ΔΙΑΧΩΡΙΣΜΟΣ ΜΟΥΣΙΚΩΝ ΠΗΓΩΝ ΜΕ ΤΕΧΝΙΚΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Διπλωματική Εργασία της Ειρήνης Γαλάνη

Επιβλέπων Καθηγητής: Κωνσταντίνος Μπερμπερίδης

Τμήμα Μηχανικών Η/Υ και Πληροφορικής
Πανεπιστήμιο Πατρών
Πάτρα, Φεβρουάριος 2022

Ευχαριστίες

Με την περάτωση της παρούσας εργασίας θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου Κωνσταντίνο Μπερμπερίδη, για την καθοδήγηση και τις υποδείξεις που μου έδωσε για την εκπόνησή της. Είμαι ευγνώμων τόσο για το χρόνο που διέθεσε για να μου δώσει εξηγήσεις πάνω στο θέμα, όσο και για τις γνώσεις που αποκόμισα καθ' όλη τη διάρκεια της άριστης συνεργασίας μας.

Επίσης, ένα μεγάλο ευχαριστώ στον Γιώργο Τζούπη (Διπλωματούχος Μηχανικός Η/Υ και Πληροφορικής Πα.Πα.) για την προθυμία και την αστείρευτη διάθεση να μου προσφέρει την πολύτιμη βοήθειά του.

Ιδιαίτερα θερμές ευχαριστίες στους γονείς μου Νίκο και Χριστίνα και στην αδερφή μου Αθηνά, για την αμέριστη και ανεκτίμητη συμπαράστασή τους όλα αυτά τα χρόνια.

Ευχαριστώ την Ευαγγελία-Ιωάννα Λυκιαρδοπούλου, την Γεωργία Νικολακοπούλου και την Κατερίνα Φιλιππoπούλου, τρεις πολύ καλές φίλες, για την συνεχή υποστήριξη και κατανόησή τους. Με το παράδειγμα που δίνουν, αποτελούν πηγή δύναμης και έμπνευσης. Επιπρόσθετα, θα ήθελα να ευχαριστήσω τους συμφοιτητές μου Βασίλη Δούρο, Μιχάλη Τσακηρίδη και Πολύκαρπο Ζαρμακούπη, με τους οποίους, καθ' όλη τη διάρκεια φοίτησής μας, αντιμετωπίσαμε κάθε δυσκολία και πρόκληση ως ομάδα.

Τέλος, πολλά μεγάλα ευχαριστώ σε όλους τους φίλους και τους αγαπημένους ανθρώπους μου, που έχουν επιλέξει να βρίσκονται δίπλα μου και να με στηρίζουν με κάθε τρόπο όλα αυτά τα χρόνια. Χωρίς αυτούς δεν θα ήμουν ο άνθρωπος που είμαι σήμερα.

Περίληψη

Η ανάκτηση μουσικής πληροφορίας (MIR) στοχεύει στην βελτίωση της λειτουργικότητας της αναζήτησης και περιήγησης του χρήστη σε μουσικές συλλογές. Ένα από τα προβλήματα της MIR, το οποίο χρησιμοποιείται σε ποικίλες εφαρμογές της, είναι ο διαχωρισμός μουσικών πηγών (MSS), ο οποίος δέχεται ως είσοδο ένα σήμα μίξης και εξάγει από αυτό τα επιμέρους σήματα πηγών, όπως της φωνής και της συνοδείας. Σε γενικό πλαίσιο, ακολουθείται μία συνήθης διαδικασία διαχωρισμού, αλλά οι προσεγγίσεις γύρω από αυτήν ποικίλουν. Κατ' επέκταση, έχουν προταθεί πολλά διαφορετικά μοντέλα διαχωρισμού μουσικών πηγών, τα αποτελέσματα των οποίων μελετώνται και συγκρίνονται βάσει ορισμένων μεθόδων αξιολόγησης. Τα τελευταία χρόνια, μοντέλα νευρωνικών δικτύων (NNs) έχουν εφαρμοστεί για τον διαχωρισμό μουσικών πηγών, συμπεριλαμβανομένων των πλήρως συνδεδεμένων NNs (FNNs), των συνελικτικών NNs (CNNs) και των αναδρομικών NNs (RNNs) καθώς και παραλλαγές αυτών. Ο λόγος είναι ότι σε αντίθεση με άλλες μεθόδους που χρειάζονται σαφή μοντέλα των πηγών για την επεξεργασία, τα νευρωνικά δίκτυα εφαρμόζουν τεχνικές βελτιστοποίησης για να εκπαιδεύσουν τα μοντέλα, χρησιμοποιώντας σύνολα δεδομένων στα οποία είναι διαθέσιμο τόσο το σήμα μίξης όσο και τα απομονωμένα σήματα πηγών. Πιο συγκεκριμένα, τα συνελικτικά νευρωνικά δίκτυα θεωρούνται από τις επικρατέστερες τεχνικές μηχανικής μάθησης που χρησιμοποιούνται στον τομέα των πολυμέσων. Ωστόσο, οι απαιτήσεις για μεγάλο βάθος στο δίκτυο, δυσκολεύουν την εκπαίδευσή τους και οδηγούν σε μείωση της απόδοσης. Για την αντιμετώπιση αυτών των προβλημάτων προτάθηκε το μοντέλο DenseNet, το οποίο είναι βασισμένο στα CNNs. Αυτή η εργασία παρουσιάζει μια σύγκριση μεταξύ δύο σύγχρονων επεκτάσεων του DenseNet όσον αφορά το πρόβλημα του διαχωρισμού φωνητικών και συνοδείας, μελετάει την ποιότητα των αποτελεσμάτων τους με τη βοήθεια μετρικών αξιολόγησης, και οδηγεί σε μία αποτίμηση της συνολικής απόδοσής τους.

Abstract

Music information retrieval (MIR) aims to improve the functionality of the user's search and browsing in music collections. One of MIR's problems used in various applications is music source separation (MSS), which takes a mixture signal as input, and extracts from it the individual source signals, such as voice and accompaniment. In general, a standard separation process is followed, however there are various different related approaches. Consequently, several diverse models of music source separation have been proposed, the results of which are studied and compared based on certain evaluation methods. In recent years, neural network models (NNs) have been implemented for the task of music source separation, including fully connected NNs (FNNs), convolutional NNs (CNNs), and recurrent NNs (RNNs), as well as their variations. The reason is that unlike other methods that require explicit source models for processing, neural networks take advantage of optimization techniques to train the models, using datasets in which both the mixing signal and the isolated source signals are

available. More specifically, convolutional neural networks are considered to be one of the most predominantly used machine learning techniques in the multimedia domain. However, considerable depth is required, making the network training difficult and leading to performance degradation. To address these problems, a new model, called DenseNet and based on CNNs, has been proposed. This paper compares two modern extensions of the DenseNet architecture regarding the problem of voice and accompaniment separation, examines the quality of their results, using evaluation metrics, and leads to an assessment of their overall performance.

ΠΕΡΙΕΧΟΜΕΝΑ

Πίνακας Ακρωνυμιών

Όρος	*Ακρωνύμιο*
Azimuth Discrimination Resynthesis	ADRess
Batch Normalization	BN
Bidirectional Long-Short Term Memory	BLSTM
Blind Source Separation Evaluation	BSSEval
Convolutional Neural Network	CNN
Deep Neural Network	DNN
Degenerate Unmixing Estimation Technique	DUET
Feed-forward fully-connected Neural Network	FNN
Generalized Wiener Filter	GWF
Gradient Descent	GD
Independent Component Analysis	ICA
Itakura Saito	IS
Kullback Leibler	KL
Long-Short Term Memory	LSTM
Multichannel Wiener Filter	MWF
MUlti Stimulus test with Hidden Reference and Anchor	MUSHRA
Music Information Retrieval	MIR
Music Source Separation	MSS
Neural Network	NN
Non-negative Matrix Factorization	NMF
Power Spectral Density	PSD
PROJection Estimation Technique	PROJET
Rectified Linear Unit	ReLU
Recurrent Neural Network	RNN
Short-Time Fourier Transform	STFT
Signal Separation Evaluation Campaign	SiSEC
Signal to Noise Ratio	SNR
Source to Artifact Ratio	SAR
Source to Distortion Ratio	SDR
Source to Interference Ratio	SIR
Stochastic Gradient Descent	SGD
Time-Frequency	TF

Πρόλογος

Η Ανάκτηση Μουσικής Πληροφορίας (MIR) έχει γνωρίσει ραγδαία ανάπτυξη τα τελευταία χρόνια, καθώς καθίσταται όλο και πιο επιτακτικό, λόγω της εξέλιξης της τεχνολογίας και του διαδικτύου, να καλυφθεί η ανάγκη ύπαρξης και υλοποίησης ενός ευρέως φάσματος τεχνικών μουσικής ανάλυσης. Το ερευνητικό πεδίο της MIR ασχολείται πρωτίστως με την εξαγωγή σημαντικών χαρακτηριστικών από τη μουσική (από το ηχητικό σήμα) και κατ' επέκταση με την ανάπτυξη διαφόρων συστημάτων αναζήτησης και ανάκτησης. Η MIR χρησιμοποιείται από μία πληθώρα εφαρμογών, από τις οποίες επωφελείται το κοινό της μουσικής βιομηχανίας, των ακροατών/καταναλωτών μουσικής και των επαγγελματιών, όπως οι καλλιτέχνες, οι δάσκαλοι μουσικής και οι μουσικολόγοι. Η MIR ασχολείται με ένα μεγάλο εύρος προβλημάτων, όπως ο διαχωρισμός μουσικών πηγών, που αποτελεί και το θέμα της παρούσας εργασίας. Στο **Κεφάλαιο 1** γίνεται μια εισαγωγή στην ανάκτηση μουσικής πληροφορίας, παρατίθενται ενδεικτικά μερικές εφαρμογές των τομέων που έχουν επωφεληθεί από αυτήν και αναφέρονται τα προβλήματα με τα οποία ασχολείται και δύο κριτήρια διαχωρισμού τους. Τέλος, γίνεται μια πρώτη αναφορά στο πρόβλημα του διαχωρισμού μουσικών πηγών και στον τρόπο με τον οποίο προσεγγίζεται το πρόβλημα αυτό από ένα σύστημα MIR.

Κατά το γενικό πρόβλημα του διαχωρισμού ηχητικών πηγών (MSS), δίνονται ένα ή περισσότερα σήματα μίξης, τα οποία, ουσιαστικά, έχουν διαμορφωθεί από συνδυασμούς των αρχικών σημάτων ήχου που παράγουν οι πηγές, και ο στόχος είναι η ανάκτηση ενός ή περισσοτέρων σημάτων από αυτά. Ο MSS αποτελεί ένα αρκετά σύνθετο πρόβλημα, λόγω της ιδιαιτερότητας της φύσης των μουσικών σημάτων. Τα μουσικά σήματα έχουν διαφορετικά φασματικά και χρονικά χαρακτηριστικά λόγω των διαφορετικών μηχανισμών παραγωγής ήχου της φωνής και των μουσικών οργάνων, τα οποία συνήθως παίζουν ταυτόχρονα, στον ίδιο τόνο και ρυθμό και παρόμοιες μελωδίες. Λόγω του πλήθους των οργάνων, δηλαδή των πηγών, που εκτελούν τα παραπάνω, συνήθως παρατηρείται επικάλυψη μεταξύ τους, με αποτέλεσμα να καθίσταται δύσκολος ο διαχωρισμός τους. Ένα ακόμα στοιχείο που έχει σημασία για το πρόβλημα του MSS είναι οι πληροφορίες που σχετίζονται με τη διαδικασία μίξης των πηγών και άρα με την τελική δομή της μίξης που παράγεται. Λαμβάνοντας υπόψιν τα παραπάνω, έχουν προταθεί από την επιστημονική κοινότητα πολλές διαφορετικές κατηγορίες μεθόδων για την υλοποίηση του MSS. Για τη σύγκριση της αποτελεσματικότητας αυτών των μεθόδων, προέκυψε η ανάγκη εύρεσης μετρικών αξιολόγησής τους. Στο **Κεφάλαιο 2**, παρατίθεται η συνήθης διαδικασία διαχωρισμού μουσικών πηγών και τα επιμέρους συστατικά στοιχεία της και στη συνέχεια παρουσιάζονται διάφοροι μέθοδοι που έχουν προταθεί για την επίλυση του προβλήματος. Τέλος, αναφέρονται οι επικρατέστερες μετρικές αξιολόγησης των μοντέλων του MSS.

Τα τελευταία χρόνια, η τάση για χρήση μεθόδων βαθιάς μάθησης (deep learning) στον τομέα της έρευνας για την ανάκτηση μουσικής πληροφορίας (MIR) και κατ' επέκταση του διαχωρισμού μουσικών πηγών, έχει ολοένα και περισσότερο ανοδική πορεία. Στο γενικό πλάισιό τους, οι προσεγγίσεις βαθιάς μάθησης προϋποθέτουν ένα δίκτυο με πολλαπλά επίπεδα, τα οποία εκπαιδεύονται

από τα δεδομένα που δέχεται στην είσοδό του. Τέτοια δίκτυα ονομάζονται βαθιά νευρωνικά δίκτυα (deep neural networks –DNNs). Οι βασικές αποφάσεις που πρέπει να παρθούν όσον αφορά ένα νευρώνικό δίκτυο, είναι, ο τρόπος σχεδίασης της δομής του, δηλαδή η αρχιτεκτονική του, και ο τρόπος εκπαίδευσης του. Οι επιλογές αυτές, παίρνονται με βάση το είδος των δεδομένων εισόδου και τις διαθέσιμες πληροφορίες για αυτά, και με βάση το πόσο αποτελεσματικά μπορούν να διεκπεραιώσουν τη λειτουργία τους αυτές οι τεχνικές, για την εκάστοτε εφαρμογή στην οποία χρησιμοποιούνται. Στο **Κεφάλαιο 3**, αρχικά αναφέρονται τα βασικά συστατικά στοιχεία της δομής ενός νευρωνικού δικτύου, ανεξαρτήτως αρχιτεκτονικής, και ακολουθεί η ανάλυση του αλγορίθμου back-propagation, ο οποίος είναι ο επικρατέστερος αλγόριθμος για την εκπαίδευση των DNNs. Τέλος, παρουσιάζονται οι αρχιτεκτονικές των συνελικτικών νευρωνικών δικτύων (convolutional neural networks –CNNs), των αναδρομικών νευρωνικών δικτύων (recurrent neural networks – RNNs) και των δικτύων Long Short Term Memory (LSTM). Αυτά τα μοντέλα δικτύων και οι διάφορες επεκτάσεις τους, είναι κυρίαρχες τεχνικές βαθιάς μάθησης για την επίλυση του προβλήματος του MSS και αποτελούν σημαντική και απαραίτητη βάση για τα δίκτυα DenseNet που χρησιμοποιούνται για τα πειράματα της παρούσας εργασίας.

Με γνώμονα το γεγονός ότι τα τελευταία χρόνια οι state-of-the-art προσεγγίσεις του προβλήματος του MSS χρησιμοποιούν τεχνικές βαθιάς μάθησης, η παρούσα εργασία εμβαθύνει στην ανάλυση και την πειραματική μελέτη τέτοιων τεχνικών. Πιο συγκεκριμένα, μία από τις πιο σύγχρονες αρχιτεκτονικές δικτύων που έχουν επιφέρει state-of-the-art αποτελέσματα είναι τα πυκνά συνδεδεμένα συνελικτικά δίκτυα (Dense convolutional Networks – DenseNets), τα οποία, γενικά, απαιτούν πολύ μικρότερο πλήθος παραμέτρων και επιτυγχάνουν μεγαλύτερη ακρίβεια αποτελεσμάτων, συγκριτικά με άλλες αρχιτεκτονικές DNN. Για την περαιτέρω μείωση των παραμέτρων και των απαιτήσεων για μνήμη αλλά και για λιγότερο χρόνο εκπαίδευσης, προτάθηκε και χρησιμοποιήθηκε το Multi-Scale DenseNet (MDenseNet) και η επέκτασή του Multi-band MDenseNet (MMDenseNet). Τέλος, με σκοπό τη βελτίωση της συνολικής απόδοσης του δικτύου, παρατηρήθηκε ότι η από κοινού χρήση των δυνατοτήτων του δικτύου LSTM και του MMDenseNet, σε μία ενιαία αρχιτεκτονική, μειώνει ακόμα περισσότερο το πλήθος των παραμέτρων και οδηγεί σε ακόμα καλύτερα αποτελέσματα. Στο **Κεφάλαιο 4** αναλύεται εκτενώς η αρχιτεκτονική του μοντέλου DenseNet, η οποία αποτελεί υπόβαθρο για τις επεκτάσεις MMDenseNet και MMDenseLSTM που παρατίθενται στη συνέχεια και θα αποτελέσουν τις μεθόδους της πειραματικής μελέτης της εργασίας.

Με βάση το πρόβλημα του διαχωρισμού μουσικών πηγών και πιο συγκεκριμένα, της απομόνωσης του σήματος της φωνής από το υπολειπόμενο σήμα συνοδείας, κατασκευάστηκαν δύο αρχιτεκτονικές για κάθε ένα από τα μοντέλα MMDenseNet και MMDenseLSTM. Οι αρχιτεκτονικές αυτές εκπαιδεύτηκαν με το σύνολο δεδομένων musdb18, το οποίο αποτελεί ένα ευρέως χρησιμοποιούμενο σύνολο δεδομένων από την επιστημονική κοινότητα, και στη συνέχεια, αξιολογήθηκαν με χρήση των μετρικών BSSEval. Τα συγκεκριμένα μοντέλα επιλέχθηκαν διότι οι

τεχνικές μηχανικής μάθησης και η χρήση νευρωνικών δικτύων έχουν συνεχή αυξητική τάση τα τελευταία χρόνια όσον αφορά το συγκεκριμένο αντικείμενο μελέτης. Τα DenseNets, εμπεριέχονται στις πιο σύγχρονες μεθόδους που αναπτύχθηκαν, και τα μοντέλα MMDenseNet και MMDenseLSTM, αποτελούν τις πιο πρόσφατες επεκτάσεις αυτών. Στο **Κεφάλαιο 5**. περιγράφονται αναλυτικά οι αρχιτεκτονικές των δύο μοντέλων και παρατίθενται και σχολιάζονται οι μετρικές αξιολόγησης των αποτελεσμάτων τους, όπως και οι τρόποι διαχείρισης που εφαρμόστηκαν για την αποφυγή πιθανών προβλημάτων. Τέλος, γίνεται μια σύγκριση των δύο μοντέλων, ώστε να παραχθούν συγκεκριμένα συμπεράσματα για την επιλογή του βέλτιστου εκ των δύο.

1 ΑΝΑΚΤΗΣΗ ΜΟΥΣΙΚΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

Η ραγδαία ανάπτυξη των εργαλείων και τεχνολογιών για χρήση, αποθήκευση και διανομή της μουσικής τα τελευταία χρόνια, έχει φέρει επανάσταση στον τρόπο με τον οποίο οι άνθρωποι έρχονται σε επαφή με τη μουσική. Σε γενικές γραμμές, το ερευνητικό πεδίο της Ανάκτησης Μουσικής Πληροφορίας (Music Information Retrieval - MIR) ασχολείται κυρίως με την εξαγωγή σημαντικών χαρακτηριστικών από την μουσική και την ανάπτυξη διαφόρων προγραμμάτων αναζήτησης και ανάκτησης, όπως η βάση περιεχομένου αναζήτηση, τα συστήματα προτάσεων μουσικής και η περιήγηση χρήστη σε μεγάλες συλλογές μουσικής [1]. Συνεπώς, η Ανάκτηση Μουσικής Πληροφορίας στοχεύει στο να καταστήσει εφικτή την πρόσβαση κάθε χρήστη στην παγκοσμίως τεράστια μουσική συλλογή, ώστε να τον διευκολύνει στο να αναζητήσει και να βρει την πληροφορία που θελει [2].

Μερικοί από τους πιο σημαντικούς λόγους για την επιτυχή ανάπτυξη της Ανάκτησης Μουσικής Πληροφορίας είναι οι παρακάτω:

(i) η ανάπτυξη τεχνικών συμπίεσης του ήχου στα τέλη της δεκαετίας του 1990,
(ii) η αύξηση της υπολογιστικής ισχύος των προσωπικών υπολογιστών, η οποία με τη σειρά της επέτρεψε στους χρήστες και τις εφαρμογές να εξάγουν μουσικά χαρακτηριστικά σε λογικό χρόνο,
(iii) η ευρεία διαθεσιμότητα των φορητών συσκευών αναπαραγωγής μουσικής και πιο πρόσφατα
(iv) η εμφάνιση εφαρμογών αναπαραγωγής μουσικής μέσω streaming όπως το Spotify, το οποίο αναφέρει την διαχείριση περισσότερων από 70 εκατομμυρίων μουσικών κομματιών και, για το δεύτερο τρίμηνο του 2021, την εξυπηρέτηση 365 εκατομμυρίων ενεργών χρηστών μηνιαία.

Κατά συνέπεια, ο μεγάλος όγκος αυτής της πληροφορίας δημιουργεί προκλήσεις όσον αφορά τη διαχείριση και την καλύτερη εξυπηρέτηση των χρηστών που με κάποιον τρόπο χρειάζεται να ανακτήσουν μέρος αυτής.

1.1 ΕΦΑΡΜΟΓΕΣ ΤΗΣ MIR

Η MIR χρησιμοποιείται από μία πληθώρα εφαρμογών, οι οποίες εφαρμόζουν συγκεκριμένα κριτήρια ομοιότητας με σκοπό την όσο το δυνατόν πιο αποτελεσματική αναζήτηση του χρήστη. Η γνώση των χαρακτηριστικών και των συστατικών στοιχείων ενός μουσικού κομματιού παίζει σημαντικό ρόλο στην μοντελοποίηση των δεδομένων και την αναβάθμιση των μεθόδων της MIR [3]. Εμπορι-

κές εφαρμογές όπως το Musicovery (musicovery.com), αναλαμβάνουν να προτείνουν μουσική στον χρήστη με βάση τα κριτήρια που αυτός εισήγαγε [4]. Πλατφόρμες όπως το Spotify (spotify.com), το Deezer (deezer.com) και το Pandora (pandora.com) έχουν προσφέρει στους χρήστες τους τεράστια πληθώρα μουσικών βιβλιοθηκών και ποικίλες ανέσεις, όπως την αναζήτηση μουσικής με βάση διαφορετικά κριτήρια όπως το είδος ή τους αγαπημένους τους καλλιτέχνες, τη δημιουργία playlists και τη συμβατότητα της πλατφόρμας με πολλές διαφορετικές συσκευές. Ακόμα, τους προτείνει μουσικές επιλογές με βάση παρελθοντικές αναζητήσεις που έχουν κάνει. Επιπλέον, τέτοιες εφαρμογές πληρώνουν τους κατόχους των πνευματικών δικαιωμάτων των μουσικών κομματιών από τις μεταδόσεις τους. Το amuse (amuse.io) είναι μια εφαρμογή που προσφέρει στους καλλιτέχνες δωρεάν διανομή της μουσικής τους σε πλατφόρμες όπως αυτές που προαναφέρθηκαν, ενώ φροντίζει και για τη μεταβίβαση των εσόδων που λαμβάνουν από αυτές. Εφαρμογές όπως το Shazam (shazam.com), στοχεύουν στην αναγνώριση του μουσικού κομματιού που ακούει ο χρήστης εκείνη την στιγμή (καθώς και του άλμπουμ και του καλλιτέχνη), λαμβάνοντας το αντίστοιχο δείγμα από το κινητό του τηλέφωνο [5], [6]. Τέλος, περιβάλλοντα, πλατφόρμες και βιντεο-παιχνίδια εκμάθησης χρησιμοποιούνται για εκπαιδευτικούς σκοπούς, όπως για παράδειγμα το Yousician και το SmartMusic [7], στα οποία ο χρήστης παίζει κάποιο μουσικό όργανο στο μικρόφωνο του υπολογιστή του και λαμβάνει άμεσα ασκήσεις και feedback όσον αφορά τις επιδόσεις του. Το Rocksmith είναι ένα βιντεο-παιχνίδι για εκμάθηση κιθάρας και μπάσου, το οποίο δίνει τη δυνατότητα στο χρήστη να εξασκείται επιλέγοντας κομμάτια της αρεσκείας του και παίζοντας παράλληλα με την εφαρμογή, ενώ στη συνέχεια αναλύει τις προτιμήσεις και την απόδοσή του και του προτείνει κατάλληλες άσκησεις [8].

1.2 ΠΡΟΒΛΗΜΑΤΑ ΤΗΣ MIR

Μερικά ενδεικτικά προβλήματα με τα οποία ασχολείται η MIR είναι:

- Αναγνώριση Μουσικής (Music Identification): Αναγνώριση ενός μουσικού κομματιού και εύρεση πληροφοριών σχετικών με αυτό. Αφορά εφαρμογές όπως το Shazam.

- Διαχωρισμός μουσικών πηγών (Music Source Separation): Εξαγωγή των σημάτων των οργάνων που περιέχονται σε ένα μουσικό κομμάτι.

- Ανίχνευση λογοκλοπής (Plagiarism Detection): Εντοπίζει την κατάχρηση μουσικής πνευματικής ιδιοκτησίας.

- Παρακολούθηση πνευματικών δικαιωμάτων (Copyright Monitoring): Παρακολουθεί τη μετάδοση της μουσικής για τυχόν παραβίαση πνευματικών δικαιωμάτων.

- Εκδόσεις (Versions): Εντοπίζει τις διάφορες εκδοχές ενός μουσικού κομματιού, δηλαδή τα remixes, τα lives, τα covers και την κανονική ηχογράφηση. Χρησιμοποιείται για την εκκαθάριση των σχεδόν διπλότυπων αποτελεσμάτων από μία βάση δεδομένων.

- Μελωδία (Melody): Εύρεση μουσικών κομματιών τα οποία περιέχουν ένα συγκεκριμένο μελωδικό τμήμα.

- Πανομοιότυπο έργο/τίτλος (Identical work/title): Ανάκτηση έργων με την ίδια μουσική σύνθεση ή τον ίδιο τίτλο τραγουδιού.

- Ερμηνευτής (Performer): Εύρεση μουσικής με κριτήριο αναζήτησης έναν συγκεκριμένο καλλιτέχνη.

- Συνθέτης (Composer): Εύρεση μουσικής με κριτήριο αναζήτησης έναν συγκεκριμένο συνθέτη.

- Προτάσεις (Recommendations): Εύρεση μουσικής η οποία να ταιριάζει με το προσωπικό προφίλ του χρήστη.

- Διάθεση (Mood): Εύρεση μουσικής χρησιμοποιώντας συναισθηματικές έννοιες, όπως χαρά, ενέργεια, μελαγχολία, χαλάρωση.

- Στυλ/Είδος (Style/Genre): Εύρεση μουσικής η οποία να ανήκει σε μία γενική κατηγορία, όπως jazz, funk, κλπ.

- Ενορχήστρωση (Instruments): Εύρεση μουσικών κομματιών με την ίδια ενορχήστρωση.

Τα ανωτέρω προβλήματα μπορούν να κατηγοριοποιηθούν με βάση την ειδικότητά τους (specificity) [5] σε τρεις μεγάλες κατηγορίες. Συστήματα υψηλής ειδικότητας προσδιορίζουν το ακριβές περιεχόμενο του ηχητικού σήματος μεμονωμένων ηχογραφήσεων. Για παράδειγμα, η Αναγνώριση Μουσικής, είναι ένα πρόβλημα υψηλής ειδικότητας. Συστήματα μέτριας ειδικότητας εξάγουν μουσικά χαρακτηριστικά υψηλού επιπέδου, όπως μελωδία. Ένα τέτοιο πρόβλημα είναι η αναζήτηση μουσικής με βάση τον ερμηνευτή ή τον συνθέτη. Τέλος, τα συστήματα χαμηλής ειδικότητας ασχολούνται με πιο αφαιρετικά χαρακτηριστικά όπως το μουσικό είδος.

Ένα άλλο κριτήριο διαχωρισμού των προβλημάτων της MIR είναι ο βαθμός ανάλυσης (granularity) ή η χρονική έκταση [1] [9]. Για την ανάκτηση ολόκληρου μουσικού κομματιού χρησιμοποιείται μεγάλος βαθμός ανάλυσης, ενώ για τον εντοπισμό συγκεκριμένων χρονικών τμημάτων χρησιμοποιείται μικρός βαθμός ανάλυσης. Για παράδειγμα, εφαρμογές Αναγνώρισης Μουσικής χρησιμοποιούν μικρό βαθμό ανάλυσης, καθώς στοχεύουν στην αναγνώριση συγκεκριμένου χρονικού τμήματος της ηχογράφησης που παίρνουν ως είσοδο, ώστε να ανακτήσουν το αντίστοιχο μουσικό κομμάτι.

1.3 ΤΟ ΠΡΟΒΛΗΜΑ ΤΟΥ ΔΙΑΧΩΡΙΣΜΟΥ ΜΟΥΣΙΚΩΝ ΠΗΓΩΝ

Ο διαχωρισμός μουσικών πηγών αποτελεί το βασικό αντικείμενο μελέτης της παρούσας εργασίας, το οποίο θα αναλυθεί εκτενέστερα στα επόμενα κεφάλαια, και είναι ένα από τα σημαντικότερα προβλήματα της Ανάκτησης Μουσικής Πληροφορίας. Οι περισσότερες δημόσια διαθέσιμες μουσικές ηχογραφήσεις (π.χ. CD, YouTube, iTunes, Spotify) διανέμονται ως μονοφωνικές ή στερεοφωνικές μίξεις. Αυτό συνήθως σημαίνει ότι πολλά μουσικά όργανα και φωνές συνυπάρχουν σε μια ηχογράφηση δύο καναλιών και επιπλέον οι πηγές πιθανώς να έχουν υποβληθεί σε επεξεργασία με την προσθήκη φίλτρων κατά τη διαδικασία της μίξης. Ο στόχος του Διαχωρισμού Μουσικών Πηγών είναι η ανάκτηση ενός ή περισσότερων σημάτων μουσικών πηγών δοσμένης μιας ορισμένης μίξης [10].

Ένα παράδειγμα της MIR που εφαρμόζει διαχωρισμό μουσικών πηγών είναι το σύστημα αναγνώρισης ερμηνευτή (singer identification). Η βασική δυσκολία αυτών των συστημάτων για επιτυχή αναγνώριση του ερμηνευτή έγγυται στις αρνητικές επιπτώσεις της συνοδείας, δηλαδή της ενορχήστρωσης. Συνεπώς, για την επιτυχή ανάκτηση της ζητούμενης μουσικής πληροφορίας, το σύστημα πρέπει να επικεντρωθεί στο φωνητικό τμήμα της ηχητικής μίξης και άρα να απομονώσει το σήμα που αντιστοιχεί στα φωνητικά από την ενορχήστρωση [11].

2 ΔΙΑΧΩΡΙΣΜΟΣ ΜΟΥΣΙΚΩΝ ΠΗΓΩΝ

Ο διαχωρισμός ηχητικών πηγών είναι ένα κυρίαρχο θέμα έρευνας της επιστημονικής κοινότητας τα τελευταία χρόνια, με το ενδιαφέρον γύρω από αυτό να αυξάνεται όλο και περισσότερο όσο εμφανίζονται καινούργιες μέθοδοι και τεχνολογίες. Τα μουσικά σήματα έχουν συγκεκριμένα χαρακτηριστικά που τα διαφοροποιούν από άλλα ηχητικά σήματα, όπως σήματα ομιλίας, και συνεπώς, ο διαχωρισμός μουσικών πηγών (Music Source Separation – MSS) είναι ακόμα πιο περίπλοκο ζήτημα, από το διαχωρισμό του σήματος θορύβου από το καθαρό σήμα ομιλίας.

Παρά τη δύσκολη φύση του, ο MSS λαμβάνει όλο και περισσότερο ενδιαφέρον, με την κοινότητα διαχωρισμού μουσικών πηγών να διεξάγει κάθε ενάμιση χρόνο, από το 2008, τον επιστημονικό διαγωνισμό SiSEC (Signal Separation Evaluation Campaign), ο οποίος αποτελεί σημείο αναφοράς για σύγκριση και αξιολόγηση όσο το δυνατόν περισσότερων μεθόδων στο θέμα του διαχωρισμού πηγών, ενώ επιπρόσθετα, παρέχει δεδομένα τα οποία η κοινότητα μπορεί να χρησιμοποιήσει για σχεδίαση και αξιολόγηση νέων μέθοδων [12].

2.1 Η ΣΥΝΗΘΗΣ ΔΙΑΔΙΚΑΣΙΑ ΔΙΑΧΩΡΙΣΜΟΥ ΜΟΥΣΙΚΩΝ ΠΗΓΩΝ

Ο ήχος δημιουργείται από τη μηχανική ταλάντωση ενός αερίου, υγρού ή ελαστικού μέσου η οποία μεταφέρει την ενέργεια της πηγής με μορφή ηχητικών κυμάτων. Ως επακόλουθο, ο ήχος καταγράφεται ως κυματομορφή, δηλαδή ως μια χρονοσειρά μετρήσεων της μετατόπισης του μικροφώνου σε απόκριση αυτών των κυμάτων πίεσης, και αναπαράγεται εάν το διάφραγμα ενός μεγαφώνου μετακινηθεί σύμφωνα με την καταγεγραμμένη κυματομορφή. Τα πολυκαναλικά σήματα αποτελούνται από πολλές κυματομορφές. Συνήθως, τα ηχητικά σήματα είναι στερεοφωνικά, δηλαδή εμπεριέχουν δύο κυματομορφές [13].

Ο MSS περιλαμβάνει την ανάλυση ενός ηχητικού σήματος $\boldsymbol{x}$ στο πεδίο του χρόνου, στις επιμέρους μουσικές πηγές $\boldsymbol{y_j}$ που το απαρτίζουν, όπου $\boldsymbol{x}$ και $\boldsymbol{y_j}$ είναι διανύσματα δειγμάτων στο χρόνο. Ο δείκτης j δηλώνει την πηγή, με $j \in \{1 \dots J\}$ και J το σύνολο των πηγών. Οι αναπαραστάσεις χρόνου-συχνότητας (time-frequency representations TF) συμβολίζονται με κεφαλαία γράμματα, με το $\boldsymbol{X}$ να δηλώνει το σύνθετο φασματογράφημα του $\boldsymbol{x}$ και το $\boldsymbol{Y_j}$ το σύνθετο φασματογράφημα της πηγής $\boldsymbol{y_j}$. Μια αναπαράσταση TF του ήχου είναι μια δισδιάστατη προβολή του ηχητικού σήματος, η οποία αναπαρίσταται στο πεδίο του χρόνου και στο πεδίο της συχνότητας ταυτόχρονα [14]. Με άλλα λόγια, είναι ένα μητρώο που κωδικοποιεί το χρονικά μεταβαλλόμενο φάσμα της κυματομορφής [13]. Τα στοιχεία του μητρώου ονομάζονται κάδοι TF (TF bins) και ουσιαστικά είναι τα διακριτά συχνοτικά και χρονικά διαστήματα $(\boldsymbol{k}, \boldsymbol{n})$ στα οποία χωρίζεται το συνολικό συχνοτικό και χρονικό εύρος του σήματος σε μια αναπαράσταση TF. Οι τιμές που ανήκουν στο εκάστοτε συχνοτικό και χρονικό διάστημα αντι-

καθίστανται από μία αντιπροσωπευτική τιμή $\boldsymbol{k}$ και $\boldsymbol{n}$ αντίστοιχα. Οι παρενθέσεις χρησιμοποιούνται για να δηλώσουν έναν μεμονωμένο κάδο TF. Το φασματογράφημα (magnitude spectrogram) της πηγής j ορίζεται ως $\boldsymbol{S_j} = |\boldsymbol{Y_j}|$, ενώ $\hat{\boldsymbol{Y}}_j$ και $\hat{\boldsymbol{S}}_j$ συμβολίζουν την εκτίμηση της αναπαράστασης TF της πηγής και την εκτίμηση του φασματογραφήματος αντίστοιχα [10]. Μία σύνοψη των βημάτων της διαδικασίας του MSS παρουσιάζεται στην **Εικ. 1**.

2.1.1 ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΣ TF

Η αναπαράσταση TF που έχει χρησιμοποιηθεί περισσότερο για τον διαχωρισμό μουσικών πηγών είναι ο Short Time Fourier Transform (STFT) [15], ο οποίος έχει το πλεονέκτημα ότι είναι υπολογιστικά αποδοτικός, αντιστρέψιμος, δηλαδή το σήμα μπορεί να αντιστραφεί πίσω στην αρχική κυματομορφή, και γραμμικός: η μίξη ισούται με το άθροισμα των πηγών στο πεδίο μετασχηματισμού, $\boldsymbol{X} = \sum_j \boldsymbol{Y_j}$ και συνεπώς, ο διαχωρισμός μπορεί να γίνει σε αυτό το πεδίο.

Πιο συγκεκριμένα, έστω $\boldsymbol{x}(\boldsymbol{n})$ ένα σήμα στο πεδίο του χρόνου. Ο STFT του σήματος ορίζεται ως:

$$X_n\left(e^{j\omega_k}\right) = X(m,\omega_k) = \sum_{m=-\infty}^{+\infty} w(n-m)x(m)e^{-j\omega_k m} \quad (1)$$

όπου $\boldsymbol{m}$ η χρονική μετατόπιση, $\boldsymbol{\omega_k}$ η συχνότητα και $\boldsymbol{w}(n)$ το παράθυρο, το οποίο προσδιορίζει το τμήμα του $x(n)$ του οποίου θα υπολογίσει τον μετασχηματισμό Fourier. Συνεπώς, ο STFT είναι μια σύνθετη συνάρτηση η οποία διαιρεί ένα σήμα μεγαλύτερου χρόνου σε μικρότερα τμήματα ίσου μήκους και στη συνέχεια υπολογίζει τον μετασχηματισμό Fourier σε κάθε ένα από αυτά τα τμήματα χωριστά [16]. Ουσιαστικά, ο STFT αποτελεί μια ακολουθία μετασχηματισμών Fourier ενός παραθυροποιημένου σήματος. Εφόσον η συχνότητα του σήματος μεταβάλλεται αρκετά κατα τη χρονική του διάρκεια, ο STFT παρέχει χρονικά τοπικές συχνοτικές πληροφορίες για αυτό. Αντίθετα, ο κλασσικός μετασχηματισμός Fourier, παρέχει πληροφορίες για τη συχνότητα του σήματος καθ' όλη τη χρονική του διάρκεια [17]. Το $|X(m,\omega_k)|$ ή το $|X(m,\omega_k)|^2$, δηλαδή το μέτρο ή η ισχύς (magnitude or power) του STFT ονομάζεται φασματογράφημα.

2.1.2 ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΩΝ ΠΗΓΩΝ (SOURCE MODELING)

Οι περισσότερες μέθοδοι MSS επικεντρώνονται στην ανάλυση του φασματογραφήματος U = |X| της μίξης. Στόχος σε αυτό το στάδιο, είναι να προκύψει μια εκτίμηση του φασματογραφήματος της μουσικής πηγής που επιθυμούμε να διαχωρίσουμε (target source).

2.1.3 ΦΙΛΤΡΑΡΙΣΜΑ (FILTERING)

Σκοπός σε αυτό το στάδιο είναι να εκτιμηθούν τα σήματα των διαχωρισμένων μουσικών πηγών. Αυτό συνήθως επιτυγχάνεται πολλαπλασιάζοντας τα στοιχεία, δηλαδή τους κάδους TF, της αναπαράστασης TF του σήματος μίξης με κάποια μάσκα TF (TF mask). Μια μάσκα TF, στη γενική της μορφή, αποτελείται από τιμές στο διάστημα [0, 1], οι οποίες καθορίζονται ανάλογα με το κατά πόσο τα περιεχόμενα κάθε κάδου TF ανήκουν στην πηγή που επιθυμούμε να διαχωρίσουμε. Η πιο κοινή μορφή μάσκας TF είναι το γενικευμένο φίλτρο Wiener (Generalized Wiener Filter – GWF) [18]. Μια ειδική περίπτωση του GWF είναι η δυαδική μάσκα (binary mask), όπου θεωρείται ότι μόνο μία πηγή έχει ενέργεια σε έναν δεδομένο κάδο TF $(\boldsymbol{k}, \boldsymbol{n})$, με συνέπεια οι τιμές της μάσκας να είναι είτε 0 είτε 1 [10]. Πιο συγκεκριμένα, η τιμή 1 υποδεικνύει ότι η ενέργεια στον συγκεκριμένο κάδο προέρχεται κυρίως από την πηγή που επιθυμούμε να διαχωρίσουμε και συνεπώς πρέπει να διατηρηθεί, ενώ η τιμή 0 υποδεικνύει ότι πρέπει να αφαιρεθεί. Δεδομένων των X και $\hat{S}_j$, ο STFT της πηγής $j = 1$ μπορεί να εκτιμηθεί χρησιμοποιώντας το GWF ως εξής:

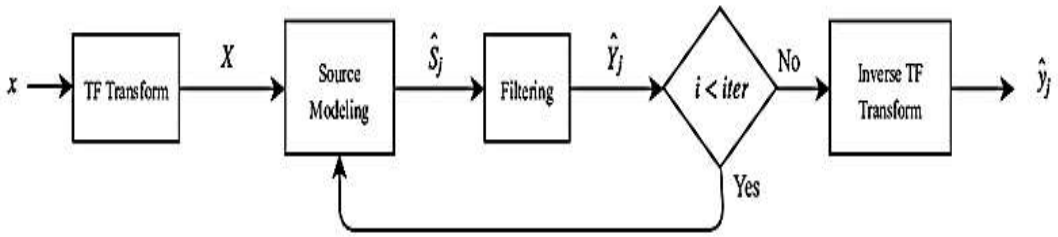
$$\hat{Y}_1(k,n) = \frac{X(k,n)\hat{S}_1(k,n)}{\sum_j \hat{S}_j(k,n)} \quad (2)$$

Η ίδια διαδικασία εφαρμόζεται για όλες τις πηγές της μίξης. Αυτό σημαίνει ότι για κάθε κάδο TF $(\boldsymbol{k}, \boldsymbol{n})$, η εκτίμηση της αναπαράστασης TF της πηγής j, υπολογίζεται από το λόγο του φασματογραφήματος του αρχικού σήματος πολλαπλασιασμένο με την εκτίμηση του φασματογραφήματος της πηγής j για τον συγκεκριμένο κάδο TF, προς το άθροισμα των φασματογραφημάτων των πηγών σε αυτό το σημείο. Ουσιαστικά, λειτουργεί σαν ένα φίλτρο που μεταβάλλεται χρονικά, για να εξασθενίσει ή να αφήσει να περάσουν οι κατάλληλες συχνότητες, ώστε να ανακτηθούν οι διαχωρισμένες πηγές.

Συνήθως, εφαρμόζεται κάποια επαναληπτική διαδικασία στα βήματα της μοντελοποίησης και του φιλτραρίσματος, έως ότου οι παράμετροι των πηγών που λαμβάνονται να είναι αρκετά καλές ώστε να οδηγήσουν σε ένα επιτυχές φιλτράρισμα.

2.1.4 ΑΝΤΙΣΤΡΟΦΟΣ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΣ (INVERSE TRANSFORM)

Το τελικό στάδιο στη διαδικασία του MSS είναι η ανάκτηση των κυματομορφών των πηγών στο πεδίο του χρόνου χρησιμοποιώντας τον αντίστροφο μετασχηματισμό της αναπαράστασης TF που εφαρμόστηκε στα προηγούμενα στάδια.



Εικ. 1: Η συνήθης διαδικασία διαχωρισμού μουσικών πηγών. Η μεταβλητή iter ισούται με τον συνολικό αριθμό επαναλήψεων και ο δείκτης i ισούται με την τρέχουσα επανάληψη.

2.2 ΠΡΟΣΕΓΓΙΣΕΙΣ ΓΙΑ ΤΟ ΔΙΑΧΩΡΙΣΜΟ ΠΗΓΩΝ

Παρακάτω αναφέρονται ενδεικτικά ορισμένες προσεγγίσεις που έχουν εφαρμοστεί για το διαχωρισμό μουσικών πηγών και κατ' επέκταση για το διαχωρισμό του κυρίαρχου (lead) σήματος, για παράδειγμα του σήματος της φωνής, από τη συνοδεία (accompaniment).

2.2.1 ΑΝΑΛΥΣΗ ΑΝΕΞΑΡΤΗΤΩΝ ΣΥΝΙΣΤΩΣΩΝ (ICA)

Μία από τις πρώτες τεχνικές που χρησιμοποιήθηκαν για την εύρεση των πηγών είναι η Ανάλυση Ανεξαρτήτων Συνιστωσών (Independent Component Analysis – ICA) [19], η οποία βασίζεται στις ακόλουθες δύο υποθέσεις:

1. Οι συνιστώσες πρέπει να είναι στατιστικά ανεξάρτητες μεταξύ τους, δηλαδή, η γνώση γύρω από μία συνιστώσα x, δεν δίνει καμία επιπλέον πληροφορία για κάποια άλλη συνιστώσα y. Αυτό, μαθηματικά μεταφράζεται ως $p(x,y) = p(x)p(y)$, όπου $p(x,y)$ η από κοινού πυκνότητα πιθανότητας των x, y και $p(x)$, $p(y)$ η συνάρτηση πυκνότητας πιθανότητας των x, y αντίστοιχα.

2. Οι συνιστώσες δεν πρέπει να έχουν Gaussian κατανομές [20].

Έστω $\boldsymbol{x}$ το διάνυσμα με στοιχεία $x_1, \dots, x_n$ τα σήματα μίξης και $\boldsymbol{s}$ το διάνυσμα με στοιχεία $s_1, \dots, s_n$ τις πηγές, οι οποίες υποθέτουμε ότι είναι στατιστικά ανεξάρτητες και δεν έχουν Gaussian κατανομή. Για κάθε σήμα μίξης x_j ισχύει ότι:

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \cdots + a_{jn}s_n \quad (3)$$

όπου a_{ji} είναι τα στοιχεία του μητρώου μίξης $\boldsymbol{A}$. Επομένως, το παραπάνω μοντέλο μίξης, μπορεί να γραφτεί διανυσματικά ως:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} \quad (4)$$

Η ICA υπολογίζει το μητρώο μίξης $\boldsymbol{A}$, ώστε στη συνέχεια να υπολογίσει το αντίστροφό του, έστω $\boldsymbol{W}$, και να πάρει τις ανεξάρτητες συνιστώσες, δηλαδή τις εκτιμήσεις των πηγών, ως εξής:

$$\hat{\boldsymbol{s}} = \boldsymbol{W}\boldsymbol{x} \quad (5)$$

Για τον υπολογισμό του W, ο αλγόριθμος εκτελεί πολλαπλές επαναλήψεις και συγκλίνει όταν το W δίνει πηγές που είναι κατά το μέγιστο μη-Gaussian [21]. Βασική προϋπόθεση για τη σωστή λειτουργία της ICA είναι οι μουσικές πηγές να είναι ίσες ή λιγότερες από τον αριθμό των καναλιών του σήματος μίξης. Ωστόσο, στα σήματα μουσικής συνήθως παρατηρείται το αντίθετο, δηλαδή, οι πηγές να είναι περισσότερες από τα κανάλια [10].

Ως αποτέλεσμα των ελλείψεων της ICA, αναπτύχθηκαν αλγόριθμοι που λειτουργούσαν στην περίπτωση που οι πηγές ήταν περισσότερες από τα κανάλια, όπως ο αλγόριθμος DUET (Degenerate Unmixing Estimation Technique) [22], ο ADRess (Azimuth Discrimination Resynthesis) [23] και ο PROJET (PROJection Estimation Technique) [24], οι οποίοι υποθέτουν ότι οι αναπαραστάσεις TF των πηγών έχουν πολύ μικρή επικάλυψη μεταξύ τους. Η υπόθεση αυτή, ενώ ισχύει σε πολύ μεγάλο βαθμό για την ομιλία, δεν ισχύει τόσο για τη μουσική. Παρόλα αυτά, έχει αποδειχθεί χρήσιμη σε αρκετές περιπτώσεις και έχει οδηγήσει σε γρήγορους αλγορίθμους που εκτελούν το διαχωρισμό σε πραγματικό χρόνο.

2.2.2 ΗΜΙΤΟΝΟΕΙΔΗ ΜΟΝΤΕΛΑ (SINUSOIDAL MODELS)

Ένα ημιτονοειδές μοντέλο αναλύει τον ήχο σε ένα σύνολο ημιτονοειδών κυμάτων χρονικά μεταβαλλόμενης συχνότητας και πλάτους [13] [25]. Το ημιτονοειδές μοντέλο προσφέρει μια σαφή αναπαράσταση ενός μουσικού σήματος, το οποίο στις περισσότερες περιπτώσεις αποτελείται από ένα σύνολο θεμελιωδών συχνοτήτων και τις αντίστοιχες αρμονικές σειρές τους [10]. Εάν οι τονικότητες που εμφανίζονται στην πηγή που θέλουμε να διαχωρίσουμε, και τα φασματικά χαρακτηριστικά των αρμονικών κάθε τόνου είναι γνωστά ή μπορούν να εκτιμηθούν, τότε τα ημιτονοειδή μοντέλα αποτελούν μια αποδοτική τεχνική για τον διαχωρισμό αρμονικών πηγών. Ωστόσο, δεδομένης της πολυπλοκότητας του μοντέλου και της πολύ λεπτομερούς γνώσης που απαιτείται να έχουμε για την πηγή που θέλουμε να διαχωρίσουμε, ώστε να οδηγηθούμε σε μια ρεαλιστική αναπαράσταση, η χρήση ημιτονοειδών μοντέλων για το πρόβλημα του MSS είναι περιορισμένη. Ωστόσο, έχουν χρησιμοποιηθεί αρκετά στο παρελθόν, σε διάφορες προσεγγίσεις, όπως αυτή του Maher [26], των Meron και Hirose για να διαχωρίσουν τα φωνητικά από τη συνοδεία πιάνου [27], των Zhang και Zhang [25] και πιο πρόσφατα των Fujihara et al. [11] [28] για αναγνώριση ερμηνευτή (singer identification).

2.2.3 ΜΟΝΤΕΛΑ ΠΑΡΑΓΟΝΤΟΠΟΙΗΣΗΣ ΦΑΣΜΑΤΟΓΡΑΦΗΜΑΤΟΣ

Μία μεγάλη ομάδα μοντέλων που ασχολούνται με τον MSS, βασίζεται σε τεχνικές παραγοντοποίησης φασματογραφήματος (spectrogram factorization techniques). Η πιο διαδεδομένη τεχνική από αυτές είναι η Μη-Αρνητική Παραγοντοποίηση Μητρώων (Non-Negative Matrix Factorization - NMF) [29], η οποία προτάθηκε για πρώτη φορά από τους Lee και Seung και βασίστηκε σε παλαιότερη εργασία του Paatero πάνω στην Παραγοντοποίηση Θετικών Μητρώων (Positive Matrix Factorization) [30]. Η NMF στοχεύει στο να παραγοντοποιήσει ένα μη-αρνητικό μητρώο σε δύο μη-αρνητικά μητρώα [31].

Όσον αφορά το πρόβλημα του MSS, εκμεταλλεύεται την υπόθεση ότι το φασματογράφημα της συνοδείας (accompaniment) ενός μουσικού κομματιού μπορεί να αναπαρασταθεί από λίγα μόνο στοιχεία [12]. Εφαρμόζεται στο μη-αρνητικό φασματογράφημα ισχύος της μίξης $\boldsymbol{U} \in \mathbb{R}^{\geq \mathbf{0}, \boldsymbol{L} \times \boldsymbol{N}}$, με σκοπό την παραγοντοποίηση του $\boldsymbol{U}$ σε γινόμενο $\boldsymbol{U} \approx \boldsymbol{WH}$, όπου $\boldsymbol{W} \in \mathbb{R}^{\geq \mathbf{0}, \boldsymbol{L} \times \boldsymbol{R}}$ είναι ένα μητρώο διανυσμάτων βάσης, το οποίο μοντελοποιεί τα φασματικά χαρακτηριστικά των πηγών, δηλαδή οι είσοδοί του αφορούν το πεδίο της συχνότητας, και $\boldsymbol{H} \in \mathbb{R}^{\geq \mathbf{0}, \boldsymbol{R} \times \boldsymbol{N}}$ είναι ένα μητρώο χρονικών ενεργοποιήσεων, δηλαδή οι είσοδοί του αφορούν το πεδίο του χρόνου. Η επιλογή της παραμέτρου R (με $R \leq L$) παίζει καθοριστικό ρόλο στα αποτελέσματα που θα πάρουμε από την παραγοντοποίηση. Εάν $R = L$, τότε τα περιεχόμενα των $\boldsymbol{W}$ και $\boldsymbol{H}$ δεν μας παρέχουν καμία ουσιαστική πληροφορία. Μειώνοντας την τιμή του R, τα $\boldsymbol{W}$ και $\boldsymbol{H}$ αρχίζουν να παίρνουν τιμές που περιγράφουν καλύτερα τα βασικά στοιχεία από τα οποία αποτελείται το $\boldsymbol{U}$. Συνεπώς, εάν επιλεχθεί μια κατάλληλη τιμή για το R, τότε είναι εφικτή η αποτελεσματική εξαγωγή των κύριων στοιχείων του φασματογραφήματος ισχύος $\boldsymbol{U}$ [32].

Η παραγοντοποίηση αντιμετωπίζεται σαν ένα πρόβλημα βελτιστοποίησης (optimization problem), όπου η απόκλιση ή το λάθος ανακατασκευής μεταξύ των $\boldsymbol{U}$ και $\boldsymbol{WH}$ ελαχιστοποιείται με τη χρήση κάποιας μετρικής απόκλισης D:

$$\min_{W,H \geq 0} D(U \parallel WH) = D\left(U \parallel \hat{U}\right) = \sum_{l=1}^{L} \sum_{n=1}^{N} d(u_{ln} \mid \hat{u}_{ln}) \quad (6)$$

όπου, $d(x \mid y)$ είναι μια βαθμωτή συνάρτηση κόστους (scalar cost function).

Η επιλογή της συνάρτησης κόστους θα πρέπει να καθορίζεται από τον τύπο των δεδομένων που θα αναλυθούν. Η πιο συνηθισμένη επιλογή είναι η Ευκλείδια απόσταση (Euclidean distance):

$$d_{EUC}(x \mid y) = \frac{1}{2}(x - y)^2 \quad (7)$$

Ωστόσο, σε εφαρμογές επεξεργασίας ήχου, συνήθως προτιμώνται οι αποκλίσεις Kullback-Leibler (KL):

$$d_{KL}(x \mid y) = x \log \frac{x}{y} - x + y \quad (8)$$

και Itakura-Saito (IS) [33]:

$$d_{IS}(x \mid y) = \frac{x}{y} - \log\frac{x}{y} - 1 \quad (9)$$

2.3 ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΜΟΝΤΕΛΩΝ MSS

Για να μπορούμε να συγκρίνουμε τις διάφορες μεθόδους και μοντέλα διαχωρισμού μουσικών πηγών ως προς την απόδοσή τους, καθίσταται αναγκαία η εύρεση μεθόδων και μετρικών αξιολόγησής τους. Η μέθοδος MUSHRA (MUlti Stimulus test with Hidden Reference and Anchor) βασίστηκε στην ανθρώπινη αντίληψη του ακροατή για την αξιολόγηση των μοντέλων. Παρόλα αυτά, η ανάγκη για την ύπαρξη αντικειμενικών μετρικών οδήγησε στην ευρεία χρήση του Blind Source Separation Evaluation (BSSEval) toolbox.

2.3.1 MULTI STIMULUS TEST WITH HIDDEN REFERENCE AND ANCHOR

Με γνώμονα την ανθρώπινη αντίληψη του ακροατή ως προς το μουσικό σήμα που φτάνει στα αυτιά του, αναπτύχθηκε η μέθοδος αξιολόγησης MUSHRA (MUlti Stimulus test with Hidden Reference and Anchor), η οποία ουσιαστικά βασίστηκε σε υποκειμενικά ακουστικά τεστ [34]. Ωστόσο, αποδείκτηκε δύσκολο να υπάρξει ένα συγκεκριμένο μέτρο αξιολόγησης ως αποτέλεσμα, διότι αυτό μεταβαλλόταν ανάλογα με την εφαρμογή που χρησιμοποιούνταν ο διαχωρισμός [13]. Επιπλέον, η διαδικασία αυτή κοστίζει τόσο σε χρόνο όσο και σε πόρους, καθώς απαιτούνται εθελοντές για τα τεστ και χρειάζεται συγκεκριμένη τεχνογνωσία για να διεξαχθούν σωστά [10]. Ένας τρόπος για να αυξηθεί ο αριθμός των συμμετεχόντων είναι η διεξαγωγή διαδικτυακών πειραμάτων. Οι συγγραφείς του [35] αναφέρουν ότι κατάφεραν να συγκεντρώσουν 530 συμμετέχοντες σε μόλις 8.2 ώρες και τα αποτελέσματα που πήραν ήταν συγκρίσιμα με αυτά που προέκυψαν στο εργαστηριακό πλαίσιο.

2.3.2 BLIND SOURCE SEPARATION EVALUATION

Το Blind Source Separation Evaluation (BSS Eval) toolbox [36] [37] είναι από τις πρώτες μεθόδους αξιολόγησης που διατέθηκαν στο ευρύ κοινό και τα αποτελέσματά του, σε ορισμένες περιπτώσεις συσχετίζονται με την ανθρώπινη αντίληψη. Αυτό έχει ως επακόλουθο να χρησιμοποιείται ευρέως μέχρι σήμερα [38] [39]. Το BSS Eval παρέχει τις ακόλουθες μετρικές ποιότητας σε decibel (dB) των αποτελεσμάτων του διαχωρισμού:

- Λόγος πηγής προς παραμόρφωση (Source to Distortion Ratio: SDR),

- Λόγος πηγής προς τις παρεμβάλλουσες πηγές (Source to Interferences Ratio: SIR),
- Λόγος πηγής προς σφάλματα (Source to Artifacts Ratio: SAR),
- Λόγος σήματος προς θόρυβο (Signal to Noise Ratio: SNR).

Οι παραπάνω μετρικές βασίζονται στην ανάλυση δεδομένης εκτίμησης $\hat{s}(t)$ μιας πηγής $s_i(t)$ ως το άθροισμα:

$$\hat{s}(t) = s_{target}(t) + e_{interf}(t) + e_{noise}(t) + e_{artif}(t) \quad (10)$$

όπου $s_{target}(t)$ είναι μια επιτρεπόμενη παραμόρφωση της πηγής που θέλουμε να διαχωρίσουμε, $e_{interf}(t)$ ένα επιτρεπόμενο σφάλμα λόγω παρεμβολής από άλλες πηγές, $e_{noise}(t)$ ένα επιτρεπόμενο σφάλμα λόγω θορύβου (ο οποίος δεν προέρχεται από τις άλλες πηγές) και $e_{artif}(t)$ το σφάλμα λόγω τεχνητού θορύβου που προκύπτει από τον αλγόριθμο διαχωρισμού [36].

Δεδομένης, συνεπώς, της παραπάνω ανάλυσης, δίνουμε τον καθολικό ορισμό των μετρικών απόδοσης [37]:

- Source to Distortion Ratio:

$$SDR := 10\log_{10}\frac{\left||s_{target}|\right|^2}{\left||e_{interf}+e_{noise}+e_{artif}|\right|^2} \quad (11)$$

- Source to Interferences Ratio:

$$SIR := 10\log_{10}\frac{\left||s_{target}|\right|^2}{\left||e_{interf}|\right|^2} \quad (12)$$

- Signal to Noise Ratio:

$$SNR := 10\log_{10}\frac{\left||s_{target}+e_{interf}|\right|^2}{||e_{noise}||^2} \quad (13)$$

- Sources to Artifacts Ratio:

$$SAR := 10\log_{10}\frac{\left||s_{target}+e_{interf}+e_{noise}|\right|^2}{\left||e_{artif}|\right|^2} \quad (14)$$

Με βάση τις παραπάνω σχέσεις, αντιλαμβανόμαστε τα εξής:

- Ο λόγος πηγής προς παραμόρφωση (SDR) καταγράφει τη συνολική ποιότητα διαχωρισμού του αλγορίθμου, δηλαδή, το πόσο καθαρό είναι το σήμα της πηγής που θέλουμε να διαχωρίσουμε από οποιασδήποτε μορφής παραμόρφωση ή θόρυβο.

- Ο λόγος πηγής προς παρεμβάλλουσες πηγές (SIR) καταγράφει την ικανότητα του αλγορίθμου να εξαλείψει τις παρεμβολές από τις άλλες πηγές και να διατηρήσει το σήμα της πηγής που θέλουμε να διαχωρίσουμε.

- Ο λόγος σήματος προς θόρυβο (SNR) καταγράφει το πόσο καθαρό είναι το σήμα από εξωγενείς θορύβους (π.χ. προσθετικός θόρυβος). Αυτού του είδους ο θόρυβος είναι ανεξάρτητος από τον θόρυβο που προσθέτουν οι υπόλοιπες πηγές στο σήμα της πηγής που θέλουμε να διαχωρίσουμε, γι' αυτό και έχουν δημιουργηθεί δύο διαφορετικές μετρικές για τον υπολογισμό τους.

- Ο λόγος πηγής προς σφάλματα (SAR) καταγράφει την ικανότητα του δικτύου να παράγει αποτελέσματα υψηλής ποιότητας, χωρίς την εισαγωγή πρόσθετων σφαλμάτων που ενδεχομένως εισάγονται κατά τη διάρκεια εκτέλεσης των βημάτων του αλγορίθμου.

Είναι σημαντικό να επισημάνουμε εδώ, ότι η ανάπτυξη νέων αντικειμενικών μετρικών αξιολόγησης των αποτελεσμάτων του MSS, σχετικών με την ανθρώπινη αντίληψη, παραμένει ένα ανοιχτό ζήτημα [40], το οποίο είναι εξαιρετικά κρίσιμο για μελλοντική έρευνα στον συγκεκριμένο τομέα.

3 ΔΙΑΧΩΡΙΣΜΟΣ ΜΕ ΜΕΘΟΔΟΥΣ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ

Τα μοντέλα νευρωνικών δικτύων, εμπνευσμένα από τον ανθρώπινο εγκέφαλο, έχουν σχεδιαστεί κατάλληλα για την αναγνώριση προτύπων εικόνας, ήχου, κειμένου, κοκ, μετεφρασμένων υπό την μορφή διανυσμάτων.

Τα βαθιά νευρωνικά δίκτυα (Deep Neural Networks – DNNs) είναι μια συλλογή από νευρώνες που συνδέονται μεταξύ τους σε μια ακολουθία πολλαπλών επιπέδων. Οι νευρώνες λαμβάνουν ως είσοδο τις ενεργοποιήσεις των νευρώνων του προηγούμενου επιπέδου, εκτελούν κάποια επεξεργασία και παράγουν μία έξοδο, η οποία θα αποτελέσει την είσοδο του επόμενου επιπέδου [41]. Οι συνδέσεις μεταξύ των νευρώνων διαφέρουν ως προς τη σημαντικότητά τους, η οποία προσδιορίζεται από τον συντελεστή βάρους. Η επεξεργασία που γίνεται σε κάθε νευρώνα, είναι ουσιαστικά η μετατροπή της αναπαράστασης των δεδομένων, ξεκινώντας από τα δεδομένα εισόδου, σε μια πιο αφαιρετική μορφή, όσο προχωράει σε υψηλότερο επίπεδο [42].

Για να χρησιμοποιηθεί ένα νευρωνικό δίκτυο πρέπει πρώτα να εκπαιδευτεί. Η μάθηση συνίσταται στον προσδιορισμό των κατάλληλων συντελεστών βάρους, ώστε το νευρωνικό δίκτυο να εκτελεί τους επιθυμητούς υπολογισμούς. Ο ρόλος των συντελεστών βάρους μπορεί να ερμηνευτεί ως αποθήκευση γνώσης, η οποία παρέχεται μέσω παραδειγμάτων. Με αυτόν τον τρόπο τα νευρωνικά δίκτυα μαθαίνουν το περιβάλλον τους και βελτιώνουν την απόδοσή τους [43].

3.1 ΒΑΣΙΚΑ ΣΥΣΤΑΤΙΚΑ ΣΤΟΙΧΕΙΑ ΕΝΟΣ ΝΕΥΡΩΝΙΚΟΥ ΔΙΚΤΥΟΥ

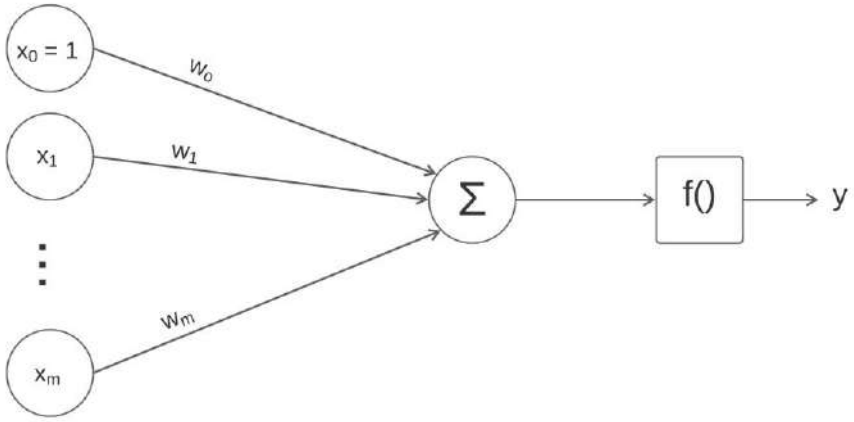
Υπάρχουν πολλά, και πολύ διαφορετικά μεταξύ τους, μοντέλα νευρωνικών δικτύων, ωστόσο όλα έχουν τα ακόλουθα βασικά συστατικά στοιχεία:

i. Τα επίπεδα ενός νευρωνικού δικτύου αποτελούνται από κόμβους, οι οποίοι ονομάζονται νευρώνες.
ii. Κάθε νευρωνικό δίκτυο χρησιμοποιεί κάποια αντικειμένικη συνάρτηση.

3.1.1 ΝΕΥΡΩΝΑΣ

Τα επίπεδα ενός νευρωνικού δικτύου αποτελούνται από κόμβους, οι οποίοι ονομάζονται νευρώνες. Ο νεύρωνας είναι η θεμελιώδης μονάδα επεξεργασίας πληροφορίας για την λειτουργία ενός νευρωνικού δικτύου. Η λειτουργία ενός νευρώνα (**Εικ. 2**) είναι να πολλαπλασιάζει τα δεδομένα που δέχεται στην είσοδο του, με ένα σύνολο βαρών, τα οποία ονομάζονται συναπτικά βάρη, και των οποίων οι τιμές μεταβάλλονται κατά τη διάρκεια εκπαίδευσης, ανάλογα με το εκάστοτε κριτήριο βελτιστοποίησης. Στη συνέχεια, τα γινόμενα αυτά αθροίζονται και το άθροισμά τους χρησιμοποιείται ως το όρισμα της συνάρτησης ενεργοποίησης του νευρώνα. Από τη συνάρτηση ενεργοποίησης προκύπτει μια νέα τιμή

στην έξοδο, η οποία καθορίζει την ενεργοποίηση ή μη-ενεργοποίηση του νευρώνα (εξού και το όνομα) και τον βαθμό στον οποίο θα επηρεάσει το τελικό αποτέλεσμα.



Εικ. 2: Η λειτουργία ενός νευρώνα.

Τα βασικά στοιχεία του νευρώνα, με βάση την παραπάνω εικόνα είναι :

- Το διάνυσμα εισόδου $\boldsymbol{x} = [x_1, x_2, \dots, x_m]^T$,
- τα συναπτικά βάρη $\boldsymbol{w_l}$, με $l = 1, \dots, m$, τα οποία αντιστοιχούν στις συνάψεις ενός πραγματικού νευρώνα,
- ένας αθροιστής $\sum_{i=0}^{m} \boldsymbol{w_i}\, \boldsymbol{x_i}$, για την πρόσθεση των επιμέρους γινομένων και
- η συνάρτηση ενεργοποίησης ***f(u)*** του νευρώνα, η οποία δέχεται ως είσοδο $\boldsymbol{u}$ την έξοδο του αθροιστή, υποδεικνύει τον εντοπισμό του προτύπου και καθορίζει την έξοδο του νευρώνα σε σχέση με τις εισόδους και τους συντελεστές βάρους.

Συνήθως υπάρχει και η πρόσθετη παράμετρος $\boldsymbol{w_0}$ του νευρώνα, που ονομάζεται πόλωση (bias), και την οποία θεωρούμε ως το βάρος που σχετίζεται με την είσοδο $\boldsymbol{x_0}$ που έχει οριστεί μόνιμα στο 1 [44]. Με βάση τα παραπάνω, η συνάρτηση που υπολογίζει ένας νευρώνας είναι η

$$y = f(\mathrm{w}^{\mathrm{T}}\mathrm{x} + w_o) \quad (15)$$

όπου το εσωτερικό γινόμενο $\mathbf{w^T x}$ δείχνει την συσχέτιση μεταξύ της εισόδου $\mathbf{x}$ και του προτύπου που περιγράφει το $\mathbf{w}$.

Συνάρτηση ενεργοποίησης

Παρακάτω παρουσιάζονται μερικές από τις κυρίαρχες συναρτήσεις ενεργοποίησης.

a) Βηματική συνάρτηση:

$$f(x) = \begin{cases} 0, & x < 0 \\ 1, & x > 0 \end{cases} \quad (16)$$

Η συνάρτηση f(x) είναι ασυνεχής εφόσον δεν ορίζεται για x = 0, και συνεπώς δεν μπορεί να είναι παραγωγίσιμη στο σημείο αυτό (**Εικ. 3(a)**).

b) **Σιγμοειδής συνάρτηση:**
Είναι μία συνεχώς παραγωγίσιμη συνάρτηση, η οποία προσεγγίζει τη βηματική, και είναι η πιο συνηθισμένη μορφή συνάρτησης ενεργοποίησης που χρησιμοποιείται στην κατασκευή νευρωνικών δικτύων. Ένα παράδειγμα σιγμοειδούς συνάρτησης είναι η λογιστική συνάρτηση, η οποία ορίζεται από τη σχέση:

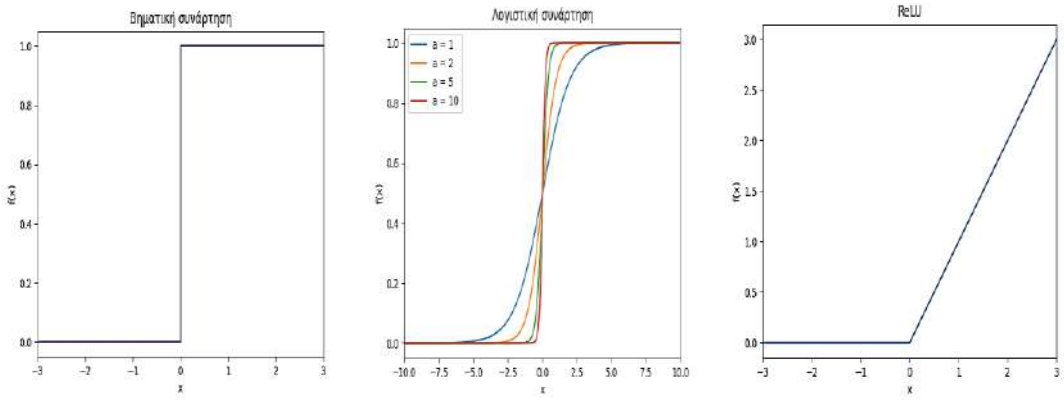
$$f(x) = \frac{1}{1 + e^{-ax}} \quad (17)$$

Μεταβάλλοντας την παράμετρο a, ουσιαστικά παίρνουμε διαφορετικές συναρτήσεις (**Εικ. 3(b)**).

c) **Rectified Linear Units (ReLU):**
Η ReLU είναι μία συνάρτηση ενεργοποίησης, την οποία εισήγαγαν οι συγγραφείς του [45], και έχει ισχυρή βιολογική και μαθηματική βάση. Από το 2011 και ύστερα είναι πλέον αποδεδειγμένο ότι συμβάλλει στην περαιτέρω βελτίωση της εκπαίδευσης των DNNs. Η ReLU ορίζεται ως:

$$f(x) = \max(0, x) \quad (18)$$

και είναι και αυτή μη-παραγωγίσιμη για x = 0 [46] (**Εικ. 3(c)**).



(a): Βηματική Συνάρτηση **(b)**: Λογιστική συνάρτηση **(c)**: ReLU

Εικ. 3: Συναρτήσεις ενεργοποίησης.

3.1.2 ΑΝΤΙΚΕΙΜΕΝΙΚΗ ΣΥΝΑΡΤΗΣΗ

Συνήθως, στα νευρωνικά δίκτυα, επιδιώκουμε την ελαχιστοποίηση της απόκλισης της εξόδου του δικτύου **ŷ(i)** από την πραγματική έξοδο **y(i)** για δεδομένη είσοδο **x(i)**, ή με άλλα λόγια, την ελαχιστοποίηση του σφάλματος. Για αυτόν το λόγο χρειάζεται να οριστεί μία αντικειμενική συνάρτηση, η οποία αναφέρεται συχνά και ως συνάρτηση κόστους (cost ή loss function) [47] [48]. Η συνάρτηση κόστους σχετίζεται άμεσα με την συνάρτηση ενεργοποίησης του επιπέδου εξόδου του νευρωνικού δικτύου και πρέπει να επιλέγεται προσεκτικά με βάση το είδος του προβλήματος που έχουμε. Παρακάτω δίνονται οι δύο κύριες αντικειμενικές συναρτήσεις που χρησιμοποιούνται στα νευρωνικά δίκτυα.

a) Τετραγωνικό Σφάλμα (Squared Error):

$$\mathcal{E}(i) = \frac{1}{2}\sum_{m=1}^{M}\left(y_m(i) - \hat{y}_m(i)\right)^2 \quad (19)$$

Το αποτέλεσμα είναι πάντα θετικό ανεξάρτητα από τις τιμές των ŷ(i) και y(i) και η ιδανική τιμή του σφάλματος θα ήταν το μηδέν (0). Συνήθως χρησιμοποιείται για προβλήματα παλινδρόμησης (regression problems), δηλαδή σε προβλήματα που η έξοδος είναι μια πραγματική ή συνεχής τιμή.

b) Διασταυρωμένη Εντροπία (Cross-Entropy):
Η διασταυρωμένη εντροπία προτιμάται ως συνάρτηση κόστους από το άθροισμα των τετραγώνων, όσον αφορά προβλήματα ταξινόμησης (classification problems), καθώς η χρήση της οδηγεί σε μειωμένο χρόνο εκπαίδευσης του δικτύου [49].

$$\mathcal{E}(i) = -\sum_{m=1}^{M} y_m(i) \ln \hat{y}_m(i) + \big(1 - y_m(i)\big) \ln(1 - \hat{y}_m(i)) \quad (20)$$

3.2 Ο ΑΛΓΟΡΙΘΜΟΣ BACK-PROPAGATION

Στόχος του αλγορίθμου back-propagation είναι η ελαχιστοποίηση της συνάρτησης κόστους, η οποία επιτυγχάνεται με την κατάλληλη ενημέρωση των συναπτικών βαρών. Η ενημέρωση των συναπτικών βαρών γίνεται υπολογίζοντας τις μερικές παραγώγους της συνάρτησης κόστους ως προς αυτά. Η σημασία της χρήσης των παραγώγων για την επίτευξη του στόχου του αλγορίθμου back-propagation είναι κρίσιμη, καθώς εξ' ορισμού, η παράγωγος μιας συνάρτησης *f*, δείχνει την τάση της τιμής της συνάρτησης (τιμή εξόδου) να μεταβάλλεται, όταν αλλάζει το όρισμά της *x* (τιμή εισόδου). Με άλλα λόγια, η παράγωγος δείχνει εάν η τιμή της συνάρτησης τείνει να αυξάνεται ή να μειώνεται. Κατά συνέπεια, οι μερικές παράγωγοι δείχνουν πόσο πρέπει να αλλάξει και προς ποια κατευθυνση (θετικη ή αρνητικη), μια παράμετρος *x*, ώστε να ελαχιστοποιηθεί η συνάρτηση *f* [50].

Ακολουθώντας το [51], υποθέτουμε δίκτυο $\boldsymbol{L}$ επιπέδων, όπου κάθε επίπεδο $\boldsymbol{l}$ απαρτίζεται από $\boldsymbol{k_l}$ νευρώνες, και έστω $l = 0$ το επίπεδο δεδομένων εισόδου και k_0 η διάστασή τους. Όλοι οι νευρώνες χρησιμοποιούν τη σιγμοειδή λογιστική συνάρτηση (17).

Για ένα σύνολο δεδομένων εκπαίδευσης έχουμε τα ζεύγη εισόδου- εξόδου $\big(\boldsymbol{x}(\boldsymbol{i}), \boldsymbol{y}(\boldsymbol{i})\big)$, *i = 1, ..., N*. Εφόσον το επίπεδο εισόδου $l = 0$ έχει k_0 νευρώνες, θα δέχεται k_0 σύνολα δεδομένων στην είσοδό του. Επομένως, το διάνυσμα εισόδου (input vector of features) είναι $\boldsymbol{x}(i) = \big[x_1(i), \dots, x_{k_0}(i)\big]^T$.
Αντίστοιχα, το διάνυσμα εξόδου είναι $\boldsymbol{y}(i) = [y_1(i), \dots, y_{k_L}(i)]^T$.

Η αντικειμενική συνάρτηση βασίζεται στην (19) και ορίζεται ως:

$$\mathcal{C} = \sum_{i=1}^{N} \mathcal{E}(i) = \frac{1}{2} \sum_{i=1}^{N} \sum_{m=1}^{M} \big(y_m(i) - \hat{y}_m(i)\big)^2 \quad (21)$$

Ενημερώνουμε τα συναπτικά βάρη, υπολογίζοντας τις μερικές παραγώγους της συνάρτησης κόστους ως προς αυτά, ως εξής:

$$w_j^l \leftarrow w_j^l + \Delta w_j^l \quad (22)$$

όπου, $\boldsymbol{w_j^l}$ τα συναπτικά βάρη του νευρώνα *j* στο επίπεδο l και

$$\Delta w_j^l = -\eta \frac{\partial \mathcal{C}}{\partial w_j^l} \quad (23)$$

όπου, $\boldsymbol{\eta}$ ονομάζεται η παράμετρος ρυθμού μάθησης (learning rate parameter) και καθορίζει το μέγεθος του βήματος ενημέρωσης σε κάθε επανάληψη, με στόχο την ελαχιστοποίηση της συνάρτησης κόστους, δηλαδή, αναπαριστά την ταχύτητα εκπαίδευσης του μοντέλου. Από την (21) έχουμε ότι $\mathcal{C} = \sum_{i=1}^{N} \mathcal{E}(i)$ και άρα, ουσιαστικά μας ενδιαφέρει να υπολογίσουμε την $\frac{\partial \mathcal{E}(i)}{\partial w_j^l}$.

Ξέρουμε ότι για την είσοδο της συνάρτησης ενεργοποίησης ενός νευρώνα ισχύει:

$\boldsymbol{u} = \sum_{i=1}^{m} w_i\, x_i$, όπου x_i είναι τα δεδομένα εισόδου του νευρώνα και w_i τα συναπτικά του βάρη. Συνεπώς, στο επίπεδο l, τα δεδομένα εισόδου του νευρώνα j, είναι ουσιαστικά, τα δεδομένα εξόδου του προηγούμενου επιπέδου, δηλαδή $x_j^l = y_r^{l-1}$. Τα συναπτικά βάρη του νευρώνα j στο επίπεδο l είναι $w_j^l = [w_{j0}^l, w_{j1}^l, \dots, w_{jk_l-1}^l]$. Επομένως, η είσοδος της συνάρτησης ενεργοποίησης του νευρώνα j στο επίπεδο l, είναι

$$u_j^l(i) = \sum_{r=0}^{k_l-1} w_{jr}^l\, y_r^{l-1}(i) \quad (24)$$

όπου, $y_0^l(i) = 1, \forall l$, δηλαδή η έξοδος του νευρώνα 0 σε κάθε επίπεδο l, ισούται με 1.

Εφαρμόζοντας τον κανόνα της αλυσίδας (chain rule), έχουμε ότι:

$$\frac{\partial \mathcal{E}(i)}{\partial w_j^l} = \frac{\partial \mathcal{E}(i)}{\partial u_j^l(i)} \frac{\partial u_j^l(i)}{\partial w_j^l} \quad (25)$$

Από την (24) προκύπτει ότι:

$$\frac{\partial u_j^l(i)}{\partial w_j^l} = y^{l-1}(i) \quad (26)$$

Επιπλέον, ο λόγος $\frac{\partial \mathcal{E}(i)}{\partial u_j^l(i)}$ είναι η τοπική μερική παράγωγος του νευρώνα j (local gradient) στο επίπεδο l, ή αλλιώς η μερική παράγωγος της σταθμισμένης εισόδου (weighted input) του νευρώνα j για κάθε επίπεδο l, και συμβολίζεται ως δ_j^l, δηλαδή:

$$\delta_j^l(i) = \frac{\partial \mathcal{E}(i)}{\partial u_j^l(i)} \quad (27)$$

Επομένως, από τις (25), (26), (27), η (23) ⇒

$$\Delta w_j^l = -\eta \sum_{i=1}^{N} \delta_j^l(i) y^{l-1}(i) \quad (28)$$

Η μέθοδος υπολογισμού της τοπικής παραγώγου $\delta_j^l(i)$ είναι η εξής: Πρώτα κάνουμε τον υπολογισμό για το επίπεδο εξόδου $l = L$, και στη συνέχεια διαδίδουμε

την πληροφορία προς τα πίσω σε όλο το δίκτυο. Επομένως, ο όρος back-propagation από τον οποίο πήρε και το όνομά του ο αλγόριθμος (back-propagation algorithm ή απλά backprop), αναφέρεται στη μέθοδο υπολογισμού των μερικών παραγώγων, ενώ κάποιος άλλος αλγόριθμος, όπως η στοχαστική κατάβαση της παραγώγου (stochastic gradient descent – SGD) εφαρμόζεται για την διαδικασία εκπαίδευσης (training) χρησιμοποιώντας αυτές τις παραγώγους [48].

Για το επίπεδο $l = L$ έχουμε ότι:
Οι έξοδοι του δικτύου, δηλαδή οι έξοδοι που παράγονται από το επίπεδο $l = L$, για κάθε νευρώνα j, είναι $y_j^L(i) = \hat{y}_j(i)$.

Από την (19) ισχύει:

$$\mathcal{E}(i) = \frac{1}{2}\sum_{m=1}^{k_L}\left(y_m(i) - \hat{y}_m(i)\right)^2 \Rightarrow$$

$$\mathcal{E}(i) = \frac{1}{2}\sum_{m=1}^{k_L}\left(f\left(u_m^L(i)\right) - y_m^L(i)\right)^2 \Rightarrow$$

$$\mathcal{E}(i) = \frac{1}{2}\sum_{m=1}^{k_L} e_m^2(i) \quad (29)$$

Άρα, η (27) ⇒

$$\delta_j^l(i) = e_j(i) f'\left(u_j^L(i)\right) \quad (30)$$

όπου *f'* είναι η παράγωγος της συνάρτησης ενεργοποίησης *f(u)*.

Για τα χαμηλότερα επίπεδα, $l < L$:
Εφαρμόζουμε τον κανόνα της αλυσίδας και έχουμε:

$$\delta_j^{l-1}(i) = \frac{\partial \mathcal{E}(i)}{\partial u_j^{l-1}(i)} \Rightarrow$$

$$\delta_j^{l-1}(i) = \sum_{r=1}^{k_l}\frac{\partial \mathcal{E}(i)}{\partial u_r^l(i)}\frac{\partial u_r^l(i)}{\partial u_j^{l-1}(i)} \Rightarrow$$

$$\delta_j^{l-1}(i) = \sum_{r=1}^{k_l}\delta_r^l(i)\frac{\partial u_r^l(i)}{\partial u_j^{l-1}(i)} \Rightarrow$$

$$\delta_j^{l-1}(i) = \sum_{r=1}^{k_l}\delta_r^l(i)\frac{\partial\left[\sum_{m=0}^{k_{l-1}} w_{rm}^l\, y_m^{l-1}(i)\right]}{\partial u_j^{l-1}(i)} \Rightarrow$$

$$\delta_j^{l-1}(i) = \sum_{r=1}^{k_l} \delta_r^l(i) \frac{\partial \left[\sum_{m=0}^{k_l-1} w_{rm}^l f\left(u_m^{l-1}(i)\right)\right]}{\partial u_j^{l-1}(i)} \Rightarrow$$

$$\delta_j^{l-1}(i) = \sum_{r=1}^{k_l} \delta_r^l(i) w_{rj}^l f'\left(u_j^{l-1}(i)\right) \Rightarrow$$

$$\delta_j^{l-1}(i) = \sum_{r=1}^{k_l} [\delta_r^l(i) w_{rj}^l] f'\left(u_j^{l-1}(i)\right) \quad (31)$$

Τέλος, πρέπει να υπολογιστεί η $f'\left(u_j^{l-1}(i)\right)$, όπου στην περίπτωση της σιγμοειδούς λογιστικής συνάρτησης, μετά από πράξεις προκύπτει ότι είναι $f'(x) = af(x)\big(1 - f(x)\big)$ (32).

Αλγόριθμος Back-propagation

1: Αρχικοποιούμε τα συναπτικά βάρη σε τυχαίες τιμές.

2: Υπολογίζουμε τις εισόδους και τις εξόδους, της συνάρτησης ενεργοποίησης, u_j^l και y_j^l, για κάθε $x(i)$.

3: Υπολογίζουμε τις τοπικές μερικές παραγώγους $\delta_j^l(i)$, ξεκινώντας από το τελευταίο επίπεδο $l = L$ και κινούμενοι προς τα πίσω.

4: Ενημερώνουμε τα συναπτικά βάρη.

5: Επαναλαμβάνουμε τα βήματα 2 έως και 4 μέχρι να ικανοποιηθεί το κριτήριο τερματισμού, δηλαδή μέχρις ότου να ελαχιστοποιηθεί η συνάρτηση κόστους.

Όπως αναφέρθηκε και προηγουμένως, για το βήμα 4, χρησιμοποιείται κάποια μέθοδος υπολογισμού της Δw_j^l και ενημέρωσης των συναπτικών βαρών. Αυτή η μέθοδος είναι η κατάβαση της παραγώγου (gradient descent), η οποία διακρίνεται στις ακόλουθες υποκατηγορίες:

- **Batch Gradient Descent:** Εφαρμόζεται όταν, για τον υπολογισμό της Δw_j^l, χρησιμοποιείται πληροφορία από ολόκληρο το σύνολο δεδομένων εκπαίδευσης. Σε αυτήν την κατηγορία ανήκει και η σχέση (28) που χρησιμοποιήσαμε παραπάνω, εφόσον αθροίζουμε για *i = 1, …, N* [52].

- **Stochastic Gradient Descent (SGD):** Είναι μια απλοποιημένη μορφή της παραπάνω μεθόδου, καθώς αντί να χρησιμοποιηθεί ολόκληρο το σύνολο δεδομένων, χρησιμοποιείται ένα μεμονωμένο, τυχαία επιλεγμένο, παράδειγμα, δηλαδή,

$$\Delta w_j^l = -\eta \delta_j^l(i) y^{l-1}(i) \quad (33)$$

  Η ενημέρωση των βαρών εξαρτάται από τα παραδείγματα που επιλέγονται τυχαία σε κάθε επανάληψη [53].

- **Mini-Batch Gradient Descent:** Αποτελεί μια παραλλαγή των δύο παραπάνω μεθόδων, καθώς χρησιμοποιεί ένα υποσύνολο παραδειγμάτων $\mathcal{B}_m$, $m < N$, από το σύνολο δεδομένων εκπαίδευσης [54].

Learning Rate (Ρυθμός Μάθησης)

Η επιλογή του ρυθμού μάθησης έχει ιδιαίτερη σημασία, καθώς καθορίζει το μέγεθος του βήματος σε κάθε επανάληψη και άρα και τον χρόνο εκπαίδευσης του συστήματος. Επιπλέον, εάν ο ρυθμός μάθησης είναι πολύ μεγάλος, η κατάβαση της παραγώγου μπορεί να αυξήσει αντί να μειώσει το σφάλμα εκπαίδευσης (training error), ενώ εάν είναι πολύ μικρός, η εκπαίδευση όχι μόνο είναι πιο αργή, αλλά μπορεί να κολλήσει μόνιμα σε μια μεγάλη τιμή σφάλματος [48]. Οι τιμές του ρυθμού μάθησης ανήκουν στο διάστημα (0, 1) και συνήθως είναι μικρότερες του 1.0 και μεγαλύτερες του 10^{-6} [55]. Μπορεί να χρησιμοποιηθεί είτε μια σταθερή τιμή για τον ρυθμό μάθησης, είτε αυτός να προσαρμόζεται ανάλογα με την απόδοση του μοντέλου κατά τη διαδικασία εκπαίδευσης (adaptive learning rate). Οι κύριες μέθοδοι adaptive learning rate είναι οι AdaGrad [56], RMSprop [57], AdaDelta [58] και Adam [59], οι οποίες βασίζονται στην μέθοδο SGD [60]. Στην παρούσα εργασία θα αναφερθούμε στην μέθοδο RMSprop.

Η RMSprop διαιρεί τον ρυθμό μάθησης και τη μερική παράγωγο των συναπτικών βαρών με τον κινητό μέσο του τετραγώνου της παραγώγου. Επομένως, αρχικά διατηρεί έναν κινητό μέσο του τετραγώνου της μερικής παραγώγου για κάθε συναπτικό βάρος w_j^l.

$$MeanSquare(w_j^l, t) = \boldsymbol{\gamma} MeanSquare(w_j^l, t-1) + (1-\boldsymbol{\gamma})\left(\frac{\partial \mathcal{C}}{\partial w_j^l}\right)^2 \quad (34)$$

όπου $\boldsymbol{\gamma}$ είναι ένας παράγοντας λήθης (συνήθως $\boldsymbol{\gamma} = 0.9$).
Στη συνέχεια, ο ρυθμός μάθησης του w_j^l διαιρείται από αυτόν τον κινητό μέσο.

$$w_j^l = w_j^l - \frac{\eta}{\sqrt{MeanSquare(w_j^l, t)}} \cdot \frac{\partial \mathcal{C}}{\partial w_j^l} \quad (35)$$

Η μέθοδος RMSprop έχει δείξει εξαιρετική προσαρμογή του ρυθμού μάθησης σε διαφορετικές εφαρμογές [61] [62].

3.3 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΩΝ ΔΙΚΤΥΩΝ

Τα τελευταία χρόνια, η επιστημονική κοινότητα έχει στρέψει το ενδιαφέρον της στη χρήση μεθόδων μηχανικής μάθησης (machine learning) για το πρόβλημα του διαχωρισμού μουσικών πηγών, καθώς υπάρχουν πλέον διαθέσιμες, αρκετά μεγάλες βάσεις δεδομένων απομονωμένων φωνητικών και της συνοδείας τους. Η βασική διαφορά των μεθόδων που επιλέγονται έγγυται στην αρχιτεκτονική του δικτύου που χρησιμοποιείται και στον τρόπο με τον οποίο εκπαιδεύεται. Έχουν προταθεί πολλές αρχιτεκτονικές στο παρελθόν, όπως τα πλήρως συνδεδεμένα δίκτυα εμπρός-τροφοδότησης (feedforward fully-connected neural networks – FNNs), τα συνελικτικά νευρωνικά δίκτυα (convolutional neural networks – CNNs), ή τα αναδρομικά νευρωνικά δίκτυα (recurrent neural networks – RNNs) και παραλλαγές αυτών, όπως τα μοντέλα long short-term memory (LSTM) [13].

3.3.1 ΣΥΝΕΛΙΚΤΙΚΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

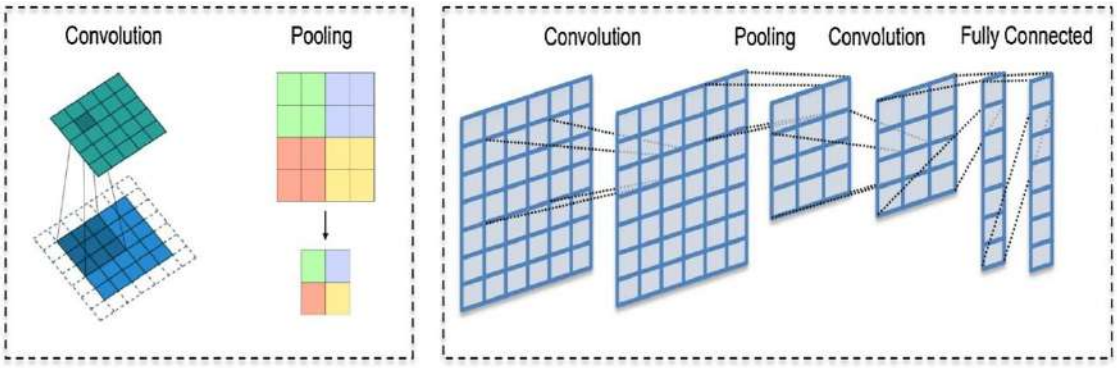
Τα συνελικτικά νευρωνικά δίκτυα έχουν σχεδιαστεί για την επεξεργασία δεδομένων υπό τη μορφή πινάκων, με αποτέλεσμα να χρησιμοποιούνται ευρέως καθώς πολλοί τύποι δεδομένων εκφράζονται υπό αυτήν τη μορφή. Μερικά παραδείγματα είναι οι πίνακες μονής διάστασης 1D για σήματα και ακολουθίες, όπως η φυσική γλώσσα, διπλής διάστασης 2D για εικόνες ή φασματογράφημα ήχου, και 3D για δεδομένα βίντεο. Οι βασικές ιδέες της αρχιτεκτονικής των CNNs είναι τέσσερις: οι τοπικές συνδέσεις, τα διαμοιραζόμενα βάρη, η συγκέντρωση (pooling) και η χρήση πολλών επιπέδων.

Η αρχιτεκτονική ενός CNN είναι δομημένη σε στάδια, τα οποία αποτελούνται από τρεις τύπους επιπέδων: δύο τύπους επιπέδων που εναλλάσσονται, τα συνελικτικά επίπεδα και τα επίπεδα συγκέντρωσης, και τέλος τα πλήρως συνδεδεμένα επίπεδα. Ένα συνελικτικό επίπεδο στοχεύει στον εντοπισμό των χαρακτηριστικών των δεδομένων εισόδου. Για την επίτευξη αυτού του σκοπού, πρώτα υπολογίζει τη συνέλιξη των δεδομένων εισόδου του επιπέδου με ένα πλήθος από φίλτρα, και στη συνέχεια, περνάει το αποτέλεσμα από μία συνάρτηση ενεργοποίησης, όπως η ReLU. Ετσι, παράγει τους λεγόμενους χάρτες χαρακτηριστικών (feature maps) [63]. Οι νευρώνες που βρίσκονται στον ίδιο χάρτη χαρακτηριστικών διαμοιράζονται τα συναπτικά βάρη μειώνοντας έτσι την πολυπλοκότητα του δικτύου διατηρώντας τον αριθμό των παραμέτρων χαμηλά [64]. Μια ιδιαιτερότητα των δικτύων CNN είναι ότι κάθε νευρώνας ενός χάρτη χαρακτηριστικών συνδέεται τοπικά μόνο με ένα υποσύνολο νευρώνων του προηγούμενου επιπέδου και άρα δέχεται μόνο ένα υποσύνολο δεδομένων στην είσοδο του.

Ένα επίπεδο συγκέντρωσης στοχεύει στο να μειώσει σταδιακά τη διάσταση της αναπαράστασης των δεδομένων και κατ' επέκταση να μειώσει περαιτέρω το πλήθος των παραμέτρων και την υπολογιστική πολυπλοκότητα του δικτύου [65]. Δέχεται ως είσοδο τους χάρτες χαρακτηριστικών του συνελικτικού επιπέδου που έχει προηγηθεί και χωρίζει κάθε έναν από αυτούς σε $m \times n$ μπλοκ. Στη συνέχεια, για κάθε μπλοκ, εφαρμόζει μία συνάρτηση συγκέντρωσης, η οποία συγχωνεύει τα χαρακτηριστικά του μπλοκ σε μία τιμή. Συνήθως χρησιμοποιείται είτε η μέση συγκέντρωση (average pooling), η οποία εξάγει τον μέσο όρο των τιμών των χαρακτηριστικών, είτε η μέγιστη συγκέντρωση (max pooling), η οποία εξάγει τη μεγαλύτερη τιμή των χαρακτηριστικών. Επομένως, παράγει εν τέλει ένα νέο χάρτη χαρακτηριστικών συνολικού μεγέθους $m \times n$ [66].

Αφότου ολοκληρωθούν οι εναλλαγές των συνελικτικών επιπέδων με τα επίπεδα συγκέντρωσης, ακολουθούν ένα ή περισσότερα πλήρως συνδεδεμένα επίπεδα, τα οποία παίρνουν όλους τους νευρώνες του προηγούμενου επιπέδου και τους συνδέουν με κάθε νευρώνα του επιπέδου τους, με σκοπό την παραγωγή καθολικής σημασιολογικής πληροφορίας. Η χρήση των πλήρως συνδεδεμένων επιπέδων στα CNNs είναι προαιρετική, καθώς ένα τέτοιο επίπεδο μπορεί να αντικατασταθεί από ένα 1x1 συνελικτικό επίπεδο [63].

Συνεπώς, ο ρόλος ενός συνελικτικού επιπέδου είναι να εντοπίζει τις από κοινού τοπικές συνδέσεις των χαρακτηριστικών από το προηγούμενο επίπεδο, ενώ ο ρόλος ενός επιπέδου συγκέντρωσης είναι να συγχωνεύσει σημασιολογικά όμοια χαρακτηριστικά σε ένα, ώστε να μειώσει τη διάσταση των δεδομένων που θα προωθηθούν στο επόμενο επίπεδο, δηλαδή το συνολικό μέγεθος της εισόδου. Τα συναπτικά βάρη των φίλτρων εκπαιδεύονται με τον αλγόριθμο back-propagation που έχουμε προ-αναφέρει [42]. Στην **Εικ. 4** απεικονίζεται η λειτουργία των επιπέδων που περιγράψαμε παραπάνω και η δομή ενός CNN.

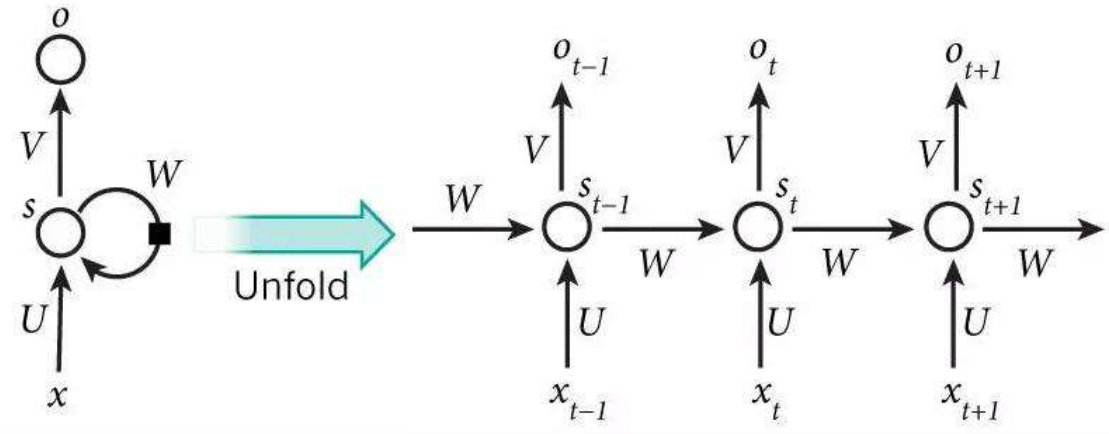


Εικ. 4: Η λειτουργία του συνελικτικού επιπέδου και του επιπέδου συγκέντρωσης, και οι μεταξύ τους συνδέσεις για την κατασκευή ενός CNN. (πηγή [67]). Αριστερά, αναπαρίσταται η τοπική σύνδεση ενός νευρώνα του χάρτη χαρακτηριστικών που παράγεται από κάποιο συνελικτικό επίπεδο, με ένα υποσύνολο νευρώνων των δεδομένων εισόδου. Ακριβώς δίπλα, αναπαρίσταται η λειτουργία ενός επιπέδου συγκέντρωσης που χωρίζει έναν χάρτη χαρακτηριστικών 4x4 σε 4 μπλοκ 2x2, στα οποία εφαρμόζει κάποια συνάρτηση συγκέντρωσης και παράγει ένα νέο χάρτη χαρακτηριστικών 2x2. Δεξιά αναπαρίσταται η αρχιτεκτονική ενός CNN.

3.3.2 ΑΝΑΔΡΟΜΙΚΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Τα αναδρομικά νευρωνικά δίκτυα επεξεργάζονται μια ακολουθία δεδομένων εισόδου ανά στοιχείο, διατηρώντας, ωστόσο, ένα διάνυσμα κατάστασης (state vector) το οποίο περιέχει πληροφορία για όλα τα προηγούμενα στοιχεία της ακολουθίας. Σε κάθε χρονική στιγμή, το δίκτυο δέχεται ένα διάνυσμα εισόδου x_t. Το πλεονέκτημα της αρχιτεκτονικής ενός RNN είναι ότι για την πρόβλεψη της εξόδου του σε μία χρονική στιγμή t_c, το δίκτυο μπορεί να αξιοποιήσει όλη την πληροφορία που έχει δεχτεί στην είσοδό του $x_t, t = \{1, 2, \dots, t_c\}$, μέχρι τη συγκεκριμένη χρονική στιγμή [68]. Επομένως, σε κάθε επανάληψη προστίθεται νέα πληροφορία στο δίκτυο, δηλαδή, σε κάθε χρονική στιγμή, κάθε επίπεδο του RNN δέχεται είσοδο, η οποία είναι συνδυασμός δεδομένων που δέχεται την τρέχουσα χρονική στιγμή και δεδομένων προηγούμενων χρονικών στιγμών, αποθηκευμένων στη μνήμη του δικτύου, στις οποίες αποκτά πρόσβαση με αναδρομή. Κατά συνέπεια η έξοδος που παράγει, είναι αποτέλεσμα αυτού του συνδυασμού [69]. Ουσιαστικά, τα RNNs χρησιμοποιούνται για την μοντελοποίηση σειριακών δεδομένων, καθώς έχουν εσωτερική μνήμη, η οποία βοηθάει το δίκτυο να έχει πρόσβαση όχι μόνο στα δεδομένα που εισέρχονται μια δεδομένη χρονική στιγμή, αλλά και στα δεδομένα πρόσφατων παρελθοντικών στιγμών. Η **Εικ. 5** αναπαριστά την λειτουργία του νευρώνα ενός RNN.

Η εκπαίδευση των RNNs εμφάνισε προβλήματα, καθώς οι μερικές παράγωγοι, κατά τη διαδικασία πίσω διάδοσης, είτε αυξάνονταν είτε συρρικνώνονταν εκθετικά σε κάθε χρονικό βήμα, με αποτέλεσμα να παρατηρείται το φαίνομενο κλιμακούμενης παραγώγου (exploding gradient) ή εξαφανιζόμενης παραγώγου (vanishing gradient) αντίστοιχα, το οποίο οδηγεί το σύστημα σε αστάθεια [48] [42]. Επιπλέον, τα εμπειρικά στοιχεία εφαρμογής των RNNs, έδειξαν ότι είναι δύσκολο να εκπαιδευτούν ώστε να αποθηκεύουν πληροφορία για πολύ μεγάλο χρονικό διάστημα, παρόλο που ακριβώς αυτός είναι ο σκοπός σχεδίασής τους [70].



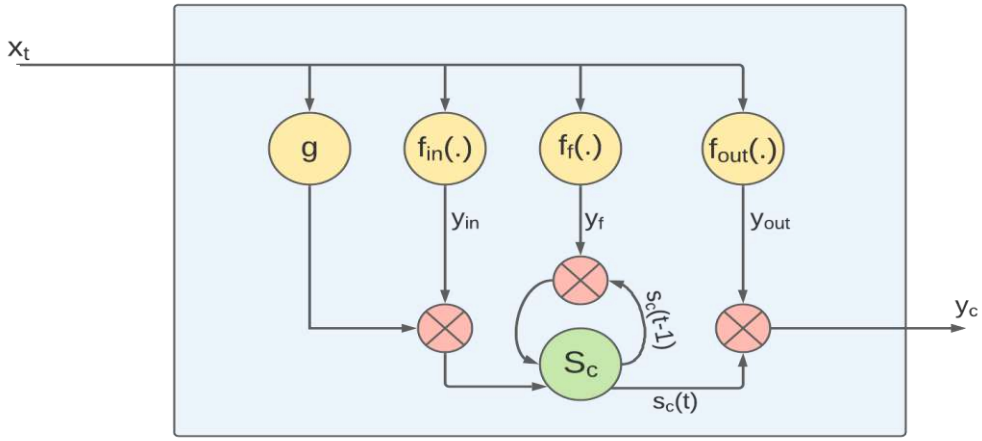
Εικ. 5: Η λειτουργία του νευρώνα ενός επιπέδου του RNN κατά τη διάρκεια μιας χρονικής ακολουθίας.

3.3.3 ΔΙΚΤΥΑ LSTM

Για την επίλυση των παραπάνω προβλημάτων προτάθηκαν τα δίκτυα LSTM, τα οποία αποδείκτηκε ότι είναι γρηγορότερα και πιο ακριβή από τα RNNs. Κάθε επίπεδο ενός δικτύου LSTM αποτελείται από ένα σύνολο αναδρομικά συνδεδεμένων μπλοκ, που ονομάζονται μπλοκ μνήμης (memory blocks) [71]. Κάθε ένα περιέχει ένα ή περισσότερα αναδρομικά συνδεδεμένα κελιά μνήμης (memory cells) και τρεις πολλαπλασιαστικές μονάδες, τις πύλες εισόδου, λήθης και εξόδου (input, forget and output gates), οι οποίες εκπαιδεύονται γύρω από το ποια πληροφορία να αποθηκεύσουν στην μνήμη (write), για πόσο πρέπει να παραμείνει στην μνήμη (reset) και πότε πρέπει να διαβαστεί (read), αντίστοιχα [72]. Κάθε μία από αυτές τις πύλες μπορεί να θεωρηθεί ως ένας τεχνητός νευρώνας, καθώς χρησιμοποιούν κάποια συνάρτηση ενεργοποίησης για να υπολογίσουν την έξοδο ενός σταθμισμένου αθροίσματος (weighted sum). Κάθε κελί μνήμης έχει στον πυρήνα του μία αναδρομικά συνδεδεμένη με τον εαυτό της μονάδα S_C, η οποία υπολογίζει ή επαναφέρει την κατάσταση του κελιού(cell state). Τα κελιά που βρίσκονται στο ίδιο μπλοκ μνήμης μοιράζονται τις ίδιες πύλες, γεγονός που έχει σαν αποτέλεσμα τη μείωση των παραμέτρων.

Στην **Εικ. 6** παρουσιάζεται η λειτουργία ενός κελιού μνήμης του δικτύου LSTM. Πιο συγκεκριμένα, έστω $\boldsymbol{x_t}$ η είσοδος του κελιού, $\boldsymbol{y_{in}}$ η έξοδος της συνάρτησης ενεργοποίησης της πύλης εισόδου, $\boldsymbol{s_c(t)}$ η κατάσταση του κελιού την χρονική στιγμή t, $\boldsymbol{y_f}$ η έξοδος της συνάρτησης ενεργοποίησης της πύλης λήθης, $\boldsymbol{y_{out}}$ η έξοδος της συνάρτησης ενεργοποίησης της πύλης εξόδου και $\boldsymbol{y_c}$ η έξοδος του κελιού [72].

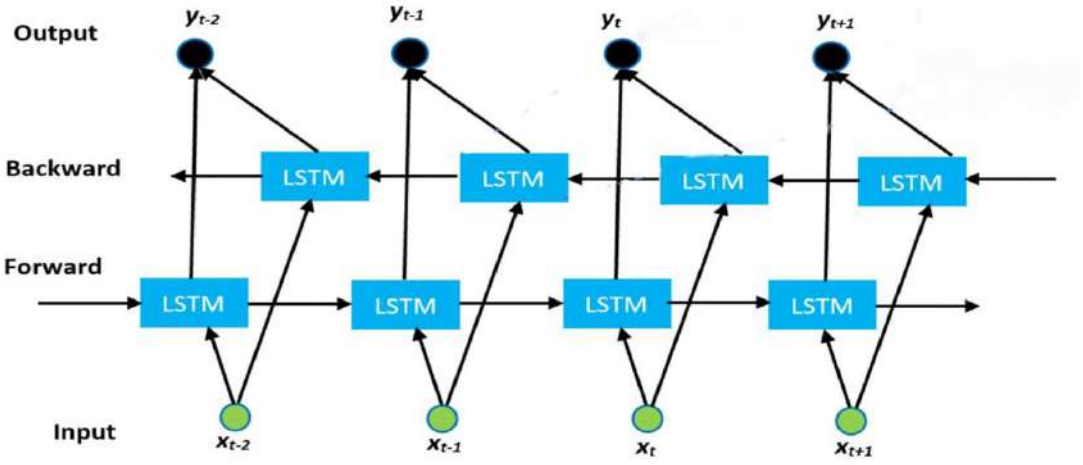
1. Η $\boldsymbol{x_t}$ συμπιέζεται από μία συνάρτηση g και το αποτέλεσμα πολλαπλασιάζεται με την $\boldsymbol{y_{in}}$.
2. Η τρέχουσα κατάσταση του κελιού $\boldsymbol{s_c(t)}$ προκύπτει ως εξής: Η προηγούμενη κατάσταση του κελιού $\boldsymbol{s_c(t-1)}$ πολλαπλασιάζεται με την $\boldsymbol{y_f}$ και το αποτέλεσμα προστίθεται με το αποτέλεσμα του βήματος 1.
3. Η έξοδος του κελιού $\boldsymbol{y_c}$ υπολογίζεται πολλαπλασιάζοντας την κατάσταση του κελιού $\boldsymbol{s_c}$ με την $\boldsymbol{y_{out}}$.



Εικ. 6: Η λειτουργία ενός κελιού του δικτύου LSTM.

Δίκτυο BLSTM (Bidirectional LSTM)

Τα δίκτυα LSTM έχουν τη δυνατότητα να διαβάσουν άρα και να διατηρήσουν πληροφορία μόνο από προηγούμενες χρονικές στιγμές. Σε ένα δίκτυο BLSTM (Bidirectional LSTM) η χρονική ροή της ακολουθίας εισόδου είναι διπλή, προς τα εμπρός (forwards) και προς τα πίσω (backwards). Για να επιτευχθεί αυτό, το BLSTM αποτελείται από δύο παράλληλα επίπεδα δικτύων LSTM τα οποία δεν συνδέονται μεταξύ τους, αλλά συνδέονται στο ίδιο επίπεδο εξόδου (**Εικ. 7**). Το ένα επίπεδο διαδίδει την πληροφορία προς τα εμπρός και το άλλο προς τα πίσω. Έτσι, η πληροφορία μπορεί να καταγραφεί τόσο σε παρελθοντικά όσο και σε μελλοντικά χρονικά πλαίσια. Επομένως, για κάθε σημείο στον χρόνο μιας δοσμένης ακολουθίας, το δίκτυο έχει όλες τις απαραίτητες πληροφορίες για όλα τα σημεία πριν και μετά από αυτό [73] [74].



Εικ. 7: Οι είσοδοι περνάνε στο δίκτυο BLSTM, το οποίο αποτελείται από δύο επίπεδα LSTM (forward layer, backward layer) και υπολογίζει τα χαρακτηριστικά εξόδου.

4 DENSE CONVOLUTIONAL NETWORK (DENSENET)

Τα DenseNets, όπως προκύπτει και από την ονομασία τους, αποτελούν μια παραλλαγή των κλασσικών CNNs, των οποίων η χρήση τείνει να έχει αυξητική τάση τα τελευταία χρόνια, λόγω των πολλαπλών πλεονεκτημάτων τους και της ικανότητάς τους να αντιμετωπίζουν προβλήματα που εμφανίζουν τα CNNs, όπως αυτό της εξαφανιζόμενης παραγώγου. Παρακάτω παρουσιάζεται αναλυτικά η αρχιτεκτονική του μοντέλου DenseNet, και οι επεκτάσεις του: MDenseNet, MMDenseNet και MMDenseLSTM.

4.1 Η ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ ΜΟΝΤΕΛΟΥ DENSENET

Η αρχιτεκτονική του DenseNet παρουσιάζεται ακολούθως. Όλα τα επίπεδα (που έχουν χάρτες χαρακτηριστικών ίδιου μεγέθους) συνδέονται απευθείας μεταξύ τους. Κάθε επίπεδο λαμβάνει πρόσθετες εισόδους από τα προηγούμενα επίπεδα και προωθεί τους δικούς του χάρτες χαρακτηριστικών στα επόμενα επίπεδα. Συνεπώς, το επίπεδο ℓ έχει ℓ εισόδους, οι οποίες αποτελούνται από τους χάρτες χαρακτηριστικών όλων των προηγούμενων επιπέδων. Οι χάρτες χαρακτηριστικών του επιπέδου ℓ προωθούνται σε όλα τα επόμενα $L - \ell$ επίπεδα. Αυτό συνεπάγεται ότι σε ένα δίκτυο L επιπέδων, θα υπάρχουν $\sum_{l=1}^{L} l = \frac{L(L+1)}{2}$ συνδέσεις. Ένα πλεονέκτημα αυτής της πυκνής συνδεσμολογίας που χαρακτηρίζει την αρχιτεκτονική του δικτύου είναι ότι απαιτεί λιγότερες παραμέτρους από τα παραδοσιακά CNNs.

Η αρχιτεκτονική των DenseNets διαφοροποιεί ρητά την πληροφορία που προστίθεται στο δικτύο με την πληροφορία που διατηρείται. Τα επίπεδα του δικτύου είναι πολύ στενά, προσθέτοντας μόνο ένα μικρό σύνολο από χάρτες χαρακτηριστικών στη συνολική γνώση του δικτύου, και διατηρούν τους υπόλοιπους χάρτες χαρακτηριστικών αμετάβλητους.

Τα DenseNets, εκμεταλλεύονται τη δυνατότητα της αρχιτεκτονικής τους να επαναχρησιμοποιεί χαρακτηριστικά, εφόσον συνδέει τους χάρτες χαρακτηριστικών διαφορετικών επιπέδων, με αποτέλεσμα να είναι πολύ αποδοτικά και εύκολο να εκπαιδευτούν [75].

4.1.1 ΠΥΚΝΗ ΔΙΑΣΥΝΔΕΣΗ

Όπως προαναφέραμε, τα DenseNets χρησιμοποιούν απευθείας συνδέσεις από οποιοδήποτε επίπεδο προς όλα τα επόμενα επίπεδα, δηλαδή το επίπεδο ℓ λαμβάνει στην είσοδό του, τους χάρτες χαρακτηριστικών $x_0, \dots, x_{l-1}$, όλων των προηγούμενων επιπέδων. Ισχύει ότι:

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) \quad (36)$$

όπου $x_0, \dots, x_{l-1}, x_l$ είναι οι έξοδοι των επιπέδων $0, 1, \dots, l-1, l$ αντίστοιχα, $[x_0, x_1, \dots, x_{l-1}]$ η σύνδεση των χαρακτηριστικών των επιπέδων $0, 1, \dots, l-1$ και $H_l(.)$ μια σύνθετη συνάρτηση με τρεις διαδοχικές πράξεις: ομαδική κανονικοποίηση (batch normalization: BN), ReLU και 3x3 συνέλιξη (Conv) [75].

Batch Normalization

Γενικά η εκπαίδευση των DNNs είναι ένα δύσκολο πρόβλημα, καθώς η κατανομή των εισόδων κάθε επιπέδου αλλάζει κατά τη διάρκειά της. Γι' αυτόν το λόγο απαιτούνται χαμηλοί ρυθμοί μάθησης και προσεκτική αρχικοποίηση των παραμέτρων, με αποτέλεσμα η διαδικασία εκπαίδευσης να επιβραδύνεται. Αυτό το φαινόμενο ονομάζεται internal covariate shift και αντιμετωπίζεται με την κανονικοποίηση των εισόδων των επιπέδων. Η μέθοδος Batch Normalization, η οποία αναλύεται παρακάτω, ουσιαστικά ενσωματώνει την κανονικοποίηση αυτή στην αρχιτεκτονική του δικτύου και την εφαρμόζει για κάθε υποσύνολο εκπαίδευσης (mini-batch). Η χρήση της BN επιτρέπει υψηλότερους ρυθμούς μάθησης και μάλιστα χωρίς να δίνεται ιδιαίτερη προσοχή στην αρχικοποίηση των παραμέτρων.

Σε ένα DNN, κάθε επίπεδο $\boldsymbol{\ell}$ δέχεται $\boldsymbol{x_l} = [x_1, \dots, x_{k_l}]^T$ εισόδους στους νευρώνες του. Κατά την εφαρμογή της μεθόδου mini-batch gradient descent, κάθε mini-batch παράγει εκτιμήσεις της μέσης τιμής και της διασποράς για κάθε $\boldsymbol{x_i}, i = 1, \dots, k_l$, ώστε αυτές οι στατιστικές κανονικοποίησης να μπορούν να συμμετέχουν στην διαδικασία back-propagation.

Σκοπός του αλγορίθμου Batch-Normalization (BN) είναι να κανονικοποιήσει τις τιμές των $\boldsymbol{x_i}$, χωρίς όμως να αλλάξει την αναπαράσταση κάθε επιπέδου στο δίκτυο. Για το λόγο αυτό εισάγει δύο παραμέτρους $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$, οι οποίες μαθαίνονται μαζί με τις αρχικές παραμέτρους του μοντέλου. Συμβολίζουμε με $\boldsymbol{x} = \boldsymbol{x_i}, i = 1, \dots, k_l$, τα δεδομένα εισόδου του νευρώνα i, $\boldsymbol{\mathcal{B}_m} = \{x_{1,\dots m}\}$ το υποσύνολο (mini-batch) μεγέθους m του x, $\widehat{\boldsymbol{x}}_{\boldsymbol{1,\dots m}}$ τις κανονικοποιημένες τιμές και $\widehat{\boldsymbol{y}}_{\boldsymbol{1\dots m}}$ τις αντίστοιχες εξόδους του αλγορίθμου BN (Batch-Normalization), ο οποίος δίνεται αναλυτικά παρακάτω:

Αλγόριθμος Batch-Normalization (BN)

Είσοδος: Το υποσύνολο $\mathcal{B}_m = \{x_{1,\dots m}\}$ και οι παράμετροι β και γ.

Έξοδος: $\{\hat{y}_i = BN_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2$$

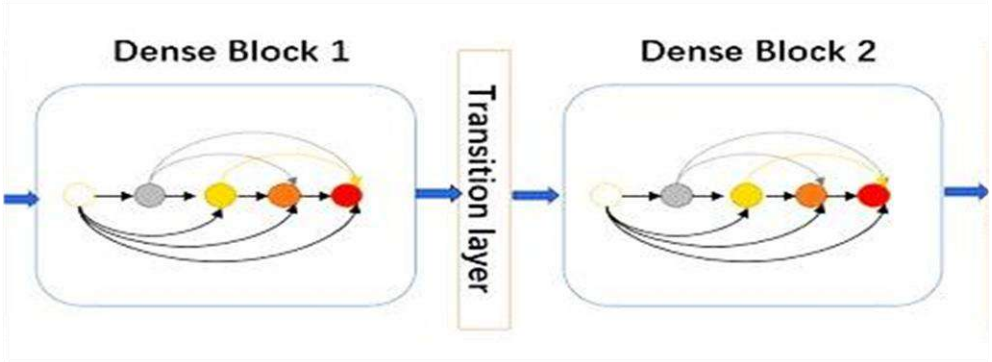
$$\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$

$$\hat{y}_i \leftarrow \gamma\hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i)$$

Στον αλγόριθμο το ϵ είναι μια μικρή σταθερά που χρησιμοποιείται για αριθμητική σταθερότητα [76].

4.1.2 ΕΠΙΠΕΔΑ ΣΥΓΚΕΝΤΡΩΣΗΣ (POOLING LAYERS)

Όταν το μέγεθος των χαρτών χαρακτηριστικών αλλάζει, η (36) δεν μπορεί να χρησιμοποιηθεί. Ωστόσο, ένα απαραίτητο κομμάτι των DenseNets είναι τα επίπεδα υποδειγματοληψίας, τα οποία αλλάζουν το μέγεθος των χαρτών χαρακτηριστικών. Για να γίνει η διαδικασία υποδειγματοληψίας πιο εύκολη, το δίκτυο χωρίζεται σε πολλαπλά πυκνά συνδεδεμένα πυκνά μπλοκ (dense blocks). Τα επίπεδα που βρίσκονται μεταξύ των dense blocks ονομάζονται επίπεδα μετάβασης (transition layers) τα οποία αποτελούνται από ένα επίπεδο ομαδικής κανονικοποίησης, ένα 1x1 συνελικτικό επίπεδο και ένα 2x2 επίπεδο μέσης συγκέντρωσης (average pooling layer) [75]. Η διάταξη αυτή αναπαρίσταται στην **Εικ. 8**. Εδώ, είναι σημαντικό να σημειωθεί ότι το τελευταίο πυκνό μπλόκ του δικτύου DenseNet δεν ακολουθείται από κάποιο επίπεδο μετάβασης, αλλά συνδέεται απευθείας με ένα επίπεδο μέσης συγκέντρωσης και ένα πλήρως συνδεδεμένο επίπεδο.



Εικ. 8: Υπο-μέρος της αρχιτεκτονικής ενός DenseNet, που αποτελείται από δύο πυκνά μπλοκ, μεταξύ των οποίων παρεμβάλλεται ένα επίπεδο μετάβασης.

4.1.3 ΡΥΘΜΟΣ ΑΝΑΠΤΥΞΗΣ

Αν κάθε H_l παράγει $\boldsymbol{k}$ χάρτες χαρακτηριστικών, τότε το επίπεδο $\boldsymbol{\ell}$ θα έχει $k_0 + k(l-1)$ χάρτες χαρακτηριστικών εισόδου, όπου k_0 είναι το πλήθος των καναλιών του επιπέδου εισόδου και k είναι το πλήθος των χαρτών χαρακτηριστικών που προσθέτει το επίπεδο και ορίζεται ως ο ρυθμός ανάπτυξης (growth rate) του δικτύου. Εφόσον κάθε επίπεδο έχει πρόσβαση σε όλους τους προηγούμενους χάρτες χαρακτηριστικών του dense block που ανήκει, και συνεπώς και στη συνολική γνώση του δικτύου, ένας σχετικά μικρός ρυθμός ανάπτυξης μπορεί να επιφέρει state-of-the-art αποτελέσματα [75].

4.1.4 ΕΠΙΠΕΔΑ ΣΥΜΦΟΡΗΣΗΣ

Τα επίπεδα συμφόρησης (bottleneck layers) χρησιμοποιούνται για να μειώσουν το πλήθος των χαρτών χαρακτηριστικών που δέχεται στην είσοδό του ένα επίπεδο και κατ'επέκταση να βελτιώσουν την υπολογιστική απόδοση, εισάγωντας μια πράξη 1x1 συνέλιξης πριν από κάθε 3x3 συνέλιξη. Το DenseNet εντάσσει τα επίπεδα συμφόρησης στην αρχιτεκτονική του, εφαρμόζοντας μια ακολουθία των πράξεων BN-ReLU-Conv(1×1)-BN-ReLU-Conv(3×3) [75].

4.1.5 ΣΥΜΠΙΕΣΗ

Είναι δυνατόν το δίκτυο να βελτιωθεί ακόμα περισσότερο εάν μειωθεί το πλήθος των χαρτών χαρακτηριστικών στα επίπεδα μετάβασης. Έστω ότι ένα πυκνό μπλοκ περιλαμβάνει $\boldsymbol{m}$ χάρτες χαρακτηριστικών. Το επίπεδο μετάβασης που ακολουθεί χρησιμοποιεί έναν συντελεστή συμπίεσης $\boldsymbol{\theta}$, όπου $0 < \theta \leq 1$, για να παράξει $\lfloor \theta_m \rfloor$ χάρτες χαρακτηριστικών στην έξοδό του. Εάν $\theta = 1$, τότε το πλήθος των χαρτών χαρακτηριστικών παραμένει αμετάβλητο [75].

4.2 MDENSENET

Στην προηγούμενη ενότητα παρουσιάστηκε το μοντέλο DenseNet και τα πλεονεκτήματά του. Ωστόσο, ένα σημαντικό μειονέκτημα του DenseNet είναι οι μεγάλες απαιτήσεις που έχει σε μνήμη, καθώς το πλήθος των συνδέσεων μεταξύ των επιπέδων αυξάνεται τετραγωνικά ως προς το βάθος του δικτύου. Αυτό είναι αρκετά περιοριστικό όσον αφορά το πρόβλημα του διαχωρισμού μουσικών πηγών, καθώς τα δεδομένα εισόδου και εξόδου πρέπει να είναι υψηλής ανάλυσης και κατ' επέκτασην οι διαστάσεις τους ιδιαίτερα μεγάλες.

Για την αντιμετώπιση του παραπάνω προβλήματος, προτάθηκε από τους συγγραφείς του [77] μια επέκταση του μοντέλου DenseNet, ένα συνελικτικό πολλαπλών μεγεθών (multi-scale) DenseNet (MDenseNet), το οποίο αποτελείται από πυκνά μπλοκ (dense blocks) με πολλαπλές αναλύσεις (**Εικ. 9**). Η είσοδος των πυκνών μπλοκ χαμηλότερης ανάλυσης παράγεται από την επαναλαμβανόμενη υποδειγματοληψία των εξόδων των προηγούμενων πυκνών μπλοκ. Στη συνέχεια, τα πυκνά μπλοκ χαμηλότερης ανάλυσης υφίστανται υπερδειγματοληψία, με σκοπό την ανάκτηση μιας υψηλότερης ανάλυσης, η οποία οδηγείται στην είσοδο των πυκνών μπλοκ υψηλότερης ανάλυσης μαζί με την έξοδο των προηγούμενων πυκνών μπλοκ με την ίδια ανάλυση.

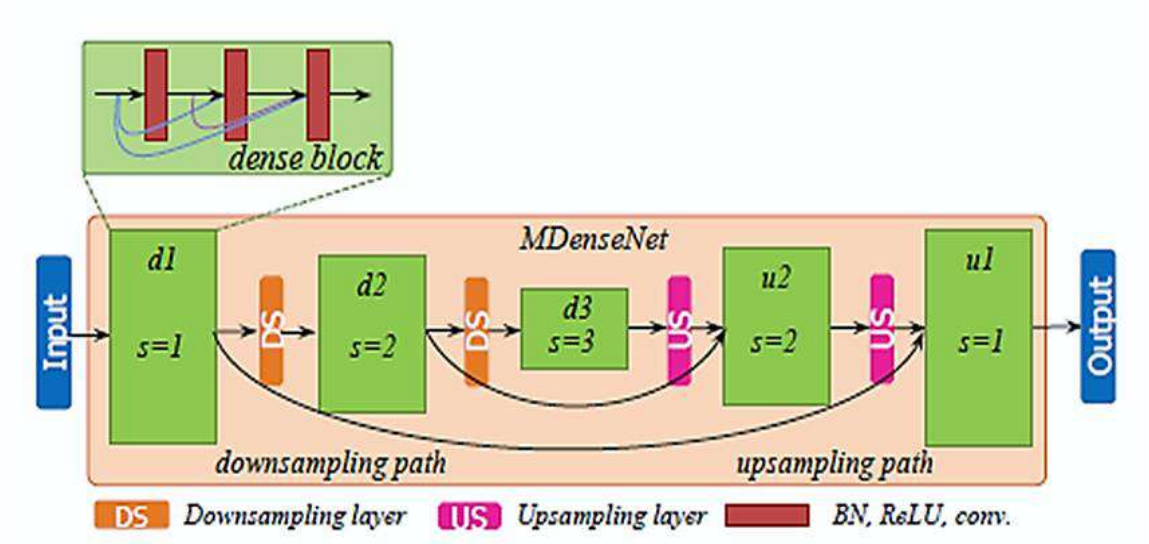
4.2.1 ΕΠΙΠΕΔΑ ΥΠΟΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΚΑΙ ΥΠΕΡΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

Τα πυκνά μπλοκ εναλλασσόμενα με τα επίπεδα υποδειγματοληψίας, διαμορφώνουν το λεγόμενο μονοπάτι υποδειγματοληψίας (down-sampling path) του δικτύου. Τα επίπεδα υποδειγματοληψίας (down-sampling layers) παράγουν χάρτες χαρακτηριστικών χαμηλότερης ανάλυσης από την ανάλυση αυτών που φτάνουν στην είσοδό τους. Η μείωση αυτή συνεπάγεται και την μείωση του υπολογιστικού κόστους, καθώς, επίσης, επιτρέπει στο δίκτυο να μοντελοποιεί πληροφορία από μεγαλύτερα χρονικά διαστήματα και πιο πλατύ εύρος συχνοτήτων.

Όπως ίσχυε για τα επίπεδα υποδειγματοληψίας, έτσι, τα πυκνά μπλοκ εναλλασσόμενα με τα επίπεδα υπερδειγματοληψίας, διαμορφώνουν το μονοπάτι υπερδειγματοληψίας (up-sampling path) του δικτύου. Τα επίπεδα υπερδειγματοληψίας (up-sampling layers) χρησιμοποιούνται για να επαναφέρουν τους χάρτες χαρακτηριστικών χαμηλής ανάλυσης, στην αρχική τους ανάλυση μέσω της πράξης ανάστροφης συνέλιξης με μέγεθος φίλτρου ίδιο με το μέγεθος συγκέντρωσης (pooling size). Αυτό επιτρέπει στο δίκτυο να επαναφέρει τις λεπτομέρειες που αφορούν τη δομή και τις διαστάσεις του φασματογραφήματος που έλαβε στην είσοδό του [77].

4.2.2 INTER-BLOCK SKIP CONNECTION

Τα επίπεδα ενός μοντέλου MDenseNet συνδέονται είτε μέσω των μονοπατιών υποδειγματοληψίας και υπερδειγματοληψίας, είτε μέσω συνδέσεων παράκαμψης μπλοκ. Οι συνδέσεις παράκαμψης μπλοκ συνδέουν απευθείας μεταξύ τους δύο πυκνά μπλοκ ίδιου μεγέθους, παρακάμπτοντας τη σειριακή ακολουθία επιπέδων. Κατά συνέπεια, έχουν τη δυνατότητα να προωθούν απευθείας τα χαρακτηριστικά εξόδου τους, παραλείποντας, δηλαδή, τη διαδικασία υποδειγματοληψίας και υπερδειγματοληψίας, που παρεμβάλλεται μεταξύ τους [77].



Εικ. 9: Multi-scale DenseNet (MDenseNet) τριών διαφορετικών μεγεθών μπλοκ.

4.3 ΤΑ ΜΟΝΤΕΛΑ MMDENSENET ΚΑΙ MMDENSELSTM

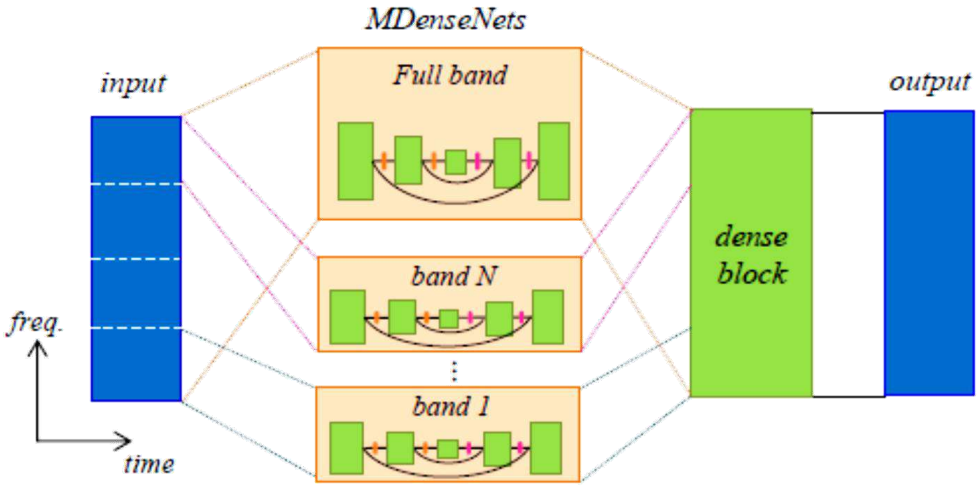
Το μοντέλο MMDenseLSTM είναι επέκταση του MMDenseNet, το οποίο με τη σειρά του αποτελεί επέκταση του μοντέλου MDenseNet που αναλύθηκε προηγουμένως. Επομένως, παρακάτω αναλύεται πρώτα το μοντέλο MMDenseNet και στη συνέχεια το μοντέλο MMDenseLSTM.

4.3.1 MMDENSENET

Τα σήματα ήχου εμφανίζουν διαφορετικές πληροφορίες σε διαφορετικές συχνοτικές ζώνες. Επομένως, ο διαχωρισμός του σήματος σε πολλαπλές συχνοτικές ζώνες οδηγεί σε πιο αποτελεσματική καταγραφή των τοπικών προτύπων. Το μοντέλο MMDenseNet (multi-band MDenseNet) υλοποιεί αυτόν το διαχωρισμό και χρησιμοποιεί ένα MDenseNet (**Εικ. 10**) για κάθε συχνοτική ζώνη. Ωστόσο, η μοντελοποίηση κάθε συχνοτικής ζώνης μεμονωμένα, εμποδίζει το δίκτυο να μοντελοποιήσει τη συνολική δομή του φασματογραφήματος του σήματος. Συνεπώς, χρησιμοποιείται ένα ακόμα MDenseNet για το συνολικό εύρος συχνοτήτων,

παράλληλα με τα υπόλοιπα, του οποίου η έξοδος συνδέεται με τις εξόδους των επιμέρους MDenseNets.

Ένα πλεονέκτημα του MMDenseNet είναι η δυνατότητα σχεδίασης κατάλληλης αρχιτεκτονικής και ανάθεσης υπολογιστικών πόρων για κάθε συχνοτική ζώνη ξεχωριστά, ανάλογα με την εκάστοτε εφαρμογή και με το πόσο σημαντική είναι κάθε ζώνη για τον διαχωρισμό της ζητούμενης πηγής από το συνολικό σήμα [77]. Επιπρόσθετα, εφόσον οι επιμέρους συχνοτικές ζώνες αποτυπώνουν τις λεπτομέρειες της δομής, το MDenseNet για το συνολικό εύρος συχνοτήτων αποτυπώνει την καθολική δομή, και συνεπώς αποτελεί ένα απλούστερο και χαμηλότερου κόστους μοντέλο.



Εικ. 10: Η αρχιτεκτονική του MMDenseNet. Οι έξοδοι των επιμέρους MDenseNets για κάθε συχνοτική ζώνη τροφοδοτούνται στο τελικό πυκνό μπλοκ, το οποίο παράγει την τελική έξοδο του δικτύου.

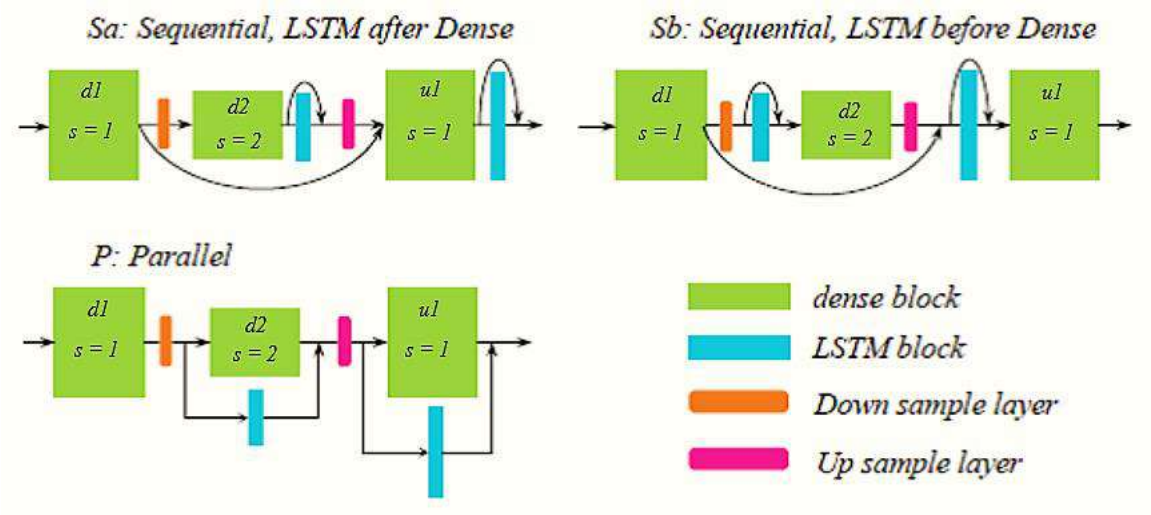
4.3.2 MMDENSELSTM

Το MMDenseLSTM είναι ουσιαστικά ο συνδυασμός του μοντέλου MMDenseNet και του LSTM σε μία ενιαία αρχιτεκτονική δικτύου. Το μπλοκ LSTM αποτελείται από μία 1x1 συνέλιξη, η οποία μειώνει το πλήθος των χαρτών χαρακτηριστικών σε 1, ακολουθείται από ένα επίπεδο BLSTM, το οποίο αντιμετωπίζει τον χάρτη χαρακτηριστικών ως μία χρονική ακολουθία δεδομένων και τέλος ένα επίπεδο εμπρός-τροφοδότησης το οποίο ανακτά την διάσταση της συχνότητας εισόδου. Έχουν προταθεί τρεις διαφορετικοί συνδυασμοί των επιπέδων των δύο δικτύων (**Εικ. 11**):

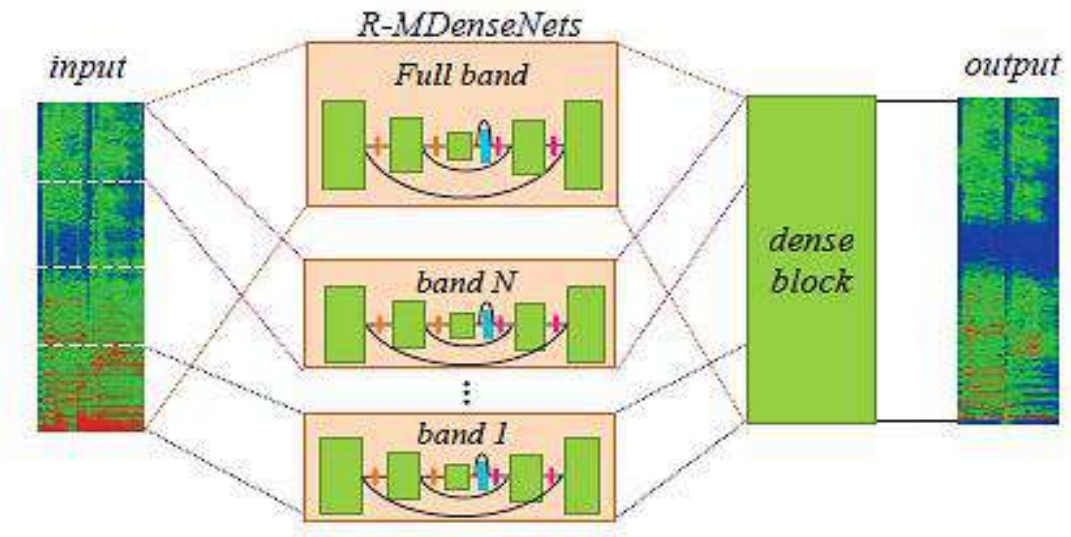
1. Το μπλοκ LSTM τοποθετείται σειριακά, μετά από κάποιο πυκνό μπλοκ κλίμακας *s*, όπως φαίνεται στην διάταξη *Sa* της **Εικ. 11**.

2. Το μπλοκ LSTM τοποθετείται σειριακά, πριν από κάποιο πυκνό μπλοκ κλίμακας *s*, όπως φαίνεται στην διάταξη *Sb* της **Εικ. 11**.
3. Το μπλοκ LSTM τοποθετείται παράλληλα με κάποιο πυκνό μπλοκ κλίμακας *s*, όπως φαίνεται στην διάταξη *P* της **Εικ. 11**.

Η προσθήκη των μπλοκ LSTM σε κάθε επίπεδο, και ειδικότερα στα πυκνα μπλοκ κλίμακας $s = 1$, οδηγεί σε μεγάλη αύξηση του συνολικού μεγέθους του δικτύου. Το πρόβλημα αυτό αντιμετωπίζεται ελαττώνοντας το πλήθος των μπλοκ LSTM στο δίκτυο. Πιο συγκεκριμένα, προστίθεται μόνο ένα μικρό πλήθος μπλοκ LSTM στο μονοπάτι υπερδειγματοληψίας, σε επίπεδα μικρής κλίμακας ($s > 1$). Στην **Εικ. 12** παρουσιάζεται η αρχιτεκτονική του δικτύου MMDenseLSTM όπως δίνεται στο [78].



Εικ. 11: Διατάξεις των μπλοκ LSTM ως προς τα πυκνά μπλοκ, σε ένα δίκτυο MMDenseLSTM. Τα μπλοκ LSTM δεν εισάγονται στο πρώτο πυκνό μπλοκ d1.



Εικ. 12: Η αρχιτεκτονική του δικτύου MMDenseLSTM.

4.4 ΘΕΜΑΤΑ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑΣ ΚΑΙ ΜΕΤΕΠΕΞΕΡΓΑΣΙΑΣ

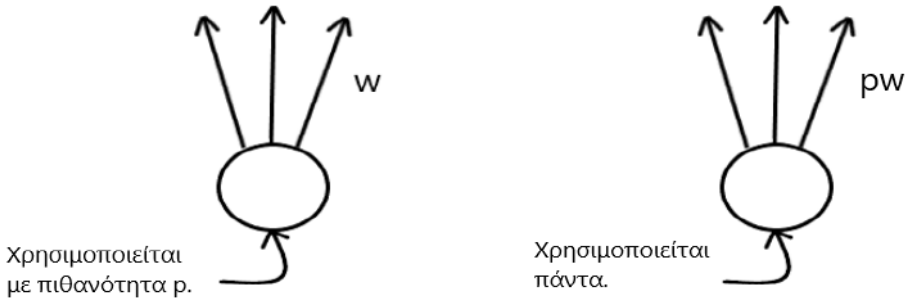
Ένα πολύ συχνό φαινόμενο στον επιστημονικό τομέα της μηχανικής μάθησης είναι η εμφάνιση ορισμένων προβλημάτων κατά τη διαδικασία εκπαίδευσης των δικτύων. Παρακάτω αναφέρονται μερικά από τα πιο σημαντικά και συνήθη προβλήματα που παρατηρούνται και οι πιθανοί τρόποι επίλυσής τους.

4.4.1 ΥΠΕΡΠΡΟΣΑΡΜΟΓΗ (OVERFITTING)

Η υπερπροσαρμογή αποτελεί βασικό πρόβλημα στις τεχνικές μηχανικής μάθησης και κατ' επέκταση στα νευρωνικά δίκτυα. Ένα μοντέλο νευρωνικού δικτύου εκπαιδεύεται σε ένα σύνολο δεδομένων εκπαίδευσης, ώστε στη συνέχεια να εφαρμόζεται σε νέα δεδομένα και να παράγει όσο το δυνατόν πιο ακριβή αποτελέσματα. Συνεπώς, κατά τη διάρκεια της εκπαίδευσης το σύστημα προσπαθεί να προσαρμοστεί στα δεδομένα με τα οποία εκπαιδεύεται για την παραγωγή ορθών αποτελεσμάτων. Ως εκ τούτου, υπάρχει ο κίνδυνος υπερπροσαρμογής του συστήματος στα συγκεκριμένα δεδομένα, δηλαδή, αντί το σύστημα να αποκτήσει τη δυνατότητα γενίκευσης βάσει των δεδομένων εκπαίδευσης, να απομνημονεύσει τις ιδιαιτερότητες των δεδομένων αυτών. Αυτό έχει ως απότοκο την χαμηλή επίδοση όταν δέχεται νέα δεδομένα ως είσοδο [79].

4.4.2 DROPOUT

Στην προηγούμενη υποενότητα έγινε μια αναφορά στο πρόβλημα της υπερπροσαρμογής στα DNNs. Μια λύση σε αυτό το πρόβλημα μπορεί να δωθεί με τη χρήση της Dropout. Η Dropout αποτελεί μια τεχνική, της οποίας η βασική ιδέα είναι η τυχαία απόρριψη ορισμένων νευρώνων (μαζί με τις συνδέσεις τους) από ένα νευρωνικό δίκτυο, κατά τη διάρκεια της εκπαίδευσής του. Ως αποτέλεσμα, δημιουργούνται αραιά δίκτυα, των οποίων το πλήθος είναι εκθετικό ως προς το πλήθος των νευρώνων στο δίκτυο. Αυτό σημαίνει ότι ένα νευρωνικό δίκτυο n νευρώνων, μπορεί να διαμορφωθεί ως μία συλλογή 2^n πιθανών αραιών δικτύων. Η απλούστερη περίπτωση της Dropout, είναι κάθε νευρώνας να διατηρείται στο δίκτυο με σταθερή πιθανότητα p. Εφαρμόζεται πολλαπλασιασμός των εξερχόμενων συναπτικών βαρών του νευρώνα με την τιμή της p, ώστε να προκύψει ένα ενιαίο νευρωνικό δίκτυο του οποίου τα συναπτικά βάρη προσεγγίζουν το μέσο όρο των βαρών των 2^n αραιών δικτύων (**Εικ. 13**). Αυτό το δίκτυο χρησιμοποιείται κατά τη διαδικασία επαλήθευσης χωρίς εφαρμογή της Dropout [80].



Εικ. 13: Αναπαράσταση ενός νευρώνα και των εξερχόμενων συναπτικών βαρών (w). ***Αριστερά*** αναπαρίσταται κατά τη διαδικασία εκπαίδευσης, ενώ ***δεξιά*** αναπαρίσταται κατά τη διαδικασία επαλήθευσης.

4.4.3 ΠΟΛΥΚΑΝΑΛΟ ΦΙΛΤΡΟ WIENER

Το πολυκάναλο φίλτρο Wiener (Multichannel Wiener Filter – MWF) εφαρμόζεται ως ένα στάδιο μετεπεξεργασίας για τη βελτίωση της απόδοσης διαχωρισμού. Η εφαρμογή του MWF ουσιαστικά διασφαλίζει ότι το άθροισμα των εκτιμήσεων των πηγών δίνει την αρχική μίξη. Μειώνει σημαντικά τις παρεμβολές από άλλες πηγές, καθώς και τον τεχνητό θόρυβο που οφείλεται στον αλγόριθμο διαχωρισμού.
Με βάση το MWF έχουμε το παρακάτω μοντέλο:

$$X(k,n) = S_i(k,n) + Z_i(k,n) \quad (37)$$

όπου $\boldsymbol{X(k,n)}$ αναπαριστά το φασματογράφημα του σήματος μίξης, $\boldsymbol{S_i(k,n)}$ το φασματογράφημα της πηγής i και $\boldsymbol{Z_i(k,n)} = \sum_{j \neq i} \boldsymbol{S_j(k,n)}$ το φασματογράφημα του υπολειπόμενου σήματος για τον $\boldsymbol{(k,n)}$ κάδο TF.

Κάθε κάδος TF των S_i είναι ένα μιγαδικό διάνυσμα που ακολουθεί Κανονική (Gaussian) κατανομή με μηδενική μέση τιμή και μητρώο συνδιασποράς $\boldsymbol{u_i(k,n)R_i(k)}$, όπου $\boldsymbol{u_i(k,n)}$ ορίζεται ως η φασματική πυκνότητα ισχύος (Power Spectral Density – PSD) και $\boldsymbol{R_i(k)}$ το χρονικά-αμετάβλητο χωρικό μητρώο συνδιασποράς, αντίστοιχα [81]. Η εκτίμηση του ελάχιστου μέσου τετραγωνικού σφάλματος των S_i είναι η εξής:

$$\hat{S}_i(k,n) = u_i(k,n)R_i(k)\left(\sum_{\forall i} u_i(k,n)R_i(k)\right)^{-1} X(k,n) \quad (38)$$

όπου,

$$u_i(k,n) = \frac{1}{2}\left\|\hat{S}_i(k,n)\right\|^2 \quad (39) \text{ και } R_i(k) = \frac{\sum_{n=1}^{N} \hat{S}_i(k,n)\hat{S}_i(k,n)^H}{\sum_{n=1}^{N} u_i(k,n)} \quad (40)$$

5 ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΛΕΤΗ

Το παρόν κεφάλαιο αποτελεί την πειραματική μελέτη της εργασίας. Σκοπός είναι η μελέτη σύγχρονων μεθόδων που να ανταποκρίνονται αποδοτικά στο πρόβλημα του διαχωρισμού μουσικών πηγών, και συγκεκριμένα στο πρόβλημα απομόνωσης του σήματος των φωνητικών από το σήμα της συνοδείας. Υλοποιήθηκαν και εκπαιδεύτηκαν δύο αρχιτεκτονικές για τα μοντέλα MMDenseNet και MMDenseLSTM, με το σύνολο δεδομένων musdb18, και στη συνέχεια αξιολογήθηκαν τα αποτελέσματά τους ως προς αυτόν το σκοπό.

5.1 ΤΟ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ MUSDB18

Το musdb18 [82] είναι ένα σύνολο δεδομένων διαχωρισμού σημάτων, το οποίο αποτελείται συνολικά από 150 μουσικά κομμάτια επαγγελματικής μίξης, από διαφορετικά είδη και περιλαμβάνει τόσο τις στερεοφωνικές μίξεις όσο και τις αρχικές πηγές (δηλαδή απομονωμένα τα ντραμς, το μπάσο, τα φωνητικά και την υπόλοιπη συνοδεία), οι οποίες χωρίζονται σε ένα υποσύνολο εκπαίδευσης (100 κομμάτια, 6.5 ώρες) και ένα υποσύνολο δοκιμής (50 κομμάτια, 3.5 ώρες). Η συνολική διάρκεια των κομματιών είναι περίπου δέκα ώρες. Όλα τα σήματα είναι στερεοφωνικά και κωδικοποιημένα στα 44.1 kHz.

Ουσιαστικά, το musdb18 λειτουργεί ως μία βάση δεδομένων αναφοράς για το σχεδιασμό και την αξιολόγηση μεθόδων διαχωρισμού πηγών, όπως αυτές που μελετώνται στην παρούσα εργασία.

Επιπλέον, έχει χρησιμοποιηθεί ως το επίσημο σύνολο δεδομένων στον διαγωνισμό SiSEC 2018, ο οποίος οργανώνεται διεθνώς από την επιστημονική κοινότητα για την αξιολόγηση αλγορίθμων διαχωρισμού πηγών.

5.2 ΛΕΠΤΟΜΕΡΕΙΕΣ ΥΛΟΠΟΙΗΣΗΣ

Το εργό των δικτύων είναι, χρησιμοποιώντας ένα σύνολο μουσικών κομματιών επαγγελματικής μίξης, να διαχωρίσουν τα φωνητικά από τη συνοδεία. Για τον σκοπό αυτό εκπαιδεύσαμε τα μοντέλα MMDenseNet και MMDenseLSTM. Κάθε μοντέλο εκπαιδεύτηκε σε δύο αρχιτεκτονικές και στη συνέχεια έγινε η σύγκριση των αποτελεσμάτων αυτών των αρχιτεκτονικών.

Κατά τον διαχωρισμό, υπολογίζεται το φασματογράφημα μίξης. Χρησιμοποιώντας το φασματογράφημα μίξης, παράγεται η εκτίμηση του φασματογραφήματος των φωνητικών, δηλαδή της πηγής που θέλουμε να διαχωρίσουμε, και η εκτίμηση της συνοδείας ως το υπόλοιπο από το φασματογράφημα μίξης, μέσω του πολυκάναλου φίλτρου Wiener. Τέλος, παράγονται τα σήματα των πηγών μετά από εφαρμογή του ISTFT.

Για τον STFT χρησιμοποιήθηκε μήκος FFT 4096 δειγμάτων με βήμα 1024 δείγματα. Το παράθυρο που χρησιμοποιήθηκε είναι το hanning. Κατά την βελτιστοποίηση χρησιμοποιήθηκε η μέθοδος RMSProp και η αντικειμενική συνάρτηση MSE.

Ορίζουμε ως:

- ***k*** τον ρυθμό ανάπτυξης, δηλαδή το πλήθος των χαρτών χαρακτηριστικών που παράγει το κάθε επίπεδο.

- ***p = 0.2*** την πιθανότητα απόρριψης νευρώνα για την Dropout.

5.2.1 MMDENSENET

Για το MMDenseNet, χρησιμοποιήθηκαν δύο αρχιτεκτονικές, οι οποίες διαφέρουν ως προς τον ρυθμό ανάπτυξης ***k*** του επιπέδου εξόδου και ως προς το μέγεθος πυρήνα της μέσης συγκέντρωσης.

Αρχιτεκτονική_1 (Large): {*k = 8, Average_pooling_kernel_size = 3x3*}

Αρχιτεκτονική_2 (Small): {*k = 4, Average_pooling_kernel_size = 2x2*}

Αρχιτεκτονική_1 (Large)

Χωρίζουμε το εύρος συχνοτήτων των δεδομένων εισόδου στη μέση με αποτέλεσμα να διαμορφώνονται τρεις συχνοτικές ζώνες για το δίκτυο: low band, high band και full band. Το δίκτυο αποτελείται από δύο κανάλια εισόδου. Αρχικά, δημιουργείται ένα πλήθος από ***M = 32*** χάρτες χαρακτηριστικών μέσω συνέλιξης με μέγεθος πυρήνα 4x3.

Low band MDenseNet:

- Αποτελείται συνολικά από 7 πυκνά μπλοκ: 4 πυκνά μπλοκ ανήκουν στο μονοπάτι υποδειγματοληψίας (το τελευταίο καλείται και bottleneck block) και 3 στο μονοπάτι υπερδειγματοληψίας. Το πρώτο πυκνό μπλοκ έχει ρυθμό ανάπτυξης k = 14 και όλα τα υπόλοιπα έχουν ρυθμό ανάπτυξης k = 16.
- Κάθε πυκνό μπλοκ αποτελείται από 4 πυκνά επίπεδα.
- Κάθε πυκνό επίπεδο αποτελείται από τρεις διαδοχικές πράξεις: ομαδική κανονικοποίηση (batch normalization: BN), ReLU και 4x3 συνέλιξη (Conv).
- Στην έξοδο κάθε πυκνού μπλοκ του μονοπατιού υποδειγματοληψίας εκτός του τελευταίου, προστίθεται ένα 3x3 επίπεδο μέσης συγκέντρωσης (average pooling layer) για την ομαλή διεκπαιρέωση της διαδικασίας υ-

ποδειγματοληψίας. Κατά συνέπεια, στην είσοδο κάθε πυκνού μπλοκ του μονοπατιού υπερδειγματοληψίας προστίθεται μια πράξη 3x3 συνέλιξης.

High band MDenseNet

- Αποτελείται συνολικά από 7 πυκνά μπλοκ: 4 πυκνά μπλοκ ανήκουν στο μονοπάτι υποδειγματοληψίας (το τελευταίο καλείται και bottleneck block) και 3 στο μονοπάτι υπερδειγματοληψίας. Ο ρυθμός ανάπτυξης των επιπέδων είναι k = 10.
- Κάθε πυκνό μπλοκ αποτελείται από 3 πυκνά επίπεδα.
- Κάθε πυκνό επίπεδο αποτελείται από τρεις διαδοχικές πράξεις: ομαδική κανονικοποίηση (batch normalization: BN), ReLU και 3x3 συνέλιξη (Conv).
- Στην έξοδο κάθε πυκνού μπλοκ του μονοπατιού υποδειγματοληψίας εκτός του τελευταίου, προστίθεται ένα 3x3 επίπεδο μέσης συγκέντρωσης (average pooling layer) για την ομαλή διεκπαιρέωση της διαδικασίας υποδειγματοληψίας. Κατά συνέπεια, στην είσοδο κάθε πυκνού μπλοκ του μονοπατιού υπερδειγματοληψίας προστίθεται μια πράξη 3x3 συνέλιξης.

Full band MDenseNet

- Αποτελείται συνολικά από 7 πυκνά μπλοκ: 4 πυκνά μπλοκ ανήκουν στο μονοπάτι υποδειγματοληψίας (το τελευταίο καλείται και bottleneck block) και 3 στο μονοπάτι υπερδειγματοληψίας. Ο ρυθμός ανάπτυξης των επιπέδων είναι k = 6.
- Κάθε πυκνό μπλοκ αποτελείται από 2 πυκνά επίπεδα, εκτός από το τελευταίο μπλοκ στο μονοπάτι υποδειγματοληψίας (bottleneck block), το οποίο αποτελείται από 4 πυκνά επίπεδα.
- Κάθε πυκνό επίπεδο αποτελείται από τρεις διαδοχικές πράξεις: ομαδική κανονικοποίηση (batch normalization: BN), ReLU και 4x3 συνέλιξη (Conv).
- Στην έξοδο κάθε πυκνού μπλοκ του μονοπατιού υποδειγματοληψίας εκτός του τελευταίου, προστίθεται ένα 3x3 επίπεδο μέσης συγκέντρωσης (average pooling layer) για την ομαλή διεκπαιρέωση της διαδικασίας υποδειγματοληψίας. Κατά συνέπεια, στην είσοδο κάθε πυκνού μπλοκ του μονοπατιού υπερδειγματοληψίας προστίθεται μια πράξη 3x3 συνέλιξης.

Dense Block Εξόδου

Το πυκνό μπλοκ εξόδου αποτελείται από 2 πυκνά επίπεδα με ρυθμό ανάπτυξης *k = 8*, τα οποία αποτελούνται από τις διαδοχικές πράξεις της ομαδικής κανονικοποίησης (batch normalization: BN), ReLU και 2x1 συνέλιξης, έτσι ώστε να παράγονται μη αρνητικά δεδομένα στην έξοδο.

Αρχιτεκτονική_2 (Small)

Για την δεύτερη αρχιτεκτονική ισχύουν οι ίδιες λεπτομέρειες υλοποίησης με την πρώτη, με τη διαφορά ότι το μέγεθος πυρήνα της μέσης συγκέντρωσης είναι 2x2 και ο ρυθμός ανάπτυξης του επιπέδου εξόδου είναι ***k = 4***.

5.2.2 MMDENSELSTM

Όπως για το MMDenseNet, έτσι και για το MMDenseLSTM χρησιμοποιήθηκαν δύο αρχιτεκτονικές, οι οποίες διαφέρουν ως προς τον ρυθμό ανάπτυξης k του επιπέδου εξόδου και ως προς το μέγεθος πυρήνα της μέσης συγκέντρωσης.

Αρχιτεκτονική_1 (Large): {*k = 8, Average_pooling_kernel_size = 3x3*}

Αρχιτεκτονική_2 (Small): {*k = 4, Average_pooling_kernel_size = 2x2*}

Αρχιτεκτονική_1 (Large)

Χωρίζουμε το εύρος συχνοτήτων των δεδομένων εισόδου σε τρεις συχνοτικές ζώνες για το δίκτυο: low band, middle band και high band. Κατά συνέπεια, μαζί με την full band έχουμε συνολικά τέσσερις συχνοτικές ζώνες. Το δίκτυο αποτελείται από δύο κανάλια εισόδου. Αρχικά, δημιουργείται ένα πλήθος από ***M = 32*** χάρτες χαρακτηριστικών μέσω συνέλιξης με μέγεθος πυρήνα 3x3.

Low band MDenseNet:

- Αποτελείται συνολικά από 7 πυκνά μπλοκ: 4 πυκνά μπλοκ ανήκουν στο μονοπάτι υποδειγματοληψίας (το τελευταίο καλείται και bottleneck block) και 3 στο μονοπάτι υπερδειγματοληψίας. Ο ρυθμός ανάπτυξης των επιπέδων είναι k = 14.
- Κάθε πυκνό μπλοκ αποτελείται από 5 πυκνά επίπεδα.
- Κάθε πυκνό επίπεδο αποτελείται από τρεις διαδοχικές πράξεις: ομαδική κανονικοποίηση (batch normalization: BN), ReLU και 3x3 συνέλιξη (Conv).
- Στην έξοδο κάθε πυκνού μπλοκ του μονοπατιού υποδειγματοληψίας εκτός του τελευταίου, προστίθεται ένα 3x3 επίπεδο μέσης συγκέντρωσης (average pooling layer) για την ομαλή διεκπαιρέωση της διαδικασίας υποδειγματοληψίας. Κατά συνέπεια, στην είσοδο κάθε πυκνού μπλοκ του μονοπατιού υπερδειγματοληψίας προστίθεται μια πράξη 3x3 συνέλιξης.
- Το μπλοκ LSTM εισάγεται σε δύο σημεία. Το πρώτο είναι πριν από το bottleneck block και το δεύτερο πριν από το δεύτερο πυκνό μπλοκ του μονοπατιού υπερδειγματοληψίας. Το μπλοκ LSTM αποτελείται από μία 1x1 συνέλιξη και από ένα επίπεδο BLSTM 64 μονάδων και τέλος ένα επίπεδο

εμπρός-τροφοδότησης το οποίο ανακτά την διάσταση των δεδομένων που έλαβε στην είσοδό του το μπλοκ LSTM.

Middle band MDenseNet:

- Αποτελείται συνολικά από 7 πυκνά μπλοκ: 4 πυκνά μπλοκ ανήκουν στο μονοπάτι υποδειγματοληψίας (το τελευταίο καλείται και bottleneck block) και 3 στο μονοπάτι υπερδειγματοληψίας. Ο ρυθμός ανάπτυξης των επιπέδων είναι k = 4.
- Κάθε πυκνό μπλοκ αποτελείται από 4 πυκνά επίπεδα.
- Κάθε πυκνό επίπεδο αποτελείται από τρεις διαδοχικές πράξεις: ομαδική κανονικοποίηση (batch normalization: BN), ReLU και 3x3 συνέλιξη (Conv).
- Στην έξοδο κάθε πυκνού μπλοκ του μονοπατιού υποδειγματοληψίας εκτός του τελευταίου, προστίθεται ένα 3x3 επίπεδο μέσης συγκέντρωσης (average pooling layer) για την ομαλή διεκπαιρέωση της διαδικασίας υποδειγματοληψίας. Κατά συνέπεια, στην είσοδο κάθε πυκνού μπλοκ του μονοπατιού υπερδειγματοληψίας προστίθεται μια πράξη 3x3 συνέλιξης.
- Το μπλοκ LSTM εισάγεται πριν από το bottleneck block. Το μπλοκ LSTM αποτελείται από μία 1x1 συνέλιξη και από ένα επίπεδο BLSTM 16 μονάδων και τέλος ένα επίπεδο εμπρός-τροφοδότησης το οποίο ανακτά την διάσταση των δεδομένων που έλαβε στην είσοδό του το μπλοκ LSTM.

High band MDenseNet:

- Αποτελείται συνολικά από 5 πυκνά μπλοκ: 3 πυκνά μπλοκ ανήκουν στο μονοπάτι υποδειγματοληψίας (το τελευταίο καλείται και bottleneck block) και 2 στο μονοπάτι υπερδειγματοληψίας. Ο ρυθμός ανάπτυξης όλων των επιπέδων είναι k = 2, εκτός από το bottleneck block το οποίο αποτελείται μόνο από το μπλοκ LSTM.
- Κάθε πυκνό μπλοκ αποτελείται από 1 πυκνό επίπεδο, εκτός από το bottleneck block το οποίο αποτελείται μόνο από το μπλοκ LSTM.
- Κάθε πυκνό επίπεδο αποτελείται από τρεις διαδοχικές πράξεις: ομαδική κανονικοποίηση (batch normalization: BN), ReLU και 3x3 συνέλιξη (Conv).
- Στην έξοδο κάθε πυκνού μπλοκ του μονοπατιού υποδειγματοληψίας εκτός του τελευταίου, προστίθεται ένα 3x3 επίπεδο μέσης συγκέντρωσης (average pooling layer) για την ομαλή διεκπαιρέωση της διαδικασίας υποδειγματοληψίας. Κατά συνέπεια, στην είσοδο κάθε πυκνού μπλοκ του μονοπατιού υπερδειγματοληψίας προστίθεται μια πράξη 3x3 συνέλιξης.
- Το μπλοκ LSTM εισάγεται πριν από το bottleneck block. Το μπλοκ LSTM αποτελείται από μία 1x1 συνέλιξη και από ένα επίπεδο BLSTM 4 μονάδων και τέλος ένα επίπεδο εμπρός-τροφοδότησης το οποίο ανακτά την διά-

σταση των δεδομένων που έλαβε στην είσοδό του το μπλοκ LSTM.

Full band MDenseNet:

- Αποτελείται συνολικά από 9 πυκνά μπλοκ: 5 πυκνά μπλοκ ανήκουν στο μονοπάτι υποδειγματοληψίας (το τελευταίο καλείται και bottleneck block) και 4 στο μονοπάτι υπερδειγματοληψίας. Ο ρυθμός ανάπτυξης των επιπέδων είναι k = 7.
- Τα πρώτα δύο πυκνά μπλοκ του μονοπατιού υποδειγματοληψίας και τα τελευταία δύο του μονοπατιού υπερδειγματοληψίας αποτελούνται από 3 πυκνά επίπεδα. Το τρίτο πυκνό μπλοκ του μονοπατιού υποδειγματοληψίας και το δεύτερο του μονοπατιού υπερδειγματοληψίας αποτελούνται από 4 πυκνά επίπεδα. Τα υπόλοιπα πυκνά μπλοκ αποτελούνται από 5 πυκνά επίπεδα.
- Κάθε πυκνό επίπεδο αποτελείται από τρεις διαδοχικές πράξεις: ομαδική κανονικοποίηση (batch normalization: BN), ReLU και 3x3 συνέλιξη (Conv).
- Στην έξοδο κάθε πυκνού μπλοκ του μονοπατιού υποδειγματοληψίας εκτός του τελευταίου, προστίθεται ένα 3x3 επίπεδο μέσης συγκέντρωσης (average pooling layer) για την ομαλή διεκπαιρέωση της διαδικασίας υποδειγματοληψίας. Κατά συνέπεια, στην είσοδο κάθε πυκνού μπλοκ του μονοπατιού υπερδειγματοληψίας προστίθεται μια πράξη 3x3 συνέλιξης.
- Το μπλοκ LSTM εισάγεται σε δύο σημεία. Το πρώτο είναι πριν από το bottleneck block και το δεύτερο πριν από το τρίτο πυκνό μπλοκ του μονοπατιού υπερδειγματοληψίας. Το μπλοκ LSTM αποτελείται από μία 1x1 συνέλιξη και από ένα επίπεδο BLSTM 64 μονάδων και τέλος ένα επίπεδο εμπρός-τροφοδότησης το οποίο ανακτά την διάσταση των δεδομένων που έλαβε στην είσοδό του το μπλοκ LSTM.

Dense Block Εξόδου

Το πυκνό μπλοκ εξόδου αποτελείται από 3 πυκνά επίπεδα με ρυθμό ανάπτυξης *k = 8*, τα οποία αποτελούνται από τις διαδοχικές πράξεις της ομαδικής κανονικοποίησης (batch normalization: BN), ReLU και 3x3 συνέλιξης, έτσι ώστε να παράγονται μη αρνητικά δεδομένα στην έξοδο.

Αρχιτεκτονική_2 (Small)

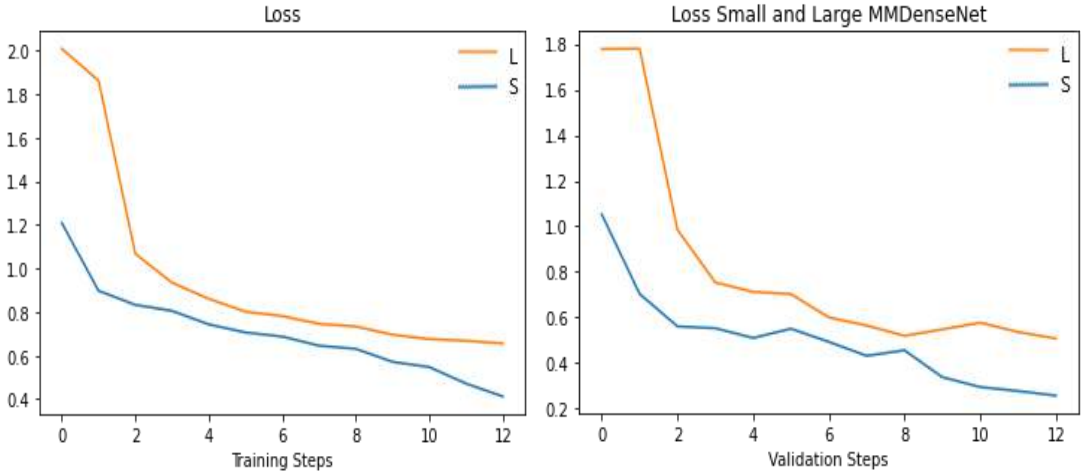
Για την δεύτερη αρχιτεκτονική ισχύουν οι ίδιες λεπτομέρειες υλοποίησης με την πρώτη, με τη διαφορά ότι το μέγεθος πυρήνα της μέσης συγκέντρωσης είναι 2x2 και ο ρυθμός ανάπτυξης του επιπέδου εξόδου είναι ***k = 4***.

5.3 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

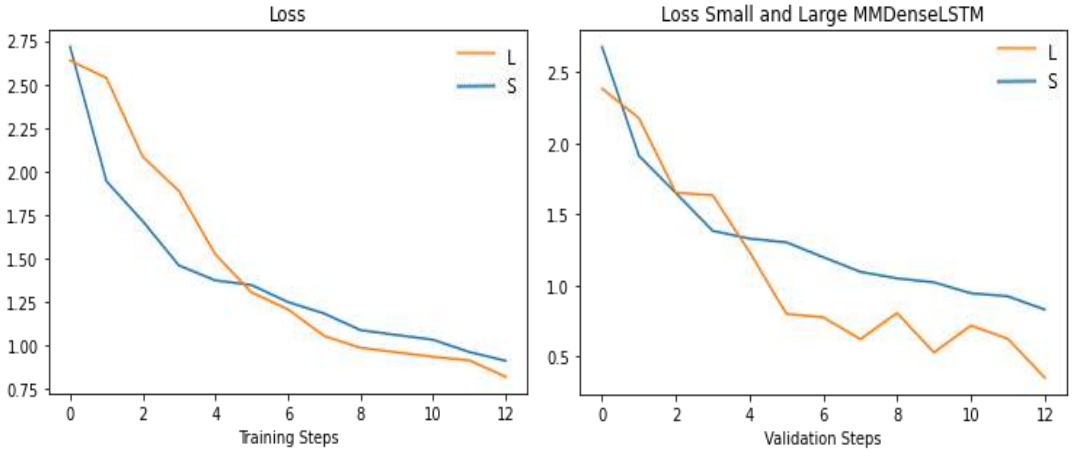
Παρακάτω παρατίθενται τα αποτελέσματα του πειράματος. Αρχικά, μελετάμε την απόδοση του διαχωρισμού με βάση τις καμπύλες μάθησης για τη διαδικασία εκπαίδευσης και επαλήθευσης των μοντέλων. Στη συνέχεια, μελετώντας τις μετρικές BSSEval, οδηγούμαστε σε συμπεράσματα σχετικά με τα αποτελέσματα διαχωρισμού των δύο αρχιτεκτονικών των μοντέλων. Τέλος, συγκρίνουμε μεταξύ τους τα δύο μοντέλα, επιβεβαιώνοντας έτσι, ότι οι πιο σύγχρονες επεκτάσεις των μοντέλων DenseNet, αποδεικνύονται περισσότερο αποτελεσματικές για το πρόβλημα του διαχωρισμού μουσικών πηγών.

5.3.1 ΚΑΜΠΥΛΕΣ ΜΑΘΗΣΗΣ

Στην (**Εικ. 14**) παρουσιάζονται οι καμπύλες μάθησης των δύο αρχιτεκτονικών του μοντέλου MMDenseNet και στην (**Εικ. 15**) παρουσιάζονται οι καμπύλες μάθησης για τις δύο αρχιτεκτονικές του μοντέλου MMDenseLSTM.



Εικ. 14: Καμπύλες μάθησης για τις αρχιτεκτονικές MMDenseNet. Η συνάρτηση κόστους (MSE) ως προς το πλήθος των επαναλήψεων (iteration steps), κατά τη διαδικασία εκπαίδευσης των φωνητικών (αριστερά) και κατά τη διαδικασία επαλήθευσης (δεξιά). Με **L** και **S** συβολίζονται η Αρχιτεκτονική_1 (Large) και η Αρχιτεκτονική_2 (Small). Ισχύει ότι 1 iteration step = 100 iterations.



Εικ. 15: Καμπύλες μάθησης για τις αρχιτεκτονικές MMDenseLSTM. Η συνάρτηση κόστους (MSE) ως προς το πλήθος των επαναλήψεων (iteration steps), κατά τη διαδικασία εκπαίδευσης των φωνητικών (αριστερά) και κατά τη διαδικασία επαλήθευσης (δεξιά). Με L και S συβολίζονται η Αρχιτεκτονική_1 (Large) και η Αρχιτεκτονική_2 (Small).

Ύστερα από κάθε επανάληψη (iteration step), υπολογίζεται η τιμή της συνάρτησης κόστους για τη διαδικασία εκπαίδευσης (*training loss*) και για τη διαδικασία επαλήθευσης (*validation loss*). Οι επαναλήψεις διακόπτονται όταν παρατηρηθεί το φαινόμενο της υπερπροσαρμογής, δηλαδή όταν οι δύο συναρτήσεις κόστους αποκτήσουν πολύ διαφορετικές τιμές και *training loss << validation loss*. Αυτή η τακτική αποφυγής της υπερπροσαρμογής ονομάζεται early stopping και ενέχει τον κίνδυνο του φαινομένου της υποπροσαρμογής, δηλαδή τον κίνδυνο να παρατηρηθούν υψηλές τιμές για τις συναρτήσεις κόστους. Ωστόσο, αυτός ο κίνδυνος εύκολα αποφεύγεται, εάν το σύνολο δεδομένων εκπαίδευσης είναι αρκετά μεγάλο, γεγονός που αυξάνει την ακρίβεια του μοντέλου. Με αυτόν τον τρόπο, αποκτούνται τα βέλτιστα αποτελέσματα, δηλαδή χαμηλές τιμές και για τις δύο συναρτήσεις κόστους και παρόμοιες μεταξύ τους.

Παρατηρούμε για το μοντέλο MMDenseNet, ότι η small αρχιτεκτονική παρουσιάζει μικρότερη τιμή MSE συγκριτικά με την large αρχιτεκτονική, δηλαδή προσαρμόζεται καλύτερα στα δεδομένα. Αντίθετα, η small αρχιτεκτονική του μοντέλου MMDenseLSTM παρουσιάζει μεγαλύτερη τιμή MSE συγκριτικά με την large αρχιτεκτονική, η οποία, κατά συνέπεια επιτυγχάνει καλύτερη προσαρμογή. Αυτό επιβεβαιώνεται στην επόμενη ενότητα, όπου αξιολογούμε την απόδοση καθεμίας αρχιτεκτονικής των δύο μοντέλων.

5.3.2 ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ BSSEVAL

Παρακάτω παρατίθενται τα αποτελέσματα των μετρικών του BSSEval toolbox, οι οποίες αναλύθηκαν στο **Κεφάλαιο 2.3.2**, για τις δύο αρχιτεκτονικές των μοντέλων MMDenseNet (**Εικ. 16**) και MMDenseLSTM (**Εικ. 17**). Το σύνολο δεδομένων musdb18, όπως έχουμε προαναφέρει, αποτελείται από κομμάτια επαγγελ-

ματικής μίξης. Αυτό σημαίνει ότι τα σήματα μίξης είναι καθαρά, δηλαδή, έχει αφαιρεθεί ο προσθετικός θόρυβος. Κατά συνέπεια, όπως ορίζεται και από το [36], όταν το σήμα θορύβου δεν είναι γνωστό, ή δεν υπάρχει προσθετικός θόρυβος, τα κριτήρια απόδοσης που υπολογίζονται είναι οι μετρικές SDR, SIR και SAR.

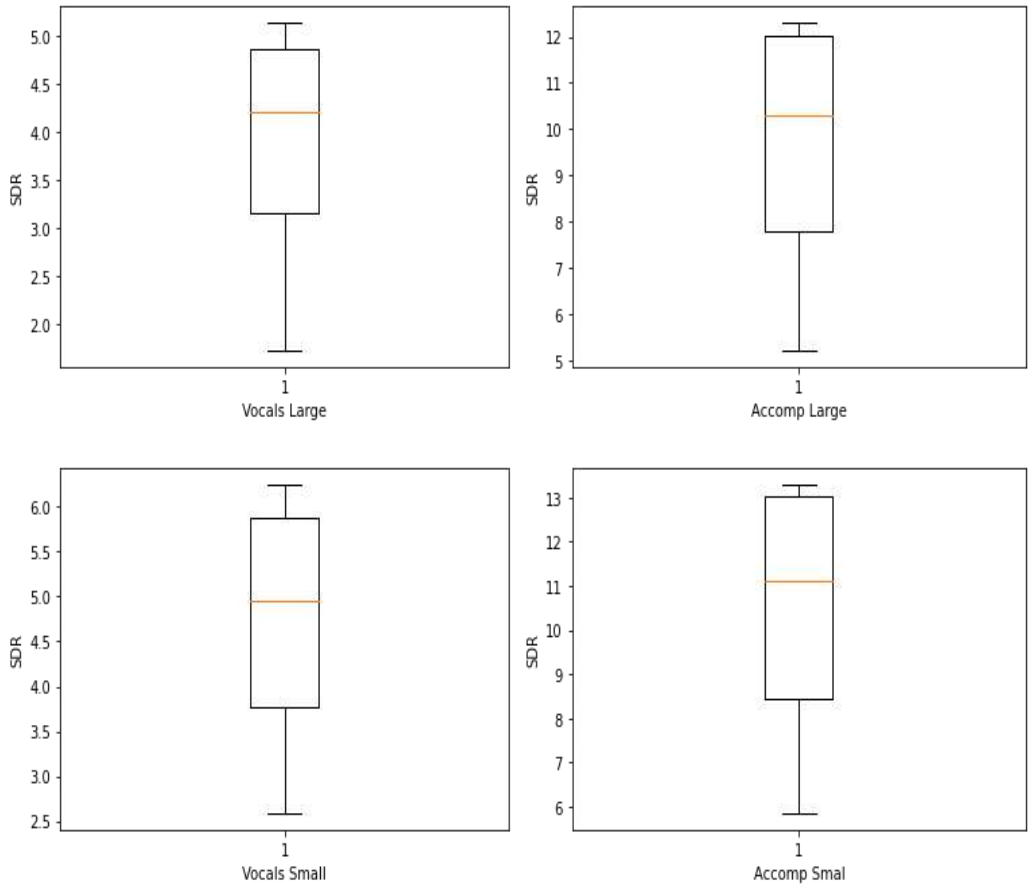
Υπενθυμίζουμε ότι:

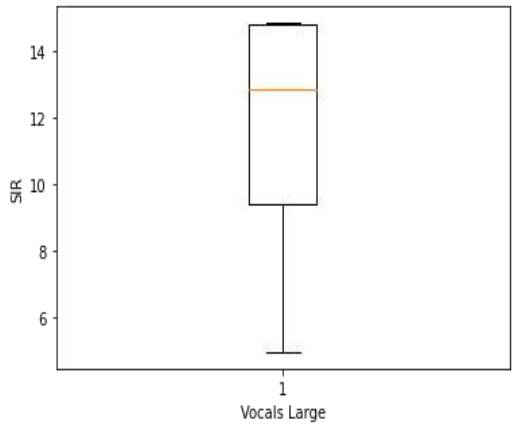
- Η μετρική **SDR** (11) αναφέρεται ως ο *λόγος πηγής προς παραμόρφωση*, ο οποίος πρακτικά υπολογίζει πόσο καθαρό είναι το σήμα από οποιαδήποτε παραμόρφωση ή θόρυβο.

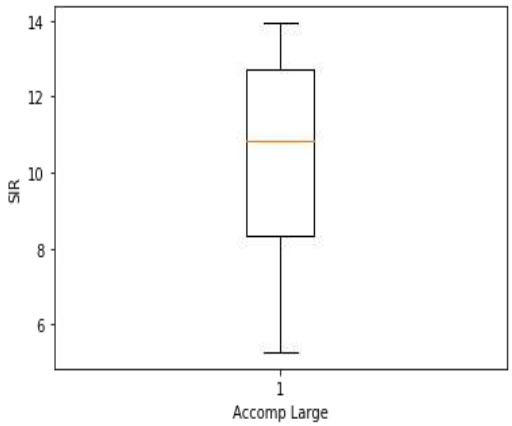
- Η μετρική **SIR** (12) αναφέρεται ως ο *λόγος πηγής προς παρεμβάλλουσες πηγές*, ο οποίος πρακτικά υπολογίζει πόσο καθαρό είναι το σήμα από παρεμβολές των υπόλοιπων σημάτων πηγών.

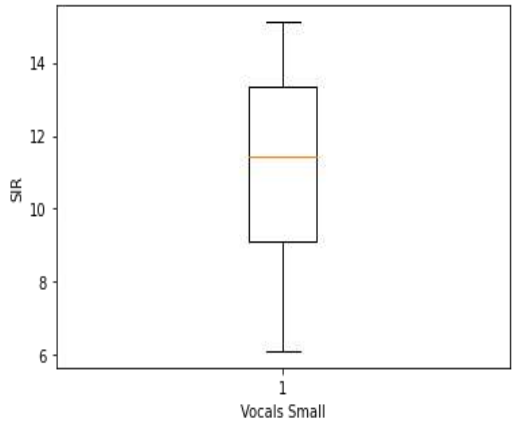
- Η μετρική **SAR** (14) αναφέρεται ως ο *λόγος πηγής προς σφάλματα*, ο οποίος πρακτικά υπολογίζει πόσο καθαρό είναι το σήμα από παρεμβολές πρόσθετων σφαλμάτων και τεχνητού θορύβου που ενδεχομένως εισάγει ο αλγόριθμος.

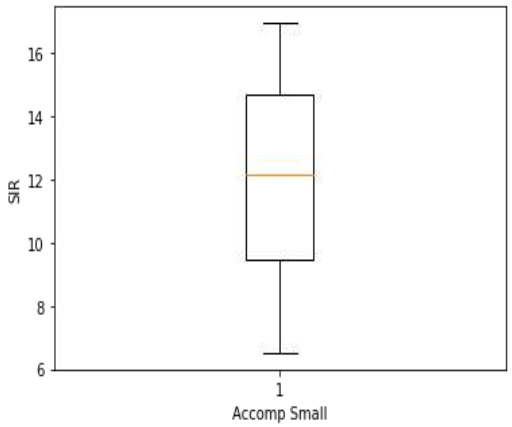
MMDenseNet

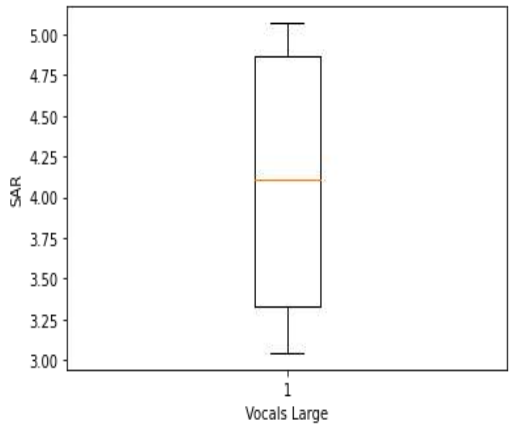


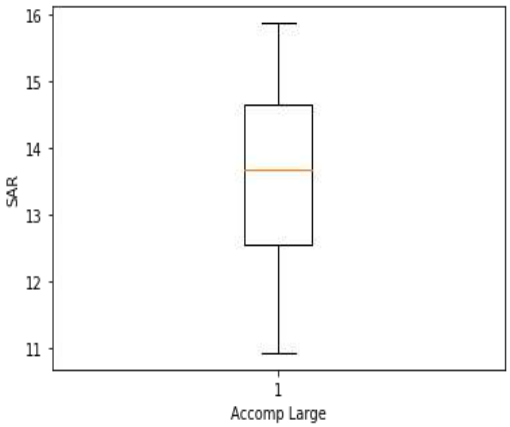

SIR
14
12
10
8
6
1
Vocals Large

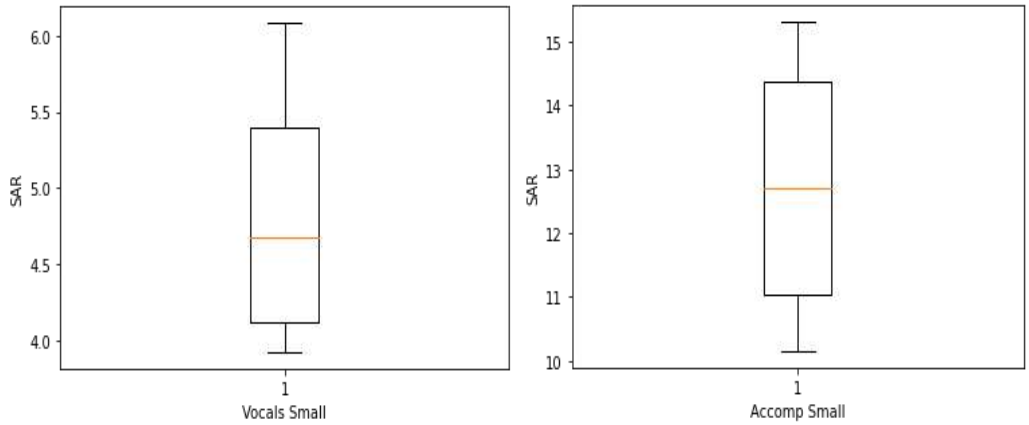

SIR
14
12
10
8
6
1
Accomp Large


SIR
14
12
10
8
6
1
Vocals Small


SIR
16
14
12
10
8
6
1
Accomp Small


SAR
5.00
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
1
Vocals Large

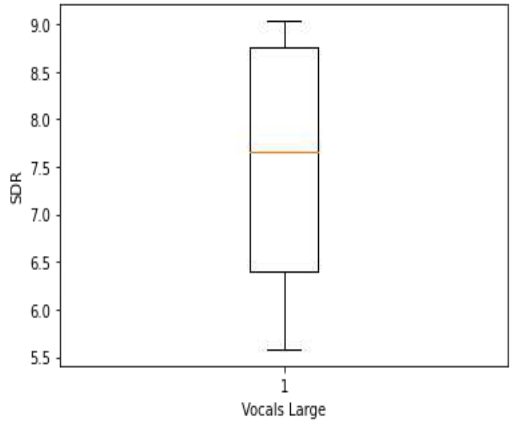

SAR
16
15
14
13
12
11
1
Accomp Large

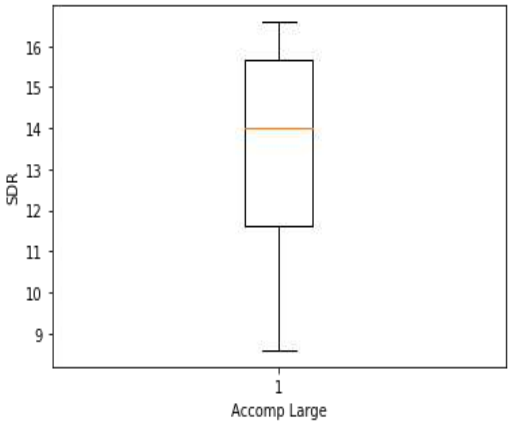


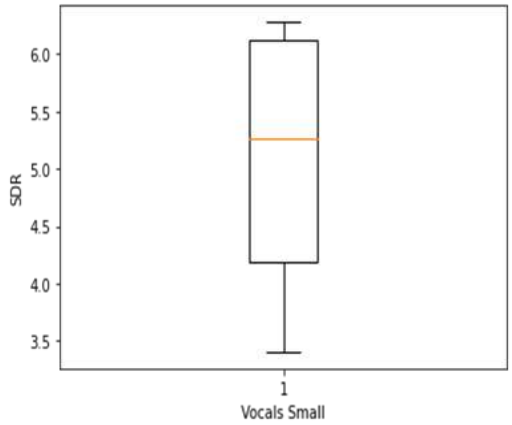
Εικ. 16: Αποτελέσματα απόδοσης του διαχωρισμού για τις αρχιτεκτονικές MMDense-Net

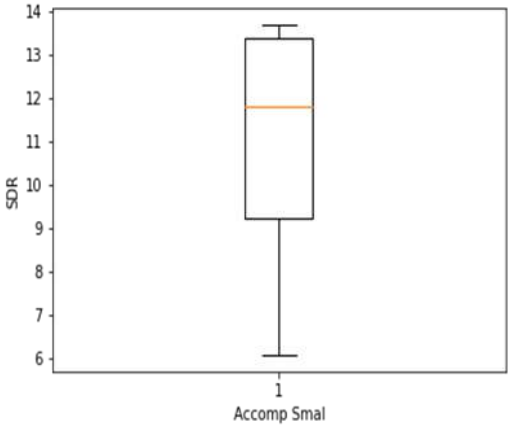
Γενικά, παρατηρούμε (**Εικ. 16**) ότι οι δύο αρχιτεκτονικές παρουσιάζουν παρόμοια απόδοση, με την αρχιτεκτονική small να εμφανίζει ελαφρώς υψηλότερες τιμές και συνεπώς καλύτερη απόδοση. Αυτό συμβαίνει διότι εκπαιδεύσαμε τις δύο αρχιτεκτονικές για τον ίδιο αριθμό επαναλήψεων, αλλά οι μεγαλύτερες αρχιτεκτονικές χρειάζονται περισσότερο training για να συγκλίνουν με μεγαλύτερη απόδοση. Η καλύτερη απόδοση SDR του σήματος συνοδείας, οφείλεται στο γεγονός ότι η μετρική SDR είναι ευαίσθητη στην χαμηλή συχνοτική ζώνη όταν περιλαμβάνει υψηλές τιμές ενέργειας [83]. Η ακριβής εκτίμηση της χαμηλής συχνοτικής ζώνης οδηγεί σε υψηλή τιμή απόδοσης SDR. Δεδομένου, όμως, ότι το σήμα της φωνής έχει σημεία με υψηλότερη ενέργεια, συγκριτικά με το σήμα συνοδείας, τα οποία είναι σημαντικό να διατηρηθούν κατά τη διαδικασία διαχωρισμού, η τιμή απόδοσης SDR των φωνητικών, είναι χαμηλότερη από αυτή της συνοδείας. Με βάση τον ορισμό της μετρικής SDR, τη σχέση (11) και εντοπίζοντας παρόμοια συμπεριφορά όσον αφορά τη μετρική SAR, συμπεραίνουμε ότι το σφάλμα λόγω τεχνητού θορύβου που προκύπτει από τον αλγόριθμο διαχωρισμού, είναι μεγαλύτερο κατά τη διαδικασία απομόνωσης των φωνητικών. Αυτό πρακτικά σημαίνει ότι, ο αλγόριθμος, στη διάρκεια εκτέλεσής του, επιχειρώντας να διατηρήσει αναλλοίωτα τα στοιχεία του σήματος της φωνής κατά τον διαχωρισμό, εισάγει κάποιο πρόσθετο σφάλμα.

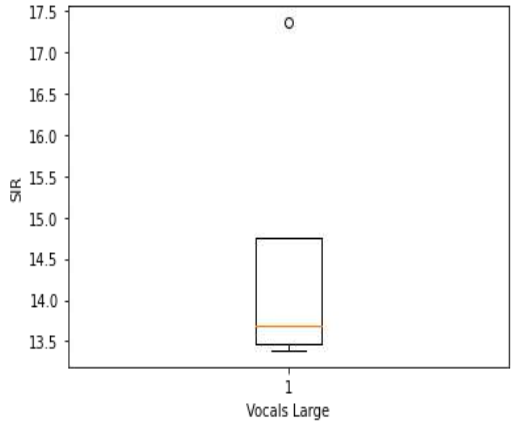
MMDenseLSTM

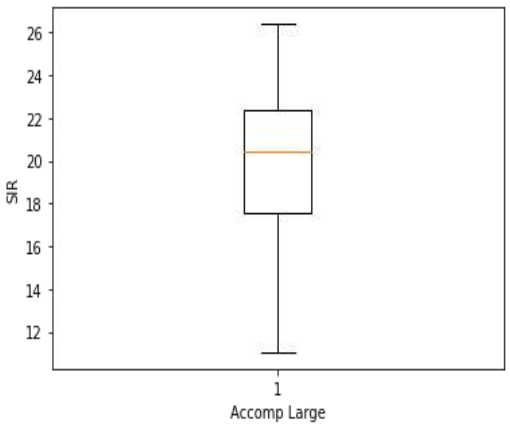


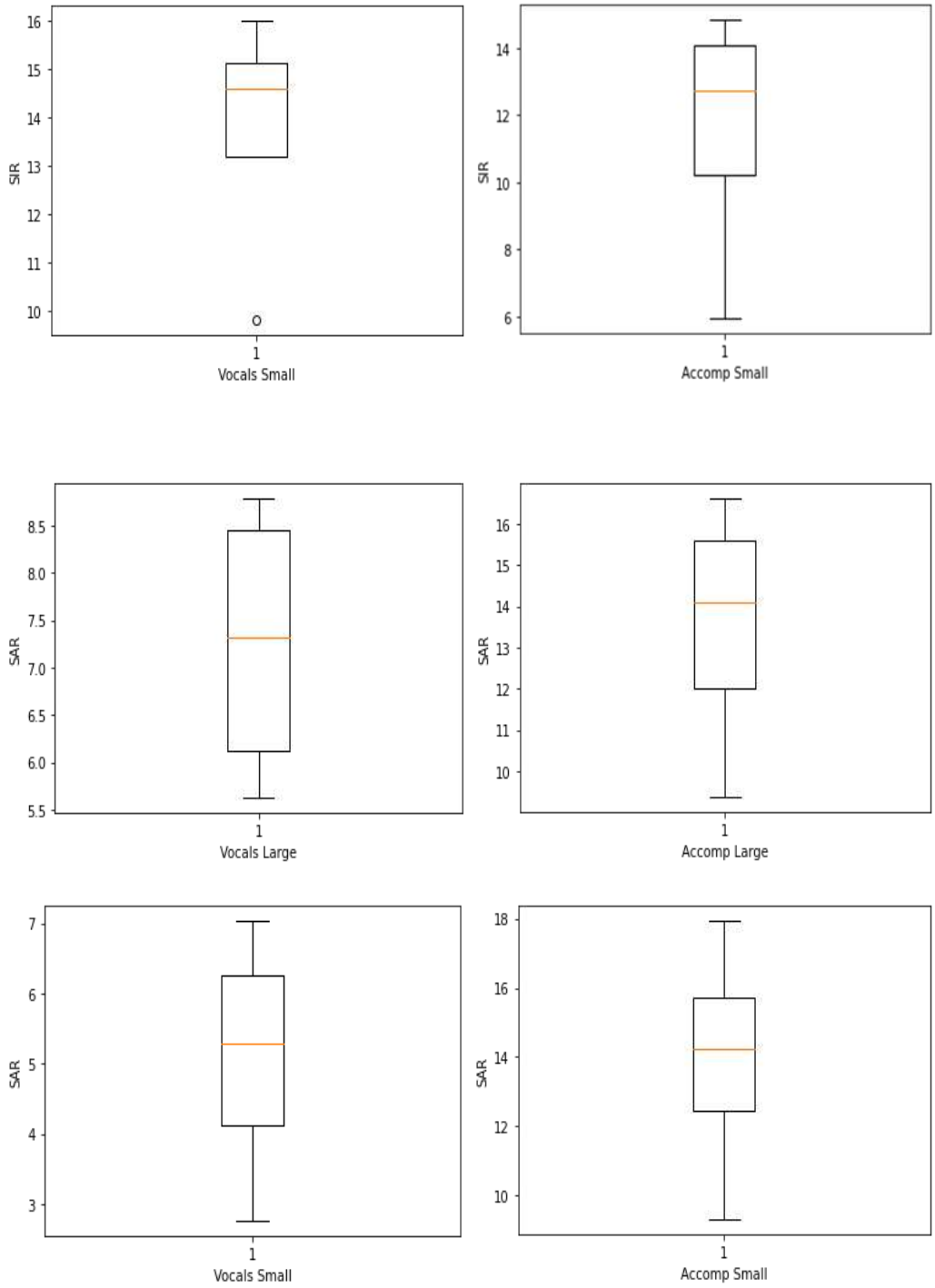












Εικ. 17: Αποτελέσματα απόδοσης του διαχωρισμού για τις αρχιτεκτονικές MMDenseLSTM.

Παρατηρούμε ότι στην Large αρχιτεκτονική, οι τιμές των μετρικών είναι υψηλότερες, γεγονός που συνεπάγεται καλύτερο διαχωρισμό (**Εικ. 17**). Τα outliers στα box plots του SIR δεν σηματοδοτούν κάποιο πρόβλημα, καθώς για μεγάλα σύνολα δεδομένων, η εμφάνιση ενός μικρού πλήθους outliers θεωρείται αναμενόμενη και δεν αποδίδεται σε κάποια ελαττωματική κατάσταση, όπως συμβαίνει στα μικρά σύνολα δεδομένων. Εξάλλου, εξ' ορισμού ένα outlier απεικονίζει ένα μεμο-

νωμένο σημείο του συνόλου δεδομένων το οποίο διαφέρει σημαντικά από τα υπόλοιπα. Γι' αυτόν το λόγο, επιλέγουμε ως αναπαράσταση για τις μετρικές μας τα box plots, των οποίων ο μέσος (median) υπολογίζεται ανεξάρτητα από τις ακραίες τιμές του συνόλου δεδομένων, και συνεπώς αντικατοπτρίζει καλύτερη τη γενική συμπεριφορά του. Για τις τιμές των SDR και SAR ισχύουν τα ίδια συμπεράσματα που αναφέρονται για το μοντέλο MMDenseNet.

5.3.3 ΣΥΓΚΡΙΣΗ ΤΩΝ ΔΥΟ ΜΟΝΤΕΛΩΝ

Για να μπορέσουμε να συγκρίνουμε ορθά τα μοντέλα MMDenseNet και MMDenseLSTM, ακολουθήσαμε για τις αρχιτεκτονικές τους, τις κατευθύνσεις των [77] και [78]. Οι δύο αυτές αναφορές αποτελούν ουσιαστικά συνέχεια η μία της άλλης, καθώς έχουν τους ίδιους συγγραφείς. Μάλιστα, στο [78] γίνεται σύγκριση του MMDenseLSTM με μερικά ακόμα μοντέλα νευρωνικών δικτύων, ένα εκ των οποίων είναι το MMDenseNet. Επομένως, παρακάτω συγκρίνουμε μεταξύ τους τα δύο μοντέλα για καθεμία από τις δύο αρχιτεκτονικές τους. Για τη σύγκριση, υπολογίστηκε, σε κάθε περίπτωση, ο μέσος (median) των μετρικών BSSEval σε dB (deciBel).

Πίνακας 1: *Σύγκριση των μετρικών BSSEval για την απομόνωση των φωνητικών στην Αρχιτεκτονική_1, με βάση το σύνολο δεδομένων musdb18.*

Αρχιτεκτονική_1 (Large) – Vocals			
Μοντέλο	SDR (dB)	SIR (dB)	SAR(dB)
MMDenseNet	4.207	12.854	4.112
MMDenseLSTM	7.666	13.682	7.312

Πίνακας 2: *Σύγκριση των μετρικών BSSEval για την απομόνωση του σήματος συνοδείας στην Αρχιτεκτονική_1, με βάση το σύνολο δεδομένων musdb18.*

Αρχιτεκτονική_1 (Large) – Accompaniment			
Μοντέλο	SDR (dB)	SIR (dB)	SAR (dB)
MMDenseNet	10.296	10.848	13.678
MMDenseLSTM	14.009	20.411	14.088

Πίνακας 3: *Σύγκριση των μετρικών BSSEval για την απομόνωση των φωνητικών στην Αρχιτεκτονική_2, με βάση το σύνολο δεδομένων musdb18.*

Αρχιτεκτονική_2 (Small) – Vocals			
Μοντέλο	SDR (dB)	SIR (dB)	SAR(dB)
MMDenseNet	4.961	11.433	4.675
MMDenseLSTM	5.265	14.596	5.283

Πίνακας 4: *Σύγκριση των μετρικών BSSEval για την απομόνωση του σήματος συνοδείας στην Αρχιτεκτονική_2, με βάση το σύνολο δεδομένων musdb18.*

Αρχιτεκτονική_2 (Small) – Accompaniment			
Μοντέλο	SDR (dB)	SIR (dB)	SAR(dB)
MMDenseNet	11.116	12.182	12.708
MMDenseLSTM	11.793	12.758	14.239

Στους πίνακες [**Πίνακας 1**- **Πίνακας 4**] παρατίθενται τα αποτελέσματα του διαχωρισμού για τις αρχιτεκτονικές των δύο μοντέλων με βάση τις μετρικές BSSEval. Παρατηρούμε ότι το μοντέλο MMDenseLSTM παρουσιάζει καλύτερη απόδοση από το MMDenseNet, σε όλες τις μετρικές, και για τις δύο αρχιτεκτονικές του. Δεδομένου ότι η μετρική SDR καταγράφει τη συνολική ποιότητα διαχωρισμού και εξαρτάται εξ' ορισμού, τόσο τα σφάλματα από παρεμβάλλουσες πηγές όσο και τα σφάλματα λόγω τεχνητού θορύβου του αλγόριθμου διαχωρισμού, η απόδοση διαχωρισμού συγκρίνεται συνήθως με βάση το SDR [83]. Συνεπώς, ο **Πίνακας 5** παρουσιάζει μια βελτίωση του SDR περίπου 3.5 dB για τις δύο πηγές (vocals, accompaniment) όσον αφορά την Αρχιτεκτονική_1 και μια μικρότερη βελτίωση της τάξης των 0.5 dB όσον αφορά την Αρχιτεκτονική_2. Επομένως, το μοντέλο MMDenseLSTM, αποδίδει καλύτερα από το MMDenseNet.

Πίνακας 5: *Βελτίωση του SDR (dB) για τις δύο αρχιτεκτονικές.*

Βελτίωση SDR (dB)			
	Vocals	Accompaniment	M.O.
Αρχιτεκτονική_1 (Large)	3.459	3.713	3.586
Αρχιτεκτονική_2 (Small)	0.304	0.677	0.491

6 ΕΠΙΛΟΓΟΣ

Το θέμα της παρούσας εργασίας ήταν ο διαχωρισμός μουσικών πηγών με τεχνικές μηχανικής μάθησης. Στα πλαίσια αυτά λοιπόν, έγινε αρχικά η παρουσίαση του αναγραφόμενου προβλήματος και η ανάλυσή του. Στη συνέχεια, ακολούθησε μια αναδρομή μεθόδων που ασχολήθηκαν εκτενώς με το πρόβλημα αυτό στο παρελθόν. Έγινε εμβάθυνση στις μεθόδους νευρωνικών δικτύων και μελετήθηκαν ορισμένα σύγχρονα μοντέλα, τα πυκνά συνελικτικά δίκτυα DenseNets. Μέσω της πειραματικής μελέτης, αναπτύχθηκαν στο σύνολο τέσσερις αρχιτεκτονικές, δύο για το μοντέλο MMDenseNet, και δύο για το μοντέλο MMDenseLSTM. Οι αρχιτεκτονικές αυτές εκπαιδεύτηκαν πάνω στο σύνολο δεδομένων musdb18 και τα αποτελέσματά τους συγκρίθηκαν και αξιολογήθηκαν χρησιμοποιώντας τις μετρικές αξιολόγησης BSSEval. Παρατηρήθηκε, ότι το μοντέλο MMDenseNet παρουσιάζει πολύ καλή απόδοση, χρησιμοποιώντας μάλιστα πολύ μικρό πλήθος παραμέτρων ($\sim 0.34 \times 10^6$). Παρόλα αυτά όμως, το μοντέλο MMDenseLSTM οδηγεί σε αισθητά καλύτερη απόδοση, ενώ ταυτόχρονα η εισαγωγή των μπλοκ LSTM στις αρχιτεκτονικές του, συνεπάγεται μια μικρή μόνο αύξηση παραμέτρων ($\sim 1.4 \times 10^6$). Επομένως, καταφέρνει και αξιοποιεί σε μία ενιαία αρχιτεκτονική, τόσο τα πλεονεκτήματα των δικτύων MMDenseNet, όσο και τα πλεονεκτήματα των δικτύων BLSTM.

6.1 ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

Η συνεχής εξέλιξη της τεχνολογίας και το αυξανόμενο ενδιαφέρον της επιστημονικής κοινότητας συνεισφέρουν στην αξιοποίηση των μεθόδων διαχωρισμού μουσικών πηγών που μελετήθηκαν στην παρούσα εργασία για μελλοντική έρευνα και χρήση σε ποικίλες εφαρμογές.

Για παράδειγμα, ο διαχωρισμός φωνητικών από τη συνοδεία μπορεί να αποτελέσει την βάση σε συστήματα αναγνώρισης μουσικής και ανίχνευσης μουσικών τμημάτων. Τέτοια συστήματα χρησιμοποιούνται από μια πληθώρα εφαρμογών, οπώς για παράδειγμα, για την εύρεση του τίτλου ή της χρονικής διάρκειας ενός μουσικού κομματιού καθώς κατά τη διαδικασία της μίξης δίνεται μεγαλύτερη ένταση στα φωνητικά από την συνοδεία. Επιπλέον, μπορούν να χρησιμοποιηθούν για την επίλυση ζητημάτων σχετικά με τα πνευματικά δικαιώματα [84] και για την αυτόματη αφαίρεση των φωνητικών ενός μουσικού κομματιού, από εφαρμογές τύπου «καραόκε» [85]. Ακόμη, η ανάπτυξη μεθόδων MSS παρέχει τη δυνατότητα χρήσης τους σε εφαρμογές αναζήτησης πολυμέσων (π.χ. βίντεο) με βάση το ηχητικό περιεχόμενό τους και στη δημιουργία εφαρμογών για κινητές συσκευές, όπως για παράδειγμα εφαρμογές κουρδίσματος μουσικών οργάνων.

Μία μελλοντική κατεύθυνση η οποία απαιτεί ακόμα σημαντικό όγκο έρευνας είναι η χρήση των προαναφερθεισών μεθόδων ως στάδιο προεπεξεργασίας για ένα αυτόματο σύστημα μίξης, καθώς είναι απαραίτητη η αξιόπιστη αναγνώριση των πραγματικών μουσικών οργάνων που συμμετέχουν σε μία ηχογράφηση,

ώστε στη συνέχεια να οδηγηθούν σε περαιτέρω επεξεργασία. Όσο περισσότερα μουσικά όργανα συμμετέχουν στην ηχογράφηση, τόσο δυσκολότερη είναι η αναγνώρισή τους. Επιπρόσθετα, η χρήση παρόμοιων οργάνων (πχ ακουστικό και ηλεκτρικό μπάσο) δυσχεραίνει ακόμα περισσότερο τη διαδικασία [86].

Τέλος, πολύ πρόσφατα, προτάθηκε η χρήση μοντέλων DenseNet για την ανίχνευση πτηνών υπό εξαφάνιση και την ταξινόμησή τους ανά είδος, με βάση το κελάηδισμά τους [87].

ΑΝΑΦΟΡΕΣ

[1] E. G. a. J. U. M. Schedl, «Music Information Retrieval: Recent Developments and Applications,» *Foundations and Trends in Information Retrieval Vol. 8, No. 2-3 ,* 2014.

[2] J. S. Downie., «Music Information Retrieval.,» *Annual Review of Information Science and Technology,* 2003.

[3] M. B. a. D. C. Kerstin Neubarth, «Associations between musicology and music information retrieval.,» σε *ISMIR*, 2011.

[4] F. R. M. Kaminskas, «Contextual music information retrieval and recommendation: State of the art and challenges,» *Computer Science Review,* 2012.

[5] C. e. al., «Content-Based Music In formation Retrieval: Current Directions and Future Challenges,» σε *Proceedings of the IEEE*, 2008.

[6] A. Wang, «The Shazam music recognition service.,» *Com. ACM,* p. 44–48, 2006.

[7] E. C. J. A. S. G. C Dittmar, «Music information retrieval meets music education,» *Dagstuhl Follow-Ups,* 2012.

[8] S. B. C. T. Anne-Marie Burns, «Learning to play the guitar at the age of interactive and collaborative Web technologies,» σε *Proceedings of the 14th Sound and Music Computing Conference*, Espoo, 2017.

[9] M. M. a. J. S. Peter Grosche, «Audio Content-Based Music Retrieval.,» *Dagstuhl Follow-Ups,* τόμ. 3, αρ. Multimodal Music Processing, 2012.

[10] D. F. A. L. M. D. P. a. F.-R. S. Estefanía Cano, «Musical Source Separation,» *IEEE Signal Processing Magazine,* January 2019.

[11] T. K. M. G. K. K. T. O. a. H. G. O. Hiromasa Fujihara, «Singer Identification Based on Accompaniment Sound Reduction and Reliable Frame Selection.,» σε *ISMIR*, London, 2005.

[12] F.-R. S. Z. R. D. K. B. R. e. a. Antoine Liutkus, «The 2016 Signal Separation Evaluation Campaign.,» σε *LVA/ICA 2017 - 13th International Conference on Latent Variable Analysis and Signal Separation*, Grenoble, 2017.

[13] A. L. F.-R. S. S. I. M. D. F. a. B. P. Zafar Rafii, «An Overview of Lead and Accompaniment Separation in Music,» *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING,* τόμ. 26, August 2018.

[14] E. I. D. a. J. J. Sejdić, «Time–frequency feature representation using energy concentration: An overview of recent advances.,» *Digital signal processing,* τόμ. 19, αρ. 1, pp. 153-183, 2009.

[15] R. J. M. a. T. F. Quatieri, «Speech analysis/synthesis based on a sinusoidal

representation,» *IEEE Trans. Acoust., Speech, Signal Process.,* Τόμ. %1 από %2ASSP-34, αρ. 4, p. 744–754, August 1986.

[16] R. L. Allen JB, «A unified approach to short-time Fourier analysis and synthesis.,» σε *Proceedings of the IEEE.*, 1977.

[17] N. Kehtarnavaz, «Short-Time Fourier Transform (STFT),» σε *Digital Signal Processing System Design (Second Edition)*, 2008.

[18] E. V. a. R. G. N. Q. K. Duong, «Under-determined reverberant audio source separation using a full-rank spatial covariance model,» *IEEE Trans. Audio Speech Language Processing,* p. 1830–1840, September 2010.

[19] J. K. a. E. O. A. Hyvärinen, «Independent Component Analysis.,» *John Willey ans Sons,* 2001.

[20] A. a. E. O. Hyvärinen, «Independent component analysis: algorithms and applications.,» *Neural networks 13,* αρ. 4-5, pp. 411-430, 1 June 2000.

[21] S. A. A. a. R. P. Sofianos, «Singing voice separation based on non-vocal independent component subtraction and amplitude discrimination.,» σε *In: Proceedings of the 13th International Conference on Digital Audio Effects*, 2010.

[22] H. S. T. W. E. D. S. Makino, «The DUET blind source separation algorithm,» *Blind Speech Separation, Signals and Communication Technology,* p. 217–241, 2007.

[23] B. L. a. E. C. D. Barry, «Real-time sound source separation using azimuth discrimination and resynthesis,» σε *Proc. 117th Audio Engineering Society (AES) Conv.*, San Francisco, 2004.

[24] A. L. a. R. B. D. FitzGerald, «Projection-based demixing of spatial audio,» *IEEE/ACM Trans. Audio Speech, Language Processing,* τόμ. 24, p. 1560–1572, September 2016.

[25] X. Serra, «Musical sound modeling with sinusoids plus noise,» *Studies on New Music Research,* p. 91–122, 1997.

[26] R. C. Maher, «An approach for the separation of voices in composite musical signals,» Champaign, IL, 1989.

[27] Y. M. a. K. Hirose, «Separation of singing and piano sounds,» σε *Proc. 5th Int. Conf. Spoken Lang. Process.*, Sydney, 1998.

[28] M. G. T. K. a. H. G. O. H. Fujihara, «A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval,» *IEEE Trans. Audio, Speech, Lang. Process.,* τόμ. 18, p. 638–648, March 2010.

[29] D. a. H. S. Lee, «Learning the parts of objects by non-negative matrix factorization,» *Nature,* pp. 788-791, 1999.

[30] P. Paatero, «Least squares formulation of robust nonnegative factor analysis,»

Chemometrics and Intelligent Laboratory Systems, pp. 23-35, 1997.

[31] D. D. L. a. H. S. Seung, «Algorithms for non-negative matrix factorization,» *Advances Neural Inform. Processing Syst.,* τόμ. 13, p. 556–562, April 2001.

[32] P. S. a. J. C. Brown, «Non-negative matrix factorization for polyphonic music transcription,» σε *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, 2003.

[33] N. B. J.-L. D. Cédric Févotte, «Nonnegative Matrix Factorization with the Itakura-Saito Divergence:With Application toMusic Analysis,» *Neural Computation,* p. 793–830, 01 March 2009.

[34] «BS.1534 : Method for the subjective assessment of intermediate quality levels of coding systems,» 2015.

[35] B. P. G. J. M. a. M. H. M. Cartwright, «Fast and easy crowdsourced perceptual audio evaluation,» σε *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, 2016.

[36] R. G. a. E. V. C. Fevotte, «BSS EVAL Toolbox User Guide Revision 2.0,» *IRISA,* 2005.

[37] R. G. a. C. F. E. Vincent, «Performance measurement in blind audio source separation,» *IEEE Trans. Audio, Speech, Lang. Process.,* τόμ. 14, p. 1462–1469, July 2006.

[38] A. S. B. P. a. A. Z. B. Fox, «Modeling perceptual similarity of audio signals for blind source separation evaluation,» σε *Proc. 7th Int. Conf. Latent Var. Anal. Signal Separation*, London, 2007.

[39] B. F. a. B. Pardo, «Towards a model of perceived quality of blind audio source separation,» σε *Proc. IEEE Int. Conf. Multimedia Expo.*, Beijing, 2007.

[40] E. M. a. A. L. U. Gupta, «On the perceptual relevance of objective source separationmeasures for singing voice separation,» σε *Proc. IEEEWorkshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, 2005.

[41] W. S. K. M. G Montavon, «Methods for interpreting and understanding deep neural networks,» *Digital Signal Processing,* 2018.

[42] Y. B. a. G. H. Yann LeCun, «Deep learning.,» *nature,* τόμ. 521, p. 436--444, 28 May 2015.

[43] Ε. Σ.ΛΥΚΟΘΑΝΑΣΗΣ, ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ - ΠΑΝΕΠΙΣΤΗΜΙΑΚΕΣ ΠΑΡΑΔΟΣΕΙΣ, ΠΑΤΡΑ, 2014.

[44] D. J. MacKay, «Information Theory, Inference and Learning Algorithms,» Cambridge University Press 2003, 2005.

[45] R. S. M. A. M. R. J. D. a. H. S. S. Richard HR Hahnloser, «Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit.,» *Nature,* τόμ. 405, 2000.

[46] A. F. Agarap., «Deep learning using rectified linear units (relu).,» arXiv preprint arXiv:1803.08375, 2018.

[47] J. Brownlee, «Loss and Loss Functions for Training Deep Learning Neural Networks,» Machine Learning Mastery, 2018.

[48] Y. B. A. C. Ian Goodfellow, Deep Learning, Cambridge, MA: MIT Press, 2016.

[49] J. Brownlee, «A Gentle Introduction to Cross-Entropy for Machine Learning,» 21 October 2019. [Ηλεκτρονικό]. Available: https://machinelearningmastery.com.

[50] S. Kostadinov, «Understanding Backpropagation Algorithm,» 8 August 2019. [Ηλεκτρονικό]. Available: https://towardsdatascience.com/understanding-backpropagation-algorithm-7bb3aa2f95fd.

[51] K. K. e. a. Sergios Theodoridis, Pattern recognition., τόμ. 19, IEEE Transactions on Neural Networks, 2008, p. 376.

[52] A. Roy, «An Introduction to Gradient Descent and Backpropagation,» 14 June 2020. [Ηλεκτρονικό]. Available: https://towardsdatascience.com/an-introduction-to-gradient-descent-and-backpropagation-81648bdb19b2.

[53] L. Bottou, «Large-Scale Machine Learning with Stochastic Gradient Descent,» σε *Proceedings of COMPSTAT'2010*, Berlin Heidelberg, 2010.

[54] N. V. Z. Y. V. F. A. G. K. R. M. W. M. J. G. Noah Golmant, «On The Computational Inefficiency Of Large Batch Sizes For Stochastic Gradient Descent,» 30 November 2018.

[55] Y. Bengio, «Practical recommendations for gradient-based training of deep architectures,» 2012.

[56] J. H. E. &. S. Y. Duchi, «Adaptive Subgradient Methods for Online Learning and Stochastic Optimization.,» *Journal of Machine Learning Research (JMLR),* 2011.

[57] T. &. H. G. Tieleman, «Lecture 6.5 - rmsprop,» COURSERA, 2012.

[58] M. D. Zeiler, «ADADELTA: An Adaptive Learning Rate Method.,» *ArXiv Preprint,* 2012.

[59] D. &. B. J. Kingma, «Adam: A Method for Stochastic Optimization.,» σε *Proceedings of International Conference on Learning Representations (ICLR).*, 2015.

[60] T. Takase, S. Oyama και M. Kurihara, «Effective neural network training with adaptive learning rate based on training loss,» *NEURAL NETWORKS,* τόμ. 101, pp. 68-78, 05 2018.

[61] N. S. K. S. G Hinton, «Neural networks for machine learning lecture 6a overview of mini-batch gradient descent,» 2012.

[62] S. K. T Kurbiel, «Training of deep neural networks based on distance measures

using RMSProp,» 2017.

[63] W. Z. K. J. M. L. S. A. S. B. L. T. W. X. W. G. C. J. C. T. Gu J, «Recent advances in convolutional neural networks.,» *Pattern Recognition,* αρ. 77, pp. 354-377, 1 May 2018.

[64] N. S. A. K. I. S. a. R. R. G.E. Hinton, « Improving neural networks by preventing co-adaptation of feature detectors,» 2012.

[65] N. R. O'Shea K, «An introduction to convolutional neural networks.,» *arXiv preprint,* 26 November 2015.

[66] W. D. C. A. N. A. Wang T, «End-to-end text recognition with convolutional neural networks.,» σε *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*, 2012.

[67] C. S. T. L. C. R. Andreas Maier, «A gentle introduction to deep learning in medical image processing,» *Zeitschrift für Medizinische Physik,* τόμ. 29, αρ. 2, pp. 86-101, May 2019.

[68] P. K. Schuster M, «Bidirectional recurrent neural networks.,» *IEEE transactions on Signal Processing.,* τόμ. 45, αρ. 11, pp. 2673-2681, November 1997.

[69] M. a. B. S. Hermans, «Training and analysing deep recurrent neural networks.,» *Advances in neural information processing systems,* τόμ. 26, pp. 190-198, 2013.

[70] Y. S. P. &. F. P. Bengio, «Learning long-term dependencies with gradient descent is difficult.,» *IEEE Trans. Neural Networks,* τόμ. 5, p. 157–166, 1994.

[71] A. a. J. S. Graves, «Framewise phoneme classification with bidirectional LSTM and other neural network architectures.,» *Neural networks,* τόμ. 18, pp. 602-610, 2005.

[72] F. A. S. N. N. &. S. J. Gers, «Learning precise timing with LSTM recurrent networks.,» *Journal of machine learning research,* pp. 115-143, August 2002.

[73] D. K. W. H. &. W. X. Kulevome, « A Bidirectional LSTM-Based Prognostication of Electrolytic Capacitor.,» *Progress In Electromagnetics Research C.,* pp. 139-152, 1 April 2021.

[74] S. Z. D. H. X. a. Y. M. Zhang, «Bidirectional long short-term memory networks for relation classification,» σε *Proceedings of th29th Pacific Asia conference on language, information and computation*, 2015.

[75] Z. L. L. V. D. M. a. K. Q. W. Gao Huang, «Densely connected convolutional networks.,» σε *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[76] S. I. a. C. Szegedy., «Batch normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,» σε *International conference on machine learning*, 2015.

[77] N. T. a. Y. Mitsufuji., «Multi-scale multi-band densenets for audio source separation.,» σε *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[78] G. N. M. Y. Takahashi N, «MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation.,» σε *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[79] T. Dietterich, «Overfitting and undercomputing in machine learning.,» *ACM computing surveys (CSUR) ,* τόμ. 27, αρ. 3, pp. 326-327, 1 September 1995.

[80] N. Srivastava, «Improving neural networks with dropout.,» 2013.

[81] S. M. P. F. G. M. E. T. K. N. T. a. Y. M. Uhlich, «Improving music source separation based on deep neural networks through data augmentation and network blending.,» σε *In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[82] L. A. F.-R. S. M. S. I. a. B. R. Rafii Zafar, «The MUSDB18 corpus for music separation,» Dec 2017. [Ηλεκτρονικό]. Available: https://doi.org/10.5281/zenodo.1117372.

[83] W.-H. H. K. a. O.-W. K. Heo, «Source separation using dilated time-frequency DenseNet for music identification in broadcast contents.,» *Applied Sciences,* τόμ. 10, αρ. 5, p. 1727, 2020.

[84] W.-H. H. K. a. O.-W. K. Heo, «Source separation using dilated time-frequency DenseNet for music identification in broadcast contents,» *Applied Sciences ,* τόμ. 10, αρ. 5, p. 1727, 2020.

[85] A. J. G. R. a. M. D. P. Simpson, «Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network.,» σε *International Conference on Latent Variable Analysis and Signal Separation*, Cham, 2015.

[86] D. a. B. K. .. Koszewski, «Musical instrument tagging using data augmentation and effective noisy data processing.,» *Journal of the Audio Engineering Society,* τόμ. 68, αρ. 1/2, pp. 57-65, 2020.

[87] C. L. T. Z. a. Y. L. H. Liu, «Bird Song Classification Based on Improved Bi-LSTM-DenseNet Network,» σε *4th International Conference on Robotics, Control and Automation Engineering (RCAE)*, 2021.

[88] S. A. a. A. R. V.-K. Amir Mosavi, «List of Deep Learning Models,» 2019.