1 Εισαγωγη

Το περιεχόμενο αυτής της ενότητας αφορά τα Βήματα 1-4 της εργασίας.

Το περιεχόμενο της δεύτερης εργαστηριακής άσκησης αποτέλεσε η υλοποίηση συστημάτων επεξεργασίας και αναγνώρισης φωνής, με εφαρμογή σε αναγνώριση μεμονωμένων λέξεων. Για το προπαρασκευαστικό μέρος της εργαστηριακής άσκησης, τα δεδομένα που χρησιμοποιήθηκαν αποτελούν εκφωνήσεις από 15 διαφορετικούς ομιλητές για καθένα εκ των ψηφίων 1-9, με εξαίρεση τα ψηφία 6 και 8, για τα οποία οι συνολικές εκφωνήσεις είναι 14. Η συχνότητα δειγματοληψίας των εκφωνήσεων είναι 16 kHz, ενώ η διάρκειά τους διαφέρει. Σε ό,τι αφορά το κύριο μέρος της εργαστηριακής άσκησης, χρησιμοποιήθηκαν τα δεδομένα του συνόλου Free Spoken Digit Dataset (FSDD), τα οποία αποτελούν 50 εκφωνήσεις καθενός εκ των ψηφίων 0-10 από 6 διαφορετικούς εκφωνητές (σύνολο 3000 εκφωνήσεων) με συχνότητα δειγματοληψίας 8 kHz.

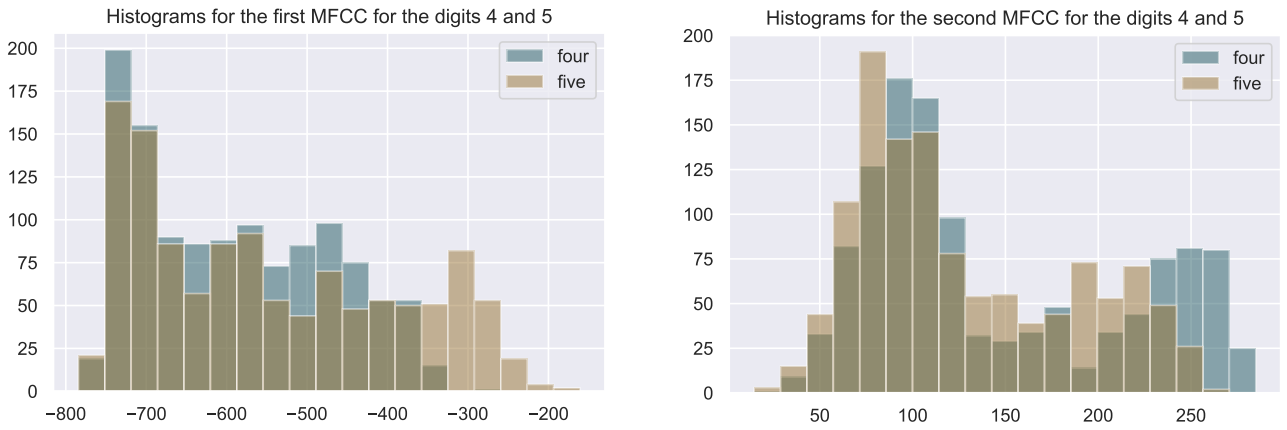
Πριν την ανάλυση των προαναφερθέντων βασικών δεδομένων, αξιοποιήθηκαν δύο διαφορετικά αρχεία, με σκοπό μια αρχική εξοικείωση με τις κυματομορφές που προκύπτουν σε δεδομένα ήχου, καθώς και την ανάλυσή τους. Τα αρχεία αυτά αντιστοιχούν σε δύο εκφωνήσεις των λέξεων “one two three”, μία από γυναίκα και μία από άνδρα, και η ανάλυσή τους (Βήμα 1) πραγματοποιήθηκε μέσω του λογισμικού Praat με στόχο την εξαγωγή της μέσης τιμής (mean) για το pitch στα φωνήενα «α», «ου» και «ι». Επιπροσθέτως, μέσω της επιλογής Formant -> Formant Listing, εξήχθησαν τα 3 πρώτα formants (F_i, $i = 1, 2, 3$) κάθε φωνήεντος. Τα αποτελέσματα παρουσιάζονται στον ακόλουθο πίνακα (Πίνακας 1.1), όπου οι δείκτες M και F δεικτοδοτούν εάν η εκφώνηση πραγματοποιήθηκε από τον άνδρα ή τη γυναίκα, αντίστοιχα.

		ου				α				ι			
		mean	F_1	F_2	F_3	mean	F_1	F_2	F_3	mean	F_1	F_2	F_3
1	M	133	530	933	2209	135	795	935	2302	-	-	-	-
	F	186	558	888	2352	178	945	1569	2998	-	-	-	-
2	M	129	304	1814	2356	-	-	-	-	-	-	-	-
	F	183	345	1661	2604	-	-	-	-	-	-	-	-
3	M	-	-	-	-	-	-	-	-	130	385	2021	2519
	F	-	-	-	-	-	-	-	-	173	327	1699	2713

Πίνακας 1.1: Μέσες τιμές του pitch και 3 πρώτα formants για τα φωνήεντα των ψηφίων 1,2,3.

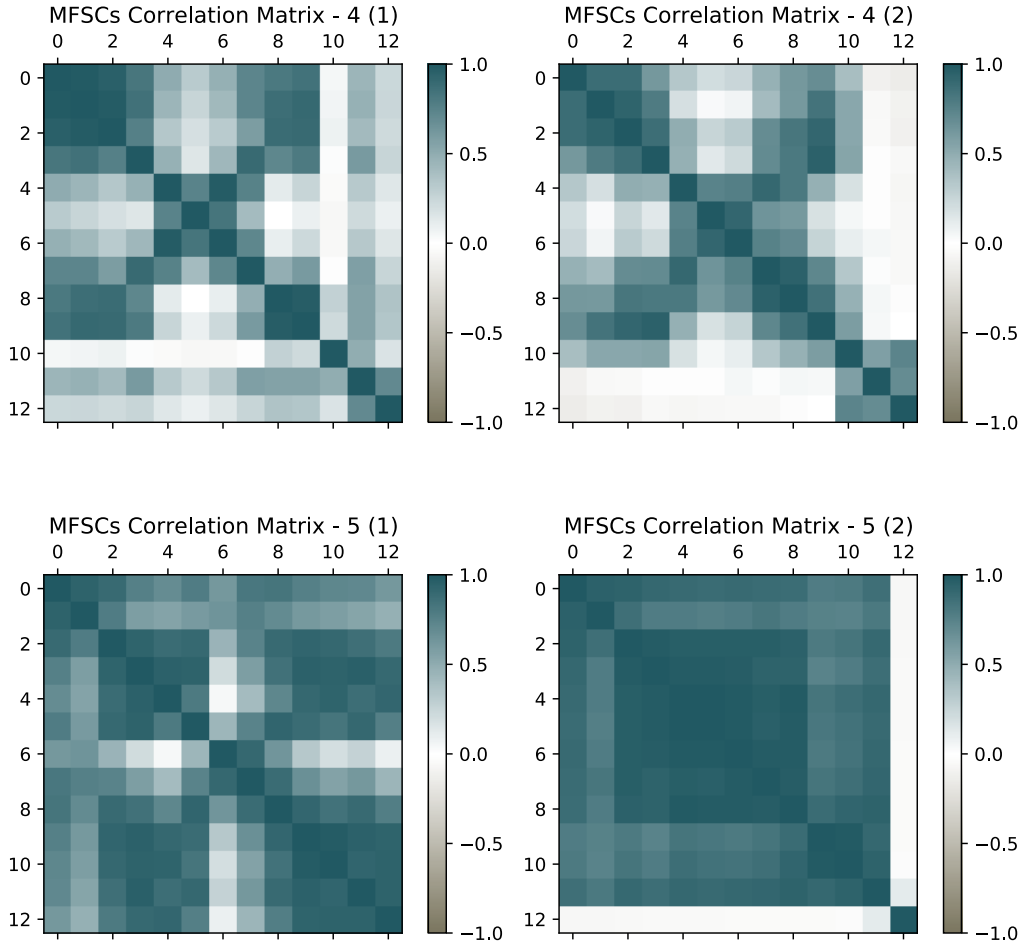
Γίνεται εμφανές πως, τόσο τα formants, όσο και το μέσο pitch, για τις εκφωνήσεις του άνδρα έχουν χαμηλότερες τιμές σε σχέση με τα αντίστοιχα της γυναίκας. Επιπλέον, ενώ το μέσο pitch είναι περίπου σταθερό για κάθε εκφωνητή στις εκφωνήσεις του «ου», τα formants F_1 και F_2 παρουσιάζουν σημαντικές αποκλίσεις.

Προχωρώντας στην ανάλυση των βασικών δεδομένων, το πρώτο βήμα (Βήμα 2) αποτέλεσε η κατασκευή μιας συνάρτησης με στόχο την ανάγνωση των δεδομένων ήχου μέσω της βιβλιοθήκης `librosa` και την εκτύπωση τριών λιστών που να περιέχουν το διακριτοποιημένο ηχητικό σήμα κάθε εκφώνησης, τον αντίστοιχο ομιλητή, καθώς και το εκφωνούμενο ψηφίο. Στη συνέχεια (Βήμα 3), για κάθε εκφώνηση εξήχθησαν 13 χαρακτηριστικά τύπου Mel-Frequency Cepstral Coefficients (MFCCs), χρησιμοποιώντας μήκος παραθύρου 25 ms και βήμα 10 ms, καθώς και οι πρώτες και δεύτερες τοπικές παράγωγοι των χαρακτηριστικών. Στα ιστογράμματα της Εικόνας 1.1 (Βήμα 4) παρουσιάζονται το 1^o (αριστερά) και το 2^o (δεξιά) MFCC, αντίστοιχα, για το σύνολο των εκφωνήσεων των ψηφίων 4 και 5. Όπως γίνεται εύκολα αντιληπτό, η επικάλυψη μεταξύ των ιστογραμμάτων που αντιστοιχούν σε ένα MFCC είναι σημαντική, παρότι τα ψηφία διαφέρουν μεταξύ τους.



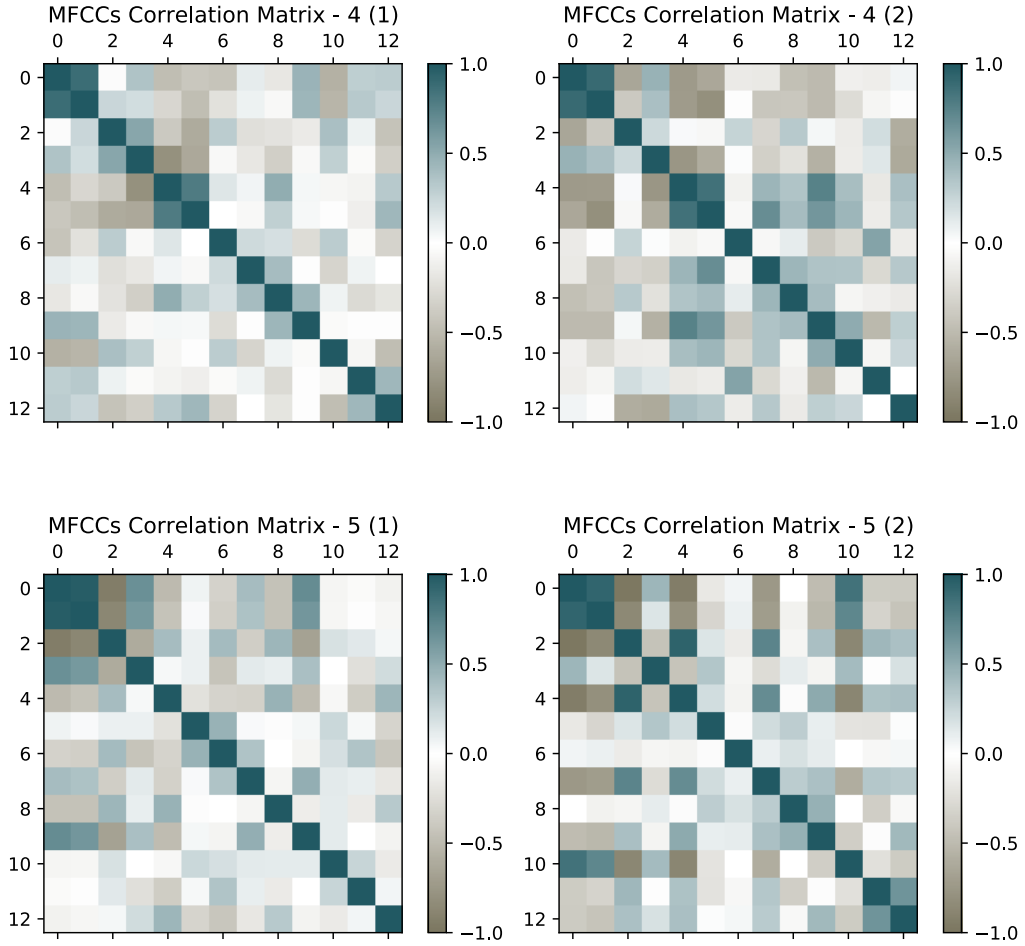
Εικόνα 1.1: Ιστογράμματα του 1ου (αριστερά) και του 2ου (δεξιά) MFCC των ψηφίων 4 και 5.

Ένα άλλο χαρακτηριστικό το οποίο μπορεί να χρησιμοποιηθεί για τη μελέτη αρχείων ήχου είναι τα Mel Filterbank Spectral Coefficients (MFSCs), τα οποία δεν είναι παρά τα MFCCs, χωρίς όμως τον τελικό διακριτό μετασχηματισμό συνημιτόνου (DCT). Επιλέγοντας δύο τυχαίες εκφωνήσεις των ψηφίων 4 και 5, υπολογίστηκαν οι πίνακες συσχέτισης για 13 MFSCs ανά εκφώνηση, οι οποίοι απεικονίζονται στην Εικόνα 1.2 υπό τη μορφή heatmaps.



Εικόνα 1.2: Πίνακες συσχέτισης για τα MFSCs δύο εκφωνήσεων των ψηφίων 4 και 5.

Τα MFSCs εμφανίζονται υψηλά συσχετισμένα (βλ. χρωματική κλίμακα), έχοντας μάλιστα σχεδόν αποκλειστικά θετική συσχέτιση. Επαναλαμβάνοντας την ίδια διαδικασία για τα MFCCs των ίδιων εκφωνήσεων, προκύπτουν οι πίνακες συσχέτισης που απεικονίζονται στην Εικόνα 1.3.



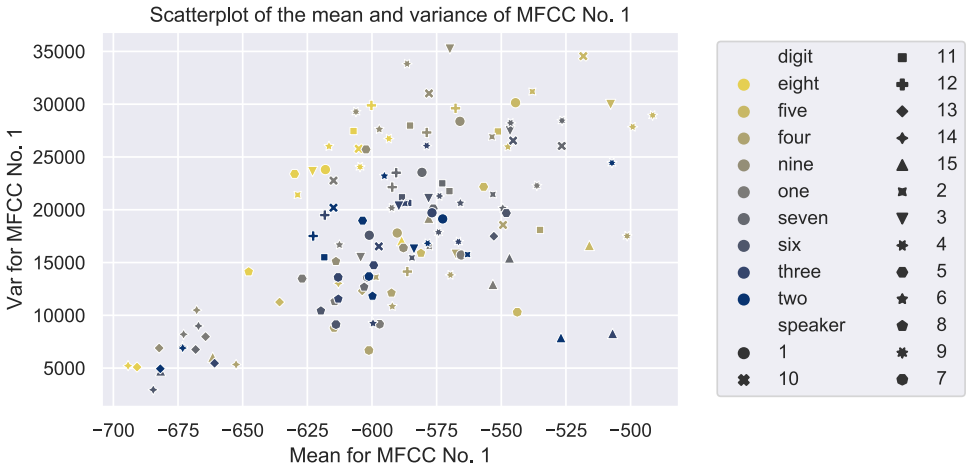
Εικόνα 1.3: Πίνακες συσχέτισης για τα MFCCs δύο εκφωνήσεων των ψηφίων 4 και 5.

Στην περίπτωση των MFCCs, με εξαίρεση τα στοιχεία της διαγωνίου - όπως εξάλλου είναι αναμενόμενο - η συσχέτιση δεν είναι τόσο έντονη, με τις ακραίες τιμές 1 και -1 να προσεγγίζονται από λίγα ζεύγη τιμών σε ολόκληρα τα heatmaps. Γίνεται, επομένως, ξεκάθαρος ο λόγος για τον οποίο τα MFCCs υπερτερούν των MFSCs: γενικά, τα χαρακτηριστικά που αξιοποιούνται για την εκπαίδευση μοντέλων πρέπει να έχουν όσο το δυνατό λιγότερη κοινή πληροφορία, ειδάλλως ορισμένα εξ αυτών ενδέχεται να περιττεύουν για την περιγραφή του εκάστοτε μοντέλου. Ένα μέτρο της κοινής αυτής πληροφορίας είναι η συσχέτιση που παρουσιάζουν μεταξύ τους τα επί μέρους χαρακτηριστικά. Έτσι, αφού η υψηλή συσχέτιση συνεπάγεται περισσότερη κοινή πληροφορία, η χαμηλή συσχέτιση που παρουσιάζουν μεταξύ τους τα διαφορετικά MFCCs τα καθιστά καταλληλότερα ως χαρακτηριστικά για την ανάλυση σημάτων ήχου. Αξίζει να αναφερθεί στο σημείο αυτό πως ένα ακόμη προτέρημα των MFCCs έναντι των MFSCs είναι πως είναι πιο συμπιέσιμα, με αποτέλεσμα 13 μόνο MFCCs να αρκούν ως πλήθος χαρακτηριστικών για την εκπαίδευση μοντέλων, τη στιγμή που το αντίστοιχο πλήθος των MFSCs που θα έπρεπε να εξαχθούν από τα δεδομένα ήχου για μια αντίστοιχη ανάλυση θα ήταν υψηλότερο.

2 Προκαταρκτικη Ταξινομηση Ψηφιων

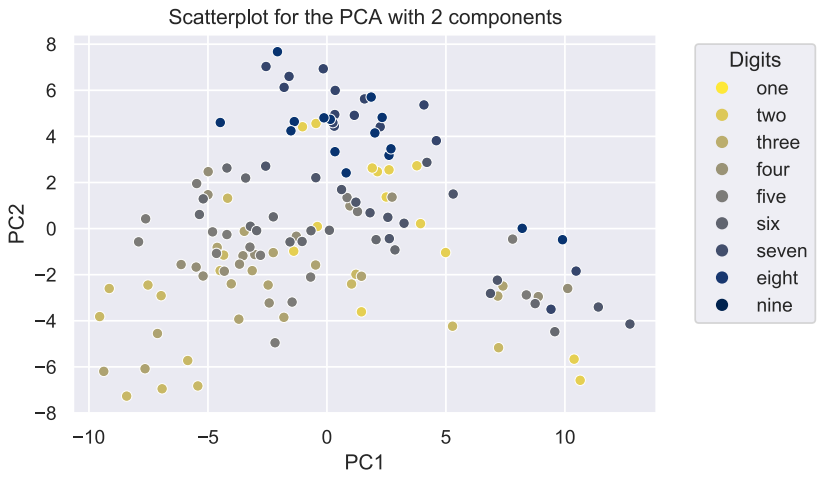
Το περιεχόμενο αυτής της ενότητας αφορά τα Βήματα 5-8 της εργασίας.

Χρησιμοποιώντας ως χαρακτηριστικά τα 13 MFCCs ανά εκφώνηση, καθώς και τις πρώτες και δεύτερες τοπικές παραγώγους, δημιουργήθηκαν για κάθε εκφώνηση 78-διάστατα διανύσματα αποτελούμενα από τις μέσες τιμές και τις διακυμάνσεις όλων των παραθύρων των 39 αυτών χαρακτηριστικών (Βήμα 5), προκειμένου να πραγματοποιηθούν ορισμένες προκαταρκτικές ταξινομήσεις.

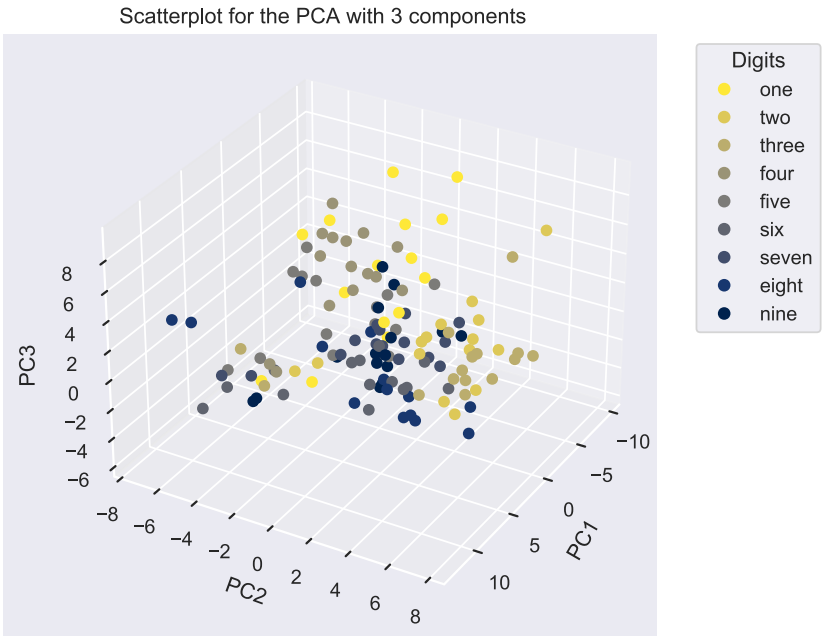


Εικόνα 2.1: Διάγραμμα διασποράς των πρώτων δύο διαστάσεων του διανύσματος χαρακτηριστικών κάθε εκφώνησης.

Στο διάγραμμα διασποράς της Εικόνας 2.1 απεικονίζονται οι πρώτες δύο διαστάσεις των διανυσμάτων αυτών. Παρότι παρατηρείται σχηματισμός συστάδων (clusters) μεταξύ σημείων που αντιστοιχούν σε κοινά ψηφία, είναι εμφανές πως ο βαθμός επικάλυψης είναι υψηλός. Προκειμένου το σύνολο των χαρακτηριστικών να μπορεί να απεικονιστεί σε ένα διάγραμμα, αξιοποιήθηκε η μέθοδος Principal Component Analysis (PCA - Βήμα 6) με σκοπό τη διαστατική μείωση του διανύσματος χαρακτηριστικών σε 2 (Εικόνα 2.2) ή σε 3 (Εικόνα 2.3) κύριες συνιστώσες.



Εικόνα 2.2: Αναγωγή των συνολικών χαρακτηριστικών σε 2 κύριες συνιστώσες.



Εικόνα 2.3: Αναγωγή των συνολικών χαρακτηριστικών σε 3 κύριες συνιστώσες.

Ενώ ορισμένες συστάδες εξακολουθούν να σχηματίζονται, η επικάλυψη μεταξύ σημείων που αντιστοιχούν σε διαφορετικά ψηφία δεν παύει να υπάρχει. Επιπλέον, η αρχική διασπορά που διατηρείται από τις προκύπτουσες συνιστώσες σε δύο διαστάσεις είναι 46.58%, ενώ σε τρεις διαστάσεις 57.29%, γεγονός που φανερώνει πως η διαστατική μείωση σε τόσο λίγες διαστάσεις οδηγεί σε σημαντική απώλεια πληροφορίας. Βάσει αυτών, θα έλεγε κανείς πως η μείωση διαστάσεων δεν είναι ιδιαίτερα επιτυχημένη.

Κρατώντας, λοιπόν, τα αρχικά διανύσματα χαρακτηριστικών ως δεδομένα για κάθε εκφώνηση, αρχικά αυτά χωρίστηκαν κατά 70%-30% σε σύνολα εκπαίδευσης - αξιολόγησης, αντίστοιχα (Βήμα 7). Στη συνέχεια, μέσω εκπαίδευσης του `StandardScaler` της `sklearn` στα δεδομένα εκπαίδευσης, δημιουργήθηκαν δύο νέα σύνολα εκπαίδευσης-αξιολόγησης, αποτελούμενα από κανονικοποιημένα δεδομένα. Για λόγους πληρότητας, όλοι οι ταξινομητές εφαρμόστηκαν τόσο στα κανονικοποιημένα δεδομένα, όσο και στα αρχικά.

Αφού πρώτα ο ταξινομητής `CustomNBClassifier` της προηγούμενης εργαστηριακής άσκησης αναπροσαρμόστηκε ώστε να μπορεί να ταξινομεί μόνο 9 ψηφία, εκπαιδεύτηκε στο σύνολο των δεδομένων εκπαίδευσης και αξιολογήθηκε στα δεδομένα αξιολόγησης με ακρίβεια 72.5%. Την ίδια ακριβώς επιτυχία είχε και ο ταξινομητής `GaussianNB` της `sklearn`, αφού, όπως αναλύθηκε και στην προηγούμενη εργαστηριακή αναφορά, οι δύο ταξινομητές ταυτίζονται για τις default τιμές των παραμέτρων τους. Σημειώνεται στο σημείο αυτό πως στην προηγούμενη εργαστηριακή άσκηση είχε μελετηθεί επίσης μια διαφορετική υλοποίηση του Naive Bayes, βασισμένη στη Βήτα κατανομή, με απόδοση παραπλήσια του τυπικού (Gaussian) Naive Bayes. Στην προκείμενη περίπτωση, παρά το εκτεταμένο grid search, ο Beta Naive Bayes δεν παρουσίασε καλά αποτελέσματα.

Εκτός των παραπάνω ταξινομητών, μελετήθηκαν επίσης οι ταξινομητές πλησιέστερου γείτονα (kNN) καθώς και οι SVM για διάφορους πυρήνες. Όπου ήταν εφικτό, προηγήθηκε grid search για την εύρεση των υπερπαραμέτρων που βελτιστοποιούσαν το cv-score. Στον Πίνακα 2.1 συνοψίζονται οι αποδόσεις όλων των ταξινομητών που εφαρμόστηκαν στα δεδομένα, τόσο ως προς την ακρίβειά τους στο σύνολο αξιολόγησης, όσο και βάσει cross-validation. Παρατηρείται πως οι ταξινομητές SVM πέτυχαν τα καλύτερα σκορ, με μεγαλύτερο αυτό του SVM sigmoid (82.7% cv-score), ενώ η κανονικοποίηση των δεδομένων όντως διαδραμάτισε σημαντικό ρόλο

στην όλη διαδικασία, αυξάνοντας θεαματικά την απόδοση της πλειοψηφίας των ταξινομητών.

	Αρχικά δεδομένα		Κανονικοποιημένα δεδομένα	
Ταξινομητής	test score	cv-score	test score	cv-score
sklearn NB (Gaussian)	72.5%	61.2%	72.5%	60.2%
Custom NB (Gaussian)	72.5%	61.2%	72.5%	60.2%
Custom NB (Beta)	-	-	25.0%	27.0%
kNN	37.5%	37.4%	80.0%	67.8%
SVM linear	57.5%	55.8%	82.5%	80.6%
SVM rbf	10.0%	10.8%	82.5%	81.7%
SVM sigmoid	10.0%	10.8%	85.0%	82.7%
SVM polynomial	65.0%	48.0%	75.0%	68.6%

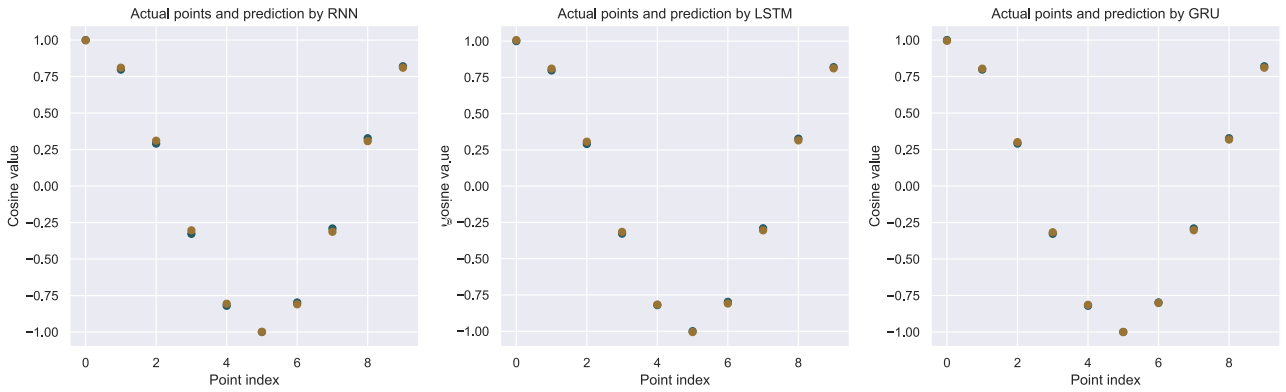
Πίνακας 2.1: Συγκεντρωτικός πίνακας απόδοσης ταξινομητών.

Το τελικό βήμα (Βήμα 8) του προπαρασκευαστικού μέρους αποτέλεσε η κατασκευή ενός Αναδρομικού Νευρωνικού Δικτύου (RNN), το οποίο δέχεται ως είσοδο ακολουθίες 10 σημείων ενός ημιτόνου συχνότητας 40 Hz και προβλέπει στην έξοδό του την αντίστοιχη ακολουθία 10 σημείων του συνημιτόνου. Για το σκοπό αυτό, αφού πρώτα δημιουργήθηκε η συνάρτηση που γεννά τις εν λόγω ακολουθίες, παρήχθησαν 10^4 δεδομένα, τα οποία χωρίστηκαν κατ' αναλογία 60%-25%-15% σε δεδομένα εκπαίδευσης, αξιολόγησης και επικύρωσης (validation) αντίστοιχα. Το να συμπεριληφθούν δεδομένα επικύρωσης κρίθηκε απαραίτητο αφού, όπως θα φανεί και στο τελικό βήμα της εργασίας, ένα βήμα επικύρωσης είναι απαραίτητο να ακολουθεί κάθε βήμα εκπαίδευσης, προκειμένου να φαίνεται εάν η πιο πρόσφατη ανανέωση των βαρών οδήγησε σε βελτίωση του μοντέλου ή όχι.

Εκτός από το απλό RNN υλοποιήθηκαν μονάδες LSTM και GRU, προκειμένου τα διάφορα αποτελέσματα να συγκριθούν μεταξύ τους. Οι συγκεκριμένες μονάδες είναι πιο διαδεδομένες, καθώς οδηγούν σε σημαντικές βελτιώσεις στο μεγαλύτερο πρόβλημα των RNNs, δηλαδή την εκθετική μείωση της παραγώγου της συνάρτησης κόστους, η οποία έχει ως αποτέλεσμα τη γρήγορη απώλεια μνήμης στο δίκτυο [Cal20, p. 554]. Στα LSTM το πρόβλημα αυτό αντιμετωπίζεται με την εισαγωγή ενός επιπλέον επιπέδου (memory cell), καθώς και πρόσθετες πύλες που ελέγχουν την είσοδο και την έξοδο της πληροφορίας από τη μνήμη, ενώ στα GRU η αρχιτεκτονική είναι παρόμοια με αυτή των LSTM, αλλά σημαντικά απλούστερη, καθιστώντας τα έτσι λιγότερο περίπλοκα, αλλά πιο εύκολα εκπαιδεύσιμα.

Τα δίκτυα εκπαιδεύτηκαν στο ίδιο σύνολο δεδομένων και για κοινό διαμοιρασμό τους σε δεδομένα εκπαίδευσης-αξιολόγησης-επικύρωσης. Η αρχιτεκτονική των δικτύων ήταν κοινή, δηλαδή 3 stacked layers κάθε τύπου και ένα hidden layer να διατηρεί πληροφορία για 64 χαρακτηριστικά. Η εισαγωγή των δεδομένων πραγματοποιούνταν σε batches των 64, με learning rate ίσο με 0.001 και με συνάρτηση κόστους την MSE, η οποία κρίθηκε κατάλληλη, αφού το πρόβλημα δεν αντιστοιχεί σε πρόβλημα ταξινόμησης, αλλά σε πρόβλημα προσέγγισης σημείων. Η εκπαίδευση κάθε δικτύου διήρκησε για 100 εποχές, στο τέλος των οποίων το validation loss ήταν $\mathcal{O}\left(10^{-5}\right)$ για το RNN, $\mathcal{O}\left(10^{-4}\right)$ για το LSTM και επίσης $\mathcal{O}\left(10^{-4}\right)$ για το GRU. Στην Εικόνα 2.4 απεικονίζονται με μπλε οι πραγματικές τιμές του συνημιτόνου από μια ακολουθία αξιολόγησης και με χρυσό οι τιμές του συνημιτόνου όπως εκτιμώνται από το αντίστοιχο δίκτυο, για την ίδια ακολουθία σημείων κάθε φορά.

Όπως γίνεται φανερό, και τα τρία δίκτυα αποδίδουν αρκετά καλά σε ό,τι αφορά την πρόβλεψη της ακολουθίας σημείων του συνημιτόνου, αφού τα μπλε και τα χρυσά σημεία επικαλύπτονται σχεδόν πλήρως. Η εικόνα αυτή είναι αντιπροσωπευτική της γενικής απόδοσης των δικτύων και δεν αφορά μόνο τη συγκεκριμένη (τυχαία επιλεγμένη) ακολουθία σημείων, αφού είναι αντίστοιχη για καθένα εκ των 2500 ακολουθιών αξιολόγησης.



Εικόνα 2.4: Πρόβλεψη των τιμών του συνημιτόνου μέσω δικτύων RNN (αριστερά), LSTM (κεντρικά) και GRU (δεξιά).

3 Κρυφα Μοντελα Markov με Gaussian Mixtures

Το περιεχόμενο αυτής της ενότητας αφορά τα Βήματα 9-13 της εργασίας.

Προχωρώντας στο κύριο μέρος της εργασίας, αρχικά εξήχθησαν για κάθε εκφώνηση του νέου συνόλου δεδομένων (FSDD) 13 MFCCs, με μήκος παραθύρου 30 ms και βήμα 15 ms. Στη συνέχεια (Βήμα 9), τα δεδομένα αυτά χωρίστηκαν σε δεδομένα εκπαίδευσης και αξιολόγησης με ποσοστό 80%-20%, αντίστοιχα. Κατόπιν, το δείγμα εκπαίδευσης χωρίστηκε περαιτέρω ώστε να δημιουργηθεί και ένα δείγμα επικύρωσης, το οποίο απαιτείται πάντα για τη ρύθμιση των εκάστοτε υπερπαραμέτρων ενός μοντέλου, οι οποίες δεν μπορούν να ρυθμιστούν με κριτήριο την απόδοση του μοντέλου στο σύνολο εκπαίδευσης, καθώς υπάρχει ο κίνδυνος υπερπροσαρμογής (overfitting). Κάθε διαχωρισμός πραγματοποιήθηκε με τέτοιο τρόπο, ώστε η αναλογία των ψηφίων να είναι κοινή σε κάθε σύνολο δεδομένων, με το σύνολο εκπαίδευσης να έχει τελικά 2000 εκφωνήσεις (200 ανά ψηφίο), το σύνολο επικύρωσης 400 (40 ανά ψηφίο) και το σύνολο αξιολόγησης 600 (60 ανά ψηφίο). Αξίζει επίσης να αναφερθεί πως κάθε φορά που γινόταν ο συγκεκριμένος διαχωρισμός, η επιλογή των εκφωνήσεων που μεταφέρονταν στο σύνολο αξιολόγησης γινόταν με εγγενώς τυχαίο τρόπο, προκειμένου τα τελικά αποτελέσματα να μπορούν να διασταυρωθούν για διάφορες διαμορφώσεις των συνόλων εκπαίδευσης-επικύρωσης-αξιολόγησης και να μην υπάρχει συστηματικότητα ως προς αυτό το διαχωρισμό.

Ο πρώτος ταξινομητής που κατασκευάστηκε για την ταξινόμηση των δεδομένων FSDD ήταν ένας ταξινομητής GMM-HMM, μέσω της βιβλιοθήκης `hmmlearn`[1]. Αφότου τα δεδομένα εκπαίδευσης χωρίστηκαν βάσει του ψηφίου στο οποίο αντιστοιχεί η κάθε εκφώνηση, εκπαιδεύτηκαν 10 ξεχωριστά HMMs, ένα για κάθε ψηφίο, στα δεδομένα του κάθε ψηφίου ξεχωριστά. Κατασκευάστηκε, έτσι, μια συλλογή από HMMs, το καθένα εκ των οποίων υπολογίζει την πιθανότητα (μέσω του λογαρίθμου της πιθανοφάνειας) μια εκχώρηση σε αυτό να αντιστοιχεί στο ψηφίο που έχει μάθει να αναγνωρίζει. Έχοντας το λογάριθμο της πιθανοφάνειας για καθένα εκ των HMMs της συλλογής, η ταξινόμηση ενός νέου δεδομένου γίνεται στο ψηφίο που αντιστοιχεί στο HMM με το μέγιστο αποτέλεσμα για το λογάριθμο της πιθανοφάνειας (maximum log-likelihood).

Σε ό,τι αφορά τη μοντελοποίηση των χαρακτηριστικών των εκφωνήσεων, αυτή πραγματοποιήθηκε από μίγματα Γκαουσιανών κατανομών (εξ ου και το GMM), με πλήθος N_g Γκαουσια-

[1] Στην παρουσίαση του εργαστηρίου της 6/12/2021 αναφέρθηκε πως είναι στη δική μας ευχέρεια η επιλογή της βιβλιοθήκης για την εκπαίδευση των GMM-HMMs και πως η `pomegranate` αποτελεί απλή σύσταση, μιας και ο βοηθητικός κώδικας είναι φτιαγμένος στα δικά της μέτρα.

νών ανά μίγμα. Το μοντέλο που επιλέχθηκε για τα HMMs (Βήμα 10) ήταν ένα μοντέλο τύπου left-right, όπου οι αρχικές πιθανότητες των καταστάσεων περιγράφονταν από το N_s-διάστατο διάνυσμα $\boldsymbol{\pi} = (1, 0, \dots, 0)$, όπου N_s το πλήθος των καταστάσεων HMM, ενώ η αρχική μορφή του πίνακα μετάβασης ήταν

$$\mathbf{A} = \begin{bmatrix} \star & \star & \dots & 0.0 & 0.0 \\ 0.0 & \star & \dots & 0.0 & 0.0 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 0.0 & 0.0 & \dots & \star & \star \\ 0.0 & 0.0 & \dots & 0.0 & 1.0 \end{bmatrix},$$

ο οποίος σε κάθε περίπτωση είναι ένας $N_s \times N_s$ πίνακας, που προβλέπει μεταβάσεις μόνο μεταξύ διαδοχικών καταστάσεων (και γι' αυτό το λόγο υπάρχουν μόνο 2 μη μηδενικά στοιχεία σε κάθε γραμμή, με εξαίρεση την τελευταία που έχει μόνο 1).

Η εκπαίδευση των HMMs (Βήμα 11) έγινε με χρήση του αλγόριθμου EM (αυτόματα από την υλοποίηση της `hmmlearn`) για πλήθος επαναλήψεων $N_{\text{iter}} = 20$, ή μέχρι να επέλθει σύγκλιση, όταν η διαφορά στη log-likelihood διαδοχικών επαναλήψεων δεν ξεπερνά την τιμή μιας παραμέτρου `tolerance` η οποία τέθηκε ίση με 0.01. Την εκπαίδευση των HMMs ακολουθούσε κάθε φορά η αξιολόγησή τους στο σύνολο επικύρωσης (Βήμα 12), μεταβάλλοντας τις υπερπαραμέτρους N_g, N_s, καθώς και την παράμετρο `covariance_type`, η οποία θέτει τη μορφή του πίνακα συνδιακύμανσης. Σε κάθε περίπτωση, φάνηκε η παράμετρος `covariance_type` να μην επηρεάζει σημαντικά την απόδοση του ταξινομητή. Σε ό,τι αφορά το πλήθος καταστάσεων HMM και το πλήθος των Γκαουσιανών ανά μίγμα, εκπαιδεύτηκαν ταξινομητές για κάθε συνδυασμό των παραμέτρων N_s και N_g, με $N_s \in \{1,2,3,4\}$ και $N_g \in \{1,2,3,4,5\}$. Τα αποτελέσματα για την απόδοση κάθε ταξινομητή[2] στο σύνολο επικύρωσης απεικονίζονται στον Πίνακα 3.1.

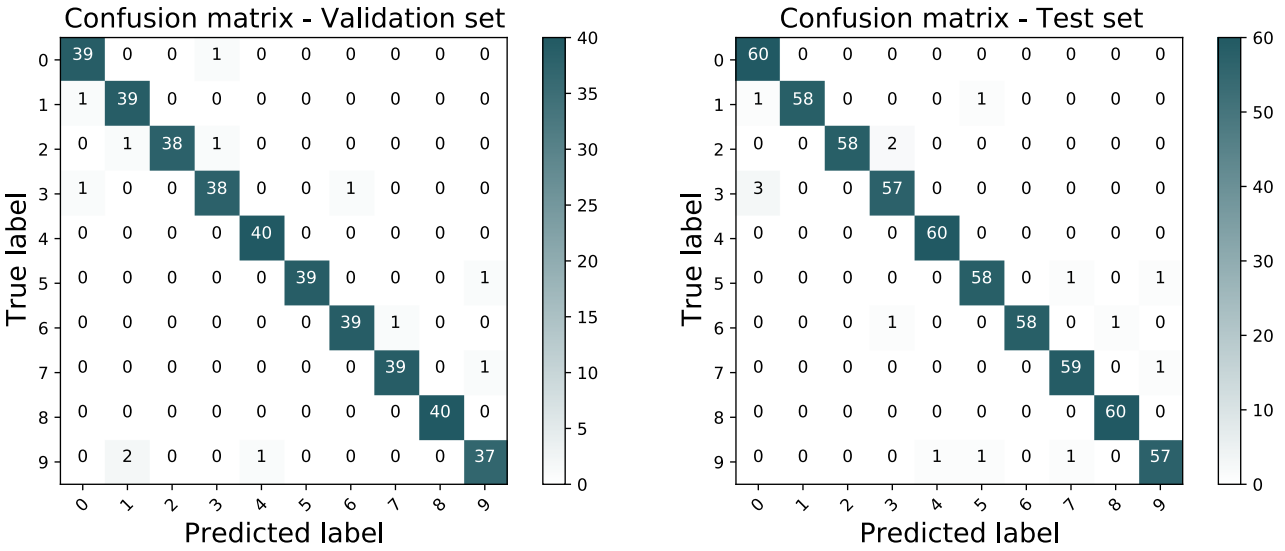
	$N_s = 1$	$N_s = 2$	$N_s = 3$	$N_s = 4$
$N_g = 1$	70.50%	68.50%	78.50%	84.75%
$N_g = 2$	80.25%	82.25%	89.50%	93.50%
$N_g = 3$	84.50%	91.75%	92.00%	95.75%
$N_g = 4$	89.25%	92.00%	95.25%	97.00%
$N_g = 5$	91.25%	96.50%	96.50%	94.75%

Πίνακας 3.1: Ακρίβεια του ταξινομητή GMM-HMM στην ταξινόμηση των ψηφίων του συνόλου επικύρωσης για διάφορους συνδυασμούς του πλήθους καταστάσεων HMM (N_s) και του πλήθους Γκαουσιανών ανά μίγμα (N_g).

Καθίσταται εμφανές πως ο συνδυασμός των N_s και N_g για τον οποίο επιτυγχάνεται η υψηλότερη ακρίβεια κατά την αξιολόγηση του ταξινομητή στο σύνολο επικύρωσης (97.00%) είναι 4 καταστάσεις HMM με 4 Γκαουσιανές ανά μίγμα. Στο αριστερό μέρος της Εικόνας 3.1 απεικονίζεται ο πίνακας σύγχυσης (Βήμα 13) του αντιστοιχεί στην ταξινόμηση του συνόλου επικύρωσης για αυτές τις τιμές των N_s και N_g. Χρησιμοποιώντας τον ίδιο ταξινομητή για το σύνολο αξιολόγησης, στο οποίο δεν είχε πρόσβαση σε οποιαδήποτε προηγούμενη φάση, η ακρίβεια που

[2] Σημειώνεται πως η εκπαίδευση του μοντέλου GMM-HMM ξεκινά με ένα βήμα συσταδοποίησης, προκειμένου να προσδιοριστεί από ποια Γκαουσιανή κατανομή έχει προέλθει καθένα εκ των σημείων εκπαίδευσης. Η υλοποίηση της `hmmlearn` πραγματοποιεί τη συσταδοποίηση αυτή μέσω του αλγορίθμου k-means, ο οποίος είναι εγγενώς μη ντετερμινιστικός. Ως εκ τούτου, είναι πιθανό η συλλογή HMMs που προκύπτει για δεδομένο σύνολο εκπαίδευσης να μην είναι πάντοτε ίδια και κατ' επέκταση τα αποτελέσματα του Πίνακα 3.1 να διαφέρουν από επανάληψη σε επανάληψη, χωρίς όμως να διαφοροποιούνται σημαντικά (οι μέγιστες μεταβολές που παρατηρήθηκαν από επανάληψη σε επανάληψη ήταν της τάξης του 3%).

σημείωσε ήταν 97.50%, με τον αντίστοιχο πίνακα σύγχυσης να απεικονίζεται στο δεξί μέρος της Εικόνας 3.1.



Εικόνα 3.1: Πίνακας σύγχυσης για τις προβλέψεις της συλλογής GMM-HMM με $N_s = 2$ και $N_g = 2$ για το σύνολο επικύρωσης (αριστερά) και το σύνολο αξιολόγησης (δεξιά).

Η ταξινόμηση των δεδομένων αξιολόγησης έγινε με αρκετά καλή ακρίβεια, αφού λανθασμένα ταξινομήθηκαν μόλις 13 από τις 600 εκφωνήσεις, με τα περισσότερα λάθη να πραγματοποιούνται για το ψηφίο 3, ο οποίος 3 φορές ταξινομήθηκε λανθασμένα ως το ψηφίο 0. Αυτό είναι ίσως λογικό, μιας και οι λέξεις «zero» και «three» περιλαμβάνουν και οι δύο το φωνήεν «ι». Κλείνοντας την ανάλυση των GMM-HMMs, σημειώνεται πως επιχειρήθηκε η εκπαίδευσή τους με χρήση επιπλέον χαρακτηριστικών, όπως οι τοπικές παράγωγοι πρώτης και δεύτερης τάξης, ή το zero-crossing rate της `librosa`. Παρ' όλα αυτά, η απόδοση του ταξινομητή φάνηκε να μειώνεται ελάχιστα (ενδεχομένως διότι το σύνολο δεδομένων δεν ήταν αρκετά μεγάλο σε σχέση με το πλήθος των χαρακτηριστικών) και αυτός ήταν ο λόγος για τον οποίο απορρίφθηκαν και χρησιμοποιήθηκαν μόνο τα αρχικά MFCCs.

4 Νευρωνικο Δικτυο LSTM

Το περιεχόμενο αυτής της ενότητας αφορά το Βήμα 14 της εργασίας.

Το τελικό βήμα της εργαστηριακής διαδικασίας αποτέλεσε η εκπαίδευση ενός νευρωνικού δικτύου μακράς βραχυπρόθεσμης μνήμης (LSTM), με σκοπό την αναγνώριση ψηφίων από το σύνολο FSDD. Όπως και στα HMMs, ένα κύριο γνώρισμα που καθιστά τα LSTMs ιδανικά για την ανάλυση σημάτων ήχου είναι το γεγονός πως μπορούν να επεξεργαστούν χαρακτηριστικά τα οποία δομούνται από ακολουθίες μεταβλητού μήκους - στην προκείμενη περίπτωση τα MFCCs, τα οποία συντίθενται από έναν μεταβλητό αριθμό frames ανά εκφώνηση, ανάλογα με τη χρονική της διάρκεια. Σε αντίθεση, όμως, με τα HMMs, τα δεδομένα πρέπει πρώτα να εισαχθούν σε dataloders, ώστε να μπορούν να οργανωθούν σε batches, επομένως για την υλοποίηση του LSTM ήταν απαραίτητη μια επιπλέον προεργασία, η οποία συνοψίζεται στα παρακάτω.

Αρχικά, τα δεδομένα κάθε συνόλου οργανώθηκαν κατά φθίνοντα αριθμό frames ανά MFCC, ούτως ώστε η πρώτη καταχώρηση να αντιστοιχεί στην εκφώνηση της οποίας τα MFCCs είχαν το μεγαλύτερο αριθμό frames, δηλαδή τα χαρακτηριστικά με το μεγαλύτερο μήκος ακολουθίας.

Κατόπιν, πραγματοποιήθηκε το λεγόμενο zero padding, δηλαδή οι ακολουθίες όλων των εκφωνήσεων συμπληρώθηκαν με μηδενικά, μέχρι το μήκος τους να ταυτίζεται με αυτό της ακολουθίας με το μεγαλύτερο μήκος. Παρ' όλα αυτά, η πληροφορία για το αρχικό μήκος κάθε ακολουθίας διατηρήθηκε σε μια λίστα, ώστε κάθε φορά που το νευρωνικό καλείται να επεξεργαστεί μια εκφώνηση να έχει την πληροφορία του πόσα στοιχεία της ακολουθίας (δηλαδή πόσα frames) πρέπει να διαβάσει και να αγνοεί έτσι τα μηδενικά - η μέθοδος γνωστή ως packing[3]. Η χρησιμότητα της διαδικασίας αυτής είναι διττή: αφενός, η ταχύτητα εκπαίδευσης του δικτύου αυξάνεται σημαντικά, αφού οι τιμές των ακολουθιών που δημιουργήθηκαν τεχνηέντως αγνοούνται a priori και, αφετέρου, το γεγονός πως τα δεδομένα εκχωρούνται σε batches με παρόμοια μήκη ακολουθιών (αφού οι εκφωνήσεις έχουν ταξινομηθεί κατά φθίνοντα αριθμό frames) βελτιστοποιεί τον τρόπο εκπαίδευσης του δικτύου. Σημειώνεται εδώ πως, όπως είναι αναμενόμενο, δοκιμές που πραγματοποιήθηκαν χωρίς packing χρειάστηκαν σημαντικά μεγαλύτερους χρόνους εκπαίδευσης.

Σε ό,τι αφορά τα υπόλοιπα χαρακτηριστικά του LSTM, κάθε εποχή εκπαίδευσής του ακολουθείται από μια διαδικασία επικύρωσης, προκειμένου να υπάρχει μια εικόνα σχετικά με το αν το μοντέλο εκπαιδεύεται κατάλληλα, ή αν οδηγείται σε υπερπροσαρμογή. Το κόστος υπολογίζεται ως το μέσο validation loss ανά εποχή εκπαίδευσης. Μέσω του αλγορίθμου early stopping[4], εάν στο τέλος μιας εποχής το μέσο validation loss είναι υψηλότερο από αυτό της προηγούμενης, δημιουργείται ένα checkpoint του μοντέλου στο τέλος της προηγούμενης εποχής. Εάν το μέσο validation loss δεν αποκτήσει νέο ελάχιστο για έναν αριθμό μετέπειτα εποχών (ο οποίος καλείται `patience`), τότε η εκπαίδευση τερματίζεται και το δίκτυο παίρνει τη μορφή που είχε στο τελευταίο checkpoint. Η διαδικασία αυτή όχι μόνο αποδεσμεύει από την ανάγκη ελέγχου των εποχών εκπαίδευσης, αλλά διασφαλίζει πως το τελικό μοντέλο θα είναι το βέλτιστο ως προς το ποσοστό υπερπροσαρμογής.

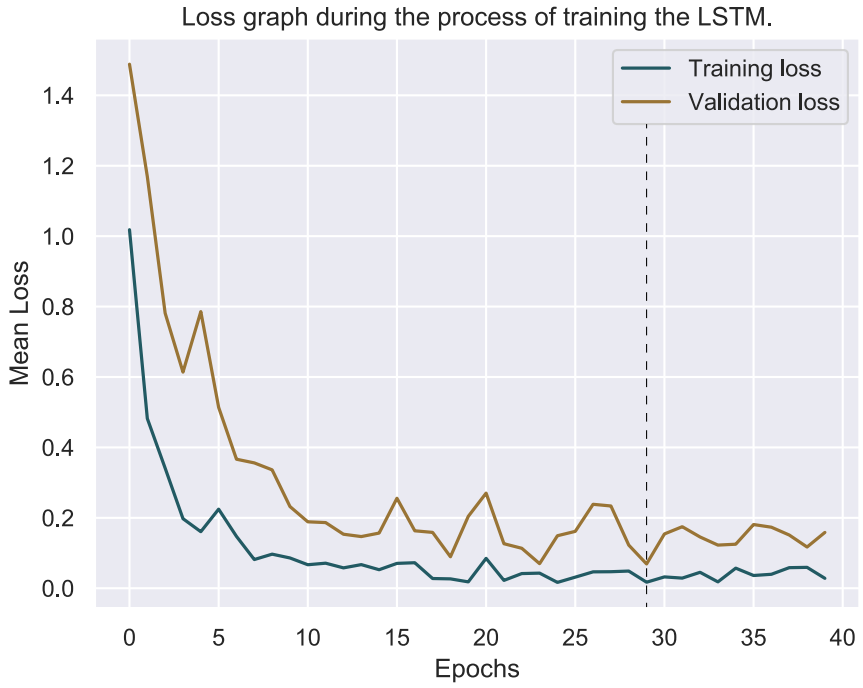
Άλλες παράμετροι που εισήχθησαν προκειμένου να μειωθεί η πιθανότητα υπερπροσαρμογής στα δεδομένα κατά την εκπαίδευση του δικτύου ήταν η πιθανότητα dropout, καθώς και η παράμετρος `weight_decay`, για L2 regularization. Η πρώτη εισαγάγει ένα dropout layer μετά από κάθε επίπεδο LSTM (με εξαίρεση το τελικό) με πιθανότητα ίση με την τιμή της παραμέτρου `dropout`. Η dropout είναι στην ουσία μια μέθοδος ομαλοποίησης, η οποία εισαγάγει ένα είδος θορύβου στη διαδικασία εκπαίδευσης, αποτρέποντας το μοντέλο από το να αναπτύξει ισχυρές συσχετίσεις κάποιων νευρώνων με άλλους, συγκεκριμένους νευρώνες. Ένα σημαντικό μειονέκτημα που έχει η διαδικασία dropout για τα LSTM (και για τα RNN γενικότερα) είναι ότι εισαγάγει την πιθανότητα να ξεχαστεί κάποια πληροφορία που δε θα θέλαμε κανονικά να ξεχαστεί και αυτός είναι και ο λόγος που δε χρησιμοποιήθηκε στο τελικό μοντέλο που θα παρουσιαστεί παρακάτω. Σε ό,τι αφορά την `weight_decay`, αυτή έρχεται αυτόματα εάν ως optimizer της συνάρτησης κόστους (εδώ η Cross Entropy Loss, εφόσον το πρόβλημα αντιστοιχεί σε ταξινόμηση) χρησιμοποιηθεί ο αλγόριθμος Adam και επίσης βοηθά στην αποφυγή της υπερπροσαρμογής, εισάγοντας ένα penalty υπό τη μορφή μιας 2-norm των βαρών, το οποίο αποτρέπει τα βάρη από το να αποκτήσουν πολύ υψηλές τιμές.

Το τελευταίο άξιο αναφοράς χαρακτηριστικό του LSTM που αναπτύχθηκε είναι η δυνατότητα bidirectionality, μέσω μιας boolean παραμέτρου `bidirectional`. Εάν η παράμετρος αυτή τεθεί σε αληθή τιμή, τότε στην ουσία το τελικό δίκτυο ισοδυναμεί με 2 LSTM εκ των

[3] Σημειώνεται στο σημείο αυτό πως χάρη στη δομή της `hmmlearn` η αντίστοιχη διαδικασία έγινε με ένα απλό concatenation όλων των frames κάθε MFCC σε μία ενιαία λίστα και με την καταχώρηση του μήκους της ακολουθίας των frames κάθε εκφώνησης σε μία επιπλέον λίστα, κατ' αντιστοιχία με όσα έγιναν για να είναι εφικτό το packing στο LSTM.

[4] Ο αλγόριθμος early stopping που αναπτύχθηκε για το LSTM βασίστηκε σε μεγάλο βαθμό στην υλοποίηση του Bjarte Mehus Sunde στο πακέτο pytorchtools.

οποίων το ένα «διαβάζει» τις ακολουθίες των χαρακτηριστικών με ανάποδη σειρά σε σχέση με το άλλο και στο τέλος πραγματοποιείται μια συγχώνευση των αποτελεσμάτων τους. Προφανώς, το χαρακτηριστικό αυτό επιφέρει αλλαγές στο εσωτερικό του δικτύου, αφού οι διαστάσεις των διάφορων επιπέδων πρέπει να διπλασιαστούν, μιας και στην πράξη έχει διπλασιαστεί ο αριθμός των LSTM. Το μεγάλο προτέρημά του είναι πως οδηγεί το δίκτυο στο να διατηρεί μνήμη όχι μόνο με φορά από το παρελθόν προς το μέλλον, αλλά και αντιστρόφως. Έτσι, εάν σε κάποιες ακολουθίες χαρακτηριστικών η σημαντική πληροφορία βρίσκεται προς το τέλος[5], το bidirectional LSTM θα την κρατήσει στη μνήμη του εξ αρχής και έτσι θα εκπαιδευτεί καλύτερα.

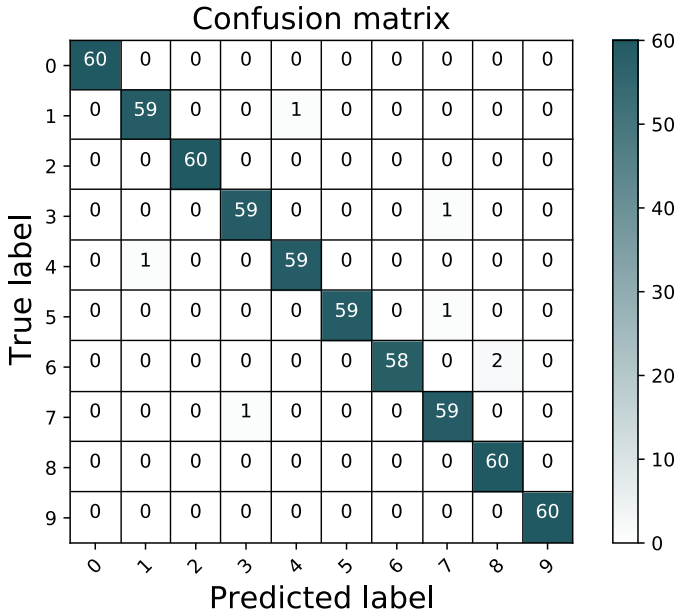


Εικόνα 4.1: Διάγραμμα εκπαίδευσης του LSTM. Με μπλε και χρυσό χρώμα φαίνονται οι καμπύλες του μέσου κόστους ανά εποχή για το σύνολο εκπαίδευσης και αξιολόγησης, αντίστοιχα, ενώ η διακεκομμένη γραμμή αντιστοιχεί στην εποχή όπου ενεργοποιείται το τελικό early stopping, σηματοδοτώντας το πέρας της εκπαίδευσης του δικτύου.

Μετά από αρκετό πειραματισμό, έγινε ξεκάθαρο πως η απόδοση του δικτύου στο σύνολο επικύρωσης ήταν αρκετά υψηλή, για διάφορους συνδυασμούς των παραμέτρων που το χαρακτηρίζουν. Το ίδιο παρατηρήθηκε και κρατώντας σταθερές τις τιμές των παραμέτρων του δικτύου και μεταβάλλοντας τα σύνολα εκπαίδευσης και επικύρωσης, ώστε να καταστεί βέβαιο πως η υψηλή απόδοση δεν οφείλεται σε ευκαιριακούς παράγοντες. Στην Εικόνα 4.1 παρατίθεται το διάγραμμα εκπαίδευσης ενός δικτύου LSTM, το οποίο σχεδιάστηκε για 40 κρυφά χαρακτηριστικά, χωρίς stacking των LSTMs (και επομένως χωρίς `dropout`), χωρίς bidirectionality, με ρυθμό εκμάθησης ίσο με $5 \cdot 10^{-3}$ και `weight_decay` $= 10^{-6}$. Ο αριθμός εποχών επιλέχθηκε υψηλός, ώστε η εκπαίδευση να ρυθμίζεται βάσει early stopping με τιμή `patience` = 10, επομένως το δίκτυο εκπαιδεύτηκε πλήρως στο τέλος της 30ης μόλις εποχής (βλ. διακεκομμένη γραμμή στην Εικόνα 4.1), μιας και για τις 10 επόμενες εποχές το validation loss δεν απέκτησε νέα ελάχιστη τιμή. Η απόδοση του δικτύου αυτού στο σύνολο των 600 δεδομένων αξιολόγησης

[5] Στην επεξεργασία φυσικής γλώσσας κάτι τέτοιο θα μπορούσε να σημαίνει μια πρόταση με μια «λέξι κλειδί» προς το τέλος, ενώ στα πλαίσια επεξεργασίας ήχου θα μπορούσε να είναι κάποιο χαρακτηριστικό φωνήεν στο τέλος της εκφώνησης.

ήταν 98.83%, δηλαδή οι λάθος ταξινομήσεις αφορούσαν μόλις 7 καταχωρήσεις, όπως φαίνεται και από τον ακόλουθο πίνακα σύγχυσης (Εικόνα 4.2).



Εικόνα 4.2: Πίνακας σύγχυσης για το LSTM στο σύνολο αξιολόγησης.

Συμπερασματικά, τόσο ο ταξινομητής της συλλογής GMM-HMMs, όσο και ο ταξινομητής νευρωνικού δικτύου LSTM σημείωσαν πολύ υψηλή απόδοση σε ό,τι αφορά την ταξινόμηση των ψηφίων του FSDD. Φαίνεται, λοιπόν, πως η χρήση της πλήρους ακολουθίας των frames ανά MFCC οδηγεί σε σημαντικά ακριβέστερα μοντέλα, σε σχέση με αυτά που εκπαιδεύονται με χαρακτηριστικά τις μέσες τιμές ή/και τις διακυμάνσεις των MFCCs, όπως έγινε στην προπαρασκευαστική φάση της εργασίας αυτής.

Αναφορες

[Cal20] O. Calin, *Deep Learning Architectures*, Springer, 2020. ISBN: 978-3-030-36721-3 Cited on p. 7

[KT08] K. Koutroumbas, S. Theodoridis, *Pattern Recognition*, Academic Press, 2008. ISBN: 978-1-59749-272-0