VOICE CONVERSION BASED ON PARAMETER TRANSFORMATION

JM Gutierrez-Arriola[♦], *YS Hsiao*[♦♦], *JM Montero*[♦], *JM Pardo*[♦], *DG Childers*[♦♦]

[♦]Grupo de Tecnología del Habla, Departamento de Ingeniería Electrónica, ETSI de Telecomunicaciones, Universidad Politécnica de Madrid. Ciudad Universitaria 28040 Madrid. Mail: juana@die.upm.es

[♦♦]Mind Machine Interaction Center. Electronic and Computer Engineer Department. University of Florida. Gainesville.

ABSTRACT

This paper describes a voice conversion system based on parameter transformation [1]. Voice conversion is the process of making one person's voice "source" sound like another person's voice "target"[2].

We will present a voice conversion scheme consisting of three stages. First an analysis is performed on the natural speech to obtain the acoustical parameters. These parameters will be voiced and unvoiced regions, the glottal source model, pitch, energy, formants and bandwidths. Once these parameters have been obtained for two different speakers they are transformed using linear functions. Finally the transformed parameters are synthesized by means of a formant synthesizer.

Experiments will show that this scheme is effective in transforming the speaker individuality. It will also be shown that the transformation can not be unique from one speaker to another but it has to be divided in several functions each to transform a certain part of the speech signal. Segmentation based on spectral stability will divide the sentence into parts, for each segment a transformation function will be applied.

1. INTRODUCTION

Since every acoustic parameter has something to do with the voice quality of a given speaker, many parametric attempts to control voice quality have been tried[3]. In this paper we will present a system that attempts to perform voice conversion transforming five sets of parameters: voice type (voiced or unvoiced), pitch contour, gain contour, glottal source model and vocal tract model (formants and bandwidths). For each parameter a linear transformation function will be defined except for the voice type that is automatically transformed by Dynamic Time Warping (DTW).

The three basic stages for voice personality transformation are the analyzer, the parameter transformer, and the synthesizer. The following sections describe each of these components in detail.

2. ANALYSIS

The aim of this system is to obtain the parameters to drive a formant based synthesizer. The parameters needed are voiced classification, gain, pitch, glottal source parameters, formant frequencies and bandwidths. The first three parameters control the excitation function that drives the filters that corresponds to the formants.

The first step is the pre-processing of the speech signal. It is normalized via dividing by the maximum amplitude and segmented into 25ms frames with a 5ms overlap. The speech signal is then filtered by a zero-phase filter, H(z), to remove the low frequency drift. This filter is given by:

$$H(z)=\frac{1-z^{-1}}{1-0.99z^{-1}}$$

After this, we perform a fixed frame LP analysis (orthogonal covariance method[5]) to get the LP coefficients and the "prediction error signal". A linear predictor of 13th order is chosen for our speech data (sampled at 10KHz). Then we classify each frame into voiced or unvoiced, and for each voiced region we extract the glottal closure instants, the glottal source model and the formants.

2.1. Resonant Tract Estimation

For the voiced configuration, the first five formants have to be estimated from the LP polynomial. One method for estimating formants is to factor the LP polynomial and to assign the appropriate roots to simulate the resonances of the vocal tract. For our analysis, a thirteenth order LP polynomial provides thirteen roots. A formant estimation procedure is applied to find which of these roots belong to the vocal tract[7]. For each root z_i with angle ϕ_i and radius r_i in the z-domain, its transfer function is given by:

$$H(z)=\frac{1}{1-r_i e^{j\phi_i} z^{-1}}$$

If the sampling frequency is Fs, the corresponding frequency and bandwidth are defined as follows:

$$Formant\_Frequency=\frac{\phi i}{2\pi}*Fs$$

$$Formant\_Bandwidth=\cos^{-1}\left(\frac{4r_i-1-r_i^2}{2r_i}\right)\Big/\pi*Fs$$

2.2. Excitation Source Model

There are two types of excitation sources in our system. One is voiced, which involves quasi-periodic vibrations of the vocal folds. The other is unvoiced, which involves the generation of turbulence noise by rapid flow of air past a narrow constriction.

For unvoiced sounds we used a stochastic codebook [4].

For the voiced parts we used a polynomial model described by:

$$P(t)=c_0+c_1\tau+c_2\tau^2+c_3\tau^3+c_4\tau^4+c_5\tau^5+c_6\tau^6 \qquad \tau=t/T$$

Where T is the pitch period. The coefficients are obtained by a polynomial fitting algorithm to fit the integral of the "prediction error signal".[4]

2.3. Control Parameters

There are three types of parameters that control the excitation functions:

- **Voiced/unvoiced classification**. Determines which synthesis scheme (voiced or unvoiced) is adopted for each synthesis frame.
- **Gain parameter**. This parameter is needed to control the intensity of the synthesized speech.
- **Pitch parameters**. The pitch period parameter is the parameter which determines the length of the glottal excitation waveform.

3.SPEECH SYNTHESIS

For the purpose of the system we use a formant-based linear prediction (FBLP) synthesizer, which is an hybrid system that uses the formant synthesis scheme to produce voiced sounds and the LP synthesis scheme to generate unvoiced sounds [1]. The vocal tract is characterized by five formants for voiced sounds and thirteenth order linear prediction coefficients for unvoiced sounds. Depending upon the classification of voiced/unvoiced sound, one of two categories of speech synthesis is used.

4.PARAMETER TRANSFORMATION

Two speakers, the source speaker and the target speaker, pronounce the same sentences and the acoustic parameters are extracted from these two signals via the analysis process. Each set of speech parameters forms a frame-based vector. The source vectors are then time-aligned with the corresponding target vectors by the dynamic time warping (DTW) algorithm. Then each parameter is converted independently by means of a linear transformation of the form X=AY+B, where X are the target parameters and Y are the source parameters. A and B are calculated by means of a linear regression algorithm. The

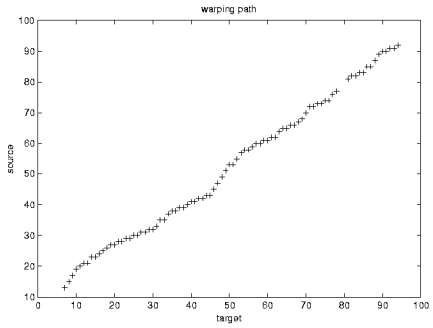

Figure 1: Result of the Dynamic Time Warping algorithm

five conversion parameters are: voice type, gain contour, pitch contour, glottal source and vocal tract (formants and bandwidths).

The parameters are transformed only for voiced regions, for unvoiced regions the target parameters are kept.

Two types of conversion have been tried. The first attempts to convert the whole utterance (sound, word or sentence) from one speaker to the other. This approach works well for short sounds or a short word. However, due to speech variability, sentences are more troublesome. For that reason we split a sentence into segments, performing voice conversion in each segment. The results for this approach work well, giving good

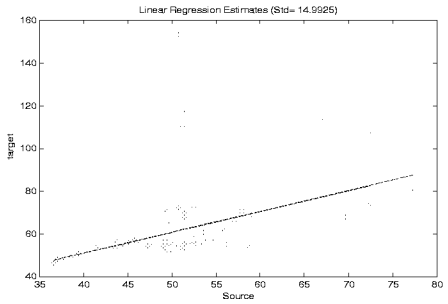

Figure 2: Result of the linear regression algorithm

quality, in that the characteristics of the target speaker are reasonably well matched.

4.1. Segmentation

Segmentation is performed over the target speech using a simple normalized measure of spectral change given by:

$$\frac{\sum_{\omega}(|S_1(\omega)|-|S_2(\omega)|)^2}{\sum_{\omega}|S_1(\omega)|^2} \geq threshold$$

S1 and S2 are two consecutive frames. If this value exceeds the threshold, then a new segment is specified.

In our system the segmentation can be chosen to be automatic or manual. For the automatic option we have three possibilities: many segments which corresponds to a threshold of 0.2, medium, which corresponds to a 0.5 threshold, and few, which corresponds to a threshold of 1. No segment of one frame is allowed and it is assigned to the closest segment.

The following figures show examples of segmentation for the sentence "we were away a year ago".

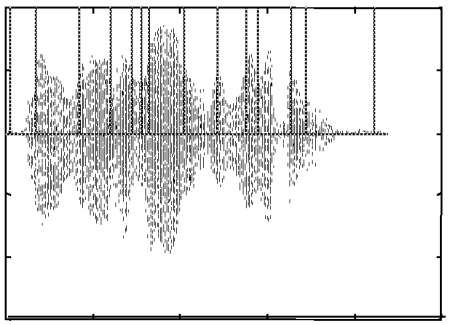

Figure 3: Segmentation with many segments

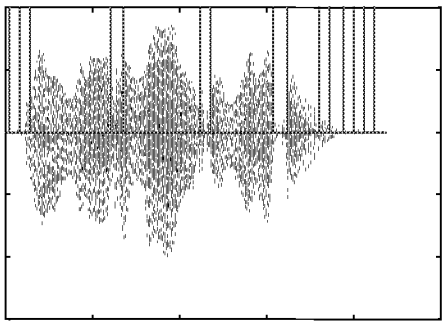

Figure 4: Segmentation with medium number of segments

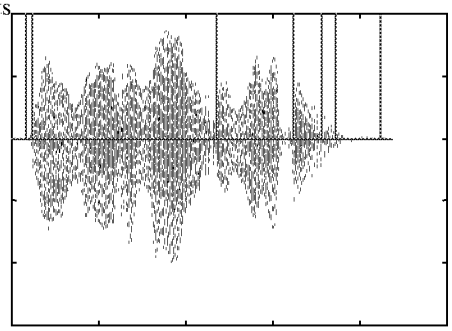

Figure 5: Segmentation with few segments.

4.2. Pitch contour conversion

In order to have a systematic point of view, the Glottal Closure Instant (GCI) sequence is transformed into the pitch contour, which is defined as the GCI vs. its pitch period.

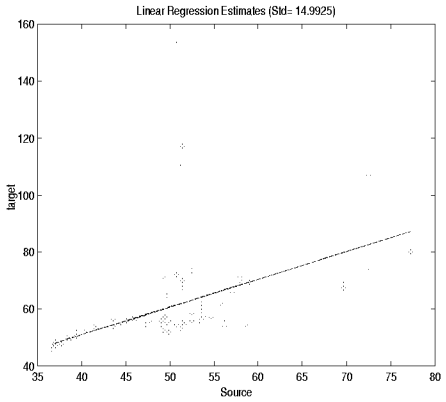

Figure 6: Correspondence between the pitch contour of two female speakers

Through the DTW result we have a correspondence between the source and the target pitch contour. We can build a correspondence point to point as shown in the figure 6 and applying a linear regression algorithm we obtain the coefficients of the equation:

$$Y = aX + b$$

Where Y will be the approximate target pitch contour, X will be the source pitch contour and a,b are the coefficients obtained by linear regression.

Once the transformation is performed the GCI sequence is rebuilt. If we don't use segmentation only a function is applied for the whole utterance. If pitch contour conversion is performed on a segment by segment basis, we need special processing in the limits of the segments. To put all the contour

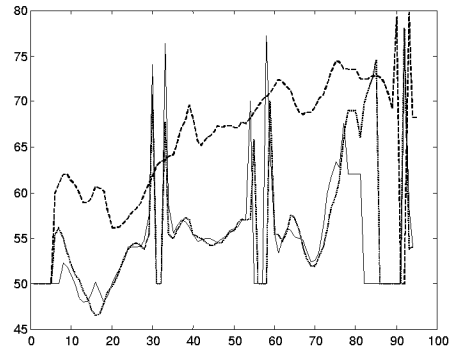

Figure 7: ..... Target pitch contour; -------- Converted without segmentation; _____ Converted with segmentation

together the only restriction is that the first GCI of the actual segment must be bigger than the last GCI of the previous one. And then the first GCI of the actual segment is fixed to the middle point between the last CGI of the previous one and the second of the current. This is to avoid discontinuities in the pitch contour and too short pitch periods. Figure 7 shows the target pitch contour converted with and without segmentation.

4.3. Gain contour conversion

The same scheme as with pitch contour conversion is applied in this case.

Gain contour conversion differs slightly from that calculated without segmentation, which is fixed at the beginning and end to zero. This can cause numerous discontinuities if segmentation is used. After the segments have been converted, then the complete gain contour is calculated. Next a median filter of length 7 is used to smooth any remaining discontinuities at segment boundaries. Figure 8 show examples of gain contour conversion with and without segmentation.

4.4. Glottal source conversion

We obtain a function for each coefficient of the source model. The Glottal source conversion is the same as that calculated without segmentation, except the calculation is performed on a segment by segment basis. For unvoiced sounds the parameters are copied from the target.

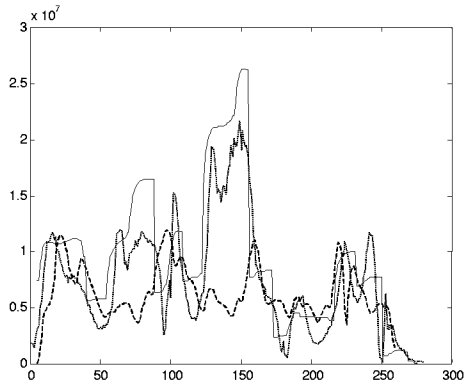

Figure 8: ..... Target gain contour; -------- Converted without segmentation; _____ Converted with segmentation

4.5. Vocal tract conversion

The reference to vocal tract conversion refers to conversion of formants and formant bandwidths.

For each formant we obtain a transformation function. The formants are converted with only a function for the whole utterance if we don't use segmentation, and with one function per segment if we apply segmentation. Bandwidths are not converted but calculated to match the desired spectrum as close as possible.[6]

5. RESULTS

Short experiments have been tried comparing the conversion with and without segmentation. In these experiments we used three sentences uttered by four speakers (two male and two female). All the experiments conclude that the transformation with segmentation was better, the improvement is more notorious when performing male-female transformation.

5.1. With and without segmentation.

In these experiments segmentation is performed manually. The segments are placed at the allophone boundaries. Four examples are illustrated: a)male-male, b)male-female, c)female-male and d)female-female.

The results show that in all cases the converted speech has a better quality and it is closer to the target when using segmentation. Segmentation seems to be more useful when converting male-female than with male-male or female-female conversions.

Another experiment converted only one parameter copying the remaining parameters from the target. This was to examine if the segmentation improved every aspect of the converted speech. Results show that the transformation with segmentation improves all parameters involved in the conversion.

5.2. How many segments?

In this experiment we study the effect of the number of segments on the converted speech. We compared the segmentation performed manually with the three automatic ones. With this experiment we found that:

- Automatic segmentation with many segments and the manual segmentation gave equivalent perceptual quality.
- The more segments we have the better quality. But also you have to calculate more coefficients.

6. CONCLUSIONS

The voice conversion algorithms are established on a speaker adaptation model that treats speaker differences as arising from a parametric transformation. The voice conversion task is then performed as the mapping between two sets of parameters. We found that the linear transformation is effective for converting single-syllable words, but less so for sentences.

We have applied a segmentation algorithm to the sentences and built a set of transformation functions for each segment. The results show that this approach is successful in transforming the speaker voice personality.

7. ACKNOWLEDGMENT

This work was supported in part by NSF grant IIS-9526049, FPU scholarship and project TIC 95-0147 Demóstenes.

8. REFERENCES

1. Hsiao, Y.S., *"Speech Synthesis Algorithms for Voice Conversion"*. PhD Dissertation, University of Florida, 1996.
2. Savic, M., Nam, IH. "Voice Personality Transformation". *Digital Signal Processing 1*: pp 107-110. 1991
3. Kuwabara, H., Sagisaka, Y. "Acoustic characteristics of speaker individuality: Control and conversion". *Speech Communication 16*: pp 165-173. 1995.
4. Childers, D. G., Hu, H.T. "Speech synthesis by glottal excited linear prediction". *J. Acoust. Soc. Am., 96(4):* pp 2026-2036. 1994.
5. Ning, T., and Whiting, S. "Power spectrum estimation via ortogonal transformation" *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.:* 2523-2526. 1990.
6. Hsiao, Y.S. and Childers, D.g: "A New Approach to Formant Estimation and Modification Based on Pole Interaction". *30th Asilomar Conf. On Signals, Systems and Computers*,Pacific grove,CA., November, 1996.
7. Hsiao, Y.S. and Childers, D.G. "A modified Root-Finding Formant Estimation Algorithm Based on LP Analysis" *Proceedings of the IASTED International Conf. On Signal and Image Processing (SIP'96):* pp30-33. 1996.