Αριστοτελειο πανεπιστημιο Θεσσαλονικης
Τμημα Ηλεκτρολογων Μηχανικων & Μηχανικων Υπολογιστων

Τεχνολογια Ηχου Και Εικονας

Αναγνώριση Συναισθήματος Από Ομιλία

Δούκας Παράσχος (7878)
Καλταμπανίδης Ιωάννης (7887)
Κωνσταντέλος Στέλιος (7910)
Λεσγίδης Ιάσων (8048)

Νοεμβριος 2019

Περιεχόμενα

1 Εισαγωγή

Η ομιλία αποτελεί τον πιο φυσικό και αποτελεσματικό τρόπο επικοινωνίας μεταξύ των ανθρώπων. Πέραν της γλωσσολογικής πληροφορίας που μεταδίδεται, το σήμα αυτό ενθυλακώνει τη συναισθηματική κατάσταση του ομιλούντος, πληροφορία απαραίτητη για την ομαλή αλληλεπίδραση ανθρώπου-υπολογιστή (**H**uman **C**omputer **I**nteraction).

Το γεγονός αυτό, μεταξύ άλλων, έχει εμπνεύσει ένα σχετικά νέο πεδίο εφαρμογής της μηχανικής μάθησης, αυτό της αναγνώρισης συναισθήματος μέσω σήματος ομιλίας (**S**peech **E**motion **R**ecognition). Το ***SER*** ορίζεται ως η αυτοματοποιημένη εξαγωγή της συναισθηματικής κατάστασης του ομιλούντος από το σήμα ομιλίας του.

Μολονότι δεν υπάρχει ένας αποδεκτός διεπιστημονικός ορισμός του συναισθήματος [1]-[3], η αναγνώριση συναισθήματος από σήμα ομιλίας στηρίζεται στον αφηρημένο πλην όμως βιωματικό και διαισθητικό ορισμό του συναισθήματος. Ως εκ τούτου, στο παρόν κείμενο, ο όρος **συναίσθημα/συναισθηματική κατάσταση** ταυτίζεται με τον λεξικολογικό όρο/ετικέτα (label) που χρησιμοποιεί το άτομο ώστε να περιγράψει το πως αισθάνεται, δίχως να αποδοθεί αυστηρός ορισμός. Αρκετοί ερευνητές συμφωνούν με την θεωρία παλέτας (palette theory), η οποία προτάθηκε στις αρχές του 1970 από τους Anscombe και Geach. Σύμφωνα με αυτήν, κάθε συναίσθημα αποτελεί συνδυασμό πρωταρχικών συναισθημάτων (primary emotions). Τα πρωταρχικά συναισθήματα είναι ο Θυμός (Anger), η Απέχθεια (Disgust), ο Φόβος (Fear), η Χαρά (Joy), η Λύπη (Sadness) και η Έκπληξη (Surprise) [4]. Στα αμέσως επόμενα χρόνια έγινε ευρέως διαδεδομένη η θεωρία των βασικών συναισθημάτων (basic emotions), η οποία καθιερώθηκε από των Ekman [5]. Η ορθότητα ή μη των παραπάνω θεωριών ξεπερνά τα όρια της αναφοράς και δε θα αναλυθεί. Ωστόσο, η πλειονότητα των προσεγγίσεων *SER* βασίζεται στα παραπάνω συναισθήματα.

Η αναγνώριση συναισθήματος από ομιλία ανάγεται σε πρόβλημα κατηγοριοποίησης (classification). Όπως κάθε πρόβλημα επιβλεπόμενης μάθησης, θα μας απασχολήσουν τρία βασικά ζητήματα που θα αναλυθούν σε ακόλουθες ενότητες. Αρχικά, γίνεται λόγος για τα διαθέσιμα σύνολα δεδομένων (dataset), τα συνήθη χαρακτηριστικά (features) και τους δημοφιλέστερους ταξινομητές/κατηγοριοποιητές (classifiers) που έχουν εξεταστεί στη βιβλιογραφία. Εν συνεχεία, παρουσιάζεται η προσωπική μας προσέγγιση στο πρόβλημα και τέλος παρατίθενται τα αποτελέσματα και ο σχολιασμός τους.

1.1 Εφαρμογές *SER*

Η αναγνώριση συναισθήματος από ομιλία βρίσκει αντίκρισμα σε πληθώρα εφαρμογών διεπιστημονικών κλάδων, όπως ενδεικτικά αναφέρονται:

- Διάγνωση διαταραχών φάσματος αυτισμού σε εφήβους και παιδιά [6].
- Προσαρμογή των πράξεων ενός ρομποτικού συστήματος, για την επίτευξη και διατήρηση θετικών συναισθηματικών καταστάσεων, ιδίως σε ηλικιωμένους. [7].
- Βελτίωση της ποιότητας συνομιλιών σε τηλεφωνικά κέντρα [8].
- Αυτόματη ρύθμιση φωτισμού σε θεατρικές παραστάσεις [9].
- Συνεισφορά στην κλινική διάγνωση κατάθλιψης και αυτοκτονικών τάσεων [10].
- Απόδοση προτεραιότητας σε κλήσεις εκτάκτου ανάγκης.

2 Κατηγορίες Συνόλων Δεδομένων

Η επιλογή συνόλου δεδομένων για την εκπαίδευση του κατηγοριοποιητή παίζει κρίσιμο ρόλο στην αποτελεσματικότητα του μοντέλου. Υπάρχουν τρεις βασικές κατηγορίες *SER dataset* [11]:

1. **Βασισμένα Σε Ηθοποιούς** (Actor Based): Τα ηχητικά δεδομένα συλλέγονται από έμπειρους ηθοποιούς που υποδύονται ένα συναίσθημα αρθρώνοντας μια πρόταση.

2. **Εκμαιευμένα** (Elicited): Το υποκείμενο εμπλέκεται σε μια συνομιλία που έχει ως στόχο την εκμαίευση των ζητούμενων συναισθημάτων. Το υποκείμενο δεν επιδιώκει να εκφράσει κάποιο συναίσθημα, αυτό γίνεται εμμέσως και εν αγνοία του μέσω του διαλόγου.

3. **Φυσικά** (Natural): Φυσικοί διάλογοι που παρατηρούνται στην καθημερινότητα, π.χ. τηλεφωνικές κλήσεις.

Στον πίνακα 1 παρουσιάζονται τα βασικότερα χαρακτηριστικά των κατηγοριών αυτών. Σημειώνεται πως τα πιο συχνά χρησιμοποιούμενα dataset είναι actor-based [12].

Κατηγορία	*Θετικά*	*Αρνητικά*
Βασισμένα Σε Ηθοποιούς	- Χρησιμοποιούνται συχνά.	- Η έκφραση του συναισθήματος υπόκειται στην υποκειμενική κρίση του ηθοποιού.
	- Τυποποιημένα.	- Δεν υπάρχει εννοιολογική συνάφεια μεταξύ των προτάσεων
	- Συγκρίσιμα αποτελέσματα.	- Έχουν επεισοδιακή φύση
	- Ευρεία γκάμα διαθέσιμων συναισθημάτων.	
	- Διαθέσιμα δεδομένα για πολλές γλώσσες.	
Εκμαιευμένα	- Πιο κοντά σε Φυσικά σύνολα δεδομένων	- Ενδεχομένως δεν είναι διαθέσιμο ευρύ φάσμα συναισθημάτων.
	- Υπάρχει εννοιολογική συνάφεια μεταξύ των προτάσεων.	- Εφόσον οι ομιλητές γνωρίζουν πως ηχογραφούνται, η ποιότητα των δεδομένων πλήττεται.
Φυσικά	- Πλέον ρεαλιστικά δεδομένα.	- Ενδεχομένως δεν είναι διαθέσιμο ευρύ φάσμα συναισθημάτων.
		- Υπόκεινται σε πνευματικά δικαιώματα.
		- Ύπαρξη θορύβου.

Πίνακας 1: *Ιδιότητες Κατηγοριών Συνόλων Δεδομένων [11]*

3 Διαθέσιμα Χαρακτηριστικά

Η εκτενής φυσιολογική και μαθηματική μελέτη της ομιλίας έχει αποδώσει περισσότερα από 1000 διαθέσιμα προς έλεγχο χαρακτηριστικά, που ακόμα και με τη χρήση τεχνικών επιλογής χαρακτηριστικών (feature selection) ξεπερνούν τα 100 [13]. Συνεπώς, ο ορισμός και η ανάλυσή τους θα ήταν μια επίπονη διαδικασία που υπερβαίνει το σκοπό της τρέχουσας αναφοράς.

Το σήμα ομιλίας συχνά διαχωρίζεται σε μικρότερα τμήματα, συνήθως σταθερού μήκους, που ονομάζονται *frames*. Τα χαρακτηριστικά που εξάγονται από κάθε *frame* ονομάζονται τοπικά (local), ενώ απεικονίσεις των τοπικών χαρακτηριστικών (όπως στατιστικές ροπές, ελάχιστο-μέγιστο κτλπ) ονομάζονται καθολικά (global) [14]. Ο συνδυασμός τοπικών και καθολικών χαρακτηριστικών σε ένα μοντέλο κατηγοριοποίησης φαίνεται να παρουσιάζει καλύτερα αποτελέσματα κατηγοριοποίησης από την χρήση ενός εκ των δύο [15], [16]. Δεδομένου πως τα χαρακτηριστικά μπορούν να διαχωριστούν σε περαιτέρω κατηγορίες [14], [17], θα αρκεστούμε στην κατηγοριοποίησή τους ανάλογα με το πεδίο εξαγωγής τους. Μερικά από τα συνήθη χαρακτηριστικά που χρησιμοποιούνται [14], [17], [18]-[22] παρατίθενται στον παρακάτω πίνακα.

Πεδίο χρόνου	Entropy, Zero Crossing Rate Energy: RMS, min, max, median, std, range, shimmer linear regression coefficients, 4th order Legendre parameters Duration: speech rate, ratio of voiced and unvoiced regions
Πεδίο Συχνότητας	Roll-off Frequency Fundamental Frequency (F_0): mean, contour slope, rate, reference line Formants: First and second formants and their bandwidths
Πεδίο Φάσματος	Centroid, Flatness, Spread, Irregularity, Kurtosis, Brightness, LPC, OSALPPC, SMC, LSMYWE
Πεδίο Σάφματος	MFCC, LPCC, OSALPCC

Πίνακας 2: *Συνήθη Χαρακτηριστικά SER*

4 Ταξινομητές

Ταξινομητές **βασισμένοι σε χαρακτηριστικά** όπως *SVM, MLP, HMM, Bayesian* αλλά και ομάδες ταξινομητών έχουν εξεταστεί ενδελεχώς στην βιβλιογραφία. Επισημαίνεται πως η σύγκριση των μοντέλων μεταξύ διαφορετικών μελετών αποτελεί κινδυνολογία, διότι τόσο τα σύνολα δεδομένων όσο και τα σύνολα χαρακτηριστικών που χρησιμοποιούνται εν γένει διαφέρουν.

Το μοντέλο *MLB* αποδίδει χειρότερα από αυτό του *k-NN* με βάση χαρακτηριστικά του τόνου φωνής (pitch) [23]. Μοντέλα *SVM, HMM* υπόσχονται ακρίβεια κατηγοριοποίησης άνω του 90% στο σύνολο δεδομένων *D.E.S* (Danish Emotional Speech) [24]. Το διάνυσμα χαρακτηριστικών για τα κρυφά μαρκοβιανά μοντέλα συντέθηκε από παραγώγους των παραμέτρων *F0, F1, MBE*. Στην ίδια μελέτη ο ταξινομητής *SVM* με συνάρτηση πυρήνα ακτινωτής βάσης (radial basis function) απέδωσε εξίσου ικανοποιητική ακρίβεια ταξινόμησης με διάνυσμα εισόδου συντελεστές *MEDC*. Μοντέλα *SVM* με πυρήνα *RBF* παρουσιάζουν καλύτερη ακρίβεια κατηγοριοποίησης απ΄ότι *MLP* τριών επιπέδων (με 16 νευρώνες στο κρυφό επίπεδο), ενώ αντίστοιχα αποτελέσματα παρουσιάζουν πιθανοτικά νευρωνικά δίκτυα [27].

Πιο πρόσφατες προσεγγίσεις χρησιμοποιούν τεχνικές βαθιάς μηχανικής μάθησης για την αντιμετώπιση του προβλήματος αναγνώρισης συναισθήματος από ομιλία. Συνελικτικά νευρωνικά δίκτυα έχουν χρησιμοποιηθεί για να εξάγουν χαρακτηριστικά από φασματόγραμμα, σε συνδυασμό με αναδρομικά νευρωνικά δίκτυα που προβαίνουν σε κατηγοριοποίηση [25, 26].

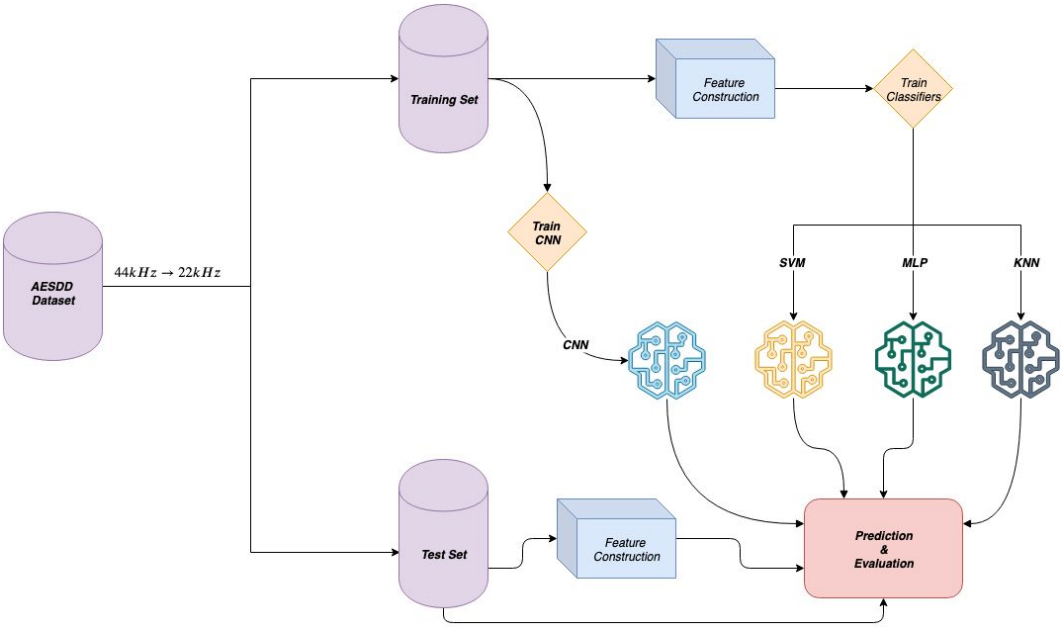
5 Μεθοδολογία

Το *dataset* που θα χρησιμοποιηθεί είναι το *AESDD* [9], [28]. Πρόκειται για ένα *actor-based dataset* πέντε συναισθημάτων (*anger, disgust, fear, happiness, sadness*) στην ελληνική γλώσσα. Το *dataset* αποτελείται από 603 προτάσεις (utterances) μέσης διάρκειας $4,1$ *sec*. Δεδομένου πως τα ακουστικά αρχεία στο *dataset* έχουν εγγραφεί με συχνότητα δειγματοληψίας $f_s = 44100\ Hz$, θα γίνει υποδειγματοληψία (resampling) με συχνότητα $f'_s = \frac{f_s}{2}$.

Σε ό,τι αφορά τους ταξινομητές βασισμένους σε χαρακτηριστικά, θα εξεταστούν *SVM*, *MLP* διαφορετικών παραμέτρων, καθώς και ο *k-NN*. Κάθε ηχητικό δείγμα χωρίζεται σε τμήματα (frames) μέσης διάρκειας 1.1 sec και από κάθε τμήμα εξάγονται τα ακόλουθα χαρακτηριστικά: 13 mfcc, rolloff frequency(30.50,70,85), zcr, rms.

Πέραν των ταξινομητών που βασίζονται σε χαρακτηριστικά, θα εξεταστεί και ένα συνελικτικό νευρωνικό δίκτυο (convolutional neural network) με είσοδο το φασματόγραμμα των σημάτων ήχου.

Τέλος, τα μοντέλα θα αξιολογηθούν και θα εξαχθούν τα μητρώα σύγχυσης (confusion matrices), καθώς και οι συνήθεις μετρικές που σχετίζονται με αυτά. Στο Σχ.1 παρουσιάζεται η ενδεικτική ροή εργασίας που θα ακολουθηθεί.



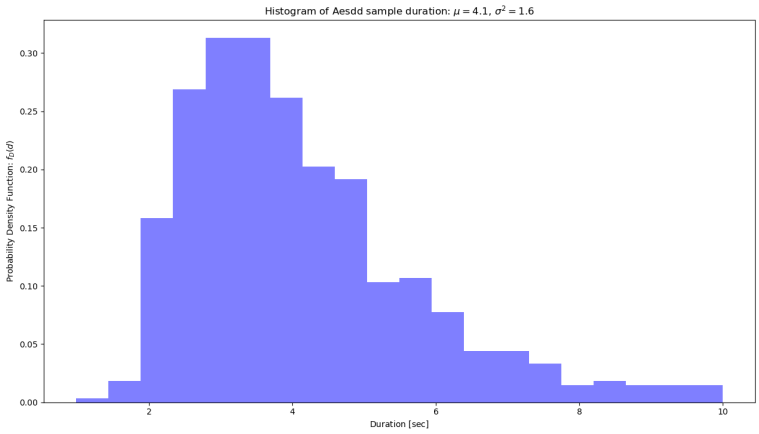
Σχήμα 1: *Ενδεικτική Ροή Εργασίας*

Όσον αφορά το τεχνικό σκέλος της εργασίας, θα χρησιμοποιηθεί η γλώσσα προγραμματισμού *python* σε συνδυασμό με τις βιβλιοθήκες *numpy, scipy, pandas, librosa, matplotlib*.

6 Αποτελέσματα Και Σχολιασμός

6.1 Συνελικτικό Νευρωνικό Δίκτυο

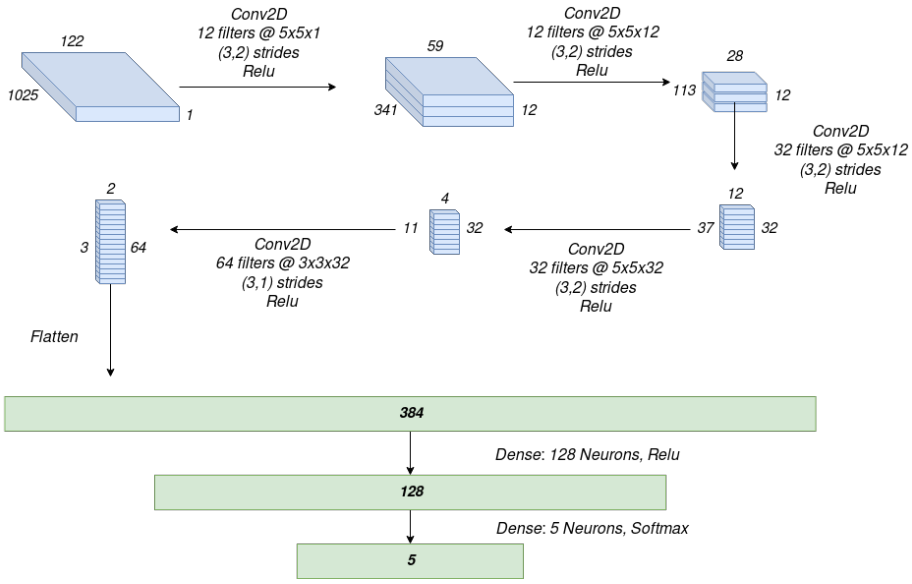
Στο σχήμα 2 παρουσιάζεται το ιστόγραμμα της χρονικής διάρκειας των δειγμάτων ήχου του συνόλου δεδομένων AESDD. Όπως έχει προαναφερθεί τα ηχητικά δείγματα υποδειγματοληπτούνται με λόγο $2:1$ από $f_s = 44.1\ kHz$ σε $f'_s = 22.050\ kHz$. Με στόχο την εξαγωγή φασματογραμμάτων ίδιας διάστασης για κάθε ηχητικό δείγμα, εφαρμόστηκε συμπλήρωση με μηδενικά (zero padding) ώστε κάθε δείγμα να έχει την ίδια χρονική διάρκεια.



Σχήμα 2: *Ιστόγραμμα Χρονικής Διάρκειας Ηχητικών Δειγμάτων AESDD*

Έτσι, εφαρμόζοντας παράθυρα 2048 δειγμάτων ($\sim 93\ ms$) με επικάλυψη 50% το φασματόγραμμα κάθε δείγματος έχει διαστάσεις μητρώου $1025\ X\ 122$.

Η αρχιτεκτονική του νευρωνικού δικτύου που επικράτησε απεικονίζεται λεπτομερώς στο σχήμα 3.

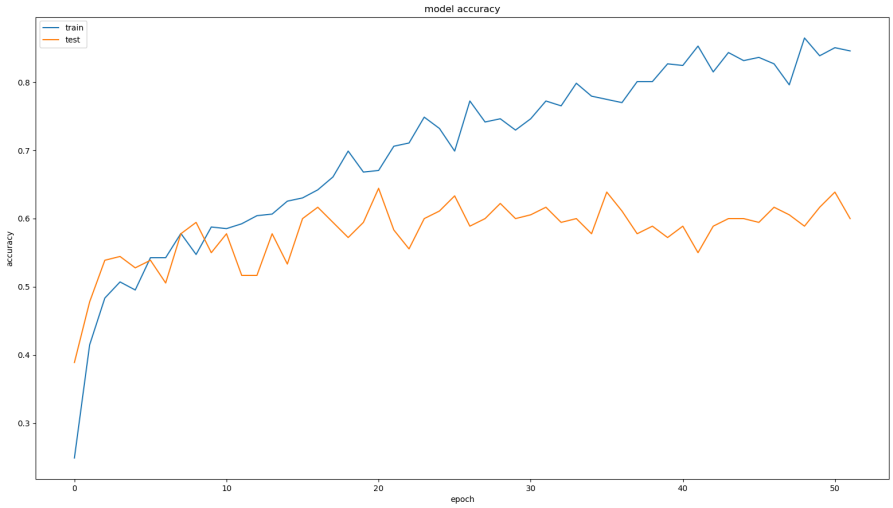


Σχήμα 3: *Αρχιτεκτονική Συνελικτικού Νευρωνικού Δικτύου*

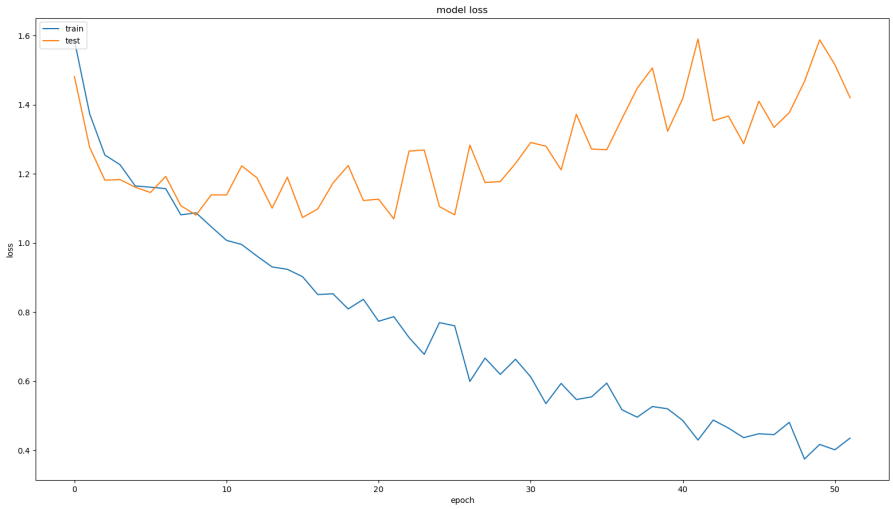
Σημειώνεται πως κατά την εκπαίδευση προστέθηκαν επίπεδα απόρριψης (dropout layers) με 0.3 πιθανότητα απόρριψης. Ως συνάρτηση κόστους επιλέχθηκε η κατηγορική διασταυρωμένη εντροπία (categorical crossentropy) και ως αλγόριθμος κατάβασης δυναμικού (gradient descent) επιλέχθηκε ο RMSprop. Το 70% του συνόλου δεδομένων ορίστηκε ως σύνολο εκπαίδευσης (training set) και το 30% ως σύνολο δοκιμής (test set).

Στο σχήμα 4 απεικονίζεται η διαδικασία εκπαίδευσης του συνελικτικού νευρωνικού δικτύου ανά εποχή. Επιλέχθηκε το μοντέλο με την μέγιστη ακρίβεια κατηγοριοποίησης **0.6444** επί του συνόλου δοκιμής.

Στο σχήμα 5 παρατίθεται το μητρώο σύγχυσης του συνελικτικού νευρωνικού δικτύου επί του συνόλου δοκιμής.



(α′) *Ακρίβεια*



(β′) *Συνάρτηση Κόστους*

Σχήμα 4: *Εκπαίδευση CNN Ανά Εποχή*

Σχήμα 5: *Μητρώο Σύγχυσης CNN*

Τα διανύσματα ακρίβειας P, ανάκλησης R προκύπτουν από το Σχ. 5, είναι:

$$P = \begin{bmatrix} 0.56 \\ 0.75 \\ 0.7667 \\ 0.3721 \\ 0.9655 \end{bmatrix}, \quad R = \begin{bmatrix} 0.8236 \\ 0.6364 \\ 0.5227 \\ 0.4444 \\ 0.8485 \end{bmatrix}$$

Αναφορές

[1] Kleinginna, Paul R., and Anne M. Kleinginna. "A categorized list of motivation definitions, with a suggestion for a consensual definition." *Motivation and emotion* 5.3 (1981): 263-291.

[2] Cabanac, Michel. "What is emotion?." *Behavioural processes* 60.2 (2002): 69-83.

[3] Mulligan, Kevin, and Klaus R. Scherer. "Toward a working definition of emotion." *Emotion Review* 4.4 (2012): 345-357.

[4] Cowie, Roddy, et al. "Emotion recognition in human-computer interaction." *IEEE Signal processing magazine* 18.1 (2001): 32-80.

[5] Ekman, Paul. "An argument for basic emotions." *Cognition & emotion* 6.3-4 (1992): 169-200.

[6] Kuusikko, Sanna, et al. "Emotion recognition in children and adolescents with autism spectrum disorders." *Journal of autism and developmental disorders* 39.6 (2009): 938-945.

[7] Castillo, José Carlos, et al. "A framework for recognizing and regulating emotions in the elderly." *International Workshop on Ambient Assisted Living*. Springer, Cham, 2014.

[8] Vaudable, Christophe, and Laurence Devillers. "Negative emotions detection as an indicator of dialogs quality in call centers." *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2012.

[9] Vryzas, Nikolaos, et al. "Speech emotion recognition for performance interaction." *Journal of the Audio Engineering Society* 66.6 (2018): 457-467.

[10] France, Daniel Joseph, et al. "Acoustical properties of speech as indicators of depression and suicidal risk." *IEEE transactions on Biomedical Engineering* 47.7 (2000): 829-837.

[11] Koolagudi, Shashidhar G., and K. Sreenivasa Rao. "Emotion recognition from speech: a review." *International journal of speech technology* 15.2 (2012): 99-117.

[12] Ververidis, Dimitrios, and Constantine Kotropoulos. "A state of the art review on emotional speech databases." *Proceedings of 1st Richmedia Conference.* 2003.

[13] Vogt, Thurid, and Elisabeth André. "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition." *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005.

[14] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." *Pattern Recognition* 44.3 (2011): 572-587.

[15] Hu, Hao, Ming-Xing Xu, and Wei Wu. "Fusion of global statistical and segmental spectral features for speech emotion recognition." *Eighth Annual Conference of the International Speech Communication Association.* 2007.

[16] Shami, Mohammad T., and Mohamed S. Kamel. "Segment-based approach to the recognition of emotions in speech." *2005 IEEE International Conference on Multimedia and Expo.* IEEE, 2005.

[17] Anagnostopoulos, Christos-Nikolaos, Theodoros Iliou, and Ioannis Giannoukos. "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011." *Artificial Intelligence Review* 43.2 (2015): 155-177.

[18] Alim, Sabur Ajibola, and Nahrul Khair Alang Rashid. "Some commonly used speech feature extraction algorithms." *From Natural to Artificial Intelligence-Algorithms and Applications* (2018).

[19] Busso, Carlos, et al. "Toward effective automatic recognition systems of emotion in speech." *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds* (2013): 110-127.

[20] Chen, Lijiang, et al. "Speech emotion recognition: Features and classification models." *Digital signal processing* 22.6 (2012): 1154-1160.

[21] Kotsakis, Rigas, George Kalliris, and Charalampos Dimoulas. "Investigation of broadcast-audio semantic analysis scenarios employing radio-programme-adaptive pattern classification." *Speech Communication* 54.6 (2012): 743-762.

[22] Vryzas, N., et al. "Augmenting Drama: A Speech Emotion-Controlled Stage Lighting Framework." *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences.* ACM, 2017.

[23] Dellaert, Frank, Thomas Polzin, and Alex Waibel. "Recognizing emotion in speech." *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96.* Vol. 3. IEEE, 1996.

[24] Lin, Yi-Lin, and Gang Wei. "Speech emotion recognition based on HMM and SVM." *2005 international conference on machine learning and cybernetics.* Vol. 8. IEEE, 2005.

[25] Lim, Wootaek, Daeyoung Jang, and Taejin Lee. "Speech emotion recognition using convolutional and recurrent neural networks." *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA).* IEEE, 2016.

[26] Satt, Aharon, Shai Rozenberg, and Ron Hoory. "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms." *INTERSPEECH*. 2017.

[27] Iliou, Theodoros, and Christos-Nikolaos Anagnostopoulos. "SVM-MLP-PNN classifiers on speech emotion recognition field-A comparative study." *2010 Fifth International Conference on Digital Telecommunications*. IEEE, 2010.

[28] Vryzas, Nikolaos, et al. "Subjective Evaluation of a Speech Emotion Recognition Interaction Framework." *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*. ACM, 2018.