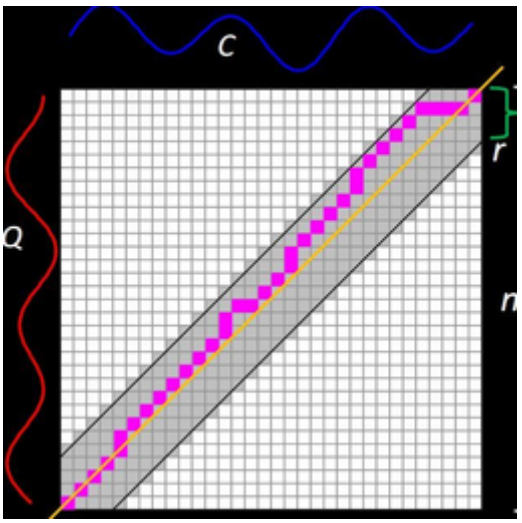


Τελική Αναφορά

Επεξεργασία Ομιλίας και Ήχου

Ιόνιο Πανεπιστήμιο
Τμήμα Πληροφορικής

Παναγιώτης Σιώλος	Π2017160	p17siol@ionio.gr
Χρήστος Μήλιος	Π2018219	p18mili1@ionio.gr

Εισαγωγή

Η εφαρμογή που υλοποιήσαμε ονομάζεται ***AuthByVoice*** και είναι ένα desktop application το οποίο κάνει ταυτοποίηση μέσω αναγνώρισης φωνής καλώντας το χρήστη να πει το ονοματεπώνυμό του.

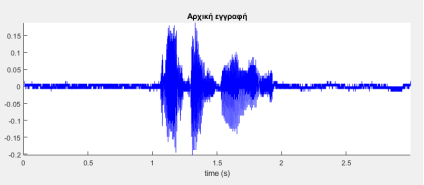
Κάθε σύστημα αναγνώρισης φωνής αποτελείται από δύο στάδια, την ***καταχώρηση*** και την ***εξακρίβωση***. Στο στάδιο καταχώρησης, η φωνή του χρήστη καταγράφεται και εξάγονται κάποια χαρακτηριστικά βάσει των οποίων σχηματίζεται ένα φωνητικό αποτύπωμα (σήμα αναφοράς). Στο στάδιο εξακρίβωσης, ένα φωνητικό δείγμα συγκρίνεται με κάποιο φωνητικό αποτύπωμα. Ανάλογα με το σύστημα αναγνώρισης, το φωνητικό αποτύπωμα μπορεί να εξαρτάται ή όχι (text-dependent και text-independent, αντίστοιχα) από το περιεχόμενο των φράσεων. Στην πρώτη περίπτωση, συγκεκριμένες λέξεις, αριθμοί ή φράσεις που προφέρει ο χρήστης συγκρίνονται με τα αντίστοιχα σήματα αναφοράς του χρήστη που περιέχουν την ίδια φράση με αυτή του δείγματος. Στη δεύτερη περίπτωση, το σύστημα ψάχνει να βρει μοναδικά φωνητικά χαρακτηριστικά στο δείγμα ελεύθερου λόγου που έχει εισάγει ο χρήστης και να βρει αντιστοιχία μέσα σε ένα σύνολο σημάτων αναφοράς πολλών χρηστών.

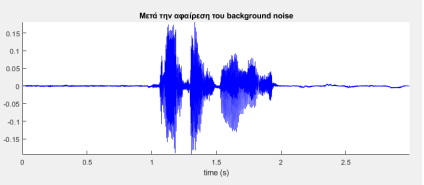
Η εφαρμογή *AuthByVoice* είναι ένα text-dependent συστήματα ταυτοποίησης, το οποίο σημαίνει ότι ένα δείγμα εισόδου πρέπει να συγκριθεί με πολλά σήματα αναφοράς για να προσδιοριστεί αν υπάρχει αντιστοίχιση.

Εγγραφή και Επεξεργασία Σήματος

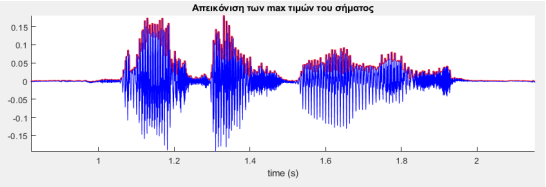
Οι προδιαγραφές των σημάτων που χρησιμοποιούμε είναι FS = 44100 Hz, channels = 1 και nbits = 16 με διάρκεια 3 δευτερόλεπτα. Αρχικά, δοκιμάσαμε με συχνότητα δειγματοληψίας 8000 Hz αλλα στην πορεία διαπιστώσαμε ότι η χρήση μεγαλύτερων τιμών της FS επέφερε μία μικρή βελτίωση στα αποτελέσματα των συγκρίσεων.

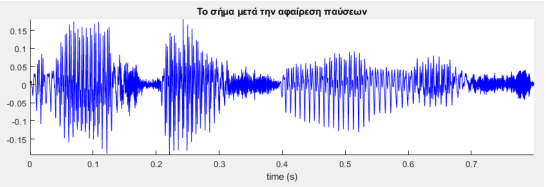
Μετά την εγγραφή, το σήμα μετατρέπεται σε πίνακα και ακολουθεί η διαδικασία εξάλειψης θορύβου από το σήμα με τη βοήθεια της συνάρτησης ***specsub*** της εργαλειοθήκης Voicebox. Η *specsub*() χρησιμοποιώντας κατάλληλες συναρτήσεις της Voicebox κάνει πλαισιοποίηση (με παράθυρο Hamming και επικάλυψη 75%), εφαρμόζει διακριτό μετασχηματισμό Fourier (DFT), αποτιμά το επίπεδο θορύβου, διορθώνει τα πλαίσια θορύβου και εφαρμόζει αντίστροφο μετασχηματισμό Fourier για να δώσει στην έξοδό της το καθαρό από θόρυβο σήμα.





Με τη συνάρτηση ***movmax*** υπολογίζονται οι μέγιστες τιμές του σήματος για ένα κυλιόμενο παράθυρο μήκους R διαδοχικών τιμών του σήματος και με ένα κατώτατο κατώφλι “κόβουμε” τις τιμές που βρίσκονται κάτω από αυτό ώστε να προκύψει το σήμα όχι μόνο χωρίς τα κενά αλλά να βρίσκεται και στη χρονική στιγμή 0.





Στη συνέχεια υπολογίζουμε για κάθε σήμα τους συντελεστές MFCC (Mel Frequency Cepstral Coefficients). Η ανάλυση cepstral βασίζεται στην εύρεση του *cepstrum*, το οποίο ορίζεται ως ο αντίστροφος DFT του μέτρου του λογάριθμου ενός DFT ενός σήματος. Το MFC είναι η αναπαράσταση της ενέργειας του φάσματος (power spectrum) ενός σήματος ήχου, που βασίζεται στο γραμμικό μετασχηματισμό συνημιτόνου ενός λογάριθμου φάσματος ισχύος σε μια μη γραμμική κλίμακα συχνότητας Mel. Οι συντελεστές MFCCs είναι αυτοί που συνθέτουν ένα MFC. Η διαφορά μεταξύ του cepstrum και του MFC είναι ότι στο MFC οι ζώνες συχνοτήτων απέχουν εξίσου στην κλίμακα Mel, προσεγγίζοντας καλύτερα την απόκριση του ανθρώπινου ακουστικού συστήματος από τις γραμμικά διαχωρισμένες ζώνες συχνοτήτων στο cepstrum. Για το σκοπό αυτό χρησιμοποιούμε τη συνάρτηση ***mfcc*** με μήκος παραθύρου 0.02*Fs και 0.013*Fs αφήνοντας τις default τιμές για τις υπόλοιπες παραμέτρους.

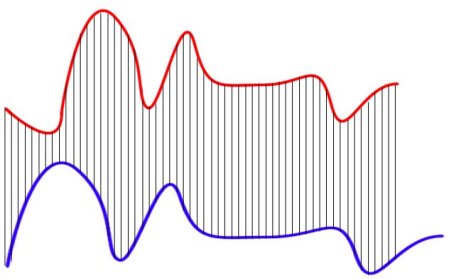
Στο αποτέλεσμα που δίνει η mfcc εφαρμόζουμε standardization (ή z-normalization) με τη συνάρτηση ***zscore*** και πλέον το σήμα είναι έτοιμο για σύγκριση. Η χρήση standardization πριν από την **DTW** αντί της κανονικοποίησης είναι σημαντική γιατί κανονικοποιούμε ξεχωριστά κάθε υποπλαίσιο του σήματος και όχι ολόκληρο το σήμα. Ο τύπος για τη standardization είναι

$$\mathbf{x_{i,new} = (x_i - mean(x)) / std(x)},$$

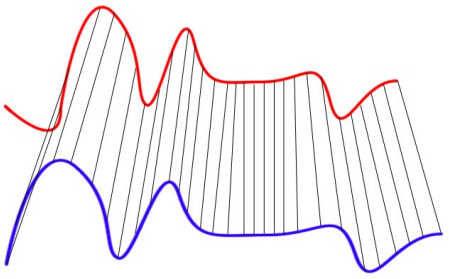
όπου **mean(x)** και **std(x)** είναι μέση τιμή και η τυπική απόκλιση του **x**.

Σύγκριση σημάτων

Η σύγκριση σημάτων προϋποθέτει ότι τα σήματα είναι ίδιας ταχύτητας, στην περίπτωσή μας όμως είναι πρακτικά αδύνατο. Ο αλγόριθμός Dynamic Time Warping έρχεται να λύσει αυτό το πρόβλημα συγκρίνοντας κάθε σημείο ενός ψηφιακού σήματος με πολλά σημεία του άλλου και αντίστροφα, για να υπολογίσει την ελάχιστη δυνατή μεταξύ τους τους διαφορά. Στο παράδειγμά μας, το μπλε και το κόκκινο σήμα ακολουθούν το ίδιο μοτίβο αλλά το μπλε είναι μεγαλύτερης διάρκειας. Με αντιστοίχιση 1:1 (ευκλείδεια απόσταση) βλέπουμε ότι όταν φτάνει σε κορυφή το κόκκινο σήμα, το μπλε καθυστερεί ή προηγείται. Αντίθετα με αντιστοίχιση 1:N και N:1 (dynamic time warping) η σύγκριση απλώνεται σε όλο το μήκος του σήματος και οι κορυφές του κόκκινου συμπίπτουν χρονικά με τις κορυφές του μπλε σήματος.



Euclidean Matching



Dynamic Time Warping Matching

Η έκδοση της Dynamic Time Warping συνάρτησης που χρησιμοποιούμε είναι η DTW[1] η οποία υπολογίζει το μονοπάτι και τον αριθμό των βημάτων k που ακολουθεί ο αλγόριθμος για να υπολογίσει την τελική απόσταση (διαφορά) μεταξύ των σημάτων και διαιρεί στο τέλος την τιμή της απόστασης με το k ώστε οι dtw τιμές να είναι πιο ευανάγνωστες.

Διαδικασία απόρριψης/αποδοχής

Η επιλογή της τιμής κατωφλίου απόρριψης/αποδοχής υπολογίζεται με βάσει το υπάρχον φωνητικό σετ δεδομένων του χρήστη που αποτελείται από 20 αρχεία. Κάθε σήμα αναφοράς ενός χρήστη συγκρίνεται με τα υπόλοιπα 19 (αφού σύγκριση ενός σήματος με τον εαυτό του δίνει απόσταση 0). Για κάθε σήμα εισόδου υπολογίζεται η μέγιστη τιμή απόστασης από τις 19 συγκρίσεις, (δοκιμάσαμε και με την ελάχιστη απόσταση και με τη μέση τιμή αλλά η μέγιστη τιμή έδωσε με διαφορά τα πιο καλά αποτελέσματα). Έτσι έχουμε ένα σύνολο 20 μέγιστων τιμών στη στήλη Max Vals και από αυτή τη στήλη υπολογίζουμε την ελάχιστη απόσταση, η οποία είναι και η τιμή κατωφλίου απόρριψης/αποδοχής.

Είσοδος / Βάση	xristos1	xristos2	xristos3	xristos4	xristos5	xristos6	xristos7	xristos8	xristos9	xristos10	xristos11	xristos12	xristos13	xristos14	xristos15	xristos16	xristos17	xristos18	xristos19	xristos20	Min Val	Max Val
DTW xristos1		3.8926	3.6451	3.0835	3.3466	3.3879	3.5242	3.5059	3.1575	3.1163	3.3564	3.4258	3.5802	3.1856	3.5097	3.1506	3.4239	2.7424	2.8609	3.2049	2.7424	3.8926
DTW xristos2	3.8926		3.5412	3.8918	3.1311	3.6682	4.2653	4.7672	3.7973	4.3486	3.7794	3.8483	3.5470	4.2347	4.5508	3.9849	4.0518	3.9116	3.8338	4.0298	3.1311	4.7672
DTW xristos3	3.6451	3.5412		3.6797	3.4978	3.2879	3.9298	4.1280	3.3770	3.9477	3.9174	3.6342	3.5111	3.8473	4.5301	3.8066	3.8663	3.6089	3.6743	3.8074	3.2879	4.5301
DTW xristos4	3.0835	3.8918	3.6797		3.1108	3.3344	3.6070	3.5154	3.1094	3.3664	3.5854	3.3855	3.5977	3.4947	3.5632	3.3590	3.5628	3.3722	3.2248	3.5132	3.0835	3.8918
DTW xristos5	3.3466	3.1311	3.4978	3.1108		3.4061	3.6393	3.6946	3.2629	3.3902	3.4041	3.1300	3.3701	3.4049	3.6357	3.6083	3.6388	3.3087	3.3306	3.4373	3.1108	3.6946
...																						
DTW xristos17	3.4239	4.0518	3.8663	3.5628	3.6388	3.6467	3.5638	3.8162	3.3295	3.3013	3.3970	3.5178	3.5656	3.5373	3.6565	3.2305		3.2681	3.4434	3.3234	3.2305	4.0518
DTW xristos18	2.7424	3.9116	3.6089	3.3722	3.3087	3.6004	3.4988	3.7656	3.2162	3.1684	3.2324	3.2808	3.4187	3.0613	3.4732	2.8412	3.2681		2.5918	2.7079	2.5918	3.9116
DTW xristos19	2.8609	3.8338	3.6743	3.2248	3.3306	3.5330	3.5701	3.6762	3.1201	3.1947	3.2435	3.1086	3.2682	2.8896	3.3803	2.8363	3.4434	2.5918		2.7143	2.5918	3.8338
DTW xristos20	3.2049	4.0298	3.8074	3.5132	3.4373	3.7188	3.5614	4.0714	3.3776	3.4312	3.2801	3.0799	3.4919	3.1656	3.6489	3.1933	3.3234	2.7079	2.7143		2.7079	4.0714
																				min of Max Val		3.6946

Συγκρίνοντας κάθε τιμή του πίνακα με την τιμή κατωφλίου προκύπτει ο παρακάτω πίνακας. Το μόνο που μένει είναι να ορίσουμε έναν αποδεκτό αριθμό **reject** (5 είναι επαρκής τιμή) για κάθε δείγμα εισόδου για τον οποίο η ταυτοποίηση θα κρίνεται θετική (granted). Ωστόσο, αν χρήστης μιλάει γρήγορα ή όχι καθαρά ή υπάρχει αρκετός θόρυβος περιβάλλοντος μπορεί να μην ταυτοποιηθεί (denied). Ο αριθμός των reject θα είναι μεγαλύτερος από 5 αλλά σχεδόν σίγουρα μικρότερος από 15 οπότε το σύστημα μπορεί να προτρέψει το χρήστη να προσπαθήσει ξανά ή να προειδοποιήσει ότι προσπαθεί να ταυτοποιηθεί ως κάποιος άλλος.

DTW xristos1	PASS	REJECT	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	1	GRANTED
DTW xristos2	REJECT	PASS	PASS	REJECT	PASS	PASS	REJECT	REJECT	REJECT	REJECT	REJECT	REJECT	PASS	REJECT	REJECT	REJECT	REJECT	REJECT	REJECT	REJECT	15	DENIED
DTW xristos3	PASS	PASS	PASS	PASS	PASS	PASS	REJECT	REJECT	PASS	REJECT	REJECT	PASS	PASS	REJECT	REJECT	REJECT	REJECT	PASS	PASS	REJECT	9	DENIED
DTW xristos4	PASS	REJECT	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	1	GRANTED
DTW xristos5	PASS	PASS	PASS	PASS	PASS	PASS	PASS	REJECT	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	1	GRANTED
...																						
DTW xristos16	PASS	REJECT	REJECT	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	2	GRANTED
DTW xristos17	PASS	REJECT	REJECT	PASS	PASS	PASS	PASS	REJECT	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	3	GRANTED
DTW xristos18	PASS	REJECT	PASS	PASS	PASS	PASS	PASS	REJECT	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	2	GRANTED
DTW xristos19	PASS	REJECT	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	1	GRANTED
DTW xristos20	PASS	REJECT	REJECT	PASS	PASS	REJECT	PASS	REJECT	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	PASS	4	GRANTED

Εργαλεία

Η υλοποίηση της εφαρμογής έγινε με τον App Designer του Matlab 2018b. Το τελικό αποτέλεσμα είναι ένα εκτελέσιμο αρχείο (.exe) με ενσωματωμένο web installer, ο οποίος θα κατεβάσει τις απαραίτητες runtime βιβλιοθήκες κατά την εγκατάσταση. Το πακέτο εγκατάστασης περιλαμβάνει την εξωτερική συνάρτηση ***DynamicTimeWarping.m***, την εξωτερική εργαλειοθήκη ***Voicebox*** καθώς επίσης και τις εργαλειοθήκες του matlab ***Audio System***, ***DSP System***, ***Data Acquisition***, ***Mapping***, ***Optimization***, ***Signal Processing***, ***Statistics and Machine Learning*** και όλες οι σχετικές με το ***MATLAB Compiler***. Τέλος, υπάρχουν και δυο αρχεία παραμέτρων (***keySet.mat*** και ***valSet.map***) με προεγκατεστημένες τις τιμές κατωφλίου για τα αντίστοιχα σετ φωνητικών σημάτων αναφοράς των υποφαινόμενων.

Παραπομπές

Link video της εφαρμογής: https://www.youtube.com/watch?v=Pd6C3FS0Ejg

Πηγές

[1] **DynamicTimeWraping.m:**
https://github.com/tanmayGIT/ICDAR-2015-DTW/blob/master/DynamicTimeWarping.m

[2] **Voicebox:**
http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[3]
https://towardsdatascience.com/dynamic-time-warping-3933f25fcdd

[4]
https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/12636/ZairaFasianou_diploma_new_v2.pdf?sequence=1&isAllowed=y

[5]
https://www.statology.org/standardization-vs-normalization/

[6]
https://www.mathworks.com/help/index.html