Speech Communication 8 (1989) 147–158
North-Holland

VOICE CONVERSION

D.G. CHILDERS and Ke WU
Dept. of Electrical Engineering, University of Florida, Gainesville, FL 32611, U.S.A.

D.M. HICKS
Dept. of Speech, University of Florida, Gainesville, FL 32611, U.S.A.

and

B. YEGNANARAYANA
Dept. of Computer Science, Indian Institute of Technology, Madras 600036, India

Received 31 March 1988
Revised 26 January 1989

Abstract. We describe some experiments in voice-to-voice conversion that use acoustic parameters from the speech of two talkers (source and target). Transformations are performed on the parameters of the source to convert them to match as closely as possible those of the target. The speech of both talkers and that of the transformed talker is synthesized and compared to the original speech. The objective of this research is to develop a model for (1) creating new synthetic voices, (2) studying factors responsible for synthetic voice quality, and (3) determining methods for speaker normalization.

Zusammenfassung. Wir beschreiben hier einige Experimente in Stimme zu Stimme Umwandlung basierend auf Sprachproduktionen von zwei Sprechern (Quelle und Ziel). Transformationen der Parameter der Quelle werden ausgeführt um dieselben so gut wie möglich dem Ziel anzupassen. Die transformierte Sprache, sowie die Sprache der beiden Sprecher wird synthetisch hergestellt und mit dem Original (Quelle) verglichen. Der Zweck dieser Forschung ist ein Model für (1) Neue synthetische Stimmen herzustellen, (2) Faktoren zu erforschen, die für die Qualität synthetischer Stimmen verantwortlich sind und (3) Methoden für Sprechernormalisering zu bestimmen.

Résumé. Nous décrivons des expériences dans le domaine de la conversion de la voix qui utilisent des paramètres acoustiques provenant de la voix de deux individus (source et cible). Des transformations sont faites sur les paramètres de la source afin de les faire coïncider autant que possible avec ceux de la cible. Le signal de parole des deux individus et celui obtenu après transformation sont synthétisés et comparés au signal original. Le but de cette recherche est de développer un modèle pour (1) créer de nouvelles voix synthétiques, (2) étudier les facteurs responsables de la qualité des voix synthétiques, et (3) déterminer des méthodes pour la normalisation de la voix de différents individus.

Keywords. Speech, synthesis, quality.

1. Introduction and purpose

For several years one aspect of our research has focused on determining factors responsible for the quality of synthetic speech (Childers and Wu, 1988). One objective is to develop speech synthesis systems that can mimic a desired voice or create a new voice with desired characteristics, as Mel Blanc did when he created such Looney Tune Characters as Bugs Bunny, Tweety Bird, Sylvester the Cat, Daffy Duck, and others (Childers, 1987). However, to date research along these lines has lacked formal models to assist researchers in these tasks. The work reported here may be considered a model for synthesizing speech from a set of acoustic parameters. The speech records of two talkers are analyzed and each record is represented by a set of acoustic

parameters. The acoustic parameters of one talker (called the source) are transformed to closely match those of the other talker (called the target). The objective of the transformation is to determine rules and factors for converting the speech of one talker to sound like that of another. This research is one model for studying factors responsible for the quality of synthetic speech, for mimicking voices (even vocal disorders), for creating new voices, and for speaker normalization.

Similar work along these lines has been attempted before by Atal and Hanauer (1971), who demonstrated the feasibility of modifying voice characteristics using an LPC vocoder. The phase vocoder (Flanagan, 1972) could do so as well. Seneff (1982) developed a system to modify independently the excitation and vocal tract filter. Her method, unlike linear predictive coding (LPC), performs the desired voice conversions or transformations in the frequency domain without explicitly extracting the fundamental frequency of voicing (pitch). Cheng and Guerin (1987) have examined controlling the asymmetry of the glottal excitation waveform for generating high quality synthesized male and female speech. Rosson and Cecala (1986) and Ladd et al. (1985) have been studying methods for creating any desired voice characteristic. Our own work has examined methods for converting the speech of a talker of one gender to sound like that of a talker of the other gender. (Yea, 1983; Childers et al., 1983a, b; Naik, 1984; Yegnanarayana et al., 1984; Childers et al., 1985a, b; Wu, 1985; Childers, 1985; Childers et al. 1987a, b; Childers and Wu, 1988). We have also examined methods for synthesizing vocal disorders (Pinto et al., 1989).

2. Method

The experiments described here address the voice conversion task. Speech tokens for the same sentence were taken from multiple speakers, both male (4) and female (1). Three sentences were used: (1) We were away a year ago; (2) Should we chase those cowboys?; (3) The boy was there when the sun rose. An analysis was performed on each sentence spoken by each speaker and parameters were obtained, including factors related to (1) vocal tract length (related to average formant locations), (2) formant bandwidths and overall spectral shape including slope, (3) energy level contour, (4) pitch contour, and (5) timing parameters related to glottal events. We used both the speech and electroglottograph (EGG) signals (Childers and Larar, 1984; Krishnamurthy and Childers, 1986). The data were analyzed and some simple rules were developed for converting the speech of one talker (source) to sound like (mimic) that of another talker (target). The source and target might be of opposite gender, might differ in age, or might have vocal disorders. Many specific synthesis parameters were varied for these voice conversion experiments. The main tasks are listed in Table 1. The key synthesis parameters and their ranges were established through head-phone listening tests using seven listeners.

3. Speech dependent analysis

Two types of signal analyses were conducted. One was the conventional pitch asynchronous, fixed frame and fixed window linear predictive coding (LPC) analysis. The other analysis method used a variable frame and window. This technique we call signal (or speech) dependent analysis, which we explain below.

To improve our analysis–synthesis system we incorporated a signal dependent analysis–synthesis feature into our system so that we could examine the effect of changing the frame size and rate, number of linear prediction coefficients (LPCs), altering the preemphasis factor, and varying the glottal excitation pulse shape. We derive features from the speech signal through a preliminary analysis. The knowledge gained is represented in the form of speech segment class information. We use five classes of segments: silence (S), unvoiced (U), voiced (V), transitions from unvoiced to silence or vice versa (TR_1), and transitions from voiced to unvoiced or silence and vice versa (TR_2). In the feature extraction analysis we use the segment class information to determine the effective frame size and number of

Table 1
Voice conversion experiments

	Varied	Selected
Preliminary synthesis task: reconstruct original speech of both source and target		
Analysis:		
Pitch asynchronous, fixed frame and window		
	1. LPC parameters	12
	2. Samples/frame	200, 300
	3. Frames/s	200
	4. Pitch contour construction	combined EGG & speech method
Variable frame and window (speech dependent analysis)		
	1. window	Gaussian & Hanning
	2. Samples/frame	Segment dependent
	3. Frame/s	Segment dependent
Excitation:	1. Waveform	Impulse Triple pulse Fant's LF DEGG
	2. Fant's parameters	
	Opening phase T_1	40–60%
	Closing phase T_2	10–13%
	(as % of pitch period)	
Voice conversion synthesis task: Male ⇔ Female		
Analysis:		
Used LPC parameters established in preliminary synthesis tasks for both pitch asynchronous and speech dependent analysis.		
Synthesis:	1. Spectral compression/expansion factors	Averaged & segment dependent
	2. Gain factors	With & without dynamic compensation
	3. Pitch compression/expansion factors	Averaged
	4. Intonation	With & without equalization
Excitation:	1. Waveform	Impulse Triple pulse Fant's LF DEGG
	2. Fant's parameters	
	Opening phase T_1	Averaged & segment dependent
	Closing phase T_2	
	(as % of pitch period)	

parameters needed to represent the frame. Our goal is to obtain 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 have applied this system to convert a male voice (source) to a female voice (target) and vice versa.

Conventional analysis–synthesis systems use a fixed frame size, frame rate and number of

parameters per frame. These parameters are typically 20 ms for the frame size, 100 frames/s and 12 linear prediction coefficients (LPC) in a linear prediction analysis–synthesis system. These values are fixed as a compromise among the conflicting requirements for temporal (frame rate) and spectral (frame size) resolution, 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 while the disadvantage of using a larger frame size is that a poorer temporal resolution is obtained, affecting the transient segments. For silent 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 incorrect 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 for each frame: 5 classes, silent (S), unvoiced (U), voiced (V), transition from unvoiced to silent or vice versa (TR_1), and transition from voiced to unvoiced or silent or vice versa (TR_2).
(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 excitation class for each frame: 4 classes, silent (S), unvoiced (U), voiced (V), and mixed (M) (a combination of voiced and unvoiced excitation).
(6) Synthesize speech using the LPCs, pitch, gain, and excitation class information for each frame.

When we applied this speech dependent analysis system to analyze a sentence we fixed the frame rate and the frame size but the *effective frame size was varied* depending on the speech segment class. To realize an effective window size, we selected either a Gaussian or Hanning window. The standard deviation of the Gaussian window may be varied depending on the desired effective frame size.

In preliminary experiments the segment class information and pitch countour were extracted manually from the speech signal. We now have automated procedures for these tasks (Krishnamurthy and Childers, 1986; Childers et al., 1988). The gain countour and the excitation class information are obtained using automated analysis programs.

4. Excitation

Five excitation model waveforms were examined: impulse, triple pulse, Fant's (Fant, 1979), LF (Fant et al., 1985), and the differentiated EGG (DEGG). Impulse excitation was applied at the instant of glottal closure. Triple pulse excitation refers to the use of three impulses for excitation which occur at the instant of glottal opening, maximum glottal opening, and glottal closure. The magnitude and polarity of these three pulses are controlled. The first is small and positive; the second is the largest and negative; and the third is in between the other two and positive. Fant's model is controlled by four parameters, the duration of the glottal opening phase (T_1), the duration of the glottal closing phase (T_2), the pitch period (T), and a closing phase slope parameter (K). The LF model uses T_1, T_2, T, magnitude, and exponential parameters. The differentiated EGG (DEGG) waveform closely resembles the differentiated Fant waveform. Furthermore, the positive and negative peaks of the DEGG occur approximately at the instants of maximum glottal opening and glottal closure, respectively. The parameters for the excitation waveforms are easily extracted from the speech and EGG signals. The pitch and gain contours are used as well.

5. Analysis and synthesis

An outline of the analysis steps is:

(1) Measure the speech and EGG data for each subject.
(2) Construct the pitch and gain contours.
(3) Identify the steady voiced segments from the EGG and speech waveforms, pitch contour and gain contour.
(4) 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.
(5) 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.
(6) Synthesize speech using the average pitch and spectral conversion factors.

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.

We determined the average characteristics of the glottal pulse shape for male and female voices and used the parameters representing these characteristics in both synthesis and voice conversions. The steps are as follows:

(1) Determine the values of T_1 and T_2 for Fant's model from the center portion of each steady segment. Determine the average value of T_1 and T_2 for each speaker.
(2) Use these values of T_1 and T_2 in synthesis and for conversion. Use the average pitch and spectral conversion factors previously derived.

Further improvement of the conversion process involves two refinements for each steady segment of voiced speech. One refinement requires the determination of the spectral compression/expansion factor for each segment. The other refinement determines the glottal pulse characteristics represented by T_1 and T_2 for each segment. The pitch conversion factor was not determined on a segment by segment basis. Instead an average pitch conversion factor was used, but the pitch contours were hand edited to smooth abrupt changes. This smoothing can 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 T_1 and T_2 separately for each segment from EEG waveforms.
(4) Make a table of segments with entries for spectral compression factors and glottal pulse parameters.
(5) Synthesize speech by using the tables.

For the above procedures we aligned the speech of the source and the target as described below. The speech tokens of each talker were classified into the five segments from which the tables in step 4 immediately above were generated. If the speech of each talker differed with respect to deletions, insertions or allophonic variations, then we made no attempt to compensate for these differences. In our experiments we were searching for rules that would transform various acoustic parameters that affected the characteristics of the "voice" but retained the intelligibility of the words spoken.

6. Rules and algorithms for voice conversion

6.1. Overview

These experiments convinced us that several analysis and synthesis factors were especially important for obtaining high quality synthetic speech. These factors include:

- Proper measurement of the spectrum, including formant locations and bandwidths, overall spectral shape and slope, and energy level. For converting one voice of one gender (source) to sound like that of another voice of the other gender (target), the formants and bandwidths of the source spectrum must be either compressed or expanded to match those of the

target spectrum. An average spectral compression–expansion factor can be derived to achieve the conversion between speakers. But some formants (and bandwidths) may need more compression–expansion than others, i.e., this factor may not be a constant across the spectral band.

- The fundamental frequency of voicing (F_0), or speech countour must be replicated accurately. This factor changes on a sentence by sentence basis.
- The glottal excitation waveshape should be as close to the original as possible if high quality speech is to be obtained. This waveshape appears to remain the same for each speaker across sentences. Waveshape parameters related to glottal timing events are perceptually relevant, e.g., duration of the open glottal interval, duration of the closed glottal interval, duration of the opening of the glottis and duration of the closing of the glottis. The glottal duty cycle (ratio of open interval to total period) is important. The ratio of the duration of the glottal opening interval to the duration of the glottal closure interval appears to remain stable from sentence to sentence for each speaker.

6.2. Voiced/unvoiced/mixed/silent (V/U/M/S) interval identification and classification

Our rules and algorithms for performing this step appear in Pinto et al. (1989) and Childers et al. (1988). We use both the speech and EGG signals.

6.3. Pitch contour transformation

One factor in voice conversion is the fundamental frequency of voicing or the pitch contour. The pitch contours of the two talkers (source and target) will generally have many differences. In our studies we have used two methods for pitch contour transformation. One method used only a simple scale factor to transform the pitch contour of the source to march the pitch contour of the target on the average. The scale factor was that multiplicative constant that converted the average pitch of the source to be the average pitch of the target. The scale factor may be less than (compression) or greater than (expansion) unity. The scaled pitch contour of the source is then used for synthesizing the converted speech. Note that the pitch contour retains its shape in every detail; it is simply rescaled. This method retains the pitch fluctuation characteristics of the source in the synthesized converted voices.

The other method simply uses the target pitch contour to synthesize the converted speech. This method requires aligning the different speaking rates of the two talkers using either a dynamic time warping algorithm or an interactive contour editing program. This method was generally preferred by listeners in the gender voice conversion experiments, leading us to conclude that pitch contour shape is an important factor responsible for the quality of synthesized speech. The implementation of the procedure for calculating the *average* pitch scale factor was described above,

6.4. Vocal tract length transformation

Voice conversion will generally involve a transformation of the formants, usually either an expansion or a compression. This transformation can affect the pitch contour and the speaking rate. For example, if the spectrum is to be compressed by a factor c, then the pitch contour, $p(i)$, is modified to $cp(i)$ and the number of samples per frame changes from N_s to cN_s. It is possible to keep the total number of frames to be synthesized the same, but the total number of samples will change by the factor c. If the speech is synthesized at the cf_s rate, then the pitch and speaking rate will remain the same, but the signal bandwidth will be reduced to $cf_s/2$, where f_s is the original sampling frequency. Note that it is not necessary to change the LPCs for each frame.

An alternate and more flexible approach to spectral expansion/compression is to alter or transform the roots of the LP polynomial in the z-plane. The advantage of this method is that the frequency location and bandwidth of each root can be controlled separately and the speech can be synthesized at the original sampling frequency, usually 10 kHz. A disadvantage of this approach is that numerical computation problems may occur if the LP coefficients become very large or

very small as may occur with a large change in the spectral expansion/compression factor. However, in voice conversion usually only small changes occur. The procedure for calculating the *average* spectral expansion/compression factor has been described above. We discuss below the rules for calculating this factor on a segment-by-segment basis.

6.5. Dynamic spectral expansion/compression

The *average* spectral expansion/compression factor described above does not provide a good quality speech synthesis for voice conversion. One reason for this is that when converting a male voice to a female voice the spectral expansion/compression factor is usually greater than unity since the female format locations are usually higher than the male formant locations. If we shift all formants uniformly, then the male voice converted to a female voice will tend to sound "metallic", since the shifted formants will now be located at a higher frequency and have more energy than the desired formants of the target female speech.

A rescaling of the formant amplitudes does not solve this problem. The higher frequency formants must be shifted less than the lower frequency formants. We developed a rule to dynamically shift the formants on a frame-by-frame basis by modifying the LPC coefficients.

Suppose LPC analysis yields a pole at $re^{j\theta}$. This pole can be moved by changing either r and/or θ. Changing the pole location may also alter the bandwidth of the resonance (or formant). To achieve spectral compression we use a constant or uniform shifting factor, i.e.

$$\theta_m = \alpha\theta_o, \tag{1}$$

where α is the average spectral compression factor computed as described above and θ_o and θ_m are the normalized original formant frequency and modified formant frequency, respectively. On the normalized scale $\theta = \pi$ corresponds to one-half the sampling frequency, $f_s/2$. For spectral expansion we use the following α in (1):

$$\alpha = 1 + (1/\pi)\,(\alpha_o - 1)\,(\pi - \theta_o(i)), \tag{2}$$

where α_o is the average spectral expansion factor computed as described previously and $\theta_o(i)$ is the ith normalized formant frequency. Equation (2) prescribes a different shift for different formants depending on the formant location. The lower the formant frequency, the farther that formant will be shifted from the origin. The higher the formant frequency, the less the formant will be shifted.

6.6. Dynamic energy compensation

Spectral expansion/compression can cause amplitude or "sound volume" distortion in the synthesized speech because the energy of each synthesized speech segment is determined by the gain factor, G, in LPC analysis as well as by the LPC coefficients. If we alter the formant locations by spectral expansion/compression, then we alter the LPC coefficients and in turn alter the energy contour. Kuwabara (1984) noted similar problems in synthesized speech.

A rule to compute the modified gain factor on a frame-by-frame basis from the original gain factor, the original and modified LPC coefficients and the original and synthesized speech signals can be developed for either the frequency domain or the time domain. The rule below is for the time domain:

$$G_m^2 = \frac{\sum_n \left[-\sum_{k=1}^{p} a_o(k)s_o(n-k) + \delta(n) \right]^2}{\sum_n \left[-\sum_{k=1}^{p} a_m(k)s_m(n-k) + \delta(n) \right]^2} G_o^2, \tag{3}$$

where G_o is the original gain factor per frame, G_m is the modified gain factor per frame, $s_o(n)$ is the original speech signal, $s_m(n)$ is the synthesized speech signal, $\{a_o(k)\}$ is the set of LPC coefficients for the original speech, $\{a_m(k)\}$ is the set of modified (spectrally shifted) LPC coefficients, and $\delta(n)$ is the unit pulse function. This rule was developed assuming impulse excitation and may not be applicable for all cases. But the listening tests told us that this rule is a good approximation for improving the quality of the synthesized speech.

Using the dynamic energy compensation technique, the scheme for speech synthesis is:
(1) Perform spectral expansion/compression

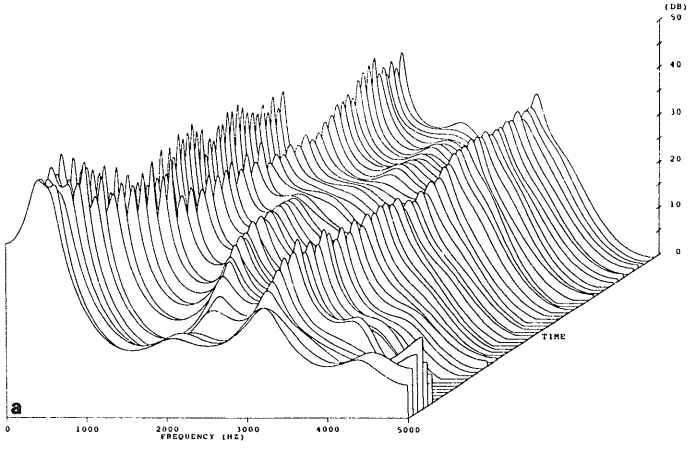
frame-by-frame. Obtain modified LPC coefficients.
(2) Compute modified gain contour from original gain contour frame-by-frame.
(3) Use modified gain contour, pitch contour and desired excitation waveform model (e.g., Fant, LF, etc.) to generate excitation signal.
(4) Use excitation signal and modified LPC coefficients to synthesize speech.

7. Discussion

The initial task in Table 1 was to obtain a sample digitized sentence for both the source and target speakers. We then synthesized each, varying the factors and parameters outlined in Table 1, finally selecting those listed. Next, the speech of the source was converted to that of the target, with the gender being varied if required. Again various factors and parameters were varied as outlined in Table 1. To achieve high quality mimicry, the source spectral factor related to vocal tract length was converted to match that of the target, while not distorting the spectrum used to mimic the target. Said another way, the formants and their bandwidths were either compressed or expanded as described in the previous section.

Replication of the spectral characteristics of a speaker's voice is essential to creating high quality synthetic speech. This was a major conclusion of our preliminary studies and agreed with earlier work by other researchers. However, the spectrum of the excitation influences the speech spectrum. For example, impulse excitation for a speech synthesizer replicates the higher formants reasonably well, but there is less excitation energy in the low frequency region of the synthetic speech. This contributes to impulse synthesized speech being judged as having lesser quality than speech synthesized using the LF or Fant's (Fant et al., 1985; Fant, 1979) excitation waveforms or the differentiated electroglottograph waveform. Another reason for using either of these three waveforms for excitation is that they all provide several glottal timing parameters such as glottal opening and closing phases.

One factor influenced by segment-by-segment analysis is the distinctive time-frequency or spectral transition patterns in the speech signal. Spectral continuity is an important factor influencing the perceived quality of the speech signal. For example, spectral continuity and conser-



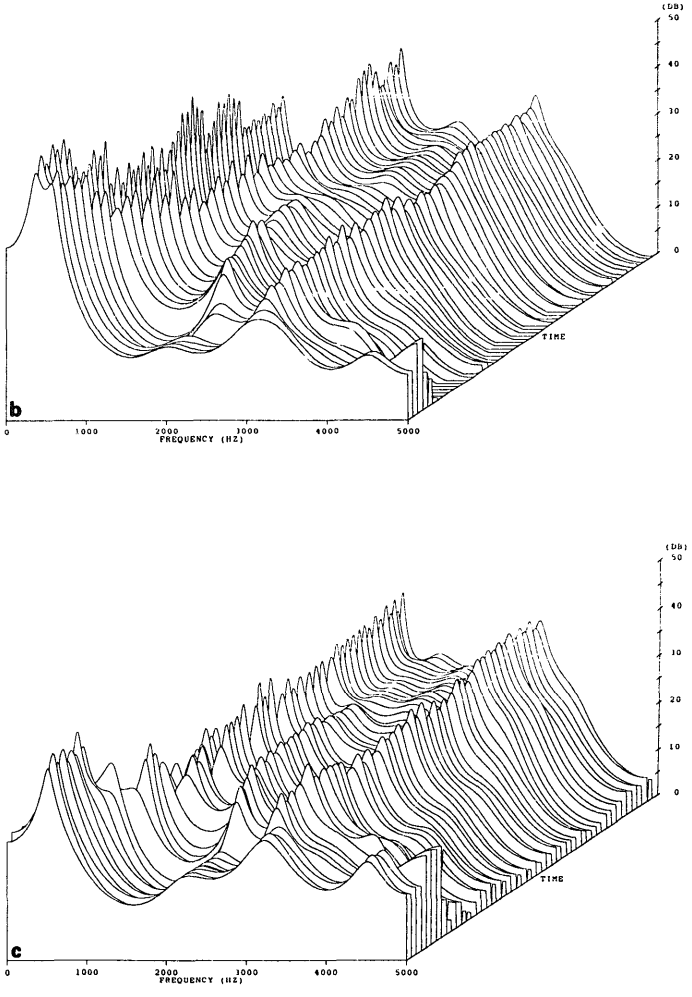


Fig. 1. Three dimensional spectrograms for the speech segment "away" taken from the sentence "We were away a year ago" for a male speaker (DMH): (a) original speech, (b) synthesized speech using Fant's model as excitation, (c) synthesized speech using impulse excitation. The separation between successive plots along the time axis is 20 ms.

vation is illustrated in Fig. 1. In Fig. 1a the spectrogram of the original male (DMH) speech is shown for the speech segment "away" taken from the sentence "We were away a year ago." The spectrum was calculated using a 10 kHz sampling rate with a non-overlapped Hamming window of 200 points. The 12 coefficient LPC envelope is plotted. Fig. 1b is the spectrogram for the synthesized speech using Fant's model as excitation. To facilitate formant tracking in the analysis stage the windowed segments were overlapped 50% (100 points). All other parameters remained as for Fig. 1a. Finally, Fig. 1c shows the synthesized speech using impulse excitation. The analysis phase did not use overlapped windows to illustrate better that formant tracking was inferior to that for Fig. 1b. Note that the first formant in Fig. 1c is greatly reduced over that shown in either Figs. 1a or 1b. This is due to less excitation energy being present in the low frequency region (Wong, 1980). The higher formants are reasonably well reproduced. There is also some degradation in formant tracking in Fig. 1c as compared to Fig. 1b, as might be expected. Because the spectrum is neither conserved nor the proper spectral transactions maintained (spectral continuity) in Fig. 1c, as compared to Fig. 1b, the synthesized speech was not judged to be as high quality.

An example of the formant tracking required to produce high quality synthetic speech is illustrated in Fig. 2 (Pinto et al., 1989). The width of the formant tracks indicates the formant bandwidths. We see that these bandwidths vary dynamically and that they are typically larger the higher the formant.

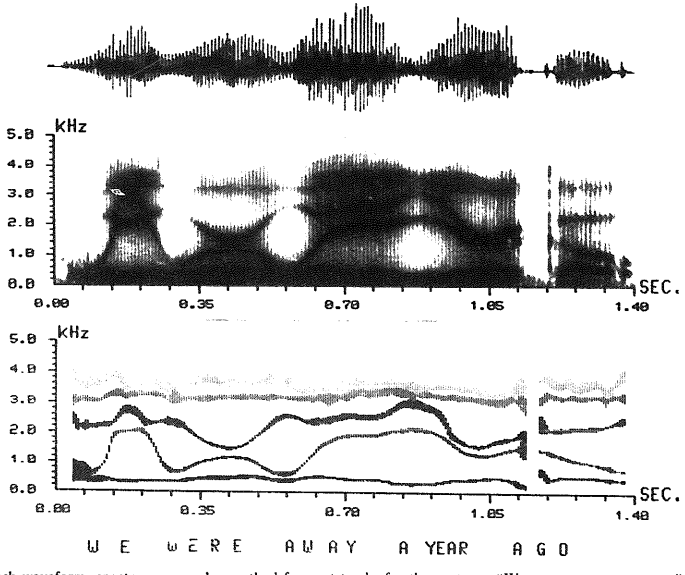


Fig. 2. Speech waveform, spectrogram, and smoothed formant tracks for the sentence "We were away a year ago" for a male speaker (DMH).

8. Conclusions

A particularly troublesome parameter is the spectral compression/expansion factor. While this parameter need not always 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 timing parameters (T_1, T_2, and duty cycle) and the pulse shape are important on an average basis. The excitation waveform parameters and waveshape appear to remain nearly the same for each speaker across sentences, with little variation in glottal vibratory patterns from sentence to sentence. Experiments in Pinto et al. (1989) that converted a normal voice to synthesized voices that were rough, breathy, or creaky (vocal fry) did not dynamically vary the glottal pulse timing parameters or pulse shape on a frame-by-frame basis. Listeners judged the synthesized voices as being of very high quality and as sounding like the voice of the original speaker but with a rough, breathy, or creaky voice. These experiments lead us to speculate that dynamic variation of the glottal pulse shape within a sentence is not necessary, at least not for the experiments we have conducted.

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

Our experiments support the previous findings of others that spectral conservation, continuity and tracking are highly correlated to the perceived quality of synthetic speech (Juang, 1984). Further, 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 electroglottograph. These timing parameters may be used to position impulses and calculate parameters for Fant's model or others. These experiments and others (Childers and Wu, 1989) have established that when actual glottal vibratory parameters are used to synthesize speech the quality of the speech is improved as judged by listeners. This is true for conventional excitations and also for the differentiated electroglottograph waveform. This is due to the fact that for the differentiated electroglottograph waveform (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 instant of glottal closure and opening where the primary and secondary excitations, respectively, normally occur.

The implication of these remarks about excitation waveform timing parameters is that we should incorporate the timing characteristics of the glottal excitation into a speech synthesizer. For efficient coding, the glottal excitation can be represented by a 4 or 5 parameter model. The parameters of the glottal excitation model can be obtained by inverse filtering the speech signal or by using the electroglottogram. If the exact excitation waveform cannot be replicated, then the duty cycle of the glottal excitation interval should be replicated in the synthesizer waveform.

Finally, we point out that this research has direct application to speaker normalization. If we are able to discover the rules by which one speaker can be converted or transformed to sound like that of another, then presumably all speakers could be converted to sound like one speaker, thus achieving speaker normalization. Our model provides a mechanism for learning these rules for a set of standard acoustic parameters.

Acknowledgement

This research was supported in part by NSF grant ECE–8413583, NIH grants NS27022 and NINCDS R01 NS17078, the University of Florida Center of Excellence Program in Information Transfer and Processing and the Mind-Machine Interaction Research Center.

References

B.S. Atal and S.L. Hanauer (1971), "Speech analysis and synthesis by linear prediction of the speech wave", *J.*

Acoust. Soc. Am., Vol. 50(2), pp. 637–655.

Y.M. Cheng and B. Guerin (1987), "Control parameters in male and female glottal sources", in *Laryngeal Function in Phonation and Respiration*, ed. by T. Baer, C. Susaki and K. Harris (College Hill Publ., San Diego), Ch. 17, pp. 219–238.

D.G. Childers, A.K. Krishnamurthy, J.J. Yea and G.P. Moore (1983a), "Laryngeal function: Assessment and role in speech analysis and synthesis", *10th Int. Congress Phonetic Sci.*, Utrecht, The Netherlands, August 1–6, pp. 833–838.

D.G. Childers, J.J. Yea and E.L. Bocchieri (1983b), "Source/vocal-tract interaction in speech and singing synthesis", presented at and appears in *Proceedings of Stockholm Music Acoustics Conference*, Stockholm, Sweden, July 28 – August 1, Vol. 1, pp. 125–141.

D.G. Childers and J. Larar (1984), "Electroglottography for laryngeal function asssessment and speech analysis", *IEEE Trans. on Biomed. Engr.*, Vol. BME-31, pp. 807–817.

D.G. Childers (1985), "Voice (as opposed to speech) synthesis", *Voice I/O Systems Applications Conference*, pp. 349–361.

D.G. Childers, A.K. Krishnamurthy, E.L. Bocchieri and J.M. Naik (1985a), "Vocal source and tract models based on speech signal analysis", in *Mathematics and Computers in Biomedical Applications*, ed. by J. Eisenfeld and C. Delisi (Elsevier Science Publ., Amsterdam), pp. 335–349.

D.G. Childers, B. Yegnanarayana and Ke Wu (1985b), "Voice conversion: Factors responsible for quality", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 19.10.1–19.10.4

D.G. Childers (1987), "Talking computers: Replacing Mel Blanc", *Computers in Mechanical Engineering*, Vol. 6(2), pp. 22–31.

D.G. Childers, Ke Wu and D.M. Hicks (1987a), "Factors in voice quality: Acoustic features related to gender", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 293–296.

D.G. Childers, Ke Wu and D.M. Hicks (1987b), "Voice conversion: A model for studying voice quality and speaker normalization", *Proc. European Conf. on Speech Technology*, Vol. 2, pp. 488–491.

D.G. Childers, M. Hahn and J.N. Larar (1988), "Silent and voiced/unvoiced/mixed excitation (4-way) classification of speech", submitted to *IEEE Trans. Acoust., Speech, Signal Process.*

D.G. Childers and Ke Wu (1989), "Some factors responsible for quality, intelligibility, and naturalness of synthetic speech", submitted to *Speech Communication*.

G. Fant (1979), "Glottal source and excitation analysis", *STL-QPSR*, pp. 85–107.

G. Fant, J. Liljencrants and Q.-G. Lin (1985), "A four parameter model of glottal flow", *STI-QPSR*, pp. 1–13.

J.L. Flanagan (1972), *Speech Analysis, Synthesis and Perception*, 2nd ed. (Springer-Verlag, New York)

B.H. Juang (1984), "On using the Itakura-Saito measures for speech coder performance evaluation", *AT&T Technical Journal*, Vol. 63(8), pp. 1477–1498.

A.K. Krishnamurthy and D.G. Childers (1986), "Two channel speech analysis", *IEEE Trans. on Acoust., Speech, Signal Process.*, Vol. ASSP-34, pp. 730–743.

H. Kuwabara (1984) "A pitch-synchronous analysis/synthesis system to independently modify formant frequencies and bandwidths for voiced speech", *Speech Communication*, Vol. 3, pp. 211–220.

D.R. Ladd, K.E.A. Silverman, F. Tolkmitt, G. Bergmann and K.R. Scherer (1985), "Evidence for the independent function of intonation contour type, voice quality, and FO range in signaling speaker affect", *J. Acoust. Soc. Am.*, Vol. 78(2), pp. 435–444.

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

N.B. Pinto, D.G. Childers and A. Lalwani (1989), "Formant speech synthesis: Improving production quality", *IEEE Trans. Acoust., Speech, Signal Process.*, to appear.

M.B. Rosson and A.J. Cecala (1986), "Designing a quality voice: An analysis of listener's reactions to synthetic voices", *Proc. Human Factors in Computing Systems*, pp. 192–197.

S. Seneff (1982), "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction", *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. ASSP–30(4), pp. 566–578.

D.Y. Wong (1980), "On understanding the quality problems of LP speech", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 208–211.

Ke Wu (11985), *A flexible speech analysis-synthesis system for voice conversion* (Master's thesis, University of Florida).

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

B. Yegnanarayana, J.M. Naik and D.G. Childers (1984), "Voice simulations: Factors affecting the quality and naturalness", *10th International Conf. on Computational Linguistics*, Proceedings Coling 84, Stanford, pp. 530–533.