MEASURING SOURCE-TRACT INTERACTION FROM SPEECH

A.S. Ananth, D.G. Childers, and B. Yegnanarayana*

Department of Electrical Engineering
University of Florida
Gainesville, Florida 32611

* Department of Computer Science
Indian Institute of Technology
Madras 600 036, India

ABSTRACT

The objective of our study is to define and measure source-tract interaction using the speech signal. Numerous researchers have conjectured that the subglottal, glottal, and supraglottal portions of our own speech production mechanism may interact, affecting the quality of our voice. Several methods for incorporating the effects of source-tract interaction into a synthetic speech model have been suggested. Speech synthesized with source-tract interaction sounds more natural than speech generated without such interaction.

Can source-tract interaction be parameterized using the speech signal and incorporated into vocoders and speech synthesizers? Can this measurement be accomplished on a pitch period by pitch period basis?

INTRODUCTION

We describe a filtering system in the modified frequency domain using group delay functions for studying formant waveforms. This system is used to analyze changes in the first formant bandwidth during a single glottal cycle. This information is used to infer the presence of source-tract interaction. The problem we are attempting to solve is essentially the same as the one which attempts to obtain good power spectral estimates using short data records.

Source-tract interaction has been conjectured to be important for synthesizing high quality, natural sounding speech [1-6]. Generally, we consider the speech production model as consisting of three cascaded sections which model respectively the subglottal, glottal, and supraglottal regions [4]. These sections are usually considered distinct and separable.

Frequently, the subglottal region model is ignored entirely, which may be a major error since during the open glottal phase, the subglottal system may damp the vocal tract filter (supraglottal region), especially the first formant. Thus, the "apparent" vocal tract filter characteristics change, e.g., the formant bandwidths increase.

When the glottis is closed, the vocal tract filter is typically thought to be best represented by an all pole model of order 10 to 14. These intra-pitch period changes, which are conjectured to occur in speech production, are almost never incorporated into a source-tract filter model, except as gross adjustments to formant bandwidths.

An example of the changes one can see in the vocal tract filter formant structure is depicted in Figure 1 [7]. The speech data were FFT analyzed. The spectra are shown for the closed glottal interval, open glottal interval, and the combined intervals. The formant structure is apparent for the closed glottal interval and somewhat less so for the combined interval, but the spectrum for the open glottal interval is markedly different. These spectra are smooth due to the few samples available for the various intervals.

There are a number of reasons why our present speech production models inadequately represent the intra-pitch period variations. A detailed knowledge of the interrelationships between the vibratory pattern of the vocal folds, the glottal area function, and the glottal volume-velocity is not available, although this is under study [8]. Until recently researchers could not easily monitor the vibratory patterns of the vocal folds during speech production. The electroglottograph measures various vocal fold events, e.g., glottal opening, glottal open phase, glottal closing, and glottal closure [8-10]. Furthermore, special filtering techniques for separating formant frequencies have recently been improved [11].

PREVIOUS RESEARCH

Fant [1-3] and his co-workers conjecture that the vocal tract system characteristics change during a single period of glottal vibration. This is due to the closed and open glottal phases. During the glottal open phase, the resonances of the vocal tract system are thought to be broader than during the closed phase. These resonance changes alter the glottal volume-velocity waveform due to source-tract loading variations during a glottal cycle. The presence of vocal tract loading on the glottal source has been confirmed only indirectly [1-5]. Synthetic speech, produced using source-tract interaction, sounds more natural than synthesis generated without such interaction [6].

CH2118-8/85/0000-1093 $1.00 © 1985 IEEE

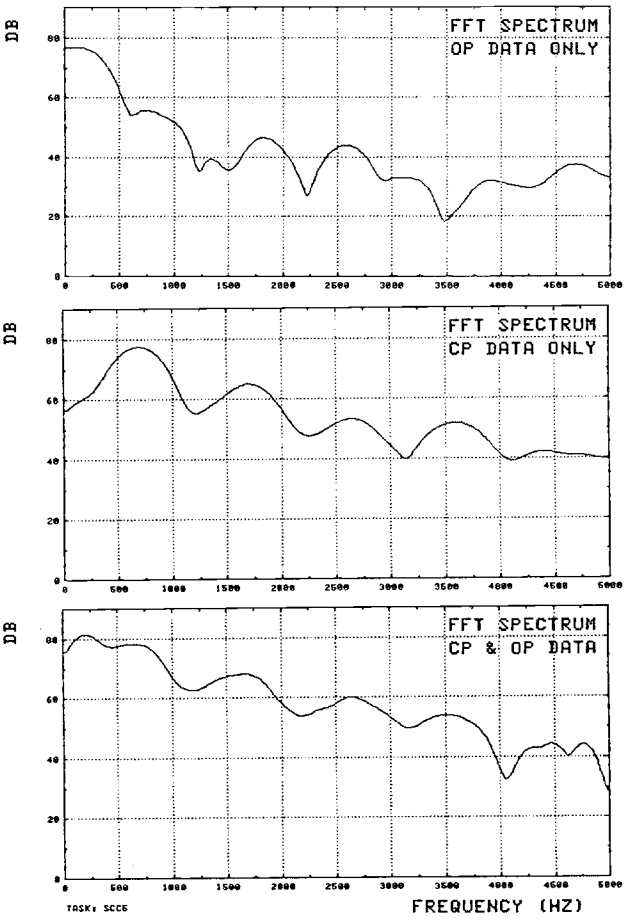

Fig. 1 Spectra for vocal tract filter for open phase (OP), closed phase (CP) and combined intervals.

Rothenburg [4-5] has studied the interaction between the first formant frequency, F_1, and the fundamental frequency of voicing, F_0. He has found that the ratio of F_1 to F_0 can serve as an index for predicting the effects of source-tract interaction.

Several parameters have been emerging from this research. Given the glottal area function and lung pressure, Fant [1-3] has been attempting to compute the glottal volume-velocity waveform. These analytical expressions decompose the volume-velocity (glottal flow) into a main pulse shape and a superimposed ripple component. The latter is conjectured to be primarily due to interaction between the glottal source volume-velocity pulse waveshape and the resonant circuit model of the first formant frequency. The slope of the glottal pulse during closure is an important factor affecting the vocal tract excitation [3-4]. The differentiated glottal pulse waveform is significant because this waveform is used most frequently as the excitation in source-tract models.

These factors have major importance in modeling vocal source and tract models. Unfortunately, these parameters are difficult to measure from the speech signal.

If source-tract interaction is important for synthetically reproducing high quality, natural sounding speech, then we need a method for measuring such interaction from the speech signal. A parametric represenation of such interaction would allow the convenient reproduction of high quality speech via synthesizers and vocoders.

A MEASURE OF SOURCE-TRACT INTERACTION

Our working definition for source-tract interaction is the variation in damping of the speech waveform which occurs as a result of variations in the glottal source characteristics during a pitch period. Investigators typically try to use the closed glottal interval to derive the filter pole positions for the vocal tract. We frequently lack the necessary information to do this analysis accurately. So we determine the vocal tract filter parameters using a glottal representation averaged over both the open and closed glottal intervals. When the closed glottal interval is known and the vocal tract filter parameters are derived using this knowledge, we rarely modify the vocal tract filter parameters during the open glottal phase. This seems contradictory since usually the closed glottal interval is shorter than the open glottal interval.

The change in damping of the first formant waveform during one pitch period is measured by our system. If the damping of the first formant ripple remains unchanged during one pitch period, then source-tract interaction is said not to take place. From a modeling point of view we may examine the envelope of the first formant ripple. The envelope of this ripple might be described using a damping factor such as e^{-at} over one pitch period, $T_i \leq t \leq T_{i+1}$, where $(T_{i+1} - T_i)$ is a pitch period. For source-tract interaction to occur, then the damping coefficient, a, must change during one pitch period. Equivalently we can measure the change in the first formant bandwidth during one pitch period. Generally, we expect an increase in damping to occur during the open glottal phase if our hypothesis about source-tract interaction is correct. When the glottis is closed we expect to see the natural response of the vocal tract, which is relatively undamped.

MEASUREMENT PROCEDURE

To analyze the output of a time-varying system, e.g., the vocal tract and vocal source, using short data records, we developed an algorithm to derive individual formant waves from speech using the properties of group delay functions [11]. We have found that we may infer the presence of source-tract interaction by examining the first formant waveform. The first formant waveform was simulated by changing the bandwidth during a single glottal vibratory cycle, i.e., from closed phase to open phase. The shapes of these simulated waves illustrate how we may

draw inferences about source-tract interaction from the first formant wave measured from an actual speech signal; see Figure 2.

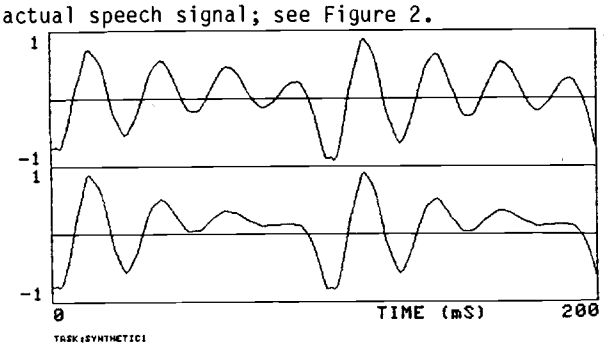

FIG. 2 Simulated first formant waveform. Upper: no bandwidth change. Lower: bandwidth change in one pitch interval.

Our analysis procedure is outlined in Figure 3. We adopted this approach over the conventional linear prediction (LP) method because the latter requires a frame size of 10ms or more to obtain proper spectral resolution. This duration is usually much longer than a pitch period and the LP coefficients derived for this analysis interval describe the average characteristics of the source and tract system. To measur a resonance damping for intervals as short as the closed or open glottal phase we must analyze speech segments as short as 1 ms, or 10 samples (at a 10 KHz sampling rate). This rules out LP analysis and the covariance method.

Our analysis procedure is broken into two steps. First, an initial estimate of the first, second, and third formant frequencies and their bandwidths is obtained using an 8th or 12th order LP analysis of a relative long data record. This LP modeling ensures an all pole structure prior to the group delay analysis. The approximate formant frequency locations are identified. The group delay function from speech is calculated and windowed for the desired formant [11]. A modified Fourier transform magnitude is derived from the truncated group delay function. This magnitude, together with the original phase is then used to estimate the desired formant waveform (first, second, or third). See Figure 3 for a schematic representation of our approach and actual processed speech data for the vowel /i/. Here we see the electroglottograph (EGG) waveform, the speech signal, and the extracted first formant waveform, all synchronized. For this discussion, glottal closure can be considered as taking place as the EGG crosses the zero axis in a negative direction, while the glottal opening occurs as the EGG crosses the axis in a positive direction.

Figure 4 presents illustrative results for simulated data, such as that shown in Figure 2.

The advantage of this method is that the magnitude characteristics of the desired formant and the phase characteristics of the entire signal are preserved. If we seek to recover the first formant, as an example, we are able to remove the effects of all other formants. With this procedure we may analyze short speech segments and estimate the changes in the desired formant frequency and bandwidth within a pitch period using a second order LP analysis.

DISCUSSION AND SUMMARY

Basically, we seek an analysis procedure which will recognize changes in system characteristics within a pitch period. This is the familiar problem of trying to obtain good power spectrum estimates using short data records.

Other analysis methods are being examined as well, such as low order (2nd order) LP and covariance analysis using longer data records. With these techniques we are looking at the residual error and the total squared error [7].

We have used a group delay filter system to remove the effects of other formants while

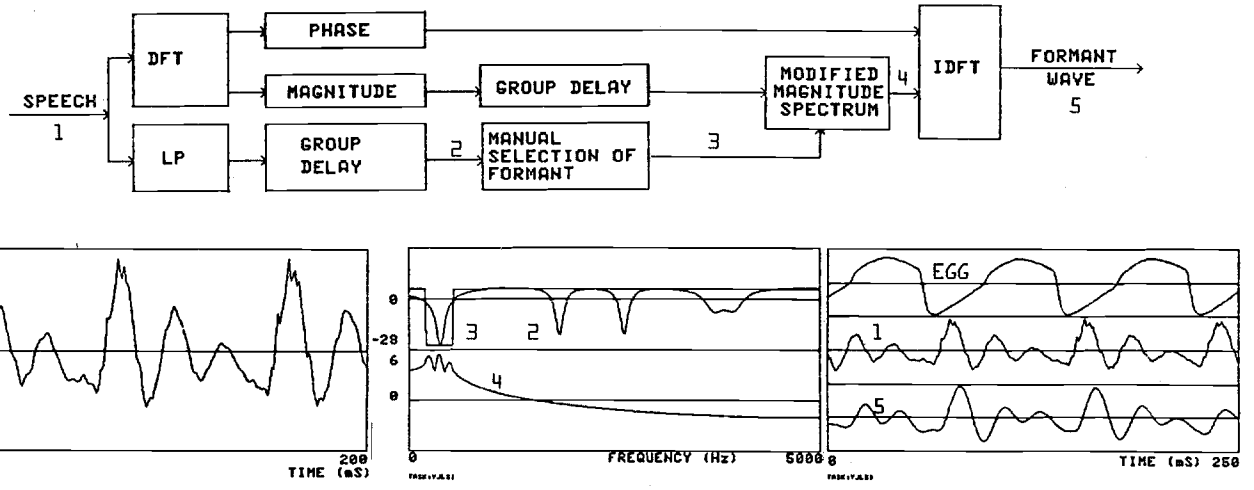

FIG. 3 Methodology for extracting formant waveforms using group delay functions. Actual speech data shown.

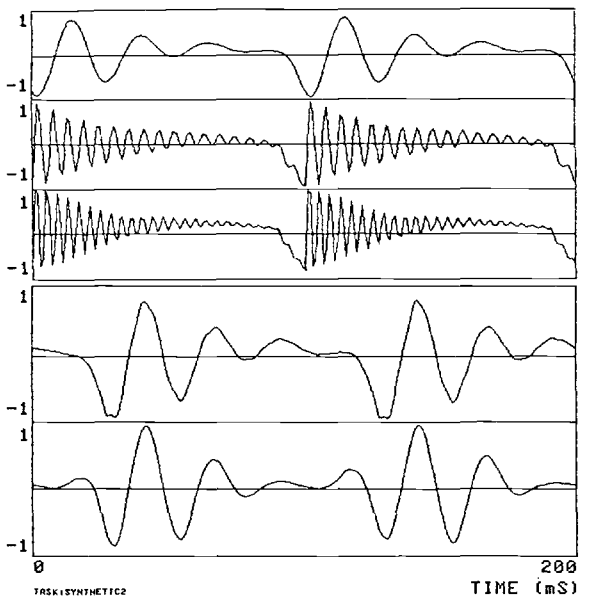

FIG. 4 Simulated data and extracted first formant wavefrom. From top to bottom
F_1: 400 Hz, ±1
F_2: 1800 Hz, ±0.5
F_3: 2570 Hz, ±0.025
$F_1 + F_2 + F_3$
Estimated F_1

investigating one formant. Usually, the presence of source-tract interaction is observed only in the first formant waveform and then only when the ratio of F_1 to F_0 is greater than 3 or 4.

Our studies to date have shown the presence of source-tract interaction in actual speech data in only a few cases. The major difficulty of analyzing speech data for these effects is the short data record length available for analysis. Simulated studies have shown that the analysis technique can work.

REFERENCES

1. G. Fant, The voice source-acoustic modeling, STL-QPSR 4/1982, pp. 28-48.

2. G. Fant, Preliminaries to analysis of the human voice source, to be published in working papers, MIT Speech Group, 1982, STL-QPSR 4/1982, pp. 1-27.

3. T.V. Ananthapadmanobha and G. Fant, Calculation of true glottal flow and its components, Speech Communication, vol. 1, pp. 167-184, 1982.

4. M. Rothenberg, An interactive model for the voice source, Ch. 12 in B.M. Bless and J.H. Abbs (Eds.), Vocal Fold Physiology, Contemporary Research and Clinical Research and Clininical Issues, College Hill Press, San Diego, 1983, pp. 155-165.

5. M. Rothenberg, Acoustic interaction between the glottal source and the vocal tract, Ch. 21, in K.N. Stevens and M. Hirano, Vocal Fold Physiology, University of Tokyo Press, 1981, pp. 303-323.

6. D.G. Childers, J.J. Yea, and E.L. Bocchieri, Source/vocal-tract interaction in speech and singing synthesis, presented at and to appear in Proceedings Stockholm Music Acoustic Conference, July 28 - August 1, 1983.

7. J.N. Larar and Y.A. Alsaka, Variability in closed phase analysis of speech, ICASSP-85, March 26-29, 1985, Tampa.

8. D.G. Childers, J.M. Naik, J.N. Larar, A.K. Krishnamurthy, and G.P. Moore, Electroglottography, speech, and ultra high-speed cinematography, paper presented at and to appear as a chapter in Vocal Fold Physiology: Physiology and Biophysics of Voice, May 4-7, 1983, University of Iowa.

9. D.G. Childers and A.K. Krishnamurthy, A critical review of electroglottography, CRC Critical Reviews in Bioengineering, (in press).

10. D.G. Childers and J.N. Larar, Electroglottography for laryngeal function assessment and speech analysis, IEEE Trans. on Biomed. Engr., vol. BME-31, December, 1984, (in press).

11. B. Yegnanarayana, D.K. Saikia, and T.R. Kishnan, Significance of group delay functions in signal reconstruction from spectrum magnitude or phase, IEEE Trans. ASSP, vol. ASSP-32, pp. 610-623, June 1984.

ACKNOWLEDGEMENT

This work was supported in part of grants NIH NS17078, NSF ECS 811 6341, University of Florida Center of Excellence, and an equipment grant from Digital Equipment Corporation.