AUGUST 13 2005

Restoration of tracheoesophageal voice with LPC resynthesis

Yingyong Qi


https://doi.org/10.1121/1.2027663
properties that are typical of this accented context only. In particular, their temporal and spectral characteristics are close to the target values that one finds in the production of carefully spoken, isolated, accented syllables. Speech generated by means of such diphones often sounds over-articulated. In unaccented syllables (and perhaps elsewhere, too), one might prefer to use "reduced" diphones, which may lead to shorter and spectrally less extreme speech sounds. Perception experiments were run in order to assess the improvement that can be achieved by also using "reduced" diphones, which are extracted from unaccented syllables. In these tests, the use of "reduced" diphones was determined by the following variables: word stress, sentence accent, word and phrase boundaries, metrical structure. [Research sponsored by the Dutch government as part of the national SPIN program "Speech analysis and synthesis."]

FF21. Including frequency-dependent losses in a time-domain, articulatory synthesizer. R. S. McGowan (Haskins Laboratories, 270 Crown Street, New Haven, CT 06511)

The Webster horn equation describes acoustic wave propagation in tubes with cylindrical symmetry. This differential equation can be derived from the linearized equations of motion for air neglecting viscosity and heat conduction. A modified horn equation including viscous loss in the boundary-layer approximation can be derived. This modified equation includes a convolution integral over time; thus the computational problem is to find an approximation to the integral. First, the integral is discretized, and then a rational Padé approximation is used to approximate the resulting infinite sum. The frequency dependence of this approximation is examined and compared to the ideal frequency dependence. [Work supported by NIH grants HD-01994 and NS-13617 to Haskins Laboratories.]

FF22. Excitation waveform extraction for pitch control in residual-excited LPC speech synthesis. K. Itoh and H. Sato (Speech and Acoustics Laboratory, NTT Human Interface Laboratories, Musashino-shi, Tokyo 180, Japan)

The residual-excited LPC is one of the most effective techniques for producing high-quality synthetic speech. However, this technique has difficulty controlling pitch frequency when applied to realizing synthesized speech with a pitch contour different from original speech or when applied to realizing arbitrary speech created with concatenated spoken units. In previous methods, pitch-period LPC residual waveforms that extracted pitch synchronously have been used to control pitch frequency. However, the extraction position ($E_p$) and window length ($E_w$) are very critical to synthesized speech quality, and cause voice quality deterioration. This paper proposes a new method of excitation waveform extraction that automatically determines $E_p$ and $E_w$ using spectral envelope distortion criteria between input and synthetic speech. Subjective evaluation experiments indicate that the pitch frequency pattern can be changed with a relatively small deterioration in quality. Application of this method to arbitrary speech synthesis will also be discussed.

FF23. Phonological rules for a semantics-to-speech system of Japanese: A rule application experiment. Shigeru Sato (Informatics Laboratory, Tohoku Institute of Technology, Yagiyama-kasumicho 35-1, Taihaku, Sendai 982, Japan)

In order to confirm validity of the phonological rules of Japanese and their application algorithms, a phonetic form generation experiment is performed in the phonological component of a semantics-to-speech system implemented in the computer. The rules are tiered according, first, to the nature of data they handle: segmental versus accentual, and, second, to their rule-applicational relevance to syntax: cyclic versus noncyclic. Successful integration of segmental and accentual phases of phonological processing is shown to be possible using the concept of recursive adjunction of a suffix to a stem. The rules were installed in the computer in a human-editable format, translated by a rule compiler into executable LISP functions, tested, and found to be valid in actual realization of synthesized speech.

FF24. On the unit set design for speech synthesis by rule using nonuniform units. Yoshinori Sagisaka (ATR Interpreting Telephony Research Laboratories, Kyoto 619-02, Japan)

The unit set design algorithm is proposed for nonuniform synthesis units using entropy measure. This algorithm enables optimal unit set building according to statistic phonotactic characteristics without using a priori linguistic knowledge. By considering each phoneme sequence unit as a state in nonuniform order Markov process, the entropy of a unit set is defined by the entropy of a Markov information source. Each unit candidate corresponds to a new state and the new state generation corresponds to the enrollment of the candidate as a unit member. Unit enrollment takes place by finding the corresponding new state that gives the greatest decrease to the entropy of the set. Applying this algorithm to Japanese using phoneme sequence statistics derived from a Japanese word dictionary, the following characteristics were observed. (1) Linguistically well-known CV-type syllables and diphthongs are selected in the very early stage. (2) Though the frequency used dyadic VC-type units are not selected in the early stage, the CVVCV-type units that contain those VC-type units are selected in the earlier stages. (3) Many early selected units coincide with the phoneme sequence contained in morphemes. Through this experiment, it has turned out that the nonuniform unit set covered phoneme sequences about two times more efficiently than the usual uniform unit set.

FF25. Robust ARMA analysis for the determination of voice source and vocal tract control parameters in speech synthesis. Johan de Veth, Louis Boves (Department of Language and Speech, Nijmegen University, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands), and Wim van Golstein Brouwers (PTT Research Neher Laboratory, Leidschendam, The Netherlands)

A cascade six pole-pair and five zero-pair synthesizer has been developed as part of a text-to-speech conversion system for Dutch. Control information for this synthesizer is derived from a.o. measurements on natural speech. A pitch synchronous robust ARMA analysis technique was developed and applied to utterances produced by a number of adult male talkers. The resulting pole-zero parameters were separated into sets pertaining to the source of the vocal tract. The vocal tract parameters were corrected in those frames where the analysis method made occasional mistakes. Analysis-resynthesis of sentence material using the corrected vocal tract parameters to control the synthesizer driven by impulse and noise excitation yielded high-quality synthetic speech. The tract parameters were then used to inverse filter the speech, to obtain the source function, that was subsequently parametrized using the Liljencrants-Fant model. It is hoped that the speech quality will be improved by replacing the impulse excitation by the controllable source model.

FF26. Restoration of tracheoesophageal voice with LPC resynthesis. Yingyong Qi (Department of Speech and Hearing Science, University of Arizona, Tucson, AZ 85721)

Four vowels, [i], [e], [a], and [u], and one diphthong [ou], produced by two male and two female tracheoesophageal speakers, were analyzed with the LPC autorecorrelation method. The vowels were synthesized by replacing the original source with an impulse train. The fundamental
frequency of the impulse train was 100 Hz for the male and 200 Hz for the female speakers. The results of an identification experiment indicated that both the vowel and the gender of the speaker can be better identified from the synthesized vowels than from the original ones. The possibility of improving the quality of tracheoesophageal speech and building prosthetic devices with the LPC technique will be discussed.

FF27. Effects of spectral smearing on speech perception. Mariken ter Keurs, Joost M. Fenton, and Reinier Plomp (Department of Otolaryngology, Free University Hospital, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands)

Connected speech presented in quiet is highly redundant. For such a condition, the resolution of spectral contrasts by the ear seems to be much larger than would be required. This suggests that the ear's high selectivity in frequency is particularly important to understanding speech in the presence of interfering sounds. The effect of reduced frequency resolution on the speech-reception threshold (SRT) for sentences in noise was investigated for eight normal-hearing subjects by simulating an auditory system with variable frequency selectivity. Signal processing was performed by short-time fast Fourier transforms (FFT), reduction of contrast in the spectral envelope without affecting the harmonic structure, and overlapping additions to reconstruct a continuous signal. The spectral envelope in the frequency region from 100 to 8000 Hz was smeared over fixed relative bandwidths of 1/8, 1/4, 1/2, 1, 2, 4, and 8 octaves. Results show that the SRT increases progressively as the spectral envelope is smeared over bandwidths exceeding the ear's critical bandwidth. In a second experiment phoneme confusions as a result of three different degrees of spectral smearing are studied in nonsense CVC syllables.

FF28. Perception of vowels by budgerigars (Melopsittacus undulatus). Robert J. Dooling, Susan D. Brown, Amy Nespor (Department of Psychology, University of Maryland, College Park, MD 20742), and John W. Hawks (Central Institute for the Deaf, St. Louis, MO 63110)

Budgerigars (parakeets) were trained using operant conditioning techniques to respond to differences between speech stimuli. Response latencies were used to construct similarity matrices and multidimensional scaling procedures were then used to produce perceptual (spatial) maps of these stimuli. For natural vowel tokens, budgerigars showed evidence of perceiving phonetic categories in spite of variation in talkers including talker gender and talker age. Experiments with synthetic vowel tokens generally confirmed and extended these findings. Multiple regression techniques revealed that the perceptual dimensions obtained from MDS were highly correlated with the frequencies of the first and second formants. These results suggest that both natural and synthetic vowels are probably perceived in similar ways by budgerigars and humans. These results have relevance for theories of speech perception and language development in humans. [Work supported by NIH.]

FF29. Articulation index importance functions for contextual speech materials. Theodore S. Bell, Donald D. Dirks, and Timothy Trine (UCLA School of Medicine, Head & Neck Surgery, Rehabilitation Building, Room 31-24, Los Angeles, CA 90024-1794)

The relative importance of one-third octave frequency bands toward the intelligibility of speech in various contexts was examined. Thirty-five young normal hearing adults heard sentences in which the final word (target stimulus) was either predictable (probability-high; PH) or unpredictable (probability-low; PL) from the context of the sentence. Sentences were presented at S/N's from -8 to +14 dB in a noise shaped to conform to the peak spectrum of the speech, and crossover frequencies as related to the articulation index (AI) were determined by successively high-pass and low-pass filtering the stimuli to bisect intelligibility. Results indicated only slight difference in 1/3 octave importance functions as the result of context (PH vs PL), although the crossover frequency was significantly different and showed no interaction attributable to the noise. The effect of context in these speech materials was related to differences in perceptual dynamic range more so than the frequency importance function. These results are contrasted to other recent studies, and methodological and theoretical aspects of parameter estimation in the AI model are discussed.

FF30. An algorithm for distinguishing between voiced stops and voiceless fricatives. LaDeana F. Weigelt, Steven J. Sadoff, and James D. Miller (Central Institute for the Deaf, 818 South Euclid Avenue, St. Louis, MO 63110)

An algorithm has previously been reported which distinguished voiceless stops from voiceless fricatives with a success rate of 96.8% [J. Acoust. Soc. Am. Suppl. 1 85, S56 (1989)]. With only slight modifications, this algorithm also makes the voiced stop/fricative distinction. Here, results are presented on the modified algorithm. The input signal is high-pass filtered (cutoff frequency of 125 Hz) and two measures of the resulting waveform are used: the rms energy and the derivative of rms energy over time (termed rate of rise, ROR). The ROR is used as the primary classifier while energy pulse duration and relative level are used to discard spurious, irrelevant peaks. Peaks in ROR are considered in order of magnitude for relevance to the stop/fricative distinction. The resulting algorithm was tested on 420 CVC tokens (three male speakers, three female speakers, three stops and an affricate [B,D,G,JI], four fricatives [Z,ZH,V, DH], and ten vowel contexts [IY, IH, EH, AE, AA, AH, AO, UH, UW, ER]) recorded in an anechoic chamber. Data from two male and two female speakers (280 tokens) were used as a training set, and the remaining data (140 tokens) were used as a test set. The overall success rate was 97.9%. [Work supported by AFOSR.]

FF31. The role of speech rate in social evaluation. Cynthia L. Crown (Department of Psychology, Xavier University, Cincinnati, OH 45207)

Prior studies concerned with the relation of speech rate to the personality variables of extraversion and to social evaluation (i.e., interpersonal perception) have primarily examined monologues. The study reported here was designed to (a) explore such dependencies in dialogues, and (b) investigate the possibility that speech rate may be influenced by differences in interpersonal attraction. The 38 female pairs who participated in the study were selected on the basis of a sociogram designed to assess their attraction to each other prior to their participation in the study. Their sociogram scores divided them into three groups: those who liked each other, those who disliked each other, and those who were unacquainted with each other. Each pair engaged in a 30-min conversation from which the speech rate of each participant was electronically determined. The result of multiple regression analyses indicated that those women who spoke more quickly perceived themselves more positively, whereas those who spoke more slowly viewed their partners more positively. However, variations in rate were related to neither interpersonal attraction nor extraversion.

FF32. Two paradigms for examining the role of phonological stress in sentence processing. David W. Gow, Jr. and Peter C. Gordon (Department of Psychology, Harvard University, Cambridge, MA 02138)

The role of phonological stress in sentence processing was studied using a syllable monitoring task, as well as a new short-term memory probe task. In both tasks, target syllables were embedded in byssylabic noun/verb homophones with syntactic category-dependent stress patterns. The stress and position within a word of target syllables were manipulated. Syntactic context was also manipulated to examine the role of syntactic constraints on the anticipation of stressed syllables. The syllable monitoring task reproduced the familiar facilitation of stressed syllable detection, and showed an effect of syntactic constraint. Stress facilitation was not found when targets were word-final. These results were interpreted in terms of stress-facilitated lexical segmentation and access processes. The short-term memory probe task introduced in this research was shown to be sensitive to stress, but not to other sentential factors that were examined. The application of this paradigm to issues in metrical phonology and short-term memory is discussed. [Work supported by AFOSR.]