SPEAKER MODIFICATION WITH LPC POLE ANALYSIS

Janet Slifka

Systems Research Laboratories, Inc.
Dayton, OH 45440
e-mail: jls@pinna.aamrl.wpafb.af.mil

Timothy R. Anderson

Armstrong Laboratory, AL/CFBA
Wright-Patterson AFB, OH 45433-7901
e-mail: tra@neural.aamrl.wpafb.af.mil

ABSTRACT

Speaker modification is the ability to change the perceived speaker identity of a recorded utterance. Basic to this is the capability to alter the vowel segments of speech. Not only do these segments comprise the majority of the voiced portion of speech but they are dominated by clearly defined acoustic parameters - formant frequencies and pitch. A method of altering the formant frequencies of vowel segments using LPC analysis/synthesis was investigated. Pole location modification based on statistical references provided individual control over formant frequencies and bandwidths but, in some transformations, lead to artifacts in the reconstructed speech.

1. INTRODUCTION

For the application of speaker modification, i.e. altering the apparent speaker identity of an utterance from the original speaker's identity to that of a target speaker, many areas must be considered. This work attempted to move as close as possible to the target speaker using modification of the vowel segments of speech. Altering the speaker identity for vowel segments requires the ability to move the formant frequencies and modify the pitch of the speech. These are acoustic qualities of the speech as opposed to speaking-style qualities. Applications for speaker modification can be found in speaker normalization for speech recognition, psychoacoustic experiments on speech perception, or as a layer within a text-to-speech system.

This work is based on residual-excited LPC analysis/synthesis [1]. With limitations, it can be said that the residual signal models the sound source and the LP polynomial models the vocal tract. The questions to be answered involved quantifying which controllable parameters describe the source and vocal tract models and how to effectively map these parameters between speakers.

Some of the complex-conjugate poles of the filter polynomial correspond to the formant resonances in voiced speech. The locations of these resonance-related poles can be moved to effect changes in the vocal tract model. Parker and Hall [5] have used a method of moving the pole locations based on a multiplicative factor for the angle and an exponential factor for the radius. Childers and Wu [2] have also applied a constant multiplicative factor based on gender transformation to the angle of the poles and Kuwabara and Tagaki [4] moved the poles by fixed percentages. This work attempts to extend the previous research by tailoring the mapping parameters to specific speakers and modifying each of the first four formants by a different criteria.

The approach presented here is outlined in the following steps.

1. A corpus of data was collected for each speaker consisting of 42 isolated words. Speech was segmented and labeled by hand to provide phoneme class labels. In all subsequent sections it is considered that each frame of speech is associated with a class label.
2. Reference parameters, in the form of frame-based averages, were collected on a per-class basis to characterize the uniqueness of each speaker in terms of the parameters of the analysis/synthesis system. Statistics for twenty vowel classes were computed. All analysis was performed at a fixed-frame rate of 10ms.
3. LPC analysis was performed on an utterance of the source speaker. This utterance was transformed toward that of the target speaker based on the average parameters. LPC synthesis using the modified residual and modified LPC filters was performed.

The following sections detail the modifications made to the LPC residual and filter.

2. RESIDUAL PITCH MODIFICATION

The primary parameter describing the source is the pitch. The series of utterances collected for each

J. Slifka is currently with the Harvard-MIT Division of Health Science and Technology, Cambridge, MA 02139, e-mail: slifka@mit.edu.

0-7803-2431-5/95 $4.00 © 1995 IEEE

speaker involved in the system were analyzed for pitch estimates [7]. Average pitch values and standard deviations were computed over all voiced frames.

The residual contains voicing and pitch information from the original signal. The residual signal was pitch-modified using a frequency scaling technique [6]. The amount of modification was based on the overall averaged voiced pitch and its standard deviation. In some cases, this introduced bandwidth alteration, e.g., to decrease pitch, warping the spectral representation in the form of a compression caused a loss in signal bandwidth. Yet, this method introduced some coarse movement of any resonant peaks present in the residual spectrum and resulted in an audibly apparent change in the transformed utterance.

Frequency warping of the residual signal introduced a change in the time duration of the residual signal. In order to prevent this change from altering the speaking rate of the final utterance, time-scale modification [3] was used to ensure that the modified residual had the same sample length as the original residual.

3. VOCAL TRACT MODEL MODIFICATION

Individual control of formant location and bandwidth was deemed as an important factor in speaker modification. Moving the formant-related pole locations individually offered the possibility of this type of control. For the pole-location manipulation of the vocal tract model, the data were analyzed at 10kHz using the autocorrelation method to estimate the LP coefficients. Analysis was performed at 10kHz to provide enough roots for analysis while keeping the number of roots manageable. Dynamic formant tracking [8] was performed for the voiced segments of speech and statistics were collected for each class on the angle and radius of the pole locations estimated to correspond to the formant resonances.

Each pole of the LP polynomial can be expressed in the form $re^{j\theta}$. In order to maintain the intraspeaker variation indicated by the standard deviation while still moving the pole toward the mean of the target speaker, a method based in zscore normalization was used to adjust the pole resonances. The frame value was normalized with statistics from the source speaker's utterances using:

$$\theta' = (\theta - \overline{\theta_{source}})/\sigma_{\theta_{source}}$$

$$r' = (r - \overline{r_{source}})/\sigma_{r_{source}}.$$

Where θ is the angle of the pole in the z-plane and r is the radius of the pole in the z-plane. The mean values are denoted by $\overline{\theta_{source}}$ and $\overline{r_{source}}$ and the stanard deviations are denoted by $\sigma_{\theta_{source}}$ and $\sigma_{r_{source}}$. The target speaker's statistics were introduced as follows:

$$\theta_{mod} = \theta' * \sigma_{\theta_{targ}} + \overline{\theta_{targ}}$$

$$r_{mod} = r' * \sigma_{r_{targ}} + \overline{r_{targ}}.$$

Poles were monitored to ensure that none were moved outside of the unit circle. Those poles that were not deemed to be associated with a formant resonance were left unmodified.

Movement of the poles can result in a change in the gain of the filter. A method suggested by Parker and Hall [5] was used to apply a scale factor to the residual to compensate for the change in filter gain. This method uses the ratio of the gain of the filter at 0Hz before and after modification as a scale factor for the residual.

4. DISCUSSION

Pole location modification yielded significant movement of the formant locations. This technique takes into account how each speaker shapes individual vowels by providing the ability to vary the amount of movement of each formant within each vowel class. Signal bandwidth was limited by the 10kHz sampling rate and bandwidth loss was possible from the frequency-scaling pitch modification of the residual. In some cases, the pole movement combined with the spectral warping of the residual signal caused the formants to be moved beyond the intended locations. This resulted in a perceived muffling of the speech signal.

The intent of this method was to move the formant-related pole locations of vowel segments based on average statistics. To determine if this had been accomplished, the corpus of words of a source speaker were transformed to a target speaker and average pole locations were computed. Tables 1 and 2 show the average, over all classes, of the difference in these locations between the source and target averages, the source and transformed averages, and the target and transformed averages. Table 1 illustrates modification from a Male to a Female speaker and Table 2 shows the reverse. Numerically all locations appear to have been moved closer to the target by the transformation. This is most noticeable in the theta parameter, which corresponds to formant frequency. Even the most informal of listening tests showed that the apparent speaker identity was altered, yet not clearly and consistently altered to the identity of the target.

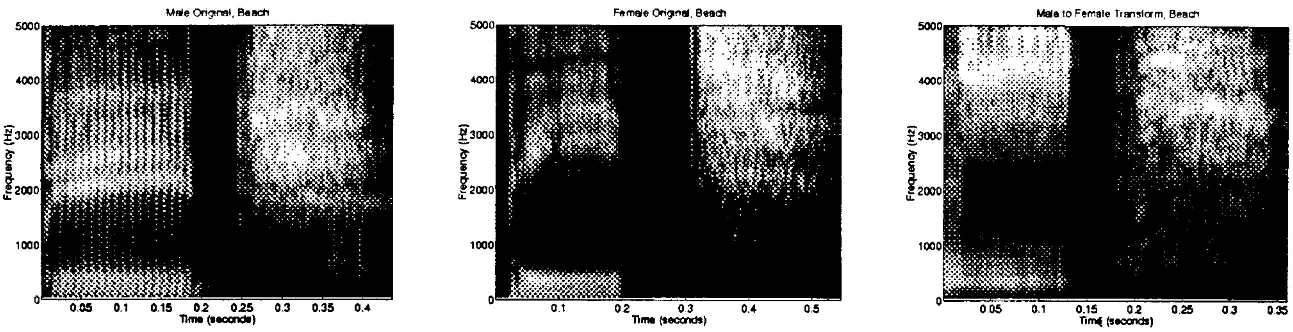

Figure 1: "Beach" spectrum for Male, Female, and Male to Female Transform.

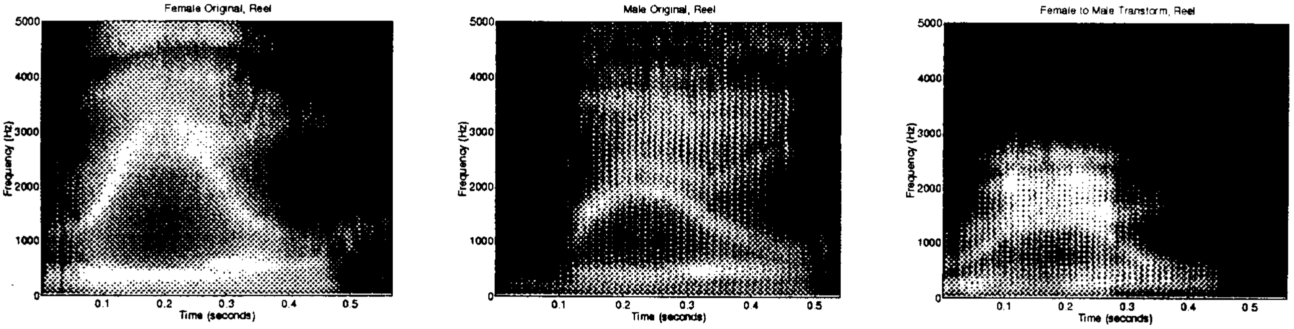

Figure 2: "Reel" spectrum for Female, Male, and Female to Male Transform.

Table 1: **Comparison of average pole locations over all classes for a Male to Female Transform.**

	ΔF1	ΔF2	ΔF3	ΔF4
Source and Target				
Theta	-0.0627	-0.3194	-0.3462	-0.4684
Radius	0.0375	0.0359	0.0501	0.0414
Source and Transformed				
Theta	-0.1089	-0.2779	-0.3854	-0.4268
Radius	0.1575	0.1218	0.0786	-0.0010
Target and Transformed				
Theta	-0.0463	0.0416	-0.0391	0.0416
Radius	0.1200	0.0859	0.0285	-0.0424

Table 2: **Comparison of average pole locations over all classes for a Female to Male Transform.**

	ΔF1	ΔF2	ΔF3	ΔF4
Source and Target				
Theta	0.0627	0.3194	0.3462	0.4684
Radius	-0.0375	-0.0359	-0.0501	-0.0414
Source and Transformed				
Theta	0.1043	0.3591	0.2850	0.4264
Radius	0.0012	0.0016	0.0536	0.0163
Target and Transformed				
Theta	0.0416	0.0396	-0.0613	-0.0421
Radius	0.0387	0.0375	0.1038	0.0578

In order to illustrate some of the results of the methods described, several spectrograms are included in which light areas indicate high energy and dark areas indicate low energy. Figure 1 shows a transformation from a male speaker to a female speaker using the word "beach." While the formants have been moved, the second and third formants are apparently fused. In Figure 2, the transformation is from a female to male with the word "reel." This illustrates the effect of bandwidth loss due to the frequency scaling pitch modification.

5. SUMMARY AND PLANS

A method of altering the formant locations of vowel segments using LPC analysis/synthesis was investigated. Pole location modification based on statistical references provided individual control over formant frequencies and bandwidths. Combined with frequency-scaling pitch-modification of the residual, the reconstructed speech showed artifacts and/or loss of signal bandwidth for some words and source/target combinations.

Current research with continuous speech is focusing on using a neural network to perform the mapping of the formant-related poles as moving the poles toward a mean value is not truly appropriate in the presence of strong co-articulation. While not fully analyzed yet, it is hoped that this method will be better able to handle such regions of speech. Also, experiments are being done to modify non-vowels frames and to develop an optimal mapping between all poles of each class. It is hoped that this will reduce artifacts and provide the means to work at a higher sampling frequency.

6. REFERENCES

[1] B. S. Atal and Suzanne L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50(2):637–55, April 1971.

[2] D.M. Hicks D.G. Childers, Ke Wu and B. Yegnanarayana. Voice conversion. *Speech Communication*, 8(2):147–58, 1989.

[3] Jr. Donald Hejna. Real-time time-scale modification of speech via the synchronized overlap-add algorithm. Master's thesis, Dept of Electrical Engineering and Computer Science, MIT, April 1990.

[4] Hisao Kuwabara and Tohru Takagi. Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method. *Speech Communication*, 10(5-6):491–5, December 1991.

[5] Sydney Parker and Geoffrey Hall. Computer modeling of voice signals for adjustable pitch and formant frequencies. In *13th Asilomar Conference on Circuits, Systems, and Computers*, pages 158–61, Nov 1979.

[6] R.W. Schafer and L.R. Rabiner. A digital signal approach to interpolation. *Proc. IEEE*, 61:692–702, June 1973.

[7] Bruce G. Secrest and George R. Doddington. An integrated pitch tracking algorithm for speech systems. In *ICASSP 83*, pages 1352–55, 1983.

[8] David Talkin. Speech formant trajectory estimation using dynamic programming with modulated transition cost. *JASA Suppl. 1*, 82:S55, 1987.