VOICE CONVERSION: FACTORS RESPONSIBLE FOR QUALITY

D.G. Childers, B. Yegnanarayana*, and Ke Wu

Department of Electrical Engineering
University of Florida
Gainesville, FL 32611

*Department of Computer Science
Indian Institute of Technology
Madras 600036, India

ABSTRACT

A flexible analysis-synthesis system with signal dependent features is described and used to realize some desired voice characteristics in synthesized speech. The intelligibility of synthetic speech appears to depend on the ability to reproduce dynamic sounds such as stops, whereas the quality of voice is mainly determined by the true reproduction of voiced segments. We describe our work in converting the speech of one speaker to sound like that of another. A number of factors are important for maintaining the quality of the voice during this conversion process. These factors are derived from both the speech and electroglottograph signals.

INTRODUCTION

For several years we have been investigating factors in speech production systems which influence the quality and intelligibility of voice reproduction [1-5]. Recently, we reported on a flexible analysis-synthesis system which was developed for creating a voice using parameters obtained from the analysis of a speech signal [5]. This system has been used for studies in time expansion and compression, pitch modifications, and spectral expansion and compression. Factors responsible for unnaturalness in synthetic speech have been and continue to be studied. The shape of the (glottal) excitation pulse has been found to determine the quality of synthetic speech to a large extent as well as the reproduction of voiced speech.

Some studies have also been conducted to assess factors responsible for loss of intelligibility in certain segments of speech. We conjecture that intelligibility of speech depends on how different segments are reproduced.

To improve our analysis-synthesis system we became convinced we needed to incorporate a signal dependent analysis-synthesis aspect into our system. Listeners perform analyses dynamically, changing their tactics as the background and speech segments fluctuate. Our system needed to be able to change the frame size and rate, number of linear prediction coefficients (LPCs), preemphasis factor, glottal excitation pulse shape, and other features. We derive features from the speech signal through a preliminary analysis. The knowledge gained is represented in the form of segment class information. In the feature extraction analysis we use the segment class information to determine the effective frame size and number of parameters needed to represent the frame. Thus we achieve a realistic representation of temporal and spectral features of different types of segments in speech. This signal dependent analysis-synthesis system produces high quality speech compared to that obtained using an analysis based on fixed frame size and a fixed number of parameters. We apply this system to converting a male voice (source) to a female voice (target) and vice versa.

SIGNAL-DEPENDENT ANALYSIS

Conventional analysis-synthesis systems use a fixed frame size, frame rate and number of parameters per frame. These parameters are typically 20 msec for the frame size, 100 frames/sec and 12 linear prediction coefficients (LPC) in a linear prediction analysis-synthesis system. These values are fixed as a compromise among the conflicting requirements such as the temporal resolution (frame rate), spectral resolution (frame size and number of LPCs), bit rate (number of LPCs), quality and intelligibility.

The disadvantage 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 silence and unvoiced segments, only the gross spectral characteristics need be represented. In fact, higher resolution through a high order LPC may produce spurious peaks, giving the perceptual impression of wrong formant locations. In a transition region from a voiced region to other regions 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 realistic representation requires a variable frame size and variable number of parameters per frame depending on the nature of the segment.

From our studies the requirements of the signal-dependent analysis-synthesis system can be summarized as follows:

1) Determine the segment class of each frame: 5 classes, silence (S), unvoiced (U), voiced (V), transistion from unvoiced to silence or vice versa (TR1), and transition from voiced to unvoiced or silence or vice versa (TR2).
2) Determine the pitch period for voiced segments.
3) Compute the LPCs for each frame.
4) Compute the gain for each frame.
5) Determine the excitition class of each frame: 4 classes, silence (S), unvoiced (U), voiced (V), and mixed (M).
6) Synthesize speech using the LPCs, pitch, gain, and excitation class information for each frame.

CH2118-8/85/0000-0748 $1.00 © 1985 IEEE

To illustrate our flexible analysis-synthesis system [5] with signal dependent analysis-synthesis we have chosen the utterance of the following sentence by a male speaker:

"Should we chase those cowboys?"

The frame rate is fixed at 200 frames/sec. The frame size is also fixed at 200 samples/sec. But the effective frame size is varied depending on the segment class. To realize an effective window size, we multiply the data with a Gaussian window. The standard deviation of the Gaussian window is varied depending on the desired effective size of the frame.

The segment class information and pitch contour are extracted manually from the speech waveform for this case. The gain contour and the excitation class information are obtained through an analysis program.

The synthetic speech generated using the signal-dependent analysis is compared with the standard method using fixed frame size and fixed number of LPCs. The recordings for comparison are listed in Recording 1.

To determine pitch and segment class information automatically we developed a new algorithm for estimating transition points precisely. The algorithm uses the Average Magnitude Difference Function (AMDF) for the speech wave for both pitch extraction as well as segment class identification [6]. To automatically identify the transition points from U to V or vice versa we developed a modified adaptive nonlinear filtering algorithm [7] which is based on the algorithm reported recently in [8]. The pitch and segment class information derived by our algorithm compares favorably with the manually derived data. But we have not yet tested the performance of our algorithm on a large data set.

VOICE CONVERSION

Preliminary studies with our system have led us to conjecture that intelligibility of speech depends on our ability to reproduce dynamic sounds like stops, whereas the quality of voice is mainly determined by the true reproduction of voiced segments. We also found that as long as the spectral envelope is adequately represented, the parameters used to represent the vocal tract system do not seem to make a significant difference in quality. Our studies on different excitation models indicate that the shape of the excitation pulse is critical and it should be close to the original if naturalness is to be obtained in the synthetic speech.

In this paper these studies are continued to determine parameters to be controlled for converting one voice to another, especially a male voice to a female voice and vice versa. At a gross level, we know that by changing the pitch and by compensating the spectrum for the length of the vocal tract, we can accomplish a crude conversion of a male voice to a female voice or vice versa. The signal features we use are derived both from the speech signal and the electroglottograph (EGG) signal [9]. The EGG may give (or may direct how to derive) from the speech signal significant characteristics of voiced or mixed voiced excitation [10]. Once we convert one voice to another satisfactorily, we should know what characterizes a given voice.

A. Data and Parameters for This Study

We report on the speech and EGG data for two speakers (one male (JL) and one female (DH)) for utterances of the following two sentences:

S1: "We were away a year ago."
S2: "Should we chase those cowboys?"

The data were collected at a 10 kHz sampling rate.

Pitch contours were derived semiautomatically using the EGG and speech waveforms [9]. The pitch period was computed for every 5 msec and the contour was smoothed using a 5 point median smoothing.

Linear prediction coefficients (LPCs) and the gain were computed using a 12th order predictor. The data samples were obtained at 200 samples per frame and a frame rate of 200 frames/sec. The gain contours were obtained from the LP residual and smoothed using a 3 point median smoothing.

The synthesis was performed in two stages [5]. In the first stage an excitation signal is derived using the pitch and gain contours and Fant's model [11] for the excitation pulse for voiced segments. The parameters of Fant's model (shown in Figure 1) were chosen as T1 = 60% and T2 = 10%. K is a slope parameter.

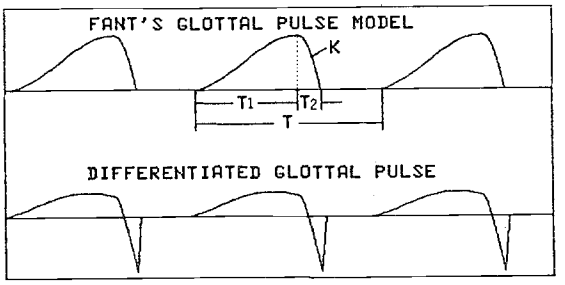

Fig. 1 Glottal Pulse Model and Differentiated Pulse Used for Synthesizer Excitation.

The excitation model is used to drive an all-pole system represented by 12 LPCs to generate synthetic speech. The original and synthetic speech of our test utterances are listed as Recording 2.

B. Conversion from One Voice to Another

This problem is approached in a phased manner by incorporating more and more features until the conversion is satisfactory. We proceed with the assumption that quality of a voice may not depend on dynamic (transient) sounds. So we do not concentrate on the specific parameters representing the transient sounds, except for the prosodic features such as duration and rate of pitch change, when converting one voice to another. All the parameters for refinement of quality of conversion are derived from the steady segments of speech.

C. Average Pitch and Vocal Tract Length Compensation

At a gross level, we identify the average pitch and vocal tract length as the two parameters to be estimated and compensated for voice conversion. We determine these averaged parameters from steady voiced segments of speech. The steps involved are:

1) Identify the steady voiced segments from the EGG and speech waveforms, pitch contour and gain contour.
2) Determine the average pitch over each of these segments and compute the overall average pitch for male and female voices separately. Determine the pitch conversion factor for converting the pitch of one voice to another.
3) Determine the ratio of the first three formants of the corresponding segments for the two speakers in each of the steady regions. Compute the average spectral compression/expansion factor for the vocal tract length compensation.
4) Synthesize speech using the average pitch and spectral conversion factors in the flexible analysis/synthesis system. Use Fant's model for voiced segments.

The average pitch conversion factor for converting the male to the female voice in our experiments was found to be 1.418 and the spectral expansion factor was found to be 1.184, while those for converting the female to the male voice were the reciprocal of these numbers.

The original and synthetic speech of our conversion experiment are listed in Recording 3.

D. Average Glottal Pulse Shape Compensation

At the next level we determine the average characteristics of the glottal pulse shape for male and female voices and use the parameters representing these characteristics in both synthesis and voice conversions. The steps are as follows:
1) Determine the values of T1 and T2 for Fant's model for the center portion of each steady segment. Determine the average value of T1 and T2 for each speaker.
2) Use these values of T1 and T2 in synthesis and for conversion. Use the average pitch and spectral conversion factors derived in Section C.

The list of recordings are given in Recording 4.

E. Detailed Analysis

Further refinement of the conversion process involves spectral and glottal pulse compensation for each steady segment of voiced speech. In order to do this the LP spectra for the corresponding steady segments in the male and female voices are studied to determine the spectral compression factor suitable for that segment. Similarly the glottal pulse characteristics represented by T1 and T2 are studied for each segment. The average pitch conversion factor derived in Section C is still used here but the pitch contours are hand edited to smooth abrupt changes. This needs to be done automatically. The steps are as follows:
1) Identify the steady voiced segments. Plot the LP spectra for male and female voices for the corresponding segments.
2) Derive the spectral compression/expansion factor for each segment separately.
3) Derive the glottal pulse parameters T1 and T2 separately for each segment from EEG waveforms. (Average values used in Recordings 5 and 6. Dynamic values used in Recording 7.)
4) Make up a table of segments with entries for spectral compression factors and glottal pulse parameters.
5) Synthesize speech by using the table entry to determine the conversion factors for each LP spectrum.

The spectral expansion/compression factor must not be applied uniformly across the spectrum, i.e., to all formants equally, otherwise one hears severe high frequency distortion in the converted synthetic speech, e.g. a male voice converted to a female voice will sound "metallic." To avoid this problem the higher frequency formants are shifted less than the lower frequency formants using a special algorithm. See Figure 2.

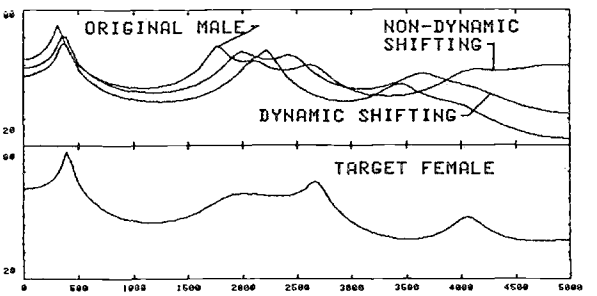

Fig. 2. Sample Spectrum Expansion (Male to Female Conversion) for One Analysis Segment.

Recording 5 presents the results for this experiment using the average values for T1 and T2.

F. Dynamic Energy Compensation

When we refined our voice conversion process to include spectral compensation for each steady segment applied non-uniformly across the band, we discovered a new problem, which we call "volume distortion." This is caused by abrupt changes occurring in the sound "volume" between each segment of the synthesized speech. As the LPCs are modified to compensate for changes in the vocal tract length, a side effect is that we change the energy level for the frame, giving rise to "volume distortion." To restore the quality of the synthesized speech one must modify the gain factor for the frame accordingly, maintaining the overall energy at the desired level within the frame. This process involves dynamically computing a modified gain contour and modified LPCs using both the source and target voice data.

Recording 6 is an example for this experiment. Finally, Recording 7 includes all of the above as well as dynamic segmental variations in T1 and T2.

SUMMARY

The tape recordings demonstrate the progress we have made toward achieving a flexible analysis-synthesis system for voice conversion.

A particularly troublesome parameter is the spectral compression/expansion factor, which must be determined for each segment analyzed. While this parameter need not be applied on a segmental basis in the synthesis the average value of this parameter is important and should be applied non-uniformly across the spectral band. This factor seems to remain the same for speakers across sentences, possibly verifying what we would guess, namely, that each individual's vocal tract length does not change greatly from sentence to sentence.

The glottal pulse parameters are also important on a segmental analysis basis. The average

pulse shape for each speaker is important and appears to remain nearly the same for each speaker across sentences. This seems to indicate little variation in glottal vibratory patterns from sentence to sentence.

The pitch contour does vary with the sentence spoken by a particular individual. Pitch and intonation do change from sentence to sentence.

These conclusions must be verified in future experiments with additional subjects speaking a wider variety of sentences.

ACKNOWLEDGEMENT

This work was supported in part by grants NIH NS17078, NSF ECS 8116341, University of Florida Center of Excellence, and an equipment grant from Digital Equipment Corporation.

TABLE OF RECORDINGS

Recording 1 - Signal dependent analysis-synthesis.

A. Original.
B. Synthetic speech using signal dependent analysis-synthesis.
C. Synthetic speech using a fixed frame size of 200 samples (frames overlapping). This method provides good spectral resolution but poor temporal resolution in transition regions.
D. Synthetic speech using a fixed frame size of 50 samples (frames overlapping). This method provides good temporal resolution but the spectra of voiced experiments is represented poorly.

The quality and intelligibility of the synthetic speech of recording B is significantly better than C and D.

Recording 2 - Typical synthesis. No conversion.
Sentence 1: "We were away a year ago."
Sentence 2: "Should we chase those cowboys?"
Male: original and synthetic
Female: original and synthetic
T1 = 60%, T2 = 10%, LPC = 12, 200 samples/frame, 200 frames per sec, use EEG & speech pitch contour every 5 ms, 5 pt. median smoother

Recording 3 - Voice conversion: average pitch and spectral compression/expansion applied to all segments. T1 and T2 fixed at 60% and 10% respectively for all segments. LPC = 12.
Male: original, synthetic, and female converted to male
Female: original, synthetic, and male converted to female

Recording 4 - Voice conversion: average pitch, average spectral compression/expansion, and average T1 and T2 applied to all segments, LPC = 12.
Male: original and synthetic: T1 = 45%, T2 = 11%, female converted to male: T1 = 45%, T2 = 11% (Sentence 1)
Female: original and synthetic: T1 = 54%, T2 = 13%, male converted to female: T1 = 54%, T2 = 13% (Sentence 1)

Recording 5 - Voice conversion: spectral compression/expansion changed for each segment, average pitch, LPC = 12.
Male: original and synthetic: T1 = 45%, T2 = 11%, female converted to male: T1 = 45%, T2 = 11% (Sentence 1)
Female: original and synthetic: T1 = 54%, T2 = 13%, male converted to female: T1 = 54%, T2 = 13% (Sentence 1)

Recording 6 - Voice conversion: same as Recording 5 but including dynamic energy compensation.

Recording 7 - Voice conversion: same as Recording 6 but including dynamic segmental variations in T1 and T2.

REFERENCES

1. J.J. Yea, The influence of glottal excitation functions on the quality of synthetic speech, Ph.D. dissertation, University of Florida, 1983.

2. J.M. Naik, Synthesis and evaluation of natural sounding speech using the linear predictive analysis-synthesis scheme, Ph.D. dissertation, University of Florida, 1984.

3. 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 of Stockholm Music Acoustics Conference, Stockholm, Sweden, July 28-Aug. 1, 1983.

4. J.J. Yea, A.K. Krishnamurthy, J.M. Naik, and D.G. Childers, Glottal sensing for speech analysis and synthesis, ICASSP-83, Boston, April 14-16, 1983, pp. 1332-1335.

5. B. Yegnanarayana, J.M. Naik, and D.G. Childers, Voice simulation: Factors affecting quality and naturalness, 10th International Conf. on Computational Linguistics, Proceeding Coling84, July 2-6, 1984, Stanford, pp. 530-533.

6. M.J. Ross, H.L. Shaffer, A. Cohen, R. Freuberg and H.J. Manley, Average magnitude difference function pitch extractor, IEEE Trans. ASSP, vol. ASSP-22, pp. 353-362, Oct. 1974.

7. C.A. Pomalaza-Raez and C.D. McGillem, An adaptive, nonlinear edge-preserving filter, IEEE Trans. ASSP, vol. ASSP-32, pp. 571-576, June 1984.

8. B. Yegnanarayana and D.G. Childers, A modification to adaptive nonlinear edge-preserving filter, (submitted).

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

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

11. G. Fant, Glottal source and excitation analysis, STL/QPSR, No. 1, 1979, pp. 85-107.