**ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ**

**ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ**

**ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ**

**Ίνεγλης Φίλιππος**

**Μεταπτυχιακός Φοιτητής**

**Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης**

Επόπτης Μεταπτ. Εργασίας: Αναπλ. Καθηγητής Αθανάσιος Μουχτάρης

**Πέμπτη, 15/12/2016, 12:00**

**Αίθουσα B108, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης**

# " Χρήση Συστοιχίας Μικροφώνων στην Αναγνώριση Φωνής "

## ΠΕΡΙΛΗΨΗ

Η Αυτόματη Αναγνώριση Ομιλίας πρωτοεμφανίστηκε το 1950. Έκτοτε έχουν γίνει πολλές προσπάθειες για την βελτίωσή της σε μονοφωνικές ηχογραφήσεις. Τα τελευταία χρόνια, πολλοί ερευνητές έχουν δείξει ενδιαφέρον στην Αυτόματη Αναγνώριση Ομιλίας και σε πολυκάναλες ηχογραφήσεις, καθώς όλο και περισσότερες συσκευές της καθημερινότητάς μας ενσωματώνουν όλο και περισσότερα μικρόφωνα. Τα μικρόφωνα αυτά τοποθετούνται σε καθορισμένες διατάξεις δίνοντάς μας την δυνατότητα να εκμεταλλευτούμε την κατευθυντικότητα του σήματος εισόδου και να επιτύχουμε καλύτερη ενίσχυση σήματος. Μερικά παραδείγματα τέτοιων συσκευών και εφαρμογών αποτελούν τα κινητά τηλέφωνα, τα tablets, οι συσκευές οικιακού αυτοματισμού όπως Amazon Echo, Google Home, οι ψηφιακοί προσωπικοί βοηθοί όπως Siri, Google Now, Cortana κ.α.

Στα πλαίσια αυτής της εργασίας, στόχος μας είναι να δημιουργήσουμε ένα σύστημα Αυτόματης Αναγνώρισης Ομιλίας συνδυασμένο με ένα σύστημα ενίσχυσης σήματος (front-end) για να επιτύχουμε τα βέλτιστα αποτελέσματα αναγνώρισης ομιλίας σε μη

ευνοϊκές συνθήκες, όπως δωμάτια με έντονη αντήχηση ή/και θόρυβο. Τα πειράματα που εκτελέσαμε περιλαμβάνουν σενάρια με στάσιμους ομιλητές, κινούμενους ομιλητές καθώς και επικαλυπτόμενους ομιλητές. Για την καλύτερη προσέγγιση του προβλήματος, χωρίσαμε τη διαδικασία σε τρεις φάσεις. Η πρώτη φάση ήταν ο πειραματισμός πάνω στα δεδομένα που χρησιμοποιήσαμε για την εκπαίδευση του ακουστικού μοντέλου. Τα ακουστικά μοντέλα που εκπαιδεύσαμε ήταν τρία. Το πρώτο ακουστικό μοντέλο εκπαιδεύτηκε με σήματα καθαρής ομιλίας, το δεύτερο με επεξεργασμένα σήματα ομιλίας και το τρίτο με τον συνδυασμό των δύο παραπάνω. Κατά τη δεύτερη φάση, δοκιμάσαμε ποικίλα συστήματα ενίσχυσης σήματος, δηλαδή τεχνικές επεξεργασίας πολυκάναλων αρχείων φωνής και τα αξιολογήσαμε με βάση τα αποτελέσματα της αναγνώρισης. Καθεμιά από τις μεθόδους επεξεργασίας πολυκάναλου σήματος που χρησιμοποιήσαμε, απαρτίζεται από δύο κύρια στοιχεία, το διαμορφωτή λοβού και το πολυκάναλο φίλτρο. Επιπλέον, προτείναμε μια μέθοδο βασισμένη στις δυαδικές μάσκες και το φίλτρο Wiener η οποία μας οδήγησε σε καλύτερα αποτελέσματα αναγνώρισης. Τα αποτελέσματα της αναγνώρισης ομιλίας έδειξαν ότι ο συνδυασμός του υπερκατευθντικού διαμορφωτή λοβού με το πολυκάναλο φίλτρο Wiener αποδίδει καλύτερα στην περίπτωση ενός ομιλητή ενώ ο ίδιος διαμορφωτής λοβού συνδυασμένος με δυαδικές μάσκες αποδίδει καλύτερα σε επικαλυπτόμενους ομιλητές. Κατά την τελευταία φάση, δημιουργήσαμε ένα ακουστικό μοντέλο το οποίο εκπαιδεύτηκε με καθαρά και επεξεργασμένα σήματα φωνής χρησιμοποιώντας ως σύστημα ενίσχυσης σήματος τις τεχνικές που αναφέραμε παραπάνω ως βέλτιστες.
Για την αξιολόγηση της απόδοσης κάθε ακουστικού μοντέλου και κάθε συστήματος ενίσχυσης σήματος, χρησιμοποιήσαμε μια μετρική η οποία είναι ευρέως γνωστή σε πειράματα αναγνώρισης ομιλίας, το ποσοστό των λάθος αναγνωρισμένων λέξεων. Η προτεινόμενη μέθοδος οδήγησε σε σημαντική βελτίωση των αποτελεσμάτων. Σημειώθηκε σχετική μείωση στο ποσοστό των λάθος αναγνωρισμένων λέξεων κατά 62,4% για στάσιμο ομιλητή, 57,9% για κινούμενο ομιλητή και 49,6% για επικαλυπτόμενους ομιλητές σε σχέση με τα αποτελέσματα αναγνώρισης των μην επεξεργασμένων σημάτων. Συγκεκριμένα, οι τροποποιήσεις που εφαρμόσαμε στο σύστημα ενίσχυσης σήματος και στις δυαδικές μάσκες στην περίπτωση των επικαλυπτόμενων ομιλητών, δηλαδή η εισαγωγή ενός συχνοτικού κατωφλίου και ένα πιο αυστηρό κριτήριο για την εφαρμογή των μασκών αυτών, οδήγησε σε σχετική βελτίωση κατά 9.9% σε σχέση με τις προεπιλεγμένες παραμέτρους.

**Ineglis Filippos**

**M.Sc. Thesis**

**Computer Science Department**

**University of Crete**

**Master's Thesis Supervisor: Associate Professor A. Mouchtaris**

**Thursday, 15/12/2016, 12:00**

**Room B108, Computer Science dept., University of Crete**

**“Incorporating Microphone Arrays into Automatic Speech Recognition”**

## ABSTRACT

The Automatic Speech Recognition (ASR) was initially introduced in the 1950s. Since then, a lot of effort has been made to improve speech recognition in single channel recordings. In the last few years, many researchers have shown interest in the combination of speech recognition and multichannel recordings, as many every day devices incorporate multiple microphones. These microphones are usually placed in specific topologies allowing us to take advantage of the directivity of the input signal and achieve more robust speech enhancement. Some examples of devices and applications are mobile phones, tablets, home automation services such as Amazon Echo and Google Home, digital personal assistants like Google Now, Siri, Cortana etc.

In the course of this thesis, we aim to create a robust ASR system combined with a front-end to improve speech recognition in challenging environments such as reverberant rooms with or without background noise. The experiments we examined included scenarios with stationary and moving speakers as well as overlapping speakers. To approach this problem, we divided it into three phases. The first phase was the experimentation on the training data for the acoustic model. Three acoustic models were trained to define the best acoustic model, one with clean speech signals, one with processed speech signals and one with the combination of the previous two training sets. During the second phase, we tested several front-ends, i.e. array processing techniques, and evaluated them in the context of their speech recognition performance. Each array processing technique consists of two main modules, a beamformer and a postfilter. In addition to that, we proposed a new front-end framework based on the binary masks and a Wiener postfilter which achieved better recognition results. The recognition results showed that the combination of a Superdirective beamformer followed by a Wiener postfilter performs better on single speaker experiments while the same beamformer combined with Binary Masks performs better on overlapping speaker experiments. The last phase was to use the outcome of the first and the second phase in order to create a robust combination of a front-end and an acoustic model.

In order to evaluate the performance of each acoustic model and each front-end, we used a common speech recognition metric known as Word Error Rate (WER). The final proposed acoustic model combined with the proposed front-end led to a significant improvement in WER in all experiments, i.e. stationary speaker, moving speaker and overlapping speakers. The relative improvement in terms of WER of the processed speech signals over the unprocessed speech signals for the three experiments is 62.4% for stationary speaker, 57.9% for moving speaker and 49.6% for overlapping speakers. In particular, the modification we proposed for the binary masks used in the front-end for

the scenarios with overlapping speakers, that is a spectral floor and a more strict criterion on the application of the binary masks, led to a relative improvement of 9.9% in WER results.