Διεπιστημονικο Συμποσιο

'πως μαθαινει ο εγκεφαλος'

Speech Recognition - Αναγνώριση Ομιλίας

Συγγραφείς:
Χρήστος Ζωνιος
Έλενα Πλατη

Πανεπιστημιο Ιωαννινων
Τμήμα Μηχανικών Η/Υ &
Πληροφορικής

Περιεχόμενα

1 Εισαγωγή

1.1 Περιγραφή του Προβλήματος

Η αναγνώριση ομιλίας είναι η διαδικασία που ακολουθεί ένα έξυπνο (νοήμον) σύστημα, όπως ο ανθρώπινος εγκέφαλος, ώστε να αναπαραστήσει την πληροφορία της ομιλίας - η οποία πρόκειται, θεμελιωδώς, για ένα ηχητικό κύμα - ως μια ακολουθία ή φράση μικρότερων δομικών στοιχείων αυτής, η οποία περιγράφει κάποια έννοια.

Αυτό έπεται ότι το πρόβλημα της αναγνώρισης ομιλίας είναι ένα πρόβλημα επικοινωνίας, μια μετάδοση γνώσης μεταξύ δύο έξυπνων συστημάτων. Στην προκειμένη περίπτωση, το κύμα μεταφέρει κάποια πληροφορία, η οποία πρέπει να αποκρυπτογραφηθεί σε δύο επίπεδα. Πρώτα σε ένα βασικό επίπεδο φωνημάτων, τα οποία έπειτα θα ενωθούν σε λέξεις, που με τη σειρά τους θα μας δώσουν τελικά μια ακολουθία. Η ακολουθία αυτή, διατηρώντας μια σαφή δομή, έχοντας αρχή και τέλος, χρησιμοποιείται για να περιγράψει την πληροφορία που μεταδόθηκε.

1.2 Μοντελοποίηση

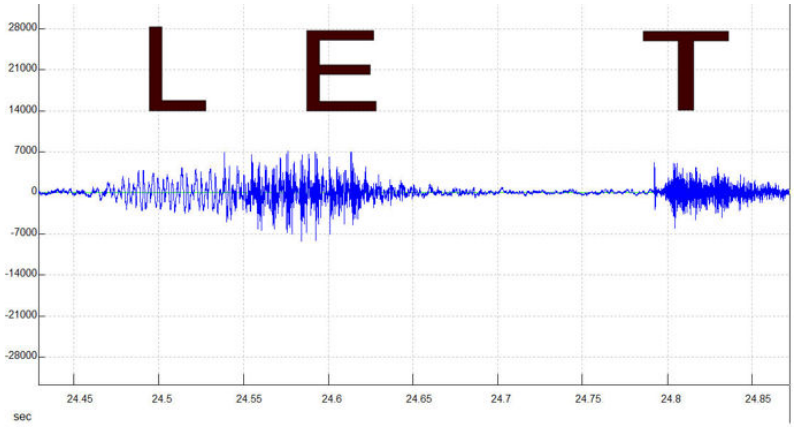
Η ακρίβεια της αναγνώρισης ομιλίας εξαρτάται από πολλούς παράγοντες, καθώς κακή ηχογράφηση, θόρυβος, επικάλυψη ήχων και λέξεις με παρόμοια προφορά ή χρήση μπορεί να προκαλέσουν σύγχυση. Άλλοι παράγοντες μπορεί να είναι η εξάρτηση ή όχι από τον ομιλητή, ο συνεχόμενος ή διακοπτόμενος λόγος ακόμα και αν η ομιλία παράγεται από αφήγηση ή αυθόρμητη παραγωγή λόγου.

Από υπολογιστική άποψη, η αναγνώριση ομιλίας είναι ένα πρόβλημα στο οποίο ένα ακουστικό μοτίβο πρέπει να αναγνωριστεί ή να ταξινομηθεί σε μια κατηγορία, η οποία θα αναπαριστά κάτι που θα αποφέρει νόημα για τον άνθρωπο. Για να αναγνωρίσει ο εγκέφαλος την ομιλία ακολουθεί μια συγκεκριμένη διαδικασία, την οποία έχουμε μοντελοποιήσει τεχνητά και παρουσιάζεται εδώ.

Εφόσον κάθε ακουστικό σήμα μπορεί να «σπάσει» σε υπό-σήματα, για να έχει μεγαλύτερη ακρίβεια, το μοντέλο θα πρέπει να κάνει τα εξής: κάθε πολύπλοκη έκφραση θα πρέπει να διαιρείται επαναληπτικά έως ότου προκύψουν πιο βασικοί, μικρότεροι και απλούστεροι ήχοι, τα φωνήματα (Σχήμα 1).

Σε αυτό το κατώτερο επίπεδο, το κάθε σύστημα θα πρέπει να έχει κανόνες που να αποφασίζουν τι αναπαριστά κάθε ήχος, και όταν αυτοί οι ήχοι θα συνδυάζονται σε περίπλοκο λόγο, τότε θα πρέπει να εφαρμόζονται εκ νέου διαφορετικοί, κατάλληλοι

κανόνες για τη σωστή αναπαράσταση της ομιλίας.



Σχήμα 1: Από ηχητικό κύμα σε φωνήματα

1.3 Ορισμός της Επιθυμητής Λύσης

Η αποκρυπτογράφηση ενός ηχητικού κύματος σε βασικούς ήχους ή φωνήματα δεν αναπαριστά λύση του προβλήματος από μόνη της. Θεωρήστε την υπόθεση πως κάποιο άτομο που βρίσκεται μακριά από εσάς, σε μια κακή σύνδεση τηλεφώνου ή σε κάποιο άλλο δωμάτιο προσπαθεί να σας πει κάτι. Είναι πολύ πιθανό να μην ακούτε καθαρά κάθε λέξη ή κάθε φώνημα, αλλά παρόλα αυτά να καταλαβαίνετε το νόημα ή ακόμα και να ανακατασκευάζετε πλήρως την φράση που σας είπε, με βάση τα αποσπάσματα που ακούσατε. Αυτή είναι η διαδικασία που μας ενδιαφέρει και θα μελετήσουμε σε αυτή την εργασία.

Πώς μετατρέπει ο εγκέφαλος μας ένα μείγμα ήχων σε μια φράση με μοναδικό νόημα; Τι ορίζουμε σαν νόημα στο πλαίσιο της αναγνώρισης ομιλίας; Μπορούμε, με τον ίδιο τρόπο, να κατηγοριοποιήσουμε τη μουσική σαν αναγνώριση ήχου και εξαγωγή νοήματος; Πώς μπορούμε, σαν μηχανικοί Τεχνητής Νοημοσύνης, να αναπαράγουμε τεχνητά αυτή τη μυστηριώδη διαδικασία του ανθρώπινου εγκεφάλου; Τέλος, ποια είναι η τεχνητή γνώση που πρέπει να αποκτήσει ένα σύστημα αναγνώρισης ομιλίας ώστε να λαμβάνει υπόψιν τα ιδιώματα, τη γραμματική και τις ιδιότητες της γλώσσας;

2 Ιστορικά στοιχεία

2.1 Πρώιμη Μελέτη

Από πολύ νωρίς ο τρόπος αναγνώρισης της ομιλίας από τον ανθρώπινο εγκέφαλο και η επεξεργασία της με αποτέλεσμα την παραγωγή δεδομένων, απασχόλησε αρκετούς ερευνητές. Πολλοί από αυτούς κατάφεραν να μιμηθούν αυτή τη λειτουργία για να δημιουργήσουν μοντέλα μετατροπής της φωνής σε κείμενο προς όφελος του ανθρώπου.

Η πρώτη προσπάθεια ξεκίνησε το 1952 όπου δημιουργήθηκε ένα σύστημα αναγνώρισης ψηφίων από έναν μόνο ακροατή, εντοπίζοντας τους σχηματισμούς στο φάσμα ισχύος κάθε άρθρωσης. Γενικά, τη δεκαετία του 1950, λόγω της περιορισμένης τεχνολογίας, υπήρχαν διαθέσιμα συστήματα που λειτουργούσαν με ένα λεξιλόγιο περίπου δέκα λέξεων. Το 1969 έγιναν επιπλέον βελτιώσεις όπως η αναγνώριση λέξεων χωρίς παύση μεταξύ τους, καθώς και αύξηση στο εύρος του λεξιλογίου που αντιλαμβάνονταν.

Έπειτα, όλο και περισσότερες χρηματοδοτήσεις άρχισαν να γίνονται και συνεπώς τα μοντέλα που δημιουργούνταν συνεχώς βελτιώνονταν με αποτέλεσμα τα μέσα του 1980 να καταφέρνουν να αναγνωρίζουν μέχρι και 20.000 λέξεις! Η προσπάθεια κάποιων να σταματήσουν να δίνουν έμφαση στον τρόπο που ο εγκέφαλος αναγνωρίζει τη φωνή και να χρησιμοποιήσουν μόνο στατιστικές τεχνικές μοντελοποίησης ήταν αμφιλεγόμενη.

Όσο, όμως, η τεχνολογία προχωρούσε και μαζί και η ταχύτητα των υπολογιστών, επιτεύχθηκε η αναγνώριση περισσότερων λέξεων και συνομιλιών ανεξάρτητα από τους ομιλητές και το θόρυβο.

Σχήμα 2: Η πρώτη συσκευή αναγνώρισης ομιλίας από την IBM στις αρχές του 1960 που έλυσε προφορικά αριθμητικά προβλήματα

2.2 Πρακτική Αναγνώριση Ομιλίας

Το 1990 είδαμε την πρώτη παραγωγή επιτυχημένων εμπορικών τεχνολογιών αναγνώρισης ομιλίας. Μια πολύ επιτυχημένη υπηρεσία ήταν η δρομολόγηση τηλεφωνικών κλήσεων χωρίς τη χρήση ανθρώπινου χειριστή το 1992. Από αυτό το σημείο και μετά, το λεξιλόγιο ενός τέτοιου τυπικού εμπορικού συστήματος ήταν μεγαλύτερο από το λεξιλόγιο του μέσου ανθρώπου.

Τέτοια συστήματα χρησιμοποιήθηκαν στο λειτουργικό σύστημα Windows XP και από την Apple, ενώ άλλα χρησιμοποιήθηκαν στη μετάδοση ειδήσεων σε παραπάνω από μία γλώσσες. Η Google υιοθέτησε επίσης τέτοια συστήματα και σήμερα η φωνητική αναζήτηση της Google υποστηρίζει πάνω από 30 διαφορετικές γλώσσες, πράγμα επίπονο ή και ακατόρθωτο να επιτευχθεί από το μέσο άνθρωπο.

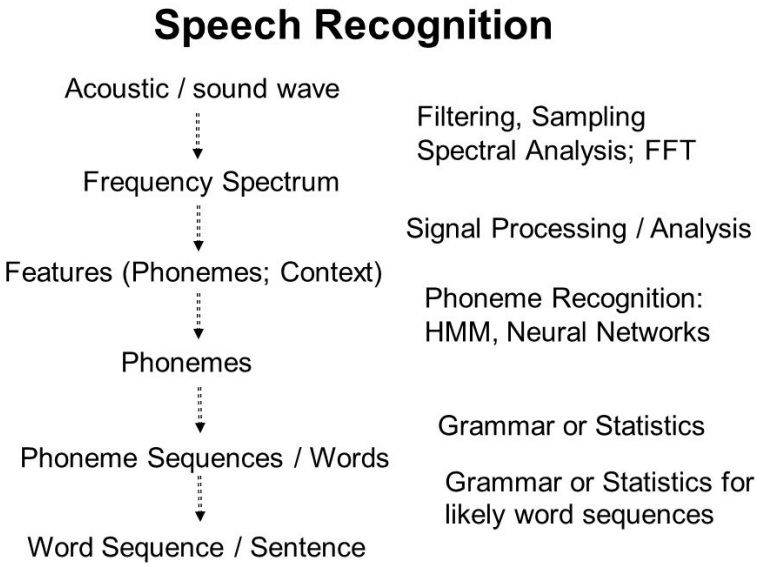
Τέλος, η Εθνική ασφάλεια των Ηνωμένων Πολιτειών χρησιμοποιεί μια μορφή αναγνώρισης ομιλίας για τον εντοπισμό λέξεων ¨κλειδιών¨ σε μακροσκελείς συζητήσεις.

2.3 Μοντέρνα Συστήματα

Από το 2000 και μετά, πέρα από τις παραδοσιακές τεχνικές, έχουν ενσωματωθεί μέθοδοι, οι οποίες αντιμετωπίζουν προβλήματα που απαιτούν μνημονικό για γεγονότα που συνέβησαν χιλιάδες διακριτά βήματα πριν. Επίσης, μέσω αυτών των νέων τεχνικών, το ποσοστό σφάλματος στην αναγνώριση μιας λέξης περιορίστηκε στο 30%.

Ακόμα και αν η χρήση μαθηματικών και στατιστικών μεθόδων για την αναγνώριση ομιλίας χρησιμοποιείται πλέον αποκλειστικά, τα αποτελέσματα πλησιάζουν όλο και περισσότερο στις λειτουργίες που επιτελεί ο ανθρώπινος εγκέφαλος, επεκτείνοντας τις λειτουργίες του και παρέχοντας μεγαλύτερη ταχύτητα στην επεξεργασία των δεδομένων.

3 Τεχνητά Μοντέλα και Μέθοδοι



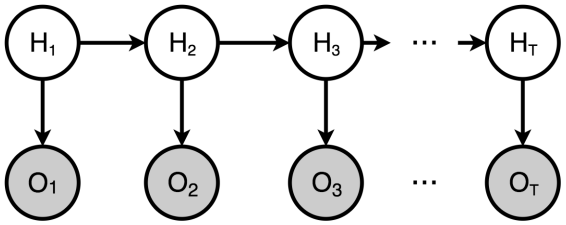
Σχήμα 3: Διαδικασία αναγνώρισης ομιλίας

3.1 Κρυμμένα Μοντέλα Markov

Τα περισσότερα μοντέρνα συστήματα αναγνώρισης ομιλίας γενικού σκοπού βασίζονται στα Κρυμμένα Μοντέλα Αλυσίδων Markov[1] (Hidden Markov Models - HMM, Σχήμα 4), τα οποία είναι στατιστικά μοντέλα που παράγουν σαν έξοδο μια ακολουθία συμβόλων ή ποσοτήτων. Τα HMM είναι πολύ διάσημα σε τέτοιες περιπτώσεις γιατί είναι απλά, μπορούν να εκπαιδευτούν αυτόματα και είναι υπολογιστικά εφικτό να χρησιμοποιηθούν.

Τα μοντέλα αυτά χρησιμοποιούνται θεωρώντας κάθε σήμα που παράγεται από την ομιλία, ως ένα μερικώς σταθερό σήμα ή ως ένα σταθερό σήμα για μικρό χρονικό διάστημα, καθώς η ομιλία σε πολύ μικρή κλίμακα χρόνου (10 χιλιοστά δευτερολέπτου) μπορεί να θεωρηθεί σταθερή διαδικασία. Σαν αποτέλεσμα παράγουν μια ακολουθία από διανύσματα N διαστάσεων και εξάγοντας κάθε ένα από αυτά κάθε 10 χιλιοστά του δευτερολέπτου. Επίσης, σε κάθε κατάσταση που αξιολογούν ένα σύντομο χρονικό διάστημα ομιλίας, παράγουν στατιστικές κατανομές, οι οποίες εκφράζουν και μια πιθανότητα για κάθε διάνυσμα που μελετάται.

Κάθε λέξη ή φώνημα θα έχει μια διαφορετική κατανομή σαν αποτέλεσμα, για ακολουθία λέξεων ή φωνημάτων, τα HMM έχουν ως αποτέλεσμα την συγκόλληση των αποτελεσμάτων των ήδη εκπαιδευμένων HMM μοντέλων για αυτές τις λέξεις.



Σχήμα 4: Αλυσίδα Markov

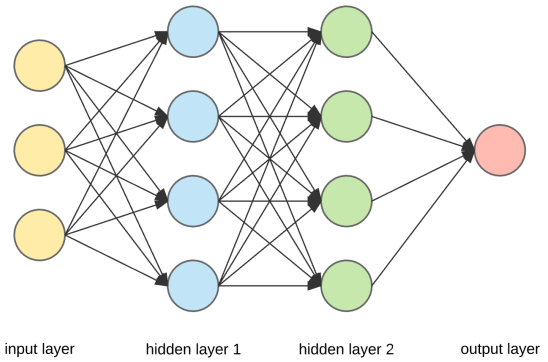
3.2 Αναγνώριση με Dynamic Time Warping (Δυναμική Χρονική Στρέβλωση)

Ο συγκεκριμένος αλγόριθμος χρησιμοποιούνταν ευρέως στο παρελθόν, ωστόσο σήμερα δεν προτιμάται. Χρησιμοποιείται στην αναγνώριση ομιλίας κυρίως σε περιπτώσεις που απαιτείται η σύγκριση δύο ακολουθιών που διαφέρουν στην ταχύτητα που παράγονται οι λέξεις κατά την ομιλία και δίνει τη δυνατότητα έτσι στον υπολογιστή να βρει ένα βέλτιστο ταίριασμα μεταξύ δύο ακολουθιών.

3.3 Νευρωνικά Δίκτυα

Μοντελοποιημένα με βάση τον εγκέφαλο, τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks, Σχήμα 5) είναι η πιο κοντινή αναπαράσταση του τρόπου μάθησης του ανθρώπινου εγκεφάλου. Καθώς η αναγνώριση ομιλίας είναι κατά βάση μια

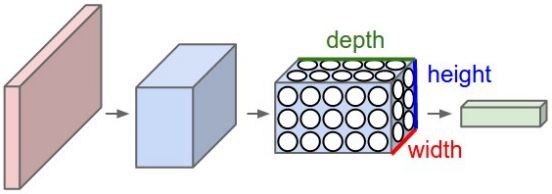
διαδικασία που μαθαίνεται από τον εγκέφαλο και εξευγενίζεται κατά τη διάρκεια της ζωής, τα νευρωνικά δίκτυα βρίσκονται στην πρώτη γραμμή της έρευνας και αποτελούν τον πλέον σύγχρονο τρόπο μοντελοποίησης της συναρπαστικής αυτής διαδικασίας. Παρακάτω, θα παραθέσουμε τις επικρατέστερες αρχιτεκτονικές νευρωνικών δικτύων για αυτόματη αναγνώριση ομιλίας.



Σχήμα 5: Πολυεπίπεδο νευρωνικό δίκτυο

3.3.1 Βαθιά Ανεστραμμένα και Συνελικτικά Νευρωνικά Δίκτυα

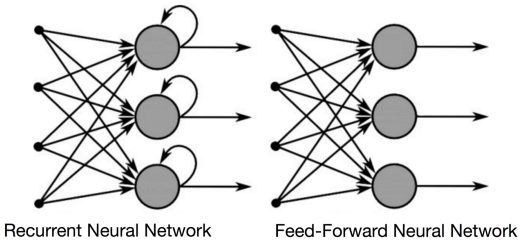
Μία από τις πλέον επικρατέστερες αρχιτεκτονικές νευρωνικών δικτύων όσον αφορά τη Μηχανική Μάθηση (Machine Learning) με χρήση ακατέργαστων (χωρίς συγκεκριμένη δομή) δεδομένων, είναι το υποπεδίο που ονομάζεται Βαθιά Μάθηση[2] (Deep Learning). Το υποπεδίο αυτό έχει καταφέρει αξιοσημείωτες επιδόσεις σε άλλους τομείς όπως η Υπολογιστική Όραση (Computer Vision), την Βιοπληροφορική, την έρευνα για νέα φάρμακα, το spam filtering κλπ. Η επικρατέστερη αρχιτεκτονική νευρωνικών δικτύων που χρησιμοποιείται στη Βαθιά Μάθηση είναι τα Συνελικτικά Νευρωνικά Δίκτυα[3] (Convolutional Neural Networks, Σχήμα 6), η οποία έχει χρησιμοποιηθεί στο πρόβλημα της αναγνώρισης ομιλίας με αξιοσημείωτες επιδόσεις (11.5% Word Error Rate).



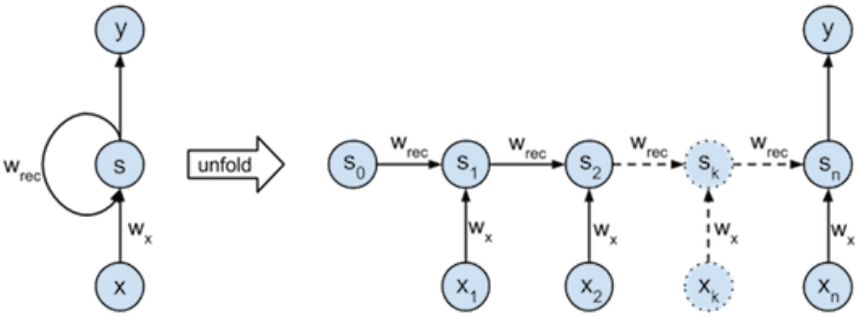
Σχήμα 6: Οπτικοποίηση ενός συνελικτικού νευρωνικού δικτύου

3.3.2 Επαναληπτικά Νευρωνικά Δίκτυα

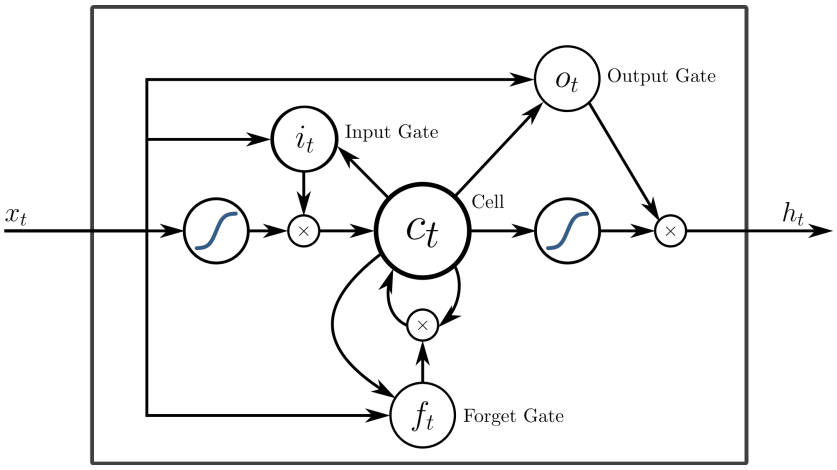
Η πλέον state-of-the-art αρχιτεκτονική που χρησιμοποιείται στην αναγνώριση ομιλίας, είναι τα Επαναληπτικά Νευρωνικά Δίκτυα[4] (Recurrent Neural Networks), τα οποία μοντελοποιούν την μνήμη που χρειάζεται (Σχήμα 7, Σχήμα 8), ώστε να υπάρχει η έννοια των συμφραζομένων, λαμβάνοντας υπόψιν του τι ειπώθηκε πριν και διορθώνοντας κενά ή λάθη που μπορεί να υπάρχουν λόγω θορύβου, ενισχύουν τη σιγουριά τους για την απόφαση που παίρνουν (η αντίστοιχη φράση, σε γραπτή μορφή). Το πιο σύνηθες είδος επαναληπτικού νευρωνικού δικτύου, είναι το μοντέλο Long-Short Term Memory (LSTM)[5] (Σχήμα 9). Το μοντέλο αυτό εισάγει (πιθανοτικά) την έννοια της λησμονιάς της προηγούμενης πληροφορίας, με αποτέλεσμα να επιτυγχάνει κατά βάση την ιδιότητα του ανθρώπινου εγκεφάλου να μην λαμβάνει πάντα υπόψιν κάτι που ίσως άκουσε (αναγνώρισε) λάθος. Οι επιδόσεις αυτού του μοντέλου είναι 5.1% Word Error Rate, σημαντικά καλύτερες από το προηγούμενο μοντέλο.



Σχήμα 7: Διαφορά μεταξύ ενός επαναληπτικού και ενός feed-forward νευρωνικού δικτύου



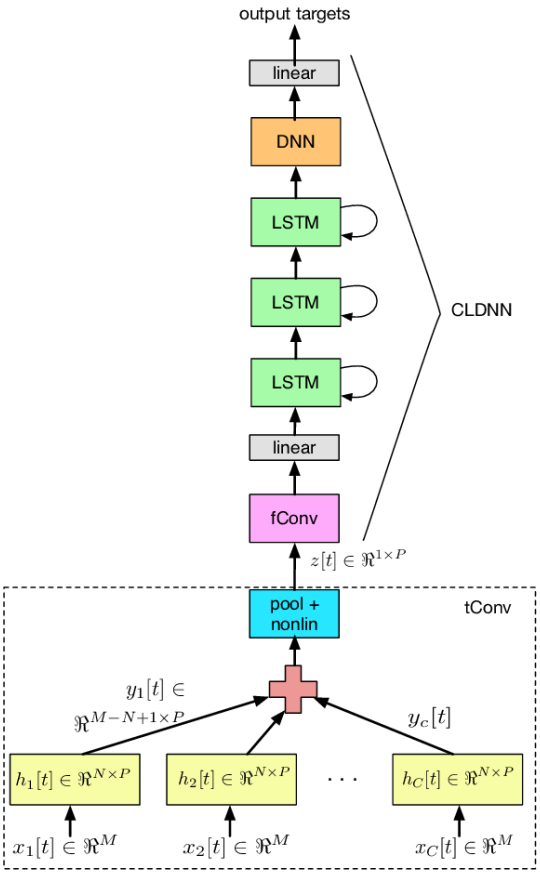
Σχήμα 8: Η αρχιτεκτονική ενός επαναληπτικού δικτύου



Σχήμα 9: Η αρχιτεκτονική ενός LSTM cell

Τελικά, αυτό που έχει επικρατήσει είναι ένας συνδυασμός των αρχιτεκτονικών, με επίπεδα συνελικτικών νευρώνων να ακολουθούνται από επίπεδα LSTM κελιών, τα οποία με τη σειρά τους καταλήγουν σε πλήρως ενωμένους νευρώνες (Σχήμα 10). Το μοντέλο αυτό προτάθηκε από τη Google και είναι το Convolutional, Long Short-Term Memory, fully connected Deep Neural Network (CLDNN)[6]. Σαν φυσική ερμηνεία, μπορούμε να θεωρήσουμε πως τα συνελικτικά επίπεδα διαχωρίζουν τις συχνότητες (διαφορετικές φωνές, όργανα, θόρυβος), έπειτα τα κελιά LSTM αναγνωρίζουν ακολουθίες και συμφραζόμενα, και τέλος, οι πλήρως ενωμένοι νευρώνες δίνουν την τελική απόφαση για το ποιά είναι η σωστή αναπαράσταση κειμένου, λαμβάνοντας υπόψιν την

πληροφορία που έχουν από τα προηγούμενα επίπεδα. Το μοντέλο αυτό έχει επιτύχει την καλύτερη απόδοση μέχρι σήμερα, 4.9% Word Error Rate.

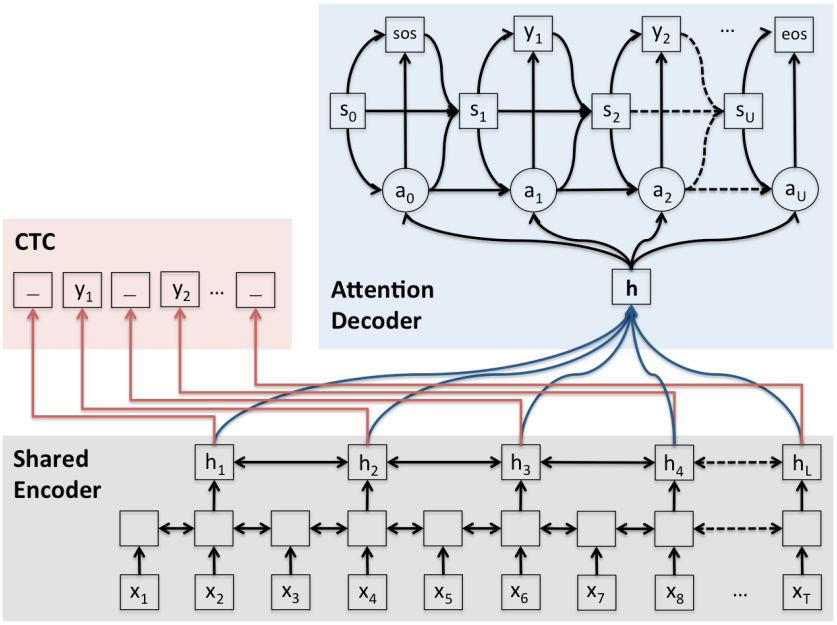


Σχήμα 10: Η αρχιτεκτονική ενός CLDNN

3.4 Αυτόματη Αναγνώριση Ομιλίας Από Άκρη Σε Άκρη

Σε αντίθεση με όλες τις υπόλοιπες προσεγγίσεις που έγιναν για την αναγνώριση ομιλίας, που χρειάζονταν διαφορετικά δεδομένα και τρόπο εκπαίδευσης για τη σωστή προφορά, κατανόηση και απόδοση ομιλίας, αυτό το μοντέλο έχει την ικανότητα να μαθαίνει από κοινού όλα τα στοιχεία που απαιτούνται για την αναγνώριση ομιλίας (Σχήμα 11). Συνεπώς, απλοποιείται η διαδικασία εκπαίδευσης και η δυνατότητα

ανάπτυξής του, καθώς η εκμάθηση κάθε λειτουργίας που περιγράψαμε ξεχωριστά, αύξανε την πολυπλοκότητα και απαιτούσε επιπλέον αποθηκευτικό χώρο.



Σχήμα 11: Η αρχιτεκτονική ενός End-To-End συστήματος αναγνώρισης ομιλίας

4 Εφαρμογές

4.1 Υγεία και Ιατρική Περίθαλψη

Στον τομέα της υγείας, τα συστήματα αναγνώρισης ομιλίας μπορούν να χρησιμοποιηθούν, αρχικά, για να διευκολύνουν την διαδικασία της ιατρικής τεκμηρίωσης και της συντήρησης των ιατρικών δεδομένων. Στόχος της χρήσης της αναγνώρισης ομιλίας στην ιατρική, είναι να γίνουν εύκολα προσβάσιμα τα ιατρικά δεδομένα και οι αναφορές μέσω των δικτύων των υπολογιστών, ούτως ώστε να γίνει πιο ακριβής η ιατρική γνωμάτευση για τους ασθενείς και να αυξηθεί η αποδοτικότητα για τους επαγγελματίες υγείας.

Η διαδικασία της αναγνώρισης ομιλίας μπορεί να γίνει με δύο τρόπους. Πρώτον, μπορεί να γίνει υπαγόρευση της γνωμάτευσης σε ένα σύστημα αναγνώρισης ομιλίας,

το οποίο θα αναπαριστά στην οθόνη το κείμενο που θα μπορεί να το τροποποιήσει και εν τέλει να το κλείσει μόνο αυτός που το υπαγόρευσε. Στην άλλη περίπτωση, τα πορίσματα θα υπαγορεύονται σε ένα ψηφιακό σύστημα υπαγόρευσης, η φωνή θα δρομολογείται σε ένα σύστημα αναγνώρισης ομιλίας και το κείμενο που παράγεται μαζί με την αρχική φωνή, πηγαίνουν σε έναν ηλεκτρονικό επεξεργαστή κειμένου, ο οποίος τελειοποιεί και αποθηκεύει το κείμενο.

Ο χρόνος ολοκλήρωσης των ιατρικών εγγράφων μειώνεται αισθητά και η ανάκτησή τους είναι άμεση. Η αναγνώριση ομιλίας μπορεί να έχει και θεραπευτική χρήση. Έρευνες έχουν δείξει πως η παρατεταμένη χρήση λογισμικού αναγνώρισης ομιλίας, σε συνδυασμό με επεξεργαστές κειμένου, βοηθούν στην αποκατάσταση μνήμης ατόμων με αρτηριοφλεβική δυσπλασία που έχουν υποβληθεί σε εκτομή.

Σχήμα 12: Η πρώτη χρήση της αναγνώρισης ομιλίας σε νοσοκομείο της Vanderbilt

Άτομα με ειδικές ανάγκες μπορούν να επωφεληθούν αρκετά από συστήματα αναγνώρισης ομιλίας. Συγκεκριμένα, για άτομα με απώλεια ή δυσκολία ακοής μπορούν να δημιουργήσουν ένα κλειστό σύστημα υποτίτλων συνομιλιών, και μπορεί να δώσουν λύση σε άτομα που δυσκολεύονται να κινήσουν τα χέρια τους, εξαιτίας επαναλαμβανόμενων τραυμάτων από στρες ή αναπηρίες που αποκλείουν τη χρήση συμβατικών συσκευών εισόδου υπολογιστών. Μπορούν, επίσης, να φανούν χρήσιμα σε άτομα με δυσλεξία που δυσκολεύονται να εκφράσουν τη σκέψη τους σε γραπτό λόγο.

Σχήμα 13: Άτομα με ειδικές ανάγκες μπορούν να επωφεληθούν αρκετά από συστήματα αναγνώρισης ομιλίας

4.2 Παιδεία και Καθημερινότητα

Η χρησιμότητα και η προσφορά των συστημάτων αναγνώρισης ομιλίας είναι μεγάλη στον τομέα της παιδείας αλλά και στην καθημερινή μας ζωή.

Όσον αφορά την εκμάθηση γλωσσών, για παράδειγμα, η αναγνώριση ομιλίας μπορεί να βοηθήσει στην εκμάθηση μιας νέας γλώσσας μαθαίνοντας τη σωστή προφορά και βοηθώντας να αναπτυχθεί ευφράδεια στον προφορικό λόγο. Επίσης, σε μαθητές με χαμηλή όραση ή τύφλωση, παρέχει τη δυνατότητα να μεταφέρουν λέξεις στον υπολογιστή και να ακούσουν τον υπολογιστή να τις αναπαράγει, όπως και να χειρίζονται έναν υπολογιστή μόνο με τη φωνή τους.

Ακόμα, μαθητές οι οποίοι υποφέρουν από σωματικές αναπηρίες ή τραυματισμούς στα άνω άκρα, μπορούν να συμβαδίζουν με τις σχολικές τους εργασίες και υποχρεώσεις. Σε περίπτωση που αντιμετωπίζουν μαθησιακές δυσκολίες, η εξέλιξή τους θα βοηθηθεί από την παραγωγή λόγου μεγαλόφωνα, κατά τη χρήση συστήματος αναγνώρισης ομιλίας.

Τέλος, τέτοια συστήματα σε συνδυασμό με ψηφιακά μαγνητόφωνα και υπολογιστή με λογισμικό επεξεργασίας κειμένου, μπορούν να βοηθήσουν στη φθορά βραχυπρόθεσμης μνήμης, σε άτομα που έχουν υποστεί εγκεφαλικά επεισόδια και κρανιοτομές.

4.3 Συστήματα Αυτοματισμού Σε Οχήματα (In-Car Systems)

Η αναγώριση ομιλίας τα τελευταία χρόνια χρησιμοποιείται πολύ σε συστήματα αλληλεπίδρασης με το αυτοκίνητο [7], αυτόνομο ή μη. Χωρίς να μετακινήσει τα χέρια του από το τιμόνι ή να πάρει τα μάτια του από το δρόμο, ο οδηγός μπορεί να ζητήσει από ένα τέτοιο σύστημα αλληλεπίδρασης οδηγίες GPS, να καλέσει κάποιον ή ακόμη και να στείλει γραπτό μήνυμα. Άλλες εφαρμογές είναι ο έλεγχος μέσων αναπαραγωγής και λειτουργίες προσωπικού βοηθού, για παράδειγμα, όταν ο οδηγός θέλει να μάθει την πρόγνωση καιρού ή τα τελευταία νέα.

Η τεχνολογία αυτή υπάρχει ήδη στο εμπόριο, με βασικούς κατασκευαστές τη Google, την Apple, καθώς και κατασκευαστικές εταιρίες αυτοκινήτων.

4.4 Μουσική

Οι τεχνολογίες που αναπτύχθηκαν για την αναγνώριση ομιλίας χρησιμοποιούνται και στη μουσική. Από τη δεκαετία του ′50, ο Ιάννης Ξενάκης ασχολήθηκε με την αυτόματη, στοχαστική σύνθεση μουσικής, χρησιμοποιώντας Μαρκοβιανά μοντέλα και αλυσίδες, εφαρμόζοντας θεωρία παιγνίων και το νόμο των Maxwell-Boltzmann-Gauss.

Όσον αφορά τον εγκέφαλο, το πεδίο αυτό μας βοηθά να κατανοήσουμε το πώς αντιλαμβανόμαστε τη μουσική και ποια είναι τα κοινά χαρακτηριστικά με την αναγνώριση της ομιλίας. Μπορεί ο εγκέφαλος να αντιληφθεί το συναίσθημα ή κάποιο νόημα μόνο μέσω του ήχου, εφόσον δεν υπάρχουν λέξεις ή συγκεκριμένες ακολουθίες; Η απάντηση είναι αρκετά περίπλοκη και ίσως υποκειμενική σε ένα μεγάλο βαθμό (ας αναλογιστούμε τα ερωτήματα: τι είναι τέχνη; μπορεί ένας υπολογιστής να δημιουργήσει αληθινή τέχνη;) αλλά το σίγουρο είναι πως, καθώς χρησιμοποιούμε μοντέλα νευρωνικών δικτύων παρόμοια με αυτά του εγκεφάλου, δε συνεπάγεται αυτό, πως τα δίκτυα αυτά μαθαίνουν με τον ίδιο τρόπο όπως εμείς; Αν ένα νευρωνικό δίκτυο εκπαιδευτεί ('μάθει') πάνω σε ένα συγκεκριμένο συνθέτη ή είδος μουσικής, έχει μεγάλη διαφορά από το πως μαθαίνει μουσική κάποιος άνθρωπος, δεχόμενος τα ίδια ακούσματα; Το μοντέλο "Performance RNN"[8] και άλλα της βιβλιοθήκης Magenta της Google μπορεί, ίσως, να σας πείσει πως παίζει κάτι που συνέθεσε ο μεγάλος Πολωνός πιανίστας Frédéric Chopin.

4.5 Άλλες Εφαρμογές

Επιγραμματικά, η αναγνώριση ομιλίας εφαρμόζεται σε πολλά άλλα προβλήματα όπως η τηλεφωνία, η αεροδιαστημική, η αυτόματη απόδοση διαλόγων και μεταφράσεων, η καταγραφή νομικών ομιλιών, η αυτοματοποίηση σπιτιών και η ρομποτική.

5 Συμπέρασμα

Σε αυτή την εργασία, περιγράψαμε τον τρόπο που ένας τεχνητός εγκέφαλος αναγνωρίζει τον ήχο και την ομιλία και τα οργανώνει σε ακολουθίες και φράσεις που έχουν ένα μοναδικό νόημα γι΄ αυτόν. Η επιστήμη των υπολογιστών περιλαμβάνει πολλούς ερευνητές που δραστηριοποιούνται και διευρύνουν τις γνώσεις μας για το πως ο ανθρώπινος εγκέφαλος αντιλαμβάνεται αυτά τα σήματα, σε συνεργασία με βιολόγους, κοινωνιολόγους, γνωστικούς επιστήμονες, παιδαγωγούς και γιατρούς.

Το πεδίο της αναγνώρισης ομιλίας φαίνεται να έχει αποκτήσει μια σημαντική θέση στην επιστήμη και ιδιαίτερα στην τεχνολογία, και αναμένεται να προσφέρει περισσότερη ασφάλεια, δυνατότητες και ευκολίες στην καθημερινή μας ζωή στο βραχυπρόθεσμο, αλλά και στο μακροπρόθεσμο μέλλον.

References

[1] X. D. Huang, Y. Ariki, and M. A. Jack, "Hidden markov models for speech recognition," 1990.

[2] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[3] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[4] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pp. 6645–6649, IEEE, 2013.

[5] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *CoRR*, vol. abs/1402.1128, 2014.

[6] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4580–4584, April 2015.

[7] C. Y. Loh, K. L. Boey, and K. S. Hong, "Speech recognition interactive system for vehicle," in *2017 IEEE 13th International Colloquium on Signal Processing its Applications (CSPA)*, pp. 85–88, March 2017.

[8] I. Simon and S. Oore, "Performance rnn: Generating music with expressive timing and dynamics." `https://magenta.tensorflow.org/performance-rnn`, 2017.