ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Μόρφη Γνωστοθέα - Βερονίκη
Μεταπτυχιακή Φοιτήτρια
Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτ. Εργασίας: **Επικ. Καθηγητής, Αθανάσιος Μουχτάρης**
Δευτέρα, 16 Φεβρουαρίου 2015, 11:00
Αίθουσα Τηλεδιάσκεψης Κ206, τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

" Ανάλυση/Σύνθεση Φωνής με τη χρήση ενός Προσαρμοστικού Αρμονικού Μοντέλου "

ΠΕΡΙΛΗΨΗ

Ένα μοντέλο παραγωγής ομιλίας το οποίο θεωρεί την ομιλία σαν το αποτέλεσμα του φιλτραρίσματος μιας κυματομορφής της γλωττιδικής διέγερσης από ένα χρονικά μεταβλητό γραμμικό φίλτρο το οποίο μοντελοποιεί τα κύρια χαρακτηριστικά της φωνητικής οδού χρησιμοποιείται ευρέως στην ψηφιακή επεξεργασία σημάτων ομιλίας. Σε πολλές εφαρμογές φωνής, δύο πιθανές καταστάσεις μπορούν να θεωρηθούν: η έμφωνη και η άφωνη. Τα μοντέλα φωνής συχνά διαχωρίζουν το φάσμα της ομιλίας σε αυτές τις δύο (ή ακόμη και περισσότερες) έμφωνες/άφωνες συχνοτικές ζώνες με τη χρήση ορίων στην συχνότητα. Ο έμφωνος λόγος μοντελοποιείται συνήθως ντετερμινιστικά στις χαμηλότερες συχνότητες, ενώ μια στοχαστική προσέγγιση χρησιμοποιείται για το ανώτερο μέρος των συχνοτήτων. Η Μέγιστη Έμφωνη Συχνότητα χωρίζει τα δύο αυτά μέρη. Ωστόσο, μπορεί να παρατηρηθεί από τους πραγματικούς μηχανισμούς παραγωγή φωνής ότι το φάσμα πλάτους της πηγής ελαττώνεται ομαλά χωρίς κάποια απότομη αλλαγή στην συχνότητα. Αναλόγως, χρειάζεται μεγάλη προσπάθεια από τη μεριά των μοντέλων πολλαπλών ζωνών για τον υπολογισμό αυτών τον ορίων. Συνεπώς, οι αλλοιώσεις που παράγονται από τις μεθόδους πολλαπλών ζωνών μπορούν να υποβαθμίσουν την ποιότητα μοντελοποίησης. Επιπλέον, ο μετασχηματισμός Fan Chirp (FChT), ο οποίος χρησιμοποιεί μια γραμμική βάση συχνοτήτων προσαρμοσμένη στις μη-στατικότητες του σήματος της φωνής, έχει επιδείξει αρμονικότητα σε υψηλότερες συχνότητες από

αυτές που παρατηρούνται συνήθως από το μετασχηματισμό Fourier (DFT). Συνεπώς, μια προσέγγιση πλήρους ζώνης είναι επιθυμητή.

Τα ημιτονοειδή και τα αρμονικά μοντέλα στοχεύουν στην αναπαράσταση ενός σήματος φωνής με ένα σετ από παραμέτρους όπως συχνότητες, πλάτη και φάσεις. Η ακρίβεια αυτών των παραμέτρων του μοντέλου είναι ένα βασικό ζήτημα. Όλα τα μοντέλα φωνής πρέπει να είναι και ακριβή και γρήγορα έτσι ώστε να αναπαριστούν το σήμα φωνής επαρκώς και να είναι ικανά να επεξεργάζονται μεγάλη ποσότητα δεδομένων σε ένα λογικό χρονικό πλαίσιο. Ως τώρα, το ημιτονοειδές μοντέλο (SM), όπου η γλωττιδική διέγερση αναπαρίσταται σαν το άθροισμα ημιτονοειδών κυμάτων, χρησιμοποιείται ευρέως σε πολλές εφαρμογές όπως ανάλυση φωνής, κωδικοποίηση και τροποποίηση φωνής. Ωστόσο, όπως δείχνουμε στις αξιολογήσεις αυτής της εργασίας, οι παράμετροι που υπολογίζονται από το SM δεν είναι τόσο ακριβείς όσο αυτές που υπολογίζονται από τα αρμονικά μοντέλα. Ακόμη, το προσαρμοστικό Σχεδόν-Αρμονικό μοντέλο (aQHM) έχει προταθεί σαν μία εναλλακτική και πιο προσαρμοστική μέθοδος ανάλυσης φωνής, η οποία χρησιμοποιεί μερικές από τις ιδιότητες τις αρμονικότητας των σημάτων. Το aQHM παρέχει περισσότερη ευελιξία από το FChT χρησιμοποιώντας ένα σετ μη-γραμμικών συναρτήσεων βάσης. Παρόλα αυτά, λόγω της υπόθεσης της aQHM, ότι το αρχικό σφάλμα των συχνοτήτων είναι περιορισμένο, μπορεί να προκληθεί σφάλμα στην αντιστοίχηση των συχνοτήτων. Ως εκ τούτου, καμία από τις μεθόδους δεν είναι κατάλληλη για μοντελοποίηση πλήρους φάσματος ενός σήματος φωνής.

Τα αρμονικά μοντέλα είχαν σχεδιαστεί αρχικά για την αναπαράσταση του ντετερμινιστικού μέρους της ομιλίας, αλλά, όπως υποδηλώνεται από την FChT, η χρήση ενός ορίου συχνότητας είναι αμφισβητήσιμη. Ως εκ τούτου, αξιοποιώντας τις ιδιότητες της aQHM, το προσαρμοστικό Αρμονικό Μοντέλο (aHM) πλήρους ζώνης μαζί με τους αντίστοιχους αλγόριθμους για τον υπολογισμό των αρμονικών μέχρι την συχνότητα Nyquist έχει προταθεί. Το aHM μοντέλο χρησιμοποιεί την λύση των Ελάχιστων Τετραγώνων (LS) στον Προσαρμοστικό Επαναληπτικό αλγόριθμο Αναμόρφωσης (AIR) έτσι ώστε να γίνει μια σωστή εκτίμηση της αναμόρφωσης της καμπύλης f_0 χωρίς τα προβλήματα λόγω σφαλμάτων στην συχνότητα. Αν και η aHM-AIR που χρησιμοποιεί την μέθοδο LS επιτρέπει μια εύρωστη εκτίμηση των αρμονικών συνιστωσών, εξαιτίας της χρήσης της LS, της λείπει η υπολογιστική αποδοτικότητα η οποία θα έκανε την χρήση της ιδανική για μεγάλες βάσεις δεδομένων.

Στην εργασία αυτή, μια μέθοδος επιλογής κορυφών (PP) προτείνεται ως αντικατάσταση της LS στον AIR αλγόριθμο. Για να ενσωματωθεί η προσαρμοστικότητα του προσαρμοστικού Αρμονικού Μοντέλου στην PP προσέγγιση, προτείνεται επιπλέον ένας προσαρμοστικός Διακριτός Μετασχηματισμός Fourier (aDFT), του οποίου η συχνοτική βάση μπορεί να ακολουθήσει πλήρως τις εναλλαγές της f_0 καμπύλης. Για να γίνει η αξιολόγηση της απόδοσης της προτεινόμενης μεθόδου, μετρήσαμε τον υπολογιστικό χρόνο και δείξαμε ότι ο αλγόριθμος έχει γίνει τέσσερις φορές πιο γρήγορος. Ακόμη, η ποιότητα της επανασύνθεσης διατηρείται σε σύγκριση με αυτή της aHM-AIR που χρησιμοποιεί την LS. Με την χρήση του σφάλματος του σήματος προς την ανακατασκευή του (SRER) και την εκτίμηση της αντιληπτικής ποιότητας της ομιλίας (PESQ), δείχνουμε ότι η ομιλία που

ανακατασκευάζεται με την χρήση της aHM-AIR με PP και aDFT διατηρεί την ποιότητα της aHM-AIR που χρησιμοποιεί την LS. Τελικά, επίσημα ακουστικά τεστ δείχνουν ότι η ομιλία που ανακατασκευάζεται από την aHM-AIR με PP και aDFT είναι παρόμοια με αυτήν που ανακατασκευάζεται από την aHM-AIR που χρησιμοποιεί την μέθοδο LS.

Morfi Gnostothea- Veroniki

M.Sc. Thesis
Computer Science Department
University of Crete
Master's Thesis Supervisor: Assistant Professor Athanasios Mouchtaris
Monday, 16/2/2015, 11:00
Room K206, Computer Science dept.,University of Crete

" Speech Analysis/Synthesis using an adaptive Harmonic Model"

ABSTRACT

A speech production model that views speech as the result of passing a glottal excitation waveform through a time-varying linear filter (the latter modeling the resonant characteristics of the vocal tract) is widely used in digital speech signal processing. In many speech applications, two possible states of the glottal excitation can be assumed: voiced or unvoiced. Voice models often split the speech spectrum into these two (or even more) voiced/unvoiced frequency bands using respective cutoff frequencies. Voiced speech is usually modeled deterministically in the lower frequencies, while a stochastic approach is used for the upper frequency part. A so-called Maximum Voiced Frequency separates the deterministic and stochastic parts. However, it can be observed from the actual voice production mechanisms that the amplitude spectrum of the voice source decreases smoothly without any abrupt frequency changes that would justify such a classification of the spectrum in deterministic and stochastic components. Accordingly, it becomes a struggle for multiband models to estimate these cutoff frequencies. Consequently, artifacts produced by multiband methods can degrade the perceived quality. Moreover, the

Fan Chirp Transformation (FChT), which uses a linear frequency basis adapted to the nonstationarities of the speech signal, has demonstrated that harmonicity is present at frequencies higher than those usually considered as voiced based on the Discrete Fourier Transform (DFT). This motivates alternative models which are based on a full-band modeling approach.

Sinusoidal and harmonic models aim to represent the speech signal with a set of parameters such as frequencies, amplitudes and phases. The accuracy and precision of the model parameters are key issues. All voice models have to be both precise and fast in order to represent the speech signal adequately and be able to process large amounts of data in a reasonable amount of time. So far, the Sinusoidal Model (SM), where the glottal excitation is represented as a sum of sine waves, has been widely used for many applications such as speech analysis, coding and modifications. However, as we show in our evaluations in this thesis, the estimated parameters are not as accurate as the ones computed by harmonic models. The adaptive Quasi-Harmonic Model (aQHM) has been proposed as an alternative and more adaptive method for speech analysis that uses some of the attributes of the harmonicity of a signal. The aQHM offers even more flexibility than the FChT by using a set of adaptive non-linear basis functions. However, due to the assumption made by aQHM, that the initial frequency tracks can have a confined error, a frequency matching problem may occur. Hence, neither method is very suitable for full-band modeling of a speech signal.

Harmonic models were initially designed for representation of the deterministic part of the speech, but, as implied by the FChT, the need of a cutoff frequency limit is questionable. Thus, exploiting the properties of aQHM, the full-band adaptive Harmonic Model (aHM) along with its corresponding algorithms for the estimation of harmonics up to the Nyquist frequency has been proposed. The aHM model uses the Least Squares (LS) solution in the Adaptive Iterative Refinement (AIR) algorithm in order to properly estimate a refinement of the f_0 curve without the problems caused by frequency errors. Even though aHM-AIR using LS allows for a robust estimation of the harmonic components, it lacks the computational efficiency that would make its use convenient for large databases, due to the use of the LS solution.

In this thesis, a Peak-Picking (PP) approach is suggested as a substitution to the LS solution used by the AIR algorithm. In order to integrate the adaptivity scheme of aHM in the PP approach, an adaptive Discrete Fourier Transform (aDFT), whose frequency basis can fully follow the variations of the f_0 curve, is also proposed. In order to evaluate the performance of the proposed method, the computational time has been calculated and an average time reduction of almost four times has been shown when comparing the proposed improvements to the original LS-based aHM-AIR algorithm. Additionally, the quality of the re-synthesis is preserved compared to the aHM-AIR using LS. With the use of Signal-To-Reconstruction-Error (SRER) and Perceptual Evaluation of Speech Quality (PESQ), we show that the speech reconstructed using aHM-AIR with PP and aDFT retains the quality of aHM-AIR using LS. Finally, formal listening tests show that the speech reconstructed by aHM-AIR with PP and aDFT is very similar to the one reconstructed by aHM-AIR using LS.