ΠΡΟΣ

1) ***Όλα τα μέλη ΔΕΠ του Τμήματος Επιστήμης Υπολογιστών***

2) ***Τους εκπροσώπους των Μεταπτυχιακών φοιτητών του Τμήματος Επιστήμης Υπολογιστών***

3) ***Την Επταμελή Εξεταστική Επιτροπή***

4) ***Όλα τα μέλη της Πανεπιστημιακής Κοινότητας***

Πρόσκληση σε Δημόσια Παρουσίαση της Διδακτορικής Διατριβής του

κ. Muhammed Shifas PV

Doctoral Dissertation Defense

Mr. Muhammed Shifas PV

Την Παρασκευή, 08/07/2022 και ώρα 12:30 μ.μ. μέσω Τηλεδιάσκεψης (zoom) https://us02web.zoom.us/j/83160557977 θα γίνει η δημόσια παρουσίαση και υποστήριξη της Διδακτορικής Διατριβής του υποψήφιου διδάκτορα του Τμήματος Επιστήμης Υπολογιστών κ. ***Muhammed Shifas PV*** με θέμα:

“ Neural Networks for the Quality and Intelligibility Enhancement of Speech”

ΠΕΡΙΛΗΨΗ

Η ομιλία είναι ο πιο αποτελεσματικός τρόπος επικοινωνίας ιδεών που δημιουργούνται στο ανθρώπινο μυαλό. Ωστόσο, η προφορική επικοινωνία στην πραγματική ζωή συχνά επηρεάζεται από τον θόρυβο στο περιβάλλον, ο οποίος μπορεί να μειώσει σημαντικά την καταληπτότητα και την αντιληπτή ποιότητα του σήματος. Τεχνικές για τη βελτίωση της επικοινωνίας έχουν προταθεί στο παρελθόν και έχουν δοκιμαστεί με επιτυχία σε σύγχρονες συσκευές όπως το Amazon Alexa, επιτρέποντάς της να λειτουργεί σε αντίξοες συνθήκες. Ο θόρυβος περιβάλλοντος μπορεί να διαταράξει τόσο τη λήψη σήματος από μια συσκευή όσο και την αντίληψη της ομιλίας από τον ακροατή. Οι τεχνικές βελτίωσης ομιλίας (SE) αναπτύσσονται για την αποκατάσταση της ομιλίας από τις θορυβώδεις παρατηρήσεις της και οι τεχνικές βελτίωσης της ακρόασης (LE) έχουν σχεδιαστεί για να βελτιώνουν την καταληπτότητα αλλάζοντας την ομιλία πριν από την έκθεσή της σε θόρυβο, καθώς η φυσικά παραγόμενη ομιλία δεν είναι πάντα πολύ κατανοητή. Ως εκ τούτου, τόσο το SE όσο και το LE είναι απαραίτητα στις σύγχρονες συσκευές για να λειτουργήσουν σε διάφορες ακουστικές συνθήκες. Συχνά τα συστήματα SE και LE λειτουργούν ως δύο ανεξάρτητες μονάδες σε σύγχρονες συσκευές, οι οποίες περιορίζουν την

απόδοσή τους. Η προσπάθεια σε αυτή τη διπλωματική εργασία είναι να συνδυαστούν οι τεχνικές βελτίωσης SE και LE ώστε να έχουμε ένα σύστημα από άκρη-σε-άκρη για εφαρμογές επικοινωνίας. Προσεγγίζουμε το πρόβλημα από τη σκοπιά των νευρωνικών δικτύων. Ως εκ τούτου, επινοήθηκαν πολλαπλές νέες αρχιτεκτονικές για SE και LE, και οι ιδέες από αυτά τα μοντέλα έχουν χρησιμοποιηθεί για την κατασκευή του τελικού συστήματος από άκρη-σε-άκρη. Τα παραδοσιακά συστήματα που βασίζονται σε στατιστικά είχαν περιορισμούς για την πλήρη μοντελοποίηση της δυναμικής της ομιλίας και του θορύβου. Τα νευρωνικά δίκτυα έχουν προκύψει ως εναλλακτική προσέγγιση για τη μοντελοποίηση δεδομένων. Ως εκ τούτου, αυτή η διατριβή επανεξετάζει τα προβλήματα SE και LE από την οπτική των νευρωνικών δικτύων.

Όσον αφορά τη βελτίωση ομιλίας (SE), έχουν εφευρεθεί τρεις νέες αρχιτεκτονικές, δύο από τις οποίες βρίσκονται στο χώρο των χαρακτηριστικών και ένα στο πεδίο της κυματομορφής. Οι αρχιτεκτονικές στο πεδίο των χαρακτηριστικών πραγματοποιούν την εργασία βελτίωσης της ομιλίας στην αναπαράσταση βραχυχρόνιου μετασχηματισμού Fourier (STFT), επομένως, είναι παραμετρικά λιγότερο περίπλοκες. Χαρακτηριστικά από τη δισδιάστατη (2D) αναπαράσταση της ομιλίας εξάγονται με τη χρήση νευρικού κυττάρου gruCNN, το οποίο βρέθηκε αποτελεσματικό στην απομόνωση θορύβων με υψηλή διακύμανση. Το μοντέλο gruCNN-SE έχει ξεπεράσει τα υπερσύγχρονα συστήματα βελτίωσης ομιλίας με τυπικά συνελικτικά νευρωνικά δίκτυα (CNN) και δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTM). Στη συνέχεια, προτείνεται μια αμφίδρομη επέκταση της ενότητας gruCNN (BigruCNN) με τη συμπερίληψη εξαρτήσεων προς τα πίσω μεταξύ των 2D πλαισίων. Επιπλέον, παρουσιάζεται ένα νέο δίκτυο πεδίου κυματομορφής με χαρακτηριστικό μοτίβο διαστολής (SE-FFTNet). Το SE-FFTNet βρέθηκε αποτελεσματικό στην εκμάθηση της στατιστικής ανομοιότητας της ομιλίας και του θορύβου σε μια θορυβώδη παρατήρηση.

Όσον αφορά τη βελτίωση της ακρόασης (LE), προτείνεται μια νέα αρχιτεκτονική παρόμοια με το WaveNet για τη βελτίωση της καταληπτότητας του ακροατή στο θόρυβο (wSSDRC). Το σύστημα wSSDRC εκτελεί τόσο φασματική διαμόρφωση (SS) όσο και συμπίεση δυναμικού εύρους (DRC) της εισόδου για βελτίωση της ευκρίνειας. Βρέθηκε ότι το μοντέλο έχει ως αποτέλεσμα μια μέση απόλυτη αύξηση καταληπτότητας 39% για κανονική ακοή και 38% για ακροατές με προβλήματα ακοής σε στάσιμο θόρυβο κατά τη διάρκεια της μη επεξεργασμένης ομιλίας.

Στη συνέχεια, προτείνεται ένα νέο σύστημα από άκρη-σε-άκρη το οποίο συνδυάζει τους στόχους του SE και του LE για να ενισχύσει την καταληπτότητα των θορυβωδών παρατηρήσεων. Το σύστημα από άκρη-σε-άκρη βρέθηκε να αυξάνει το ποσοστό σωστών λέξεων-κλειδιών των ακροατών σε στάσιμο θόρυβο από 2,5\% σε 60\% στην είσοδο SNR 0 dB και από περίπου 10\% σε 75\% σε SNR εισόδου 5 dB, σε σύγκριση με την μη επεξεργασμένη ομιλία, ενώ ξεπερνούσε σημαντικά το σύστημα με διαδοχική εφαρμογή της SE ακολουθούμενη από LE.

Επιβλέπων: Καθηγητής, Ιωάννης Στυλιανού

ABSTRACT

Speech is the most effective way to communicate ideas generated in human minds. However, spoken communication in real life is often affected by noise in the surroundings which can substantially reduce the intelligibility and perceived quality of the signal. Techniques to enhance the communication have been proposed in the past and successfully tested in modern engines like Amazon Alexa, allowing it to operate in adverse conditions. The ambient noise can disrupt both signal acquisition by a device as well as speech perception by the listener. Speech enhancement (SE) techniques are developed to restore speech from its disrupted observations, and listening enhancement (LE) techniques are designed to improve the perceived intelligibility by altering the speech before its presentation in noise as the naturally produced speech is not always very intelligible.

Often SE and LE systems are operated as two independent modules in modern devices , which limit their performance. The effort in this thesis is to combine the SE and LE enhancement techniques to have an end-to-end system for communication applications. We approach the problem from the neural networking perspective. As such, multiple novel architectures for SE and LE were invented, and the concepts from those models have been used to build the final end-to-end system.

Regarding speech enhancement (SE), three new architectures have been invented; two of which are in the feature domain and one in the waveform domain. The feature domain architectures formulate the enhancement task in the short-time Fourier transform (STFT) representation of speech, therefore, are parametrically less complex. Features from the two-dimensional (2D) representation of speech are extracted with the use gruCNN neural cell, which is found effective in isolating noises with high variance. The gruCNN-SE model has outperformed state-of-the-art speech enhancement systems with standard convolution (CNN) and long short-term memory (LSTM) cells. Subsequently, a bidirectional extension of gruCNN module (BigruCNN) is proposed with the inclusion of backward dependencies among the 2D frames. Besides, a novel waveform domain network with a characteristic dilation pattern (SE-FFTNet) is presented. The SE-FFTNet is found efficient in learning the statistical dissimilarity of speech and noise in a noisy observation.

Regarding listening enhancement (LE), a novel WaveNet-like architecture to improve the listener's intelligibility in noise (wSSDRC) is proposed. The wSSDRC system performs both spectral shaping (SS) and dynamic range compression (DRC) of the input for intelligibility enhancement. The model is found to produce a median absolute intelligibility boost of 39% for normal hearing and 38% for hearing-impaired listeners in stationary noise over the unprocessed speech.

Subsequently, a novel end-to-end system which combines the objectives of SE and LE is proposed to enhance the intelligibility of noisy observations. The end-to-end system was found to increase the listeners' keyword correct rate in stationary noise from 2.5% to 60% at 0 dB input SNR, and from about 10% to 75% at 5 dB input SNR, compared with the unprocessed speech, while substantially outperforming the modular setup with SE followed by LE.

Supervisor: Professor, Yannis Stylianou

Antonis Argyros

Chairman

Department of Computer Science

*Σε εφαρμογή του Γενικού Κανονισμού 2016/679 του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου της 27 Απριλίου 2016 για την προστασία των φυσικών προσώπων έναντι της επεξεργασίας των δεδομένων προσωπικού χαρακτήρα και για την ελεύθερη κυκλοφορία των δεδομένων αυτών, το Πανεπιστήμιο της Κρήτης σας ενημερώνει ότι: Η παρουσίαση της Διδακτορικής Διατριβής της υποψήφιας διδάκτορας του Τμήματος κ. Muhammed Shifas PV με τίτλο «*Neural Networks for the Quality and Intelligibility Enhancement of Speech*", η οποία θα πραγματοποιηθεί 08/07/2022 και ώρα 12:30 μ.μ., θα βιντεοσκοπηθεί. Το βίντεο θα χρησιμοποιηθεί μόνο για τις ανάγκες της δημόσιας παρουσίασης της εν λόγω διδακτορικής διατριβής ενώπιον της Επταμελούς Επιτροπής και ανοιχτού ακροατηρίου. Όποιος πρόκειται να συμμετάσχει στην πιο πάνω δημόσια παρουσίαση δηλώνει ότι γνωρίζει την πιο πάνω συλλογή των προσωπικών του δεδομένων και την χρήση αυτών και συναινεί. Εάν δεν επιθυμείτε τη βιντεοσκόπησή σας, μπορείτε να επικοινωνήσετε με τον κ. Ν. Τσατσάκη (+30 2810 393524), Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης.*