



**VOICE CONVERSION: A MODEL FOR STUDYING VOICE QUALITY AND SPEAKER NORMALIZATION**

D.G. Childers<sup>1</sup>, Ke Wu<sup>1</sup>, and D.M. Hicks<sup>2</sup>

**ABSTRACT**

This paper describes a number of speech analysis and synthesis factors that are important for synthesizing speech of high quality, i.e., that sounds natural. We have considered such factors as those related to 1) the synthesis model, 2) objective measures of quality including spectral replication, continuity, and tracking, 3) glottal excitation waveforms and parameters including source-tract interaction, jitter and shimmer, and 4) speech analysis, e.g., window shapes and sizes and the accurate identification of voiced/unvoiced/silent segments and fundamental frequency. We have tested three synthesizers (LPC, formant and articulatory) and present conclusions from both formal and informal listener evaluations for the LPC and formant synthesizers.

**INTRODUCTION**

Our analysis-synthesis system uses two channels (refs 1 and 2). One channel is the acoustic signal, while the other is the electroglottogram (EGG). This system has been used to study factors related to voice quality (as opposed to speech intelligibility). One data model we have examined has been the parameter sets extracted from the acoustic and EGG signals that are related to a speaker's gender (male/female voice characteristics) (ref 3). We have parameterized male speech and determined factors needed to transform (convert) this speech to sound like that of a female talker. The reverse process has been studied as well. Several factors important for this transformation process are pitch conversion, warping of the frequency spectrum (spectral expansion and compression), proper specification of the formant frequencies and bandwidths, and factors related to excitation waveform and timing. We have also examined the parametric representation of speech and EGG signals for the automatic recognition of male and female talkers (ref 3).

Besides considering the factors responsible for speech quality related to the choice of speech synthesizers, we describe spectral factors including conservation, continuity and tracking. The quality (naturalness) of synthetic speech is enhanced when the excitation waveform incorporates glottal vibratory information. Glottal events such as instants of opening and closure and their durations are important and can be measured from the EGG. These timing parameters can be used to position impulses and calculate parameters for various excitation model waveforms. When actual glottal vibratory parameters are used to synthesize speech, its quality is improved over the speech

<sup>1</sup> Dept. of Electrical Engineering, University of Florida, Gainesville, FL 32611

<sup>2</sup> Dept. of Speech, University of Florida, Gainesville, FL 32611

synthesized using conventional vocoder excitations. We have also found that the differentiated EGG waveform produces the best results in the LPC synthesizer when compared to other excitations. As we point out in the paper, the time domain characteristics of the glottal excitation waveform are more important than its spectral characteristics. So we need new analysis tools to extract these parameters from the speech and auxiliary signals, such as the EGG.

### SYNTHESIS MODEL FACTORS

Our remarks are limited to three synthesizers, formant, articulatory, and linear predictive coding. We have found that the major factors affecting the quality of speech synthesis are:

- \* formant locations and bandwidths (formant and articulatory synthesizers)
- \* pole positions and segment (frame) coupling (LPC synthesizer)
- \* excitation waveshape (all three synthesizers)
- \* source-tract interaction simulation (all three synthesizers)
- \* analysis factors:
  - \*  $F_0$  estimation
  - \* voiced/unvoiced/silent decisions
  - \* pitch synchronous methods

### OBJECTIVE MEASURES

Some factors emerging as important concepts for objective measures of speech quality are spectral continuity (the smoothing of abrupt changes that may occur in the speech energy level from analyzed segment to segment), a spectral distortion measure between two speech segments, and stable distortion statistics (ref 4).

**Spectral conservation** or the replication of the spectral features of a speech segment is important for high quality speech synthesis. We have found this to be the case in our voice conversion model, i.e., converting the speech of one gender to sound like that of the other gender (ref 5). To achieve high quality mimicry, the spectral factor related to vocal tract length compensation must not distort the final spectrum used to mimic the desired voice. The formants and their bandwidths must be either compressed or expanded when translating from the source spectrum to the target (desired) spectrum. This spectral compression/expansion factor for converting one voice to another must not be applied uniformly across the entire spectrum, i.e., to all formants equally, otherwise, one will hear high frequency distortion in the converted synthetic speech, e.g., a male voice converted to a female voice will sound "metallic." To avoid this phenomenon, the higher formants should be shifted less than the lower formants.

**Spectral continuity** may become disrupted during analysis by discontinuities in the gain function or LPC coefficients. To restore the quality of the synthesized speech, one must 1) smooth the spectral compression/expansion factor across the entire sentence, maintaining overall spectral continuity and 2) modify the gain factor for each frame accordingly. The overall energy must be maintained at the desired level within the frame. This process involves dynamically computing the spectral compression/expansion factors and a modified gain contour using both the source and target speech data.

**Spectral tracking** is important for spectral conservation and continuity. Pitch synchronous covariance analysis improves spectral tracking as does a signal dependent analysis-synthesis technique (ref 5). The reasons for this are as follows. Conventional analysis-synthesis systems do not vary the frame size, frame rate and number of parameters per frame size. These values are fixed as a compromise among the conflicting requirements of temporal resolution (frame rate), spectral resolution (frame size and number of LPCs), bit rate (number of LPCs), quality and intelligibility. The draw back of using a smaller frame size is that a poorer spectral resolution is obtained which affects voiced segments and the disadvantage of using a larger frame size is that a poorer temporal resolution is obtained which affects the transient segments. For silent and unvoiced segments, only the gross spectral characteristics need be represented. In fact, higher resolution through the use of a high order LPC model may produce spurious peaks, giving the perceptual impression of incorrect formant locations. In a transition region from a voiced region to another region or vice versa, a small analysis frame size and high spectral resolution would be required to track the formant transitions. In an analysis using a fixed frame size and a fixed number of parameters, all the segments are represented alike. Whereas a more idealistic representation requires a variable number of parameters per frame depending on the nature of the segment.

The pitch synchronous covariance method (ref 6) has been found to give very good if not the best formant frequency and bandwidth estimates for spectral tracking (ref 2). We have looked at two algorithms using the EGG and speech signals. One algorithm uses one pitch period-long frames; the other uses a frame duration corresponding to the closed glottal interval. The covariance method applied to the glottal closed phase interval provides the smoothest spectral formant contour. The fixed frame autocorrelation method yields a "noisy" contour. While the covariance method using one pitch period-long frame gives a contour intermediate between these two methods (ref 2). A pitch-synchronous LPC analysis-synthesis system similar to ours has also been tried by Kuwabara (ref 7).

Another approach is to use a system that can change the frame size and rate, number of LPCs, pre-emphasis factor, glottal excitation pulse shape, and other features (ref 5).

#### **GLOTTAL SOURCE RELATED FACTORS**

Voiced/unvoiced/silent interval detection is important for speech synthesis, as is voicing fundamental frequency estimation. To improve our procedures we use both the EGG and speech signals. The synthesizer excitation waveform should be close to the original pulse shape if naturalness is to be obtained in the synthetic sample. Source-tract interaction should be included as well as the proper amount of jitter and shimmer.

## OTHER ANALYSIS-SYNTHESIS RELATED FACTORS

Other parameters that affect the quality of the synthetic speech are the sampling rate, analysis-synthesis frame size and frame rate, types of analysis windows used, number of formants, and for the articulatory synthesizer the number of vocal tract sections to be used.

## RESULTS

Through formal and informal listening tests we have concluded that 1) replication of the spectral pattern of the speech is important, 2) the pitch period must be replicated accurately, 3) the glottal excitation pulse parameters are critical, including source-tract interaction, and 4) pitch synchronous covariance analysis or signal-dependent analysis improves the quality of the synthetic speech.

## DISCUSSION

Previous evidence, further substantiated by our result, supports the notion that the quality (naturalness) of synthetic speech is enhanced when the excitation waveform incorporates glottal vibratory information. The glottal events such as instant of opening and closure and their durations can be measured with the EGG signal. These timing parameters can be used to position impulses or calculate parameters for various excitation models. We have established that when actual glottal vibratory parameters are used to synthesize speech, the quality of the synthesized speech is improved. This is true not only for conventional excitations, but also for the differentiated EGG waveform which was found to produce the best results (in the LPC synthesizer) when compared with speech synthesized using other excitations. The latter result is due to the fact that 1) the excitation energy is distributed over the pitch period, 2) the speaker's jitter is included as a natural by-product, 3) the excitation spectrum is relatively flat, and 4) the peaks in the excitation waveform are located at the instants of glottal closure and opening where the primary and secondary excitations, respectively, normally occur.

Our results should prove useful for improving speaker independent speech recognition systems since we believe that an automated system for distinguishing male and female talkers can be implemented using this information.

## REFERENCES

1. D.G. Childers and J.N. Larar, IEEE Trans. Biomed. Engr., BME-31, 807 (1984)
2. A.K. Krishnamurthy and D.G. Childers, IEEE Trans. Acoust., Speech and Signal processing, ASSP-34, 730 (1986)
3. D.G. Childers, K. Wu and D.M. Hicks, ICASSP, 1, 293 (1987)
4. B.H. Juang, AT&T Tech. J., 63, 1477 (1984)
5. D.G. Childers, Proc. Voice I/O Systems Appl. Conf., 349 (1985)
6. S. Chandra and W.C. Lin, IEEE Trans. Acoust., Speech and Signal Processing, ASSP-22, 403 (1974)
7. H. Kuwabara, Speech Comm., 3, 211 (1984)