ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Σισαμάκη Ειρήνη
Μεταπτυχιακή Φοιτήτρια

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης
Επόπτης Μεταπτ. Εργασίας: Καθηγητής, Ι. Στυλιανού

Παρασκευή, 25/10/2019, 14:00
Αίθουσα Κ206, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

"Από-άκρη-σε-άκρη Νευρωνική σύνθεση ομιλίας από κείμενο για την Ελληνική Γλώσσα"

ΠΕΡΙΛΗΨΗ

Σύνθεση ομιλίας από κείμενο (TTS) είναι η αυτόματη μετατροπή του γραπτού λόγου σε προφορικό. Τα συστήματα σύνθεσης ομιλίας από κείμενο παίζουν σημαντικό ρόλο στη διάδραση ανθρώπου-υπολογιστή. Η συνενωτική σύνθεση ομιλίας και η στατιστική παραμετρική σύνθεση ομιλίας ήταν οι μέθοδοι που εφαρμόστηκαν για δεκαετίες.
Στην εποχή της Βαθιάς Μάθησης τα από-άκρη-σε-άκρη συστήματα έχουν βελτιώσει δραματικά την ποιότητα της συνθετικής ομιλίας. Ο στόχος αυτής της εργασίας είναι η υλοποίηση ενός νευρωνικού από-άκρη-σε-άκρη συστήματος σύνθεσης ομιλίας από κείμενο, για την ελληνική γλώσσα. Η αρχιτεκτονική νευρωνικού δικτύου του Tacotron-2 χρησιμοποιείται για σύνθεση ομιλίας κατευθείαν από κείμενο. Το σύστημα αποτελείται από ένα αναδρομικό από-ακολουθία-σε-ακολουθία δίκτυο πρόβλεψης χαρακτηριστικών, που αντιστοιχίζει ενσωματώσεις χαρακτήρων σε φασματογράμματα κλίμακας Μελ που ακολουθείται από ένα τροποποιημένο μοντέλο WaveNet, που λειτουργεί ως συνθεσάϊζερ ομιλίας για να συνθέσει κυματομορφές στο πεδίο του χρόνου από αυτά τα ακουστικά χαρακτηριστικά. Η ανάπτυξη συστημάτων σύνθεσης ομιλίας από κείμενο για μια δεδομένη γλώσσα είναι μια σημαντική πρόκληση και απαιτεί μεγάλη

ποσότητα ηχογραφήσεων υψηλής ποιότητας. Γι' αυτό αυτά τα συστήματα είναι διαθέσιμα μόνο για τις πιο ευρέως ομιλούμενες γλώσσες. Σε αυτή την εργασία περιγράφονται πειράματα με διάφορες γλώσσες και βάσεις δεδομένων που είναι ελεύθερα διαθέσιμες. Μια ελληνική βάση δεδομένων, αρχικά δημιουργημένη για αναγνώριση ομιλίας, μας δόθηκε από το Ινστιτούτο Επεξεργασίας Λόγου. Στο πρώτο μας πείραμα χρησιμοποιήθηκαν μόνο 3 ώρες ηχογραφήσεων στα Ελληνικά. Έπειτα η τεχνική της προσαρμογής γλώσσας εφαρμόστηκε, χρησιμοποιώντας 3 ώρες Ελληνικά και 18 ώρες Ισπανικά. Επίσης εφαρμόσαμε την προσαρμογή ομιλητή για να παράγουμε ομιλία με συγκεκριμένους ομιλητές από τη βάση δεδομένων μας. Το σύστημά μας για τα Ελληνικά μπορεί να συνθέτει καλής ποιότητας ομιλία με πολύ φυσική προσωδία. Μια αξιολόγηση με ένα ακουστικό τεστ με 30 εθελοντές έδωσε Μέσο Βαθμό Προτίμησης 3.15 στο μοντέλο μας και 3.82 στις ηχογραφήσεις.

Sisamaki Eirini
M.Sc. Thesis

Computer Science Department
University of Crete
Master's Thesis Supervisor: Professor, I. Stylianou

Friday, 25/10/2019, 14:00
Room K206, Computer Science Dept., University of Crete

“End-to-End Neural based Greek Text-to-Speech Synthesis”

ABSTRACT

Text-to-speech (TTS) synthesis is the automatic conversion of written text to spoken language. TTS systems play an important role in natural human-computer interaction. Concatenative speech synthesis and statistical parametric speech synthesis were the prominent methods used for decades. In the era of Deep learning, end-to-end TTS systems have dramatically improved the quality of synthetic speech. The aim of this work was the implementation of an end-to-end neural based TTS system for the Greek Language. The neural network architecture of Tacotron-2 is used for speech synthesis directly from text. The system is composed of a recurrent sequence-to-sequence feature

prediction network that maps character embeddings to acoustic features, followed by a modified WaveNet model acting as a vocoder to synthesize time-domain waveforms from the predicted acoustic features. Developing TTS systems for any given language is a significant challenge and requires large amount of high quality acoustic recordings. Because of this, these systems are only available for the most commonly and widely spoken languages. In this work, experiments are described for various languages and databases which are freely available. A Greek database, initially created for speech recognition, has been obtained from ILSP (Institute for Language and Speech Processing). In our first experiment, only 3 hours of recorded speech in Greek have been used. Then the technique of language adaptation has been applied, using 3 hours in Greek and 18 hours in Spanish. We also have applied speaker adaptation in order to produce speech with specific speakers from our database. Our TTS system for Greek can generate good quality of speech with very natural prosody. An evaluation with a listening test by 30 volunteers gave a score in MOS (Mean Opinion Score) of 3.15 to our model and 3.82 to the original recordings.