Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Σηματων, Ελεγχου και Ρομποτικης

Χρήση βαθιών νευρωνικών δικτύων και πολλαπλοτήτων για την εκπαίδευση ακουστικού μοντέλου για αυτόματη αναγνώριση φωνής

Διπλωματικη Εργασια

του

ΙΩΑΝΝΗ Μ. ΧΑΛΚΙΑΔΑΚΗ

Επιβλέπων: Αλέξανδρος Ποταμιάνος
Αναπλ. Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2016

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Χρήση βαθιών νευρωνικών δικτύων και πολλαπλοτήτων για την εκπαίδευση ακουστικού μοντέλου για αυτόματη αναγνώριση φωνής

Διπλωματικη Εργασια

του

ΙΩΑΝΝΗ Μ. ΧΑΛΚΙΑΔΑΚΗ

Επιβλέπων: Αλέξανδρος Ποταμιάνος
Αναπλ. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 10η Ιουνίου 2016.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Αλέξανδρος Ποταμιάνος	Πέτρος Μαραγκός	Shrikanth Narayanan
Αναπλ. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Professor U.South. California

Αθήνα, Ιούνιος 2016

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

.....

Ιωάννης Μ. Χαλκιαδάκης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Ο στόχος της παρούσας μελέτης ήταν να μελετήσει αρχιτεκτονικές βαθιών νευρωνικών δικτύων οι οποίες έχουν λάβει τεράστια προσοχή κατά τη διάρκεια των τελευταίων ετών, λόγω της επιτυχίας τους σε εφαρμογές που ενδιαφέρουν την επιστημονική κοινότητα μηχανικής μάθησης.

Το πεδίο εφαρμογής που επιλέξαμε ήταν η αυτόματη αναγνώριση φωνής, δεδομένου ότι οι περισσότερες ανακαλύψεις στα βαθιά νευρωνικά δίκτυα παρουσιάστηκαν για πρώτη φορά σε εφαρμογές αναγνώρισης φωνής. Επιπλέον, υιοθετήσαμε μια προσέγγιση με χρήση πολλαπλοτήτων για την βελτίωση του κριτηρίου εκπαίδευσης του δικτύου. Η ιδέα (Tomar και Rose, 2014) είναι ότι αν καταφέρουμε να διατηρήσουμε, μέσω του δικτύου, τις σχέσεις των δεδομένων εισόδου που επιβάλονται από τη δομή της πολλαπλότητας, θα μάθουμε μια πιο ακριβή και εύρωστη κατανομή των κλάσεων φωνημάτων που βρίσκονται στα δεδομένα εισόδου. Ο αλγόριθμος που θα διατηρήσει τις σχέσεις των δεδομένων εισόδου που επιβάλονται από τη δομή της πολλαπλότητας, χρησιμοποιεί τις κλάσεις φωνημάτων και τις αποστάσεις μεταξύ των χαρακτηριστικών της φωνής για να μάθει την υποκείμενη πολλαπλότητα.

Αρχικά δίνουμε μια εισαγωγή στο χώρο της αυτόματης αναγνώρισης ομιλίας με βαθιά νευρωνικά δίκτυα. Στη συνέχεια περιγράφουμε λεπτομερώς τη δουλειά που πραγματοποιήσαμε, τον τρόπο που ενσωματώσαμε την πολλαπλότητα στο βαθύ νευρωνικό δίκτυο, καθώς και τις προκλήσεις που αντιμετωπίσαμε κατά τη διάρκεια της εργασίας. Τέλος, παρουσιάζονται τα πειραματικά αποτελέσματα και επακόλουθες παρατηρήσεις.

Επιπλέον πληροφορίες για την εργασία μπορούν να βρεθούν στο repository
https://ychalkiad@bitbucket.org/ ychalkiad/lpda.git.

Λέξεις Κλειδιά

βαθιά νευρωνικά δίκτυα, μηχανική μάθηση, μάθηση πολλαπλοτήτων, κοντινότεροι γείτονες, αυτόματη αναγνώριση φωνής, αυτόματη αναγνώριση φωνής συνεχούς λόγου και μεγάλου

λεξιλογίου, ακουστικό μοντέλο, υβριδικό ακουστικό μοντέλο

Abstract

The goal of the current project was to study deep architectures of neural networks which have received tremendous attention during the past few years, because of their success in tasks of interest to the machine learning community.

The application field that we selected was automatic speech recognition, given that most breakthroughs in deep learning have first occurred in speech recognition tasks. In addition, we adopted a manifold approach for the regularization of the training criterion of the network. The idea (Tomar and Rose, 2014) is that, if we manage to maintain the manifold-constrained relationships of speech input data through the network, we will learn a more accurate and robust against noise distribution over speech units. The algorithm that will maintain the manifold-imposed relations uses classes of speech units and distances between speech features to learn the underlying manifold.

We first give an introduction to the area of automatic speech recognition with deep neural networks and then describe in detail the manifold regularized network we built, the way we incorporated the manifold criterion in the deep neural network as well as challenges we faced during development. Finally, experimental results and subsequent remarks are given.

Extra information about the project can be found in https://ychalkiad@bitbucket.org/ychalkiad/lpda.git.

Keywords

deep neural networks, machine learning, manifold learning, manifold regularization, graph embedding framework, intrinsic graph, penalty graph, approximate nearest neighbors, kd-trees, coordinate patch, automatic speech recognition, large vocabulary continuous speech recognition, acoustic modeling, tandem acoustic modeling, hybrid acoustic modeling, Theano, Julia, Kaldi

Περιεχόμενα

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

Στα επόμενα κεφάλαια θα παρουσιάσουμε τη διαδικασία που απαιτείται για την εκπαίδευση και ενσωμάτωση ενός νευρωνικού δικτύου με πολλά επίπεδα (deep neural network - DNN) σε ενα σύστημα αναγνώρισης φωνής.

Θα περιγράψουμε τις προπαρασκευαστικές ενέργειες που απαιτούνται για την εξαγωγή των χαρακτηριστικών που θα παρουσιαστούν στο δίκτυο, θα συνεχίσουμε με την ενσωμάτωση του DNN στο σύστημα ASR και του manifold όρου όπως περιγράφεται στο [82], και τελικά θα παρουσιάσουμε τα πειραματικά αποτελέσματα που αποκτήθηκαν με το νέο σύστήμα.

Κεφάλαιο 2

Βαθιά Νευρωνικά Δίκτυα και Αναγνώριση Φωνής

2.0.1 DNN και ακουστικό μοντέλο

Υπάρχουν δύο βασικοί τρόποι με τους οποίους μπορούμε να χρησιμοποιήσουμε βαθιά νευρωνικά δίκτύα στο ακουστικό μοντέλο [92]:

- η *υβριδική* προσέγγιση, όπου υπολογίζουμε μέσω του DNN την πιθανότητα παρατήρησης ενός διανύσματος ακουστικών χαρακτηριστικών που χρησιμοποιείται στο Κρυφό Μαρκοβιανό Μοντέλο ενός συστήματος αυτόματης αναγνώρισης φωνής. Το Κρυφό Μαρκοβιανό Μοντέλο έχει προηγουμένως εκπαιδευτεί με ένα μοντέλο Γκαουσιανών κατανομών (Gaussian Mixture Model)

- η *συνδυαστική* προσέγγιση, όπου εξάγουμε ένα μετασχηματισμό των χαρακτηριστικών εκπαίδευσης από ένα από τα επίπεδα του DNN και τα παρουσιάζουμε σαν είσοδο σε ένα συμβατικό σύστημα αναγνώρισης φωνής GMM/HMM.

Στην τρέχουσα εργασία χρησιμοποιήσαμε την υβριδική προσέγγιση, την οποία παρουσιάζουμε στην συνέχεια.

Υβριδική προσέγγιση

Το σύστημα DNN/HMM που είναι το αποτέλεσμα της υβριδικής προσέγγισης συνδυάζει τη δύναμη του DNN, δηλαδή, παραστατική δύναμή του, με τα ωφέλη του HMM, δηλαδή, την ικανότητα μοντελοποίησης πληροφορίας με χρονική συνοχή (sequential information).

Η συνδυαστική χρήση νευρωνικών δικτύων και HMM ξεκίνησε από τα τέλη της δεκαετίας του 1980 και τις αρχές της δεκαετίας του 1990, ωστόσο είχαν εφαρμοστεί μόνο σε εργασίες μικρού λεξιλογίου. Η έρευνα στον τομέα αναστήθηκε και πάλι αφότου τα νευρωνικά δίκτυα παρουσίασαν την ισχυρή παραστατική τους δύναμη και απέδωσαν καλά σε εφαρμογές αναγνώρισης συνεχούς ομιλίας.

Σε αυτά τα συστήματα είναι η δυναμική/χρονική πληροφορία του σήματος φωνής μοντελοποιείται με τα HMM και οι πιθανότητες παρατήρησης υπολογίζονται μέσω του νευρωνικού δικτύου: κάθε νευρώνας εξόδου έχει εκπαιδευτεί για να εκτιμηθεί η πιθανότητα εκπομπής του κομματιού του σήματος που παρουσιάστηκε στην είσοδο από μια κατάσταση του HMM μοντέλου ενός φωνήματος. Από μαθηματική άποψη η έξοδος του DNN διαμορφώνεται ως εξής:

$$P(q_t = s|\mathbf{x}_t), \forall s \in [1, \mathcal{S}]$$

όπου s είναι η κατάσταση του HMM και $\mathbf{x}_t$ είναι το διάνυσμα εισόδου των ακουστικών χαρακτηριστικών. Στα πιο πρόσφατα συστήματα DNN μοντελοποιούμε απευθείας *σενονες*, δηλαδή δεμένες HMM καταστάσεις τριφώνων. Αυτό έχει βελτιώσει όχι μόνο τις επιδόσεις ταξινόμησης του δικτύου, αλλά έχει δύο επιπλέον πλεονεκτήματα: πρώτον, ένα σύστημα DNN/HMM μπορεί να κατασκευαστεί από ένα υπάρχον GMM/HMM απαιτώντας μόνο ελάχιστες τροποποιήσεις, και δεύτερον, κάθε εξέλιξη στη μοντελοποίηση φωνημάτων από ένα GMM/HMM μπορεί εύκολα να ενσωματωθεί στο σύστημα DNN/HMM, δεδομένου ότι η βελτίωση θα αντανακλά άμεσα στις μονάδες εξόδου του δικτύου.

Δεδομένου ότι το HMM απαιτεί την πιθανότητα $p(\mathbf{x_t}|q_t)$ αντί απευθείας την έξοδο του νευρωνικού κατά τη διαδικασία αποκωδικοποίησης, η έξοδος DNN πρέπει να μετατραπεί ως εξής:

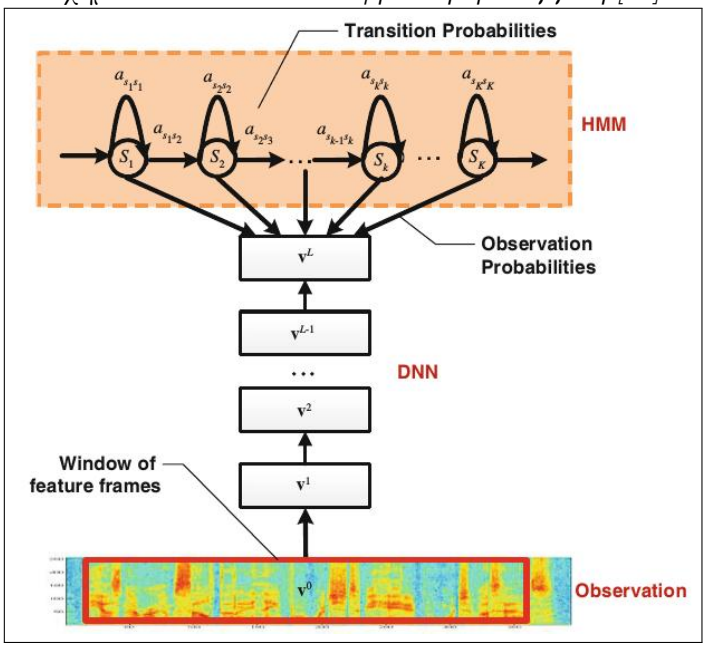
$$p(\mathbf{x_t}|q_t = s) = \frac{p(q_t = s|\mathbf{x}_t)p(\mathbf{x}_t)}{p(s)}$$

όπου $p(s)$ είναι η εκ των προτέρων πιθανότητα του κάθε senone που υπολογίζεται από το σύνολο εκπαίδευσης και ορίζεται ως:

$$p(s) = \frac{T_s}{T}$$

όπου T είναι ο συνολικός αριθμός των δεδομένων εκπαίδευσης, T_s ο αριθμός των δεδομένων που κατηγοριοποιούνται ως s, ενώ η $p(\mathbf{x_t})$ δεν επηρεάζει την αποκωδικοποίηση και μπορεί

Σχήμα 2.1: *DNN-HMM υβριδική προσέγγιση [92]*



να αγνοηθεί. Η συμβολή της εκ των προτέρων πιθανότητας στην ακρίβεια της αναγνώρισης δεν είναι μεγάλη αλλά μπορεί να είναι σημαντική στη μείωση του προβλήματος biasing των κλάσεων.

Τα πιο σημαντικά σημεία για να είναι επιτυχής η εφαρμογή των βαθιών δικτύων στην αναγνώριση φωνής συνοψίζονται στα εξής:

- το βάθος του δικτύου, δηλαδή πόσα ενδιάμεσα επίπεδα υπολογισμού υπάρχουν. Πειραματικά έχει επιβεβαιωθεί ότι η απόδοση βαθιών δικτύων είναι καλύτερη από αυτή ρηχών, ακόμα κι αν ο συνολικός αριθμός νευρώνων είναι ο ίδιος

- η είσοδος στο δίκτυο ενός παραθύρου διανυσμάτων χαρακτηριστικών και όχι μεμονωμένων διανυσμάτων, αφού ένα βαθύ δίκτυο μπορεί να αξιοποιήσει πληροφορία συσχετισμένων (correlated) διανυσμάτων

- η χρήση δεμένων τριφωνικών καταστάσεων σαν έξοδο του νευρωνικού, η οποία μειώνει τον κίνδυνο υπερεκπαίδευσης του δικτύου

2.0.2 Εκπαίδευση του αρχικού GMM/HMM

Το πρώτο βήμα πριν τη χρήση ενός DNN στην αυτόματη αναγνώριση φωνής είναι η εκπαίδευση ενός βασικού GMM/HMM συστήματος από το οποίο θα πάρουμε τις ετικέτες (labels)

Πίνακας 2.1: PFile format

Sentence	Frame	Feature Vector	Label
0	0	[0.2, 0.3, 0.5, 1.4, 1.8, 2.5]	10
0	1	[1.3, 2.1, 0.3, 0.1, 1.4, 0.9]	179
1	0	[0.3, 0.5, 0.5, 1.4, 0.8, 1.4]	32

που θα χρησιμοποιηθούν για την εκπαίδευση του δικτύου. Για το σκοπό αυτό χρησιμοποιήσαμε το σύστημα Kaldi και τις 2000 συντομότερες φράσεις του συνόλου δεδομένων της Wall Street Journal.

Η δομή που χρησιμοποιήσαμε για να αποθηκεύσουμε τα δεδομένα εισόδου ήταν τα PFiles που αναπτύχθηκαν στο Berkeley για εφαρμογές μηχανικής μάθησης. Η μορφή ενός τέτοιου αρχείου συνοψίζεται στον πίνακα .

Κεφάλαιο 3

Ενσωματώνοντας το DNN

Η γλώσσα προγραμματισμού που επιλέχθηκε για την εργασία ήταν η Python και η βιβλιοθήκη για βαθιά νευρωνικά δίκτυα Theano.

3.0.1 Δεδομένα εισόδου

Οι 39-διαστάσεων MFCC (συμπεριλαμβανομένης της ενέργειας και πρώτης και δεύτερης παραγώγου) επιλέχθηκαν ως χαρακτηριστικά εισόδου. Όπως έχουμε ήδη δει, τα βαθιά νευρωνικά αποδίδουν καλύτερα όταν κάθε διάνυσμα εισόδου χαρακτηριστικών παρουσιάζεται με ένα παράθυρο γειτονικών διανυσμάτων. Ως εκ τούτου, και ακολουθώντας τις συμβουλές από εδώ [82], παρουσιάσαμε στο δίκτυο ένα παράθυρο 9 διανυσμάτων συνολικά: κάθε διάνυσμα μαζί με 4 γειτονικά στην αριστερή και δεξιά πλευρά.

3.0.2 Αρχιτεκτονική και εκπαίδευση του δικτύου

Κριτήριο εκπαίδευσης

Το κριτήριο εκπαίδευσης που χρησιμοποιήθηκε ήταν η συνάρτηση εντροπίας (cross entropy), λόγω της ταχύτερης και πιο εύρωστης σύγκλιση της.

Αλγόριθμος εκπαίδευσης και μέγεθος batch

Ο αλγόριθμος που χρησιμοποιείται για την εκπαίδευση είναι ο αλγόριθμος οπισθοδιάδοσης (back-propagation). Η συνάρτηση κόστους ελαχιστοποιείται χρησιμοποιώντας minibatch gradient descent για την αναζήτηση στο χώρο βαρών. Το μέγεθος του minibatch παρέμεινε σταθερό κατά τη διάρκεια της εκπαίδευσης και διάφορα μεγέθη δοκιμάστηκαν με το πιο επιτυχημένο να είναι ένα μέγεθος των 256 διανυσμάτων.

Ρυθμός μάθησης και ορμή

Η προσεκτική επιλογή του ρυθμού μάθησης είναι σημαντική για τη σύγκλιση και την ταχύτητας σύγκλισης της εκπαίδευσης. Στη δική μας προσέγγιση, διαφορετικοί ρυθμοί μάθησης δοκιμάστηκαν και δύο στρατηγικές χρησιμοποιήθηκαν για μείωση τους: είτε μείωση κατά το ήμισυ είτε εκθετική μείωση κάθε φορά που το ποσοστό σφάλματος επικύρωσης αυξανόταν.

Ένας όρος ορμής συμπεριλήφθηκε στην ανανέωση των παραμέτρων του δικτύου, ελπίζοντας να βελτιώσει την ταχύτητα σύγκλισης και να σταθεροποιήσει τον αλγορίθμο οπισθοδιάδοσης, όπως κι έγινε. Ο όρος ορμής ενσωματώθηκε ως εξής:

$$v_{t+1} = \mu v_t + (1 - \mu)\epsilon \nabla f(\theta_t)$$

$$\theta_{t+1} = \theta_t + v_{t+1}$$

όπου ϵ είναι ο ρυθμός μάθησης, μ είναι η ορμή, f είναι η συνάρτηση εκπαίδευσης και θ είναι οι παράμετροι που ανανεώνονται. Επιπλέον, η ορμή μ σταδιακά αυξανόταν με το πέρασμα των εποχών εκπαίδευσης.

Αρχιτεκτονική του δικτύου

Το δίκτυο που χρησιμοποιείται για το ακουστικό μοντέλο είναι ένα τυπικό perceptron πολλαπλών επιπέδων με τουλάχιστον τέσσερα κρυμμένα επίπεδα. Δοκιμάσαμε διαφορετικού μεγέθους επίπεδα, αλλά όλα είχαν πάντα το ίδιο μέγεθος. Δεν δοκιμάστηκαν αρχιτεκτονικές με μορφή πυραμίδας, δεδομένου ότι δεν βελτιώνουν την απόδοση σύμφωνα με τη βιβλιογραφία. Το επίπεδο εισόδου είχε 117 διαστάσεις «4+1+4)×13 χαρακτηριστικά). Το επίπεδο εξόδου ήταν ένα softmax επίπεδο για τον σκοπό της ταξινόμησης και η αρχιτεκτονική του είναι κάτι που χρειάζεται περισσότερη προσοχή.

Οι στόχοι της εκπαίδευσης και κατά συνέπεια οι διαστάσεις του επιπέδου εξόδου εξαρτώνται από το σύστημα αποκωδικοποίησης που χρησιμοποιείται. Τα περισσότερα state-of-the-art συστήματα αναγνώρισης ομιλίας χρησιμοποιούν ως στόχους ταξινόμησης κωδικούς που αντιστοιχούν σε κατανομές συνάρτησης πιθανότητας, για παράδειγμα το Kaldi, χρησιμοποιεί *pdf-ids* για την αποκωδικοποίηση και συνεπώς, επειδή θα εισάγουμε το μοντέλο μας στο ίδιο σύστημα, χρησιμοποιήσαμε τους ίδιους στόχουν κατά την εκπαίδευση.

Κεφάλαιο 4

Ενσωμάτωση του όρου πολλαπλότητας

Το έργο της παρούσας εργασίας ήταν να εισαγάγει τις ιδέες που περιγράφονται στο **;;** σε μια εφαρμογή αναγνώρισης συνεχούς και μεγάλου λεξιλογίου ομιλίας, η οποία χρησιμοποιεί κομμάτι του συνόλου δεδομένων της Wall Street Journal.

Στην παρούσα εργασία επιλέξαμε την υβριδική προσέγγιση, σε αντίθεση με τους συγγραφείς της αρχικής εργασίας στην οποία βασιστήκαμε.

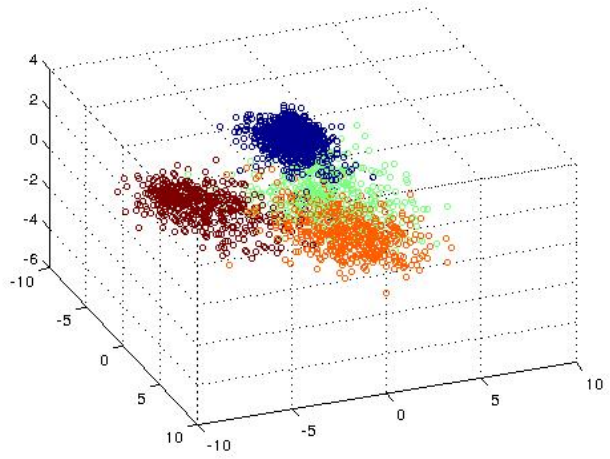
4.0.1 Ανακαλύπτοντας την δομή της πολλαπλότητας

Όπως και οι περισσότεροι αλγόριθμοι μάθησης πολλαπλοτήτων, η προσέγγιση που περιγράφεται στο [82] και [81] αρχίζει με την κατασκευή δύο γράφων αναπαράστασης γειτονιών που θα περιγράφουν τις σχέσεις μεταξύ των διανυσμάτων δεδομένων, όπως αυτές περιορίζονται πάνω στην πολλαπλότητα.

Ο ένας γράφος ($\mathbf{W}_{int}$) κατασκευάζεται λαμβάνοντας υπόψη μόνο γείτονες που έχουν την ίδια σήμανση (ετικέτα) με το εκάστοτε σημείο, ενώ ο δεύτερος ($\mathbf{W}_{pen}$) περιέχει μόνο τους γείτονες που ανήκουν σε διαφορετική φωνητική τάξη (και άρα έχουν διαφορετική σήμανση).

Οι ετικέτες που χρησιμοποιούνται για την κατασκευή των γράφων είναι υψίστης σημασίας για την επιτυχία της ανακάλυψης της πολλαπλότητας και την βελτιωμένη απόδοση του συστήματος. Θα υπέθετε κανείς ότι οι ετικέτες είναι οι ίδιες που το νευρωνικό θα χρησιμοποιήσει κατά τη διάρκεια της εκπαίδευσης. Ωστόσο, αυτό είναι σωστό μόνο εάν ο αποκωδικοποιητής ομιλίας που θα χρησιμοποιηθεί χρησιμοποιεί τις τριφωνικές καταστάσεις των Κρυφών Μαρκοβιανών Μοντέλων. Στην περίπτωσή μας, όπου χρησιμοποιείται ο αποκωδικοποιητής

Σχήμα 4.1: *LPDA, 3D projection, kpen=kint = 600, Rint=850,Rpen=3000, 2.5k data, phone-HMMstate label*



του Kaldi, ο οποίος βασίζεται σε αναγνωριστικούς κωδικούς των Γκαουσιανών κατανομών, οι οποίοι ήταν οι στόχοι για την εκπαίδευση του DNN, η εν λόγω υπόθεση δεν ισχύει.

Αυτό οφείλεται στο γεγονός ότι οι συγκεκριμένες ετικέτες δεν έχουν καμία φυσική έννοια πάνω στην πολλαπλότητα των φωνητικών μονάδων και, επιπλέον, δεν υπάρχει μια ένα-προς-ένα απεικόνιση μεταξύ των φυσικών τριφωνικών καταστάσεων και των Γκαουσιανών κατανομών. Ως εκ τούτου, δεν είναι σε θέση να βοηθήσουν στη διάκριση μεταξύ των φωνητικών κατηγοριών.

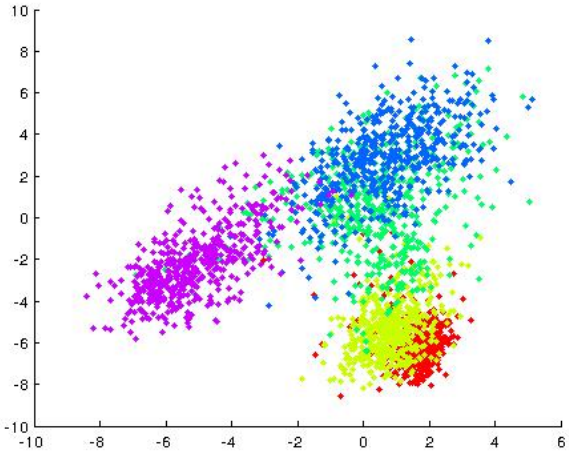
Οι εναλλακτικές ετικέτες που θα μπορούσαμε να χρησιμοποιήσουμε ήταν είτε τα φωνήματα είτε οι φυσικές τριφωνικές καταστάσεις των HMM. Επιλέξαμε τις τελευταίες, και με ένα υποσύνολο των δεδομένων εκπαίδευσης παραστήσαμε γραφικά το διαχωρισμό που επιτυγχάνεται από την εκμάθηση της πολλαπλότητας:

Κατά συνέπεια, μόνο για την κατασκευή των γράφων, χρησιμοποιήσαμε τα ζευγάρια (φώνημα-κατάσταση HMM) ως ετικέτες, και κατά τη διάρκεια της εκπαίδευσης του νευρωνικού για την υβριδική προσέγγιση χρησιμοποιήσαμε τους Γκαουσιανούς κωδικούς σαν στόχους.

Για τα βάρη που περιγράφουν τις σχέσεις μεταξύ των διανυσμάτων δεδομένων που χρησιμοποιούνται, έγινε ο εξής υπολογισμός με βάση τον οποίο συμπληρώθηκε η κατασκευή των γράφων:

$$w_{ij}^{int} = \begin{cases} e^{\frac{-\|x_i - x_j\|^2}{\rho}} & \text{ιφ } C(x_i) = C(x_j),\ e(x_i, x_j) = 1 \\ 0 & \text{αλλιώς} \end{cases}$$

Σχήμα 4.2: *LPDA, 2D projection, kpen=kint = 600, Rint=850,Rpen=3000, 2.5k data, phone-HMMstate label*



$$w_{ij}^{pen} = \begin{cases} e^{\frac{-\|x_i - x_j\|^2}{\rho}} & \text{ιφ } C(x_i) \neq C(x_j),\ e(x_i, x_j) = 1 \\ 0 & \text{αλλιώς} \end{cases}$$

όπου ρ είναι η παράμετρος του heat kernel, $C(x_i)$ η ετικέτα του διανύσματος δεδομένων x_i και $e(x_i, x_j) = 1$ σημαίνει ότι το x_i είναι γείτονας του x_j.

Προσθέτοντας τον όρο της πολλαπλότητας στο κριτήριο εκπαίδευσης του δικτύου έχουμε:

$$\mathcal{F}(\mathbf{W}; \mathbf{Y}) = \frac{1}{N} \sum_{i=1}^{N} \left\{ V(\mathbf{x}_i, \mathbf{t}_i, f) + \gamma \sum_{j=1}^{2k} \|\mathbf{y}_i - \mathbf{y}_j\|^2 w_{ij} \right\}$$

όπου V είναι το κριτήριο εκπαίδευσης που χρησιμοποιείται για το σύστημά μας, δηλαδή η συνάρτηση cross entropy, και το υπόλοιπο είναι ο όρος που επιβάλλει η πολλαπλότητα: g είναι το βάρος υπό το οποίο συμμετάσχει στο τελικό κόστος, $\mathbf{y}_i$ είναι η έξοδος του δικτύου που αντιστοιχεί στο παράθυρο εισόδου $\mathbf{x}_i$, $\mathbf{y}_j$ είναι η έξοδος του δικτύου που αντιστοιχεί στο γείτονα $\mathbf{x}_j$, $j = 1, ..., K$ συμπεριλαμβανομένων των γειτόνων και από τους δύο γράφους, και w_{ij} είναι το βάρος που περιγράφει τη σχέση μεταξύ $\mathbf{x}_i$ και $\mathbf{x}_j$ και είναι η διαφορά των αντίστοιχων εγγραφών στους δύο γράφους:

$$w_{ij} = w_{ij}^{int} - w_{ij}^{pen}$$

Με την προσθήκη του παραπάνω όρου, η παράγωγος του κριτηρίου εκπαίδευσης η οποία

θα συμμετάσχει στην ανανέωση των παραμέτρων του δικτύου, γίνεται:

$$\nabla_{\Theta_{n,m}} \mathcal{F} = \nabla_{\Theta_{n,m}} V + C \sum_{j=1}^{K} w_{ij}(y_{i,m} - y_{j,m}) \left(\frac{\partial y_{i,m}}{\partial \theta_{n,m}} - \frac{\partial y_{j,m}}{\partial \theta_{n,m}} \right)$$

Κεφάλαιο 5

Πειραματικά αποτελέσματα και συνεισφορές

Στην ενότητα αυτή θα παρουσιάσουμε τα αποτελέσματα των πειραμάτων που εκτελέστηκαν και θα συνοψίσουμε ορισμένες παρατηρήσεις. Τα αποτελέσματα συνοψίζονται στους πίνακες 5.1 (σύστημα εκπαιδευμένο με μονόφωνα) και 5.2 (σύστημα εκπαιδευμένο με τρίφωνα).

Η βάση για τη σύγκριση είναι το σύστημα GMM/HMM Kaldi, εκπαιδευμένο στις 2000 συντομότερες προτάσεις του συνόλου δεδομένων WSJ και αξιολογήθηκαν για τα δύο συνοδευτικά σύνολα δεδομένων: *dev93* και *eval92*.

Η ακρίβεια αποκωδικοποίησης που επιτυγχάνεται με το σύστημα τριφώνων είναι 76.15% στο *dev93* και 82.62% στο *eval92*, ενώ το σύστημα μονοφώνων επιτυγχάνει αντίστοιχα 64.87% και 74.45% για τα δύο σύνολα δοκιμών.

Πίνακας 5.1: Ακρίβεια αποκωδικοποίησης μονοφωνικού συστήματος (σε ποσοστό επί τοις εκατό %)

Σύστημα	**Ακρίβεια στο eval92**	**Ακρίβεια στο dev93**
GMM/HMM (Kaldi)	74.45	64.87
5x600 sigmoid	78.84	69.6
4x1024 sigmoid	79.83	69.34
4x1024 tanh	77.30	67.97
5x1024 ReLU + dropout	80.08	71.12
5x1024 sigmoid	78.5	68.97
5x1024 sigmoid + manifold	80.56	70.23

Πίνακας 5.2: Ακρίβεια αποκωδικοποίησης τριφωνικού συστήματος (σε ποσοστό επί τοις εκατό %)

Σύστημα	**Ακρίβεια στο eval92**	**Ακρίβεια στο dev93**
GMM/HMM (Kaldi)	82.62	76.15
5x1024 ReLU + dropout	86.87	80.00
5x1024 ReLU + manifold	87.19	79.27
5x1024 ReLU + dropout + manifold	88.06	81.81
5x1024 sigmoid	85.11	78.43
5x1024 sigmoid + manifold	86.66	80.38

5.1 Συνεισφορές

Οι κύριες συνεισφορές του έργου για την τρέχουσα έρευνα μπορούν να συνοψιστούν ως εξής:

- Κατασκευάσαμε ένα ακουστικό μοντέλο για αυτόματη αναγνώριση ομιλίας, χρησιμοποιώντας ένα βαθύ νευρωνικό δίκτυο, που είχε εκπαιδευτεί επιβάλλοντας περιορισμούς στη δομή του ακουστικού χώρου. Χρησιμοποιήσαμε αυτό το ακουστικό μοντέλο σε συνδυασμό με το Kaldi για την αναγνώριση της ομιλίας. Απ΄ όσο γνωρίζουμε τη στιγμή της συγγραφής, παρόμοιες προσεγγίσεις για την ακουστική μοντελοποίηση δεν έχουν χρησιμοποιήσει παρόμοιους περιορισμούς στο νευρωνικό. Η αρχική εργασία πάνω στην οποία βασιστήκαμε χρησιμοποιεί ένα βαθύ δίκτυο για την εξαγωγή εύρωστων χαρακτηριστικών και στη συνέχεια τα χρησιμοποιεί για να εκπαιδεύσει ένα ακουστικό μοντέλο. Αντ΄ αυτού, εμείς έχουμε εκπαιδεύσει ένα βαθύ δίκτυο που συμμετέχει άμεσα στην εκπαίδευση του ακουστικού μοντέλου, χωρίς την ανάγκη μακρόχρονου και επίπονου πειραματισμού, προκειμένου να εξαχθούν καλά χαρακτηριστικά και στη συνέχεια να τα χρησιμοποιηθούν στην αναγνώριση φωνής.

- Διαπιστώσαμε ότι το dropout λειτουργεί καλύτερα ως μέθοδος κανονικοποίησης από τους περιορισμούς πάνω στην πολλαπλότητα για την εκπαίδευση ενός ακουστικού μοντέλου βασισμένου σε DNN. Αυτό ήταν ένα θέμα υπό έρευνα στην αρχική εργασία, ωστόσο, πρέπει να σημειωθεί ότι αυτό διαπιστώθηκε χρησιμοποιώντας μια διαφορετική προσέγγιση (hybrid vs. tandem) για τη χρήση βαθιών δικτύων στην αναγνώριση φωνής. Αυτό που παρατηρήθηκε ήταν ότι, αν και τα συστήματα που χρησιμοποιούν τις δύο μεθόδους παρουσιάζουν περίπου την ίδια ακρίβεια αποκωδικοποίησης, το σύστημα με dropout είχε μεγαλύτερο σφάλμα στο σύνολο εκπαίδευσης το οποίο σημαίνει ότι κατάφερε να αποκτήσει τις ίδιες δυνατότητες αποκωδικοποίησης έχοντας μάθει λιγότερο

θόρυβο από το σύνολο εκπαίδευσης.

Επιπλέον, το dropout και οι μέθοδοι κανονικοποίησης πάνω στην πολλαπλότητα μπορούν να συνδυαστούν και να αλληλοσυμπληρωθούν, το οποίο παρείχε τα καλύτερα αποτελέσματα στο έργο μας.

- Έχουμε ξεκινήσει μια προσπάθεια να συνδυάσουμε αρθρωτά χαρακτηριστικά με το ακουστικό μοντέλο που κατασκευάσαμε.

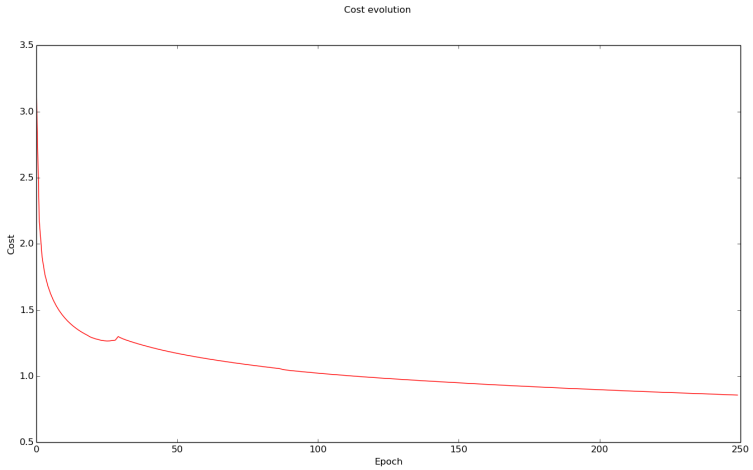
  Η ιδέα που αναπτύσσουμε βασίζεται στην ανακάλυψη μιας αντιστοιχίας από τον ακουστικό χώρο στο χώρο άρθρωσης (articulatory inversion) και στη συνέχεια πίσω στον ακουστικό χώρο, και στη χρήση των νέων χαρακτηριστικών στην εκπαίδευση ενός ακουστικού μοντέλου. Τα χαρακτηριστικά άρθρωσης έχουν αποδειχθεί αποτελεσματικά στην αναγνώριση φωνής όταν χρησιμοποιούνται συμπληρωματικά με τα παραδοσιακά ακουστικά χαρακτηριστικά, κι ελπίζουμε ότι μια νέα αναπαράσταση που προέρχεται από ένα autoencoding δίκτυο του οποίου το τμήμα κωδικοποίησης εκτελεί articulatory inversion και το τμήμα αποκωδικοποίησης γυρίζει πίσω στον ακουστικό χώρο, θα είναι επιτυχής στην αναγνώριση φωνής χωρίς την ανάγκη για συμπληρωματικά ακουστικά χαρακτηριστικά. Τα πρώτα αποτελέσματα χρησιμοποιώντας ένα μικρό dataset (MNGU0) είναι ενθαρρυντικά για τις εκτιμήσεις μας.

Παράρτημα Α΄

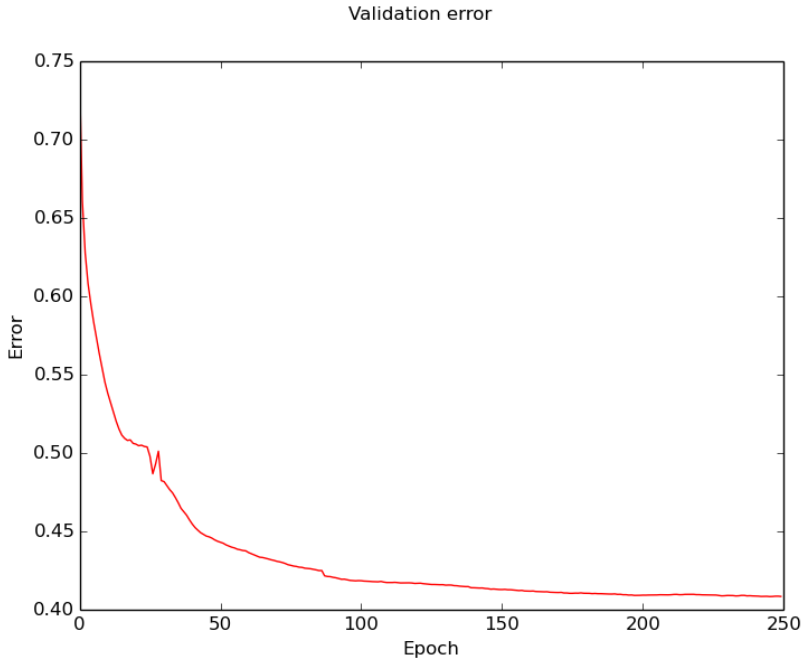
Στιγμιότυπα από την εκπαίδευση του δικτύου

Σε αυτό το κεφάλαιο περιλαμβάνονται στιγμιότυπα από την εκπαίδευση των διαφόρων δικτύων που χρησιμοποιήσαμε, δηλαδή η εξέλιξη στο χρόνο της συνάρτησης κόστους και του σφάλματος πάνω στο σύνολο των δεδομένων επαλήθευσης..

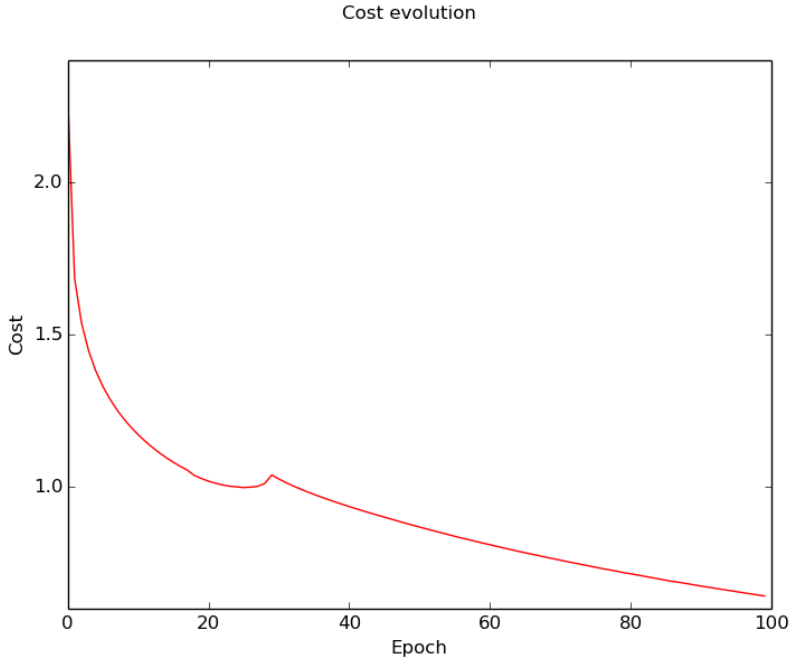
Σχήμα Α΄.1: *Monophone DNN, 5x600, sigmoid*



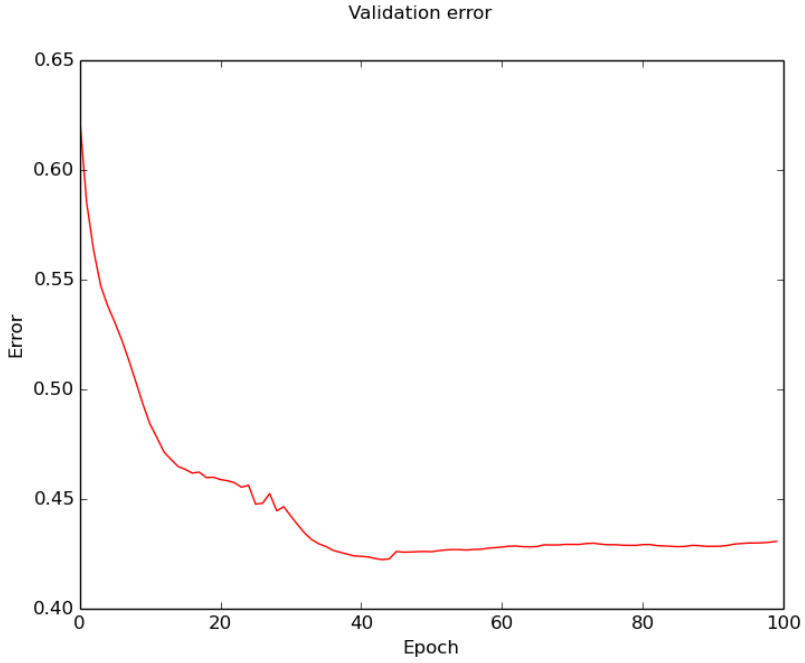
Σχήμα Α΄.2: *Monophone DNN, 5x600, sigmoid*



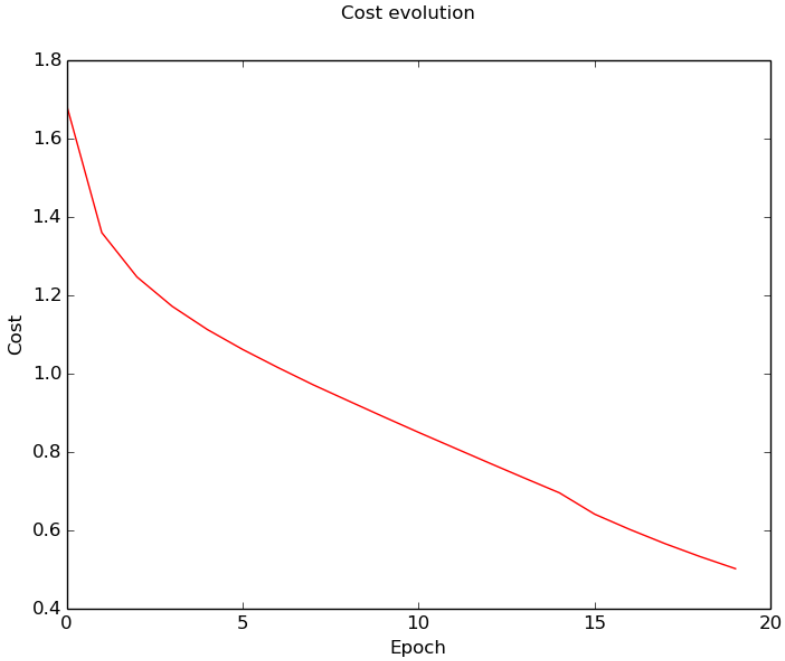
Σχήμα Α΄.3: *Monophone DNN, 4x1024, sigmoid*



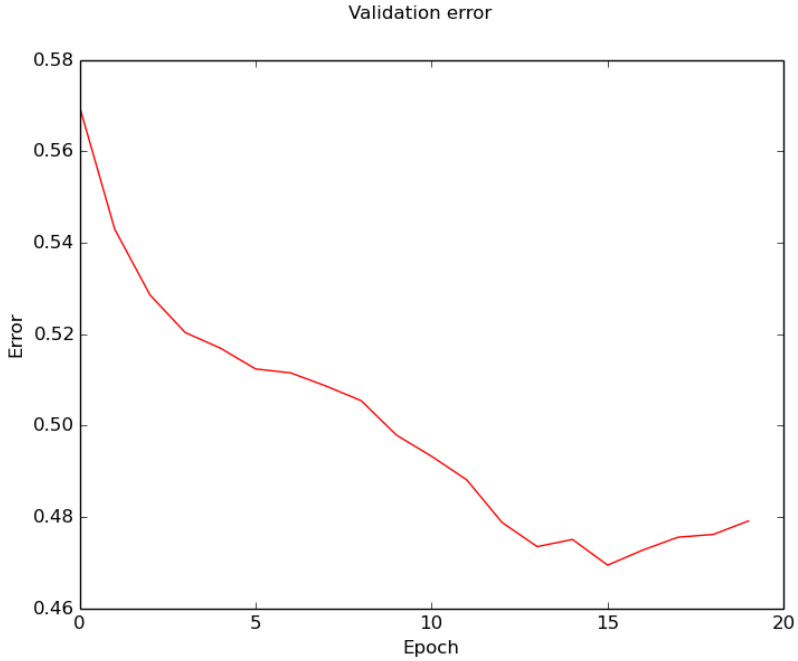
Σχήμα Α΄.4: *Monophone DNN, 4x1024, sigmoid*



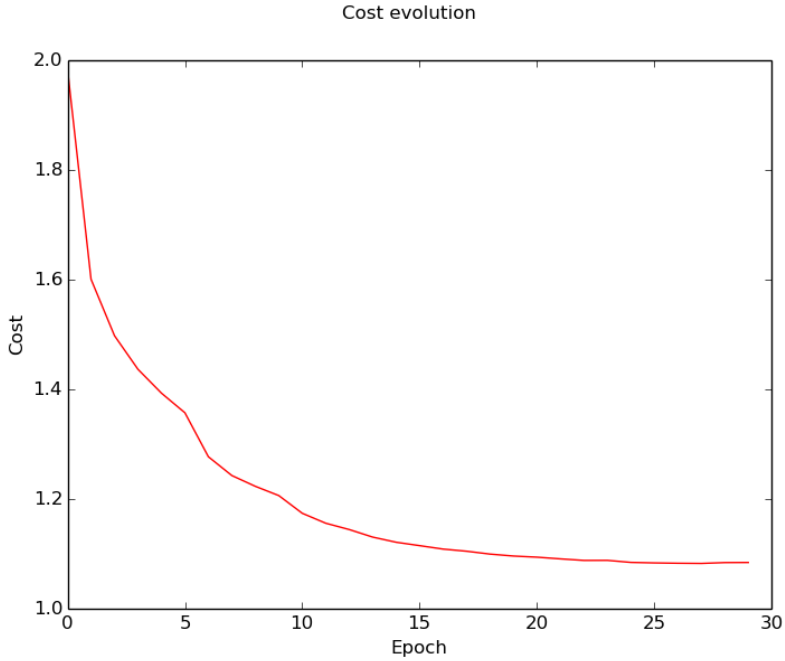
Σχήμα Α΄.5: *Monophone DNN, 4x1024, tanh*



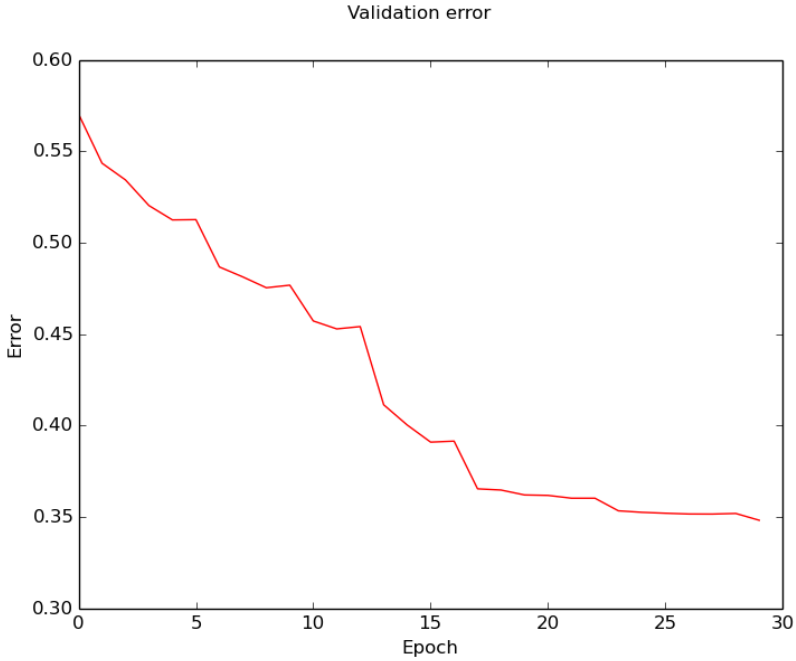
Σχήμα Α΄.6: *Monophone DNN, 4x1024, tanh*



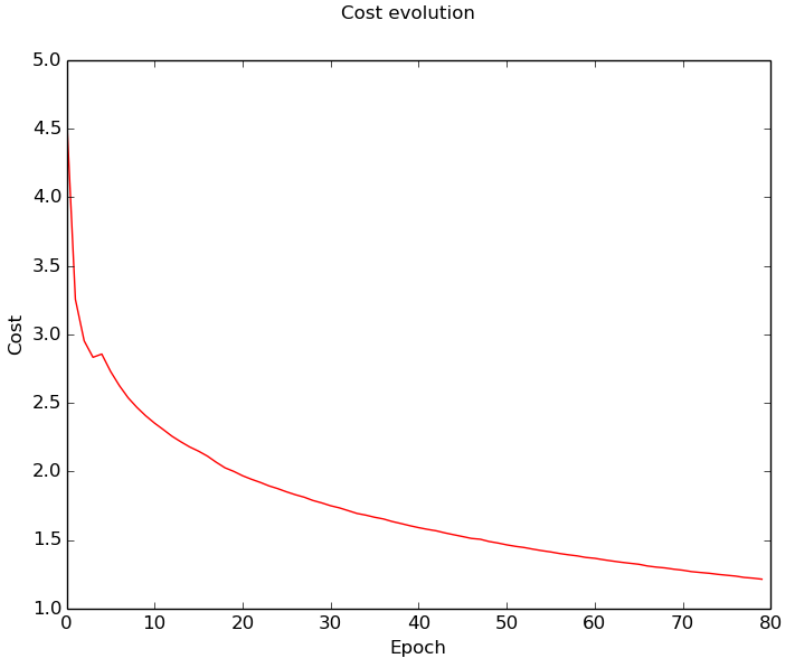
Σχήμα Α΄.7: *Monophone DNN, 4x1024, ReLU*



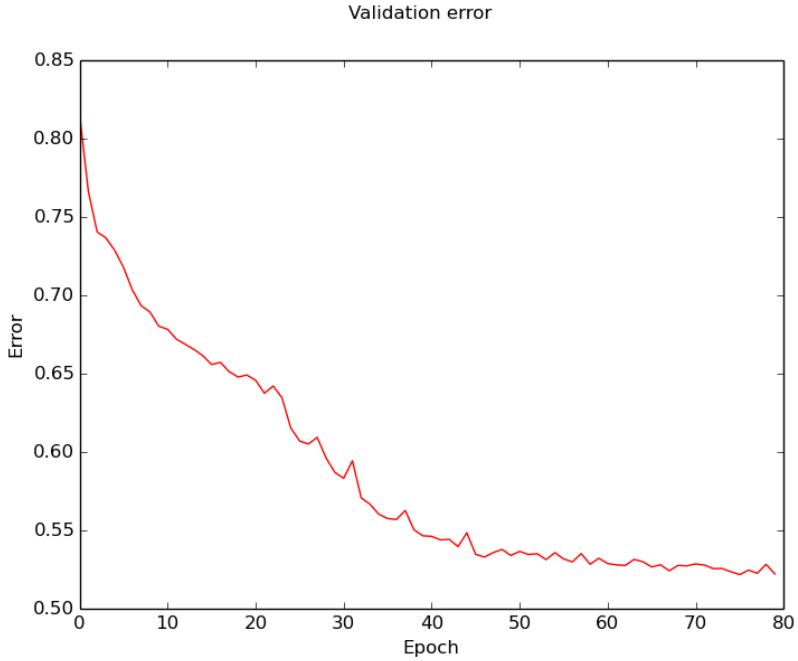
Σχήμα Α΄.8: *Monophone DNN, 4x1024, ReLU*



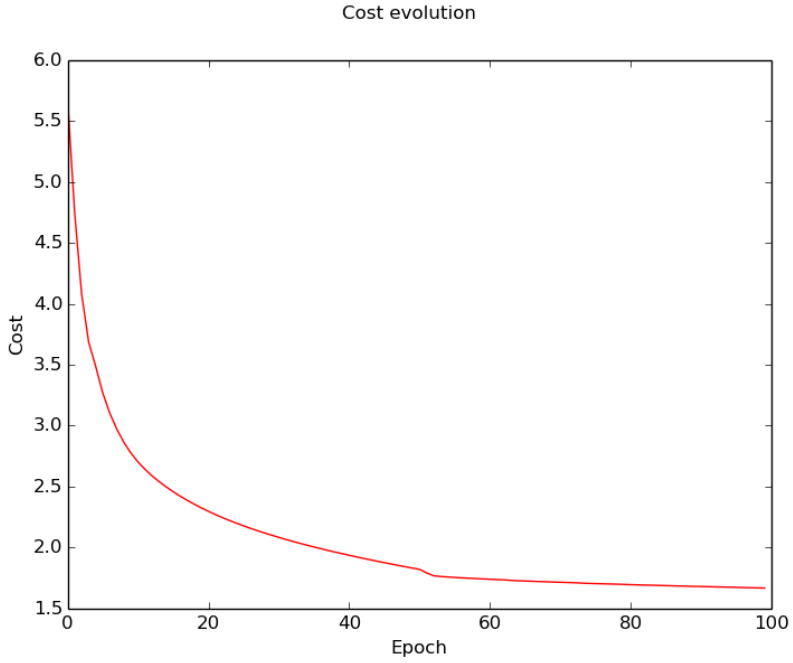
Σχήμα Α΄.9: *Triphone DNN, 6x2048, ReLU*



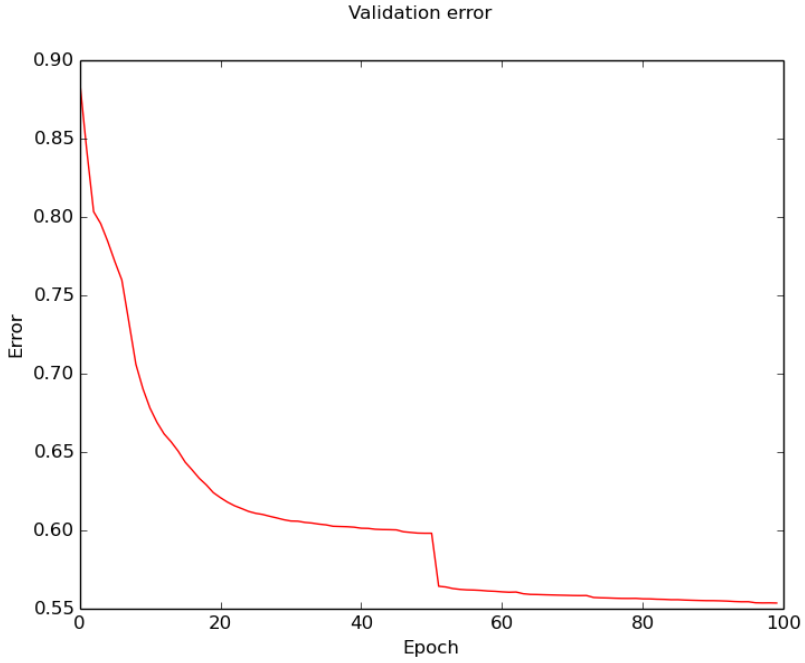
Σχήμα Α΄.10: *Triphone DNN, 6x2048, ReLU*



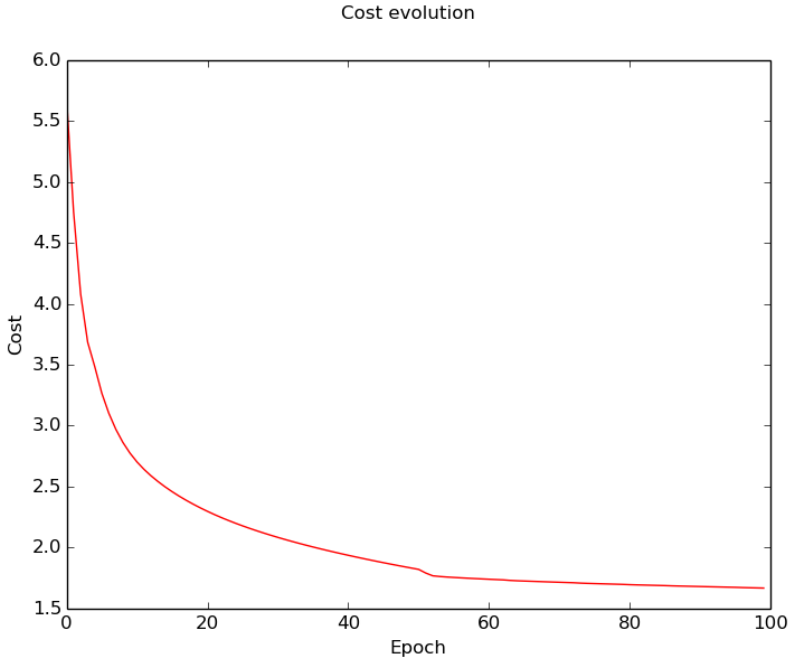
Σχήμα Α′.11: *Triphone DNN, 5x1024, sigmoid*



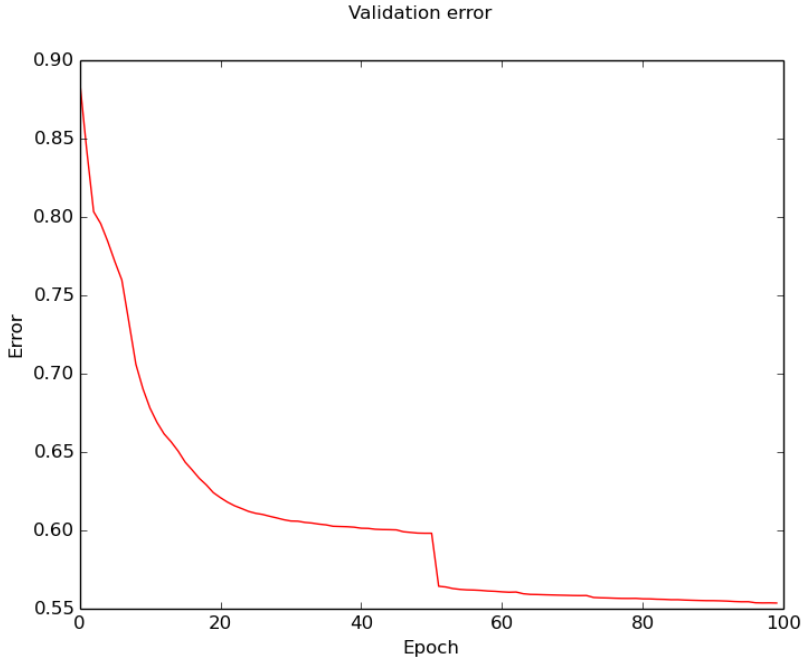
Σχήμα Α′.12: *Triphone DNN, 5x1024, sigmoid*



Σχήμα Α΄.13: *Triphone DNN, 5x1024, sigmoid*



Σχήμα Α΄.14: *Triphone DNN, 5x1024, sigmoid*



Βιβλιογραφία

[1] http://deeplearning.net/software/theano/. χ.χ.

[2] http://julialang.org/. χ.χ.

[3] http://kaldi-asr.org/. χ.χ.

[4] http://kaldi.sourceforge.net/dnn1.html. χ.χ.

[5] https://catalog.ldc.upenn.edu/docs/ldc94s13a/wsj1.txt. χ.χ.

[6] https://catalog.ldc.upenn.edu/ldc93s6a. χ.χ.

[7] https://code.google.com/archive/p/pfile-utilities/. χ.χ.

[8] https://github.com/bvlc/caffe/issues/109. χ.χ.

[9] https://github.com/juliageometry/kdtrees.jl. χ.χ.

[10] https://github.com/juliastats/clustering.jl. χ.χ.

[11] https://github.com/naxingyu/kaldi-nn/tree/master/src/nnet. χ.χ.

[12] https://martin-thoma.com/what-are-pfiles/. χ.χ.

[13] https://www.cise.ufl.edu. χ.χ.

[14] http://www.danielpovey.com/files/lecture1.pdf. χ.χ.

[15] http://www.danielpovey.com/files/lecture2.pdf. χ.χ.

[16] http://www.danielpovey.com/files/lecture3.pdf. χ.χ.

[17] http://www.danielpovey.com/files/lecture4.pdf. χ.χ.

[18] John McKenna Andrew Errity. An investigation of manifold learning for speech analysis, χ.χ.

[19] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard και Yoshua Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[20] Richard Ernest Bellman. *Dynamic Programming*. Dover Publications, Incorporated, 2003.

[21] Yoshua Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, 2009.

[22] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. *CoRR*, αβς/1206.5533, 2012.

[23] Yoshua Bengio, Aaron Courville και Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013.

[24] Yoshua Bengio, Réjean Ducharme, Pascal Vincent και Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, 2003.

[25] Ψοσηυα Βενγιο, Ιαν Θ. Γοοδφελλοω και Ααρον δυριλλε. Δεεπ λεαρνινγ. Βοοκ ιν πρεπαρατιον φορ ΜΙΤ Πρεσς, 2015.

[26] Yoshua Bengio, Jean Francois Paiement και Pascal Vincent. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. Στο *In Advances in Neural Information Processing Systems*, σελίδες 177–184. MIT Press, 2003.

[27] J. L. Bentley. Multidimensional binary search trees in database applications. *IEEE Trans. Softw. Eng.*, 5(4):333–340, 1979.

[28] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley και Yoshua Bengio. Theano: a CPU and GPU math expression compiler. Στο *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 2010.

[29] Jeff A. Bilmes. Buried markov models: a graphical-modeling approach to automatic speech recognition. *Computer Speech and Language*, 17(2.3):213 – 231, 2003.

[30] Mikael Boden. A guide to recurrent neural networks and backpropagation, 2001.

[31] L. Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 2005.

[32] George E. Dahl, Tara N. Sainath και Geoffrey E. Hinton. Improving deep neural networks for LVCSR using rectified linear units and dropout. Στο *ICASSP*, σελίδες 8609–8613, 2013.

[33] Sanjoy Dasgupta και Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.

[34] A. P. Dempster, N. M. Laird και D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[35] Li Deng και Dong Yu. Deep learning: Methods and applications. *Found. Trends Signal Process.*, 7(3&‘8211·4):197–387, 2014.

[36] Richard O. Duda, Peter E. Hart και David G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[37] Daniel P. W. Ellis, Rita Singh και Sunil Sivadas. Tandem acoustic modeling in large-vocabulary recognition. Στο *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001, 7-11 May, 2001, Salt Palace Convention Center, Salt Lake City, Utah, USA, Proceedings*, σελίδες 517–520, 2001.

[38] Xavier Glorot και Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. Στο *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10). Society for Artificial Intelligence and Statistics*, 2010.

[39] M. R. Hestenes και E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49:409–436, 1952.

[40] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, 2002.

[41] Geoffrey E. Hinton. A practical guide to training restricted boltzmann machines. Στο *Neural Networks: Tricks of the Trade (2nd ed.)*Gregoire Montavon, Genevieve B. Orr και Klaus Robert Müller, επιμελητές, τόμος 7700 στο *Lecture Notes in Computer Science*, σελίδες 599–619. Springer, 2012.

[42] Sepp Hochreiter και Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[43] Kurt Hornik, Maxwell Stinchcombe και Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366, 1989.

[44] A. Jansen και P. Niyogi. A geometric perspective on speech sounds. Τεχνική Αναφορά υπ. αρίθμ., 2005.

[45] Daniel Jurafsky και James H. Martin. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.

[46] S. Kirkpatrick, C. D. Gelatt και M. P. Vecchi. Optimization by simulated annealing. *SCIENCE*, 220(4598):671–680, 1983.

[47] E. Kreyszig. *Introductory Functional Analysis With Applications*. Wiley Classics Library. John Wiley & Sons, 1978.

[48] Yann Lecun, Leon Bottou, Yoshua Bengio και Patrick Haffner. Gradient-based learning applied to document recognition. Στο *Proceedings of the IEEE*, σελίδες 2278–2324, 1998.

[49] Yann LeCun, Léon Bottou, Genevieve B. Orr και Klaus Robert Müller. Effiicient backprop. Στο *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*, σελίδες 9–50, London, UK, UK, 1998. Springer-Verlag.

[50] D. C. Liu και J. Nocedal. On the limited memory bfgs method for large scale optimization. *Math. Program.*, 45(3):503–528, 1989.

[51] Michael Mahoney. Algorithms for massive datasets analysis, lecture1 notes, 2009.

[52] Yajie Miao. Kaldi+pdnn: Building dnn-based ASR systems with kaldi and PDNN. *CoRR*, αβς/1401.6984, 2014.

[53] Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky και Sanjeev Khudanpur. Recurrent neural network based language model. Στο *INTERSPEECH*Takao Kobayashi, Keikichi Hirose και Satoshi Nakamura, επιμελητές, σελίδες 1045–1048. ISCA, 2010.

[54] Vikramjit Mitra, Ganesh Sivaraman, Hosung Nam, Carol Y. Espy-Wilson και Elliot Saltzman. Articulatory features from deep neural networks and their role in speech recognition. Στο *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, σελίδες 3017–3021, 2014.

[55] Mehryar Mohri, Fernando Pereira και Michael Riley. Weighted finite-state transducers in speech recognition, 2001.

[56] Andrew William Moore. Efficient memory-based learning for robot control. Τεχνική Αναφορά υπ. αρίθμ. ΥCAM-CL-TP-209, University of Cambridge, Computer Laboratory, 1990.

[57] Nelson Morgan και Herve Bourlard. Continuous speech recognition using multilayer perceptrons with hidden Markov models, 1990.

[58] S Narayanan, A Toutios, V Ramanarayanan, A Lammert, J Kim, S Lee, K Nayak, YC Kim, Y Zhu, L Goldstein και others. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc). *The Journal of the Acoustical Society of America*, 136(3):1307, 2014.

[59] Michael Nielsen. Neural Networks and deep learning,. , 2015.

[60] Douglas B. Paul και Janet M. Baker. The design for the wall street journal-based csr corpus. Στο *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, σελίδες 357–362, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[61] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer και Karel Vesely. The kaldi speech recognition toolkit. Στο *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[62] Daniel Povey, Mirko Hannemann, Gilles Boulianne, Lukas Burget, Arnab Ghoshal, Milos Janda, Martin Karafiat, Stefan Kombrink, Petr Motlicek, Yanmin Qian, Korbinian Riedhammer, Karel Vesely και Ngoc Thang Vu. Generating exact lattices in the wfst framework. Στο *ICASSP*, σελίδες 4213–4216. IEEE, 2012.

[63] L. Rabiner και B. Juang. An introduction to hidden markov models. *IEEE Acoutics, Speech and Signal Processing Magazine*, 3:4–16, 1986.

[64] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. Στο *PROCEEDINGS OF THE IEEE*, σελίδες 257–286, 1989.

[65] Abdelrahman Mohamed, George Dahl και Geoffrey Hinton. Deep belief networks for phone recognition, χ.χ.

[66] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot και Yoshua Bengio. Contracting auto-encoders: Explicit invariance during feature extraction. Στο *In Proceedings of the Twenty-eight International Conference on Machine Learning (ICML11*, 2011.

[67] Tara N. Sainath, Brian Kingsbury, Abdelrahman Mohamed και Bhuvana Ramabhadran. Learning filter banks within a deep neural network framework. Στο *ASRU*, σελίδες 297–302. IEEE, 2013.

[68] Lawrence K. Saul και Sam T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res.*, 4:119–155, 2003.

[69] Frank Seide, Gang Li και Dong Yu. Conversational speech transcription using context-dependent deep neural networks. Στο *in Proc. Interspeech 2011*, σελίδες 437–440, χ.χ.

[70] Thomas Serre, Gabriel Kreiman, Minjoon Kouh, Charles Cadieu, Ulf Knoblich και Tomaso Poggio. A quantitative theory of immediate visual recognition. *PROG BRAIN RES*, σελίδες 33–56, 2007.

[71] Fei Sha και Lawrence K. Saul. Analysis and extension of spectral methods for nonlinear dimensionality reduction. Στο *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, σελίδες 784–791, New York, NY, USA, 2005. ACM.

[72] Darius Silingas και Laimutis Telksnys. Specifics of hidden markov model modifications for large vocabulary continuous speech recognition. *Informatica, Lith. Acad. Sci.*, 15(1):93–110, 2004.

[73] Ingmar Steiner, Korin Richmond, Ian Marshall και Calum Gray. The magnetic resonance imaging subset of the mngu0 articulatory corpus. *Journal of the Acoustical Society of America*, 131(2):ΕΛ106–ΕΛ111, 2012.

[74] S. S. Stevens, J. Volkmann και E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

[75] Joshua B. Tenenbaum, Vinde Silva και John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319, 2000.

[76] Vikrant Singh Tomar και Richard C. Rose. Application of a locality preserving discriminant analysis approach to ASR. Στο *11th International Conference on Information Science, Signal Processing and their Applications, ISSPA 2012, Montreal, QC, Canada, July 2-5, 2012*, σελίδες 103–107, 2012.

[77] Vikrant Singh Tomar και Richard C. Rose. A correlational discriminant approach to feature extraction for robust speech recognition. Στο *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, σελίδες 555–558, 2012.

[78] Vikrant Singh Tomar και Richard C. Rose. Efficient manifold learning for speech recognition using locality sensitive hashing. Στο *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, σελίδες 6995–6999, 2013.

[79] Vikrant Singh Tomar και Richard C. Rose. Locality sensitive hashing for fast computation of correlational manifold learning based feature space transformations. Στο *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, σελίδες 1776–1780, 2013.

[80] Vikrant Singh Tomar και Richard C. Rose. Noise aware manifold learning for robust speech recognition. Στο *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, σελίδες 7087–7091, 2013.

[81] Vikrant Singh Tomar και Richard C. Rose. A family of discriminative manifold learning algorithms and their application to speech recognition. *IEEE/ACM Transactions on Audio, Speech & Language Processing*, 22(1):161–171, 2014.

[82] Vikrant Singh Tomar και Richard C. Rose. Manifold regularized deep neural networks. Στο *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, σελίδες 348–352, 2014.

[83] Asterios Toutios και Konstantinos Margaritis. Acoustic-to-articulatory inversion of speech: A review. *Proceedings of the International 12th TAINN*, 2003.

[84] Zoltán Tüske, Pavel Golik, Ralf Schlüter και Hermann Ney. Acoustic modeling with deep neural networks using raw time signal for LVCSR. Στο *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, σελίδες 890–894, 2014.

[85] Benigno Uria, Iain Murray, Steve Renals και Korin Richmond. Deep architectures for articulatory inversion. Στο *Proc. Interspeech*, Portland, Oregon, USA, 2012.

[86] Benigno Uria, Steve Renals και Korin Richmond. A deep neural network for acoustic-articulatory speech inversion, χ.χ.

[87] Nakul Verma. Mathematical advances in manifold learning, 2008.

[88] Alexander Waibel, Toshiyuki Hanazawa, Geofrey Hinton, Kiyohiro Shikano και Kevin J. Lang. Readings in speech recognition. κεφάλαιο Πηονεμε Ρεϛογνιτιον Υσινγ Τιμε-δελαψ Νευραλ Νετωορκς, σελίδες 393–404. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[89] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong Jiang Zhang, Qiang Yang και Stephen Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):40–51, 2007.

[90] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev και P. Woodland. The htk book. *Cambridge University Engineering Department*, 3, 2002.

[91] Steve Young. Acoustic modelling for large vocabulary continuous speech recognition. *Cambridge University Engineering Department*, 2002.

[92] Dong Yu και Li Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, Incorporated, 2014.

[93] Dong Yu, Michael L. Seltzer, Jinyu Li, Jui-Ting Huang και Frank Seide. Feature learning in deep neural networks - A study on speech recognition tasks. *CoRR*, αβς/1301.3605, 2013.