθύνσεις μελλοντικής έρευνας

Με βάση τα αποτέσματα των διάφορων αρχιτεκτονικών και συνόλων δεδομένων που υλοποιήσαμε, καταλήγουμε στα παρακάτω συμπεράσματα, και επιπρόσθετα προτείνουμε τρόπους βελτίωσης των διαφόρων μεθόδων διαχωρισμού ηχητικών πηγών, αλλά και βελτίωση των υπαρχόντων συνόλων δεδομένων.

7.1 Προτάσεις βελτίωσης των υπαρχόντων συνόλων δεδομένων

Στα διάφορα μοντέλα που υλοποιήσαμε, παρατηρήσαμε καλύτερη απόδοση σε κάποια είδη μουσικής από κάποια άλλα. Παραδείγματος χάριν, τα rock τραγούδια και οι pop μπαλάντες, διαχωρίζονται πιο αποτελεσματικά από τα house τραγούδια. Επιπλέον, είναι δεδομένο ότι τα state-of-the-art συστήματα διαχωρισμού ηχητικών πηγών που πετυχαίνουν εξαιρετικές αποδόσεις, οφείλονται περισσότερο στην ύπαρξη μεγαλύτερων συνόλων δεδομένων και όχι τόσο στις διαφοροποιήσεις των αρχιτεκτονικών. Θεωρούμε επομένως ότι πρέπει:

- Να προστεθεί μεγαλύτερη ποικιλία ειδών μουσικών κομματιών στα διαθέσιμα σύνολα δεδομένων.
- Να προστεθούν πολλά περισσότερα μουσικά κομμάτια σε μορφή STEM, για μάθηση με επιτήρηση.
- Να γίνεται όσο το δυνατόν καλύτερος φασματικός διαχωρισμός μεταξύ των φωνητικών, του μπάσου, των ντραμς κ.λπ. κατά το στάδιο της επαγγελματικής μίξης σε στούντιο. Έτσι θα γίνει πιο αποδοτική η μάθηση με επιτήρηση.

Τέλος, όπως αναφέραμε ξανά στο εδάφιο (3.5.5), οι ήδη υπάρχουσες μέθοδοι επάυξησης δεδομένων, για μουσικά σύνολα δεδομένων, αποφέρουν μικρή βελτίωση

στην απόδοση του διαχωρισμού. Θα πρέπει να βρεθούν τρόποι πιο αποτελεσματικής επαύξησης των μουσικών συνόλων δεδομένων, διότι στην πράξη τα διαθέσιμα δεδομένα για μάθηση με επιτήρηση, είναι περιορισμένα.

7.2 Προτάσεις βελτίωσης στο πλαίσιο της ψηφιακής επεξεργασίας σήματος

Είναι δεδομένο ότι τα διάφορα state-of-the-art συστήματα διαχωρισμού ηχητικών πηγών, στη μεγαλύτερη πλειοψηφία τους κάνουν υποδειγματοληψία από τα 44.1kHz στα 22.05kHz, 16kHz ή ακόμη και στα 8kHz [40], [44], [48]. Για κάποιους ανθρώπους όμως, η υποδειγματοληψία στα 16kHz ή 8kHz, κάνει αρκετά αισθητή την απώλεια πληροφορίας, πράγμα το οποίο αποτελεί μεγάλο μειονέκτημα. Προτείνουμε τα εξής:

- Υποδειγματοληψία του σήματος από τα 44.1kHz στα 22.05kHz ή 16kHz το οποίο αποδεδειγμένα μειώνει την υπερπροσαρμογή και βελτιώνει τον διαχωρισμό. Αυτό επιβεβαιώνεται και από τα δικά μας αποτελέσματα. Η υλοποίηση στα 44.1kHz του εδαφίου (6.3.4) έκανε περισσότερη υπερπροσαρμογή από την υλοποίηση στα 22.05kHz του εδαφίου (6.3.2).
- Με υποδειγματοληψία στα 22.05kHz έχουμε μια άνω συχνότητα αποκοπής στα 11.025kHz, ενώ με υποδειγματοληψία στα 16kHz έχουμε μια άνω συχνότητα αποκοπής στα 8kHz. Θα μπορούσαμε να καλύψουμε αυτήν την χαμένη πληροφορία της άνω ζώνης συχνοτήτων, χρησιμοποιώντας ξανά μεθόδους βαθιάς μάθησης για επέκταση του εύρους ζώνης(Bandwidth Extension), κάνοντας πρόβλεψη του άνω μέρους του φάσματος μέχρι τα 20kHz [47].

Βιβλιογραφία

[1] E.C. Cherry. "Some experiments on the recognition of speech, with one and with two ears". In: *Journal of the Acoustic Society of America* 25 (1953), pp. 975–979.

[2] F. Rosenblatt. "The perceptron: A probabilistic model for information storage and organization in the brain." In: *Psychological Review* 65.6 (1958), pp. 386–408. ISSN: 0033-295X. DOI: 10.1037/h0042519. URL: http://dx.doi.org/10.1037/h0042519.

[3] James Cooley and John Tukey. "An Algorithm for the Machine Calculation of Complex Fourier Series". In: *Mathematics of Computation* 19.90 (1965), pp. 297–301.

[4] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323.6088 (Oct. 1986), pp. 533–536. URL: http://dx.doi.org/10.1038/323533a0.

[5] Kurt and Hornik. "Approximation capabilities of multilayer feedforward networks". In: *Neural Networks* 4.2 (1991), pp. 251–257. DOI: 10.1016/0893-6080(91)90009-T. URL: http://www.sciencedirect.com/science/article/pii/089360809190009T.

[6] Michael A. Casey and Alex Westner. "Separation of Mixed Audio Sources By Independent Subspace Analysis." In: *ICMC*. Michigan Publishing, 2000. URL: http://dblp.uni-trier.de/db/conf/icmc/icmc2000.html#CaseyW00.

[7] Richard O Duda, David G Stork, and Peter E Hart. *Pattern classification*. New York; Chichester: Wiley, 2000.

[8] Daniel D. Lee and H. Sebastian Seung. "Algorithms for Non-negative Matrix Factorization". In: *In NIPS*. MIT Press, 2000, pp. 556–562. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.7566.

[9] Nikolaos Mitianoudis and Michael E. Davies. "Audio source separation of convolutive mixtures." In: *IEEE Trans. Speech Audio Process.* 11.5 (2003), pp. 489–497. URL: http://dblp.uni-trier.de/db/journals/taslp/taslp11.html#MitianoudisD03.

[10] E. Vincent, R. Gribonval, and C. Févotte. "Performance Measurement in Blind Audio Source Separation". In: *IEEE Transactions on Speech and Audio Processing* (2006).

[11] Tuomas Virtanen. "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria." In: *IEEE Trans. Speech Audio Process.* 15.3 (2007), pp. 1066–1074. URL: http://dblp.uni-trier.de/db/journals/taslp/taslp15.html#Virtanen07.

[12] Róisín Loughran et al. "The Use of Mel-frequency Cepstral Coefficients in Musical Instrument Identification." In: *ICMC*. Michigan Publishing, 2008. URL: http://dblp.uni-trier.de/db/conf/icmc/icmc2008.html#LoughranWOO08.

[13] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis." In: *Neural Computation* 21.3 (2009), pp. 793–830. URL: http://dblp.uni-trier.de/db/journals/neco/neco21.html#FevotteBD09.

[14] Arnaud Dessein, Arshia Cont, and Guillaume Lemaitre. "Real-time Polyphonic Music Transcription with Non-negative Matrix Factorization and Beta-divergence." In: *ISMIR*. Ed. by J. Stephen Downie and Remco C. Veltkamp. International Society for Music Information Retrieval, 2010, pp. 489–494. ISBN: 978-90-393-53813. URL: http://dblp.uni-trier.de/db/conf/ismir/ismir2010.html#DesseinCL10.

[15] Hiromasa Fujihara et al. "A Modeling of Singing Voice Robust to Accompaniment Sounds and Its Application to Singer Identification and Vocal-Timbre-Similarity-Based Music Information Retrieval." In: *IEEE Trans. Speech Audio Process.* 18.3 (2010), pp. 638–648. URL: http://dblp.uni-trier.de/db/journals/taslp/taslp18.html#FujiharaGKO10.

[16] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256. URL: http://proceedings.mlr.press/v9/glorot10a.html.

[17] Zhouchen Lin, Minming Chen, and Yi Ma. "The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices." In: *CoRR* abs/1009.5055 (2010). URL: http://dblp.uni-trier.de/db/journals/corr/corr1009.html#LinCM10.

[18] Bob L. Sturm, Marcela Morvidone, and Laurent Daudet. "Musical instrument identification using multiscale Mel-frequency cepstral coefficients." In: *EUSIPCO*. IEEE, 2010, pp. 477–481. URL: http://dblp.uni-trier.de/db/conf/eusipco/eusipco2010.html#SturmMD10.

[19] Matthias Mauch et al. "Timbre and Melody Features for the Recognition of Vocal Activity and Instrumental Solos in Polyphonic Music." In: *ISMIR*. Ed. by Anssi Klapuri and Colby Leider. University of Miami, 2011, pp. 233–238. ISBN: 978-0-615-54865-4. URL: http://dblp.uni-trier.de/db/conf/ismir/ismir2011.html#MauchFYG11.

[20] Lawrence R. Rabiner and Schafer W. *Theory and Applications of Digital Speech Processing*. Pearson, 2011, pp. 44–45. ISBN: 9780136034285.

[21] Emilia Gómez et al. "Predominant Fundamental Frequency Estimation vs Singing Voice Separation for the Automatic Transcription of Accompanied Flamenco Singing." In: *ISMIR*. Ed. by Fabien Gouyon et al. FEUP Edições, 2012, pp. 601–606. ISBN: 978-972-752-144-9. URL: http://dblp.uni-trier.de/db/conf/ismir/ismir2012.html#GomezCSBCM12.

[22] G. Hinton et al. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups". In: *IEEE Signal Process. Mag.* 29.6 (2012), pp. 82–97.

[23] Po-Sen Huang et al. "Singing-voice separation from monaural recordings using robust principal component analysis." In: *ICASSP*. IEEE, 2012, pp. 57–60. ISBN: 978-1-4673-0046-9. URL: http://dblp.uni-trier.de/db/conf/icassp/icassp2012.html#HuangCSH12.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

[25] Ian J. Goodfellow et al. "Maxout Networks". In: *CoRR* abs/1302.4389 (2013). URL: http://dblp.uni-trier.de/db/journals/corr/corr1302.html#abs-1302-4389.

[26] José R. Zapata and Emilia Gómez. "Using voice suppression algorithms to improve beat tracking in the presence of highly predominant vocals." In: *ICASSP*. IEEE, 2013, pp. 51–55. URL: http://dblp.uni-trier.de/db/conf/icassp/icassp2013.html#ZapataG13.

[27] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting." In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958. URL: http://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf.

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

[29] Brian McFee, Eric J. Humphrey, and Juan Pablo Bello. "A Software Framework for Musical Data Augmentation." In: *ISMIR*. Ed. by Meinard Müller and Frans Wiering. 2015, pp. 248–254. ISBN: 978-84-606-8853-2. URL: http://dblp.uni-trier.de/db/conf/ismir/ismir2015.html#McFeeHB15.

[30] Meinard Müller. *Fundamentals of Music Processing - Audio, Analysis, Algorithms, Applications*. Springer, 2015, pp. 1–480. ISBN: 978-3-319-21945-5.

[31] Andrew J. R. Simpson, Gerard Roma, and Mark D. Plumbley. "Deep Remix: Remixing Musical Mixtures Using a Convolutional Deep Neural Network." In: *CoRR* abs/1505.00289 (2015). URL: http://dblp.uni-trier.de/db/journals/corr/corr1505.html#SimpsonRP15a.

[32] Peter Knees and Markus Schedl. *Music Similarity and Retrieval - An Introduction to Audio- and Web-based Strategies*. Vol. 36. The Information Retrieval Series. Springer, 2016, pp. 1–254. ISBN: 978-3-662-49720-3.

[33] Pritish Chandna et al. "Monoaural Audio Source Separation Using Deep Convolutional Neural Networks." In: *LVA/ICA*. Ed. by Petr Tichavský et al. Vol. 10169. Lecture Notes in Computer Science. 2017, pp. 258–266. ISBN: 978-3-319-53546-3. URL: http://dblp.uni-trier.de/db/conf/ica/ica2017.html#ChandnaMJG17.

[34] François Chollet. *Deep Learning with Python*. Manning, Nov. 2017. ISBN: 9781617294433.

[35] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. 2017. ISBN: 9780262035613 0262035618. URL: https://www.worldcat.org/title/deep-learning/oclc/985397543&referer=brief_results.

[36] Kin Wah Edward Lin et al. "Sinusoidal Partials Tracking for Singing Analysis Using the Heuristic of the Minimal Frequency and Magnitude Difference." In: *INTERSPEECH*. Ed. by Francisco Lacerda. ISCA, 2017, pp. 3038–3042. URL: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2017.html#LinASL17.

[37] Zafar Rafii et al. *The MUSDB18 corpus for music separation*. Dec. 2017. DOI: 10.5281/zenodo.1117372. URL: https://doi.org/10.5281/zenodo.1117372.

[38] Stefan Uhlich et al. "Improving music source separation based on deep neural networks through data augmentation and network blending." In: *ICASSP*. IEEE, 2017, pp. 261–265. ISBN: 978-1-5090-4117-6. URL: http://dblp.uni-trier.de/db/conf/icassp/icassp2017.html#UhlichPGEKTM17.

[39] Konstantinos Drossos et al. "Harmonic-Percussive Source Separation with Deep Neural Networks and Phase Recovery." In: *IWAENC*. IEEE, 2018, pp. 421–425. ISBN: 978-1-5386-8151-0. URL: http://dblp.uni-trier.de/db/conf/iwaenc/iwaenc2018.html#DrossosMMV18.

[40] Kin Wah Edward Lin et al. "Singing Voice Separation Using a Deep Convolutional Neural Network Trained by Ideal Binary Mask and Cross Entropy." In: *CoRR* abs/1812.01278 (2018). URL: http://dblp.uni-trier.de/db/journals/corr/corr1812.html#abs-1812-01278.

[41] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. "The 2018 Signal Separation Evaluation Campaign." In: *LVA/ICA*. Ed. by Yannick Deville et al. Vol. 10891. Lecture Notes in Computer Science. Springer, 2018, pp. 293–305. ISBN: 978-3-319-93764-9. URL: http://dblp.uni-trier.de/db/conf/ica/ica2018.html#StoterLI18.

[42] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot. *Audio Source Separation and Speech Enhancement*. Wiley, 2018, pp. 1–504. ISBN: 978-1-119-27989-1.

[43] Alice Cohen-Hadria, Axel Roebel, and Geoffroy Peeters. "Improving singing voice separation using Deep U-Net and Wave-U-Net with data augmentation." In: *EUSIPCO*. IEEE, 2019, pp. 1–5. ISBN: 978-9-0827-9703-9. URL: http://dblp.uni-trier.de/db/conf/eusipco/eusipco2019.html#Cohen-HadriaRP19.

[44] Alexandre Défossez et al. "Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed." In: *CoRR* abs/1909.01174 (2019). URL: http://dblp.uni-trier.de/db/journals/corr/corr1909.html#abs-1909-01174.

[45] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. "When does label smoothing help?" In: *NeurIPS*. Ed. by Hanna M. Wallach et al. 2019, pp. 4696–4705. URL: http://dblp.uni-trier.de/db/conf/nips/nips2019.html#MullerKH19.

[46] Fabian-Robert Stöter et al. "Open-Unmix - A Reference Implementation for Music Source Separation." In: *J. Open Source Softw.* 4.41 (2019), p. 1667. URL: http://dblp.uni-trier.de/db/journals/jossw/jossw4.html#StoterULM19.

[47] Mathieu Lagrange and Félix Gontier. "Bandwidth Extension of Musical Audio Signals With No Side Information Using Dilated Convolutional Neural Networks." In: *ICASSP*. IEEE, 2020, pp. 801–805. ISBN: 978-1-5090-6631-5. URL: http://dblp.uni-trier.de/db/conf/icassp/icassp2020.html#LagrangeG20.

[48] David Samuel, Aditya Ganeshan, and Jason Naradowsky. "Meta-learning Extractors for Music Source Separation." In: *CoRR* abs/2002.07016 (2020). URL: http://dblp.uni-trier.de/db/journals/corr/corr2002.html#abs-2002-07016.

[49] *An overview of activation functions used in neural networks*. URL: https://adl1995.github.io/an-overview-of-activation-functions-used-in-neural-networks.html.

[50] *CS231n Convolutional Neural Networks for Visual Recognition*. URL: https://cs231n.github.io/neural-networks-1/.

[51] *Early Stopping with PyTorch to Restrain your Model from Overfitting*. URL: https://mc.ai/early-stopping-with-pytorch-to-restrain-your-model-from-overfitting/.

[52] *fast.ai Forums*. URL: https://forums.fast.ai/t/deep-learning-with-audio-thread/38123/265.

[53] *Support Vector Machines*. URL: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

Συντομογραφίες - Αρτικόλεξα - Ακρωνύμια

ΔΠΘ Δημοκρίτειο Πανεπιστήμιο Θράκης

CNN Convolutional Neural Networks

LSTM Long Short Term Memory

RNN Recurrent Neural Networks

IBM Ideal Binary Mask

CPU Central Processing Unit

GPU Graphical Processing Unit

RAM Random Access Memory

SGD Stochastic Gradient Descent

MLP Multilayer Perceptron

ReLU Rectified Linear Unit

CT Continuous-Time

DT Discrete-Time

DTFT Discrete Time Fourier Transform

DFT Discrete Fourier Transform

FFT Fast Fourier Transform

STFT Short Time Fourier Transform

MFCC Mel Frequency Cepstral Coefficient

ASR Automatic Speech Recognition

SDR Source to Distortion Ratio

SIR Source to Interference Ratio

SNR Source to Noise Ratio

SAR Source to Artifacts Ratio

NMF Non-Negative Matrix Factorization

ISA Independent Subspace Analysis

ICA Independent Component Analysis

rPCA Robust Principal Component Analysis

BSS Blind Source Separation

SiSEC Signal Separation Evaluation Campaign

DAW Digital Audio Workstation

Απόδοση ξενόγλωσσων όρων

Audio Source Separation - Διαχωρισμός ηχητικών πηγών

Deep Learning - Βαθιά μάθηση

Pattern Recognition - Αναγνώριση προτύπων

Machine Learning - Μηχανική μάθηση

Supervised Learning - Μάθηση με επιτήρηση

Unsupervised Learning - Μάθηση χωρίς επιτήρηση

Self-supervised Learning - Ιδιο-επιτηρούμενη μάθηση

Reinforcement Learning - Ενισχυτική μάθηση

Music Information Retrieval - Μουσική εξόρυξη πληροφορίας

Feedforward Neural Networks - Προσο-τροφοδοτούμενα νευρωνικά δίκτυα

Convolutional Neural Networks - Συνελικτικά νευρωνικά δίκτυα

Recurrent Neural Networks - Επανατροφοδοτούμενα νευρωνικά δίκτυα

Mixture - Ηχητικό σήμα προερχόμενο από μίξη

Accompaniment - Μουσική συνοδεία

Classification - Ταξινόμηση

Regression - Παλινδρόμηση

Pixel-Wise Classification - Ταξινόμηση εικονοστοιχείων

Image Segmentation - Κατάτμηση εικόνας

Binary Cross-Entropy - Δυαδική διασταυρωμένη εντροπία

Negative Log-Likelihood - Αρνητική λογαριθμική πιθανοφάνεια

Ideal Binary Mask - Ιδανική δυαδική μάσκα

Soft Mask - Χαλαρή μάσκα

Label - Ταμπέλα

Spectrogram - Φασματογράφημα

Log-Frequency Representation - Λογαριθμική-συχνοτική αναπαράσταση

Digital Signal Processing - Ψηφιακή επεξεργασία σήματος

Audio Signal Processing - Ψηφιακή επεξεργασία σήματος του ήχου

Training Set - Σύνολο εκπαίδευσης

Test Set - Σύνολο δοκιμής

Objective Function - Αντικειμενική συνάρτηση(ή συνάρτηση κόστους σε περίπτωση ελαχιστοποίησης της)

Gradient-based Learning - Μάθηση με χρήση βαθμίδας

Learning Rate - Ρυθμός μάθησης

Optimization - Βελτιστοποίηση

Training Error - Σφάλμα εκπαίδευσης

Test Error - Σφάλμα στο σύνολο δοκιμής

Generalization Error - Σφάλμα γενίκευσης

Underfitting - Υποπροσαρμογή

Overfitting - Υπερπροσαρμογή

Capacity - Χωρητικότητα

Maximum Likelihood Estimation - Εκτίμηση μέγιστης πιθανοφάνειας

Layer - Επίπεδο του νευρωνικού δικτύου

Activation Function - Συνάρτηση ενεργοποίησης

Forward Propagation - Εμπρόσθια διάδοση

Regularization - Ομαλοποίηση

Dataset - Σύνολο δεδομένων

Dataset Augmentation - Επαύξηση συνόλου δεδομένων

Pitch - Ακουστική αίσθηση στην οποία ένας ακροατής αντιστοιχεί μουσικούς τόνους σε σχετικές θέσεις σε μια μουσική κλίμακα, βασιζόμενος πρωτίστως στην αντίληχη της συχνότητας μια δόνησης

Pitch Shifting - Ολίσθηση του pitch

Time Stretching - Χρονική παραμόρφωση

Dynamic Range Compression - Συμπίεση δυναμικού εύρους

Label Smoothing - Προσθήκη θορύβου στις ταμπέλες

Feature Map - Χάρτης χαρακτηριστικών

Pooling - Ομαδοποίηση

Stride - Βήμα απόφασης

Zero Padding - Προσθήκη μηδενικών

Valid Convolution - Έγκυρη συνέλιξη

Same Convolution - Ταυτόσιμη συνέλιξη

Full Convolution - Πλήρης συνέλιξη

Sampling - Δειγματοληψία

Quantization - Κβαντισμός

Aliasing - Το φαινόμενο που προκαλεί τα διαφορετικά σήματα να γίνονται δυσδιάκριτα όταν γίνεται δειγματοληψία

Distortion - Παραμόρφωση

Bitrate - Ο αριθμός των bits ανά δευτερόλεπτο που μεταδίδονται σε ένα ψηφιακό σύστημα

Continuous-Time Signal - Σήμα συνεχούς χρόνου

Discrete-Time Signal - Σήμα διακριτού χρόνου

Bandlimited Signal - Σήμα περιορισμένου εύρους ζώνης

Discrete Time Fourier Transform - Μετασχηματισμός Fourier διακριτού χρόνου

Discrete Fourier Transform - Διακριτός μετασχηματισμός Fourier

Fast Fourier Transform - Γρήγορος μετασχηματισμός Fourier

Short Time Fourier Transform - Βραχυχρόνιος μετασχηματισμός Fourier

Window Function - Συνάρτηση παραθύρου

Discrete Cosine Transform - Διακριτός μετασχηματισμός συνημιτόνου

Singlechannel - Μονοκαναλικό

Multichannel - Πολυκαναλικό

Point Source - Σημειακή πηγή

Diffuse Source - Διασκορπιστική πηγή

Source to Distortion Ratio - Λόγος πηγής προς παραμόρφωση

Source to Interference Ratio - Λόγος πηγής προς παρεμβολή

Source to Noise Ratio - Λόγος πηγής προς θόρυβο

Source to Artifacts Ratio - Λόγος πηγής προς παράσιτα

Mixing Process - Η διαδικασία της μίξης

State of the Art - Κορυφαία και σύγχρονη μέθοδος

Blind Source Separation - Τυφλός διαχωρισμός πηγών

Independent Component Analysis - Ανεξάρτητη ανάλυση υποχώρων

Independent Component Analysis - Ανεξάρτητη ανάλυση συνιστωσών

Robust Principal Component Analysis - Εύρωστη κύρια ανάλυση συνιστωσών

Bandwidth Extension - Επέκταση του εύρους ζώνης