Reconstruction of Speech from Whispers

Robert W. Morris*, Mark A. Clements

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Abstract

This paper investigates a method for the real-time reconstruction of normal speech from whispers. This system could be used by aphonic individuals as a voice prosthesis. It could also provide improved verbal communication when normal speech is not appropriate. The normal speech is synthesized using the mixed excitation linear prediction model. Differences between whispered and phonated speech are discussed and methods for estimating the parameters of this model from whispered speech for real-time synthesis are proposed. This includes modification of the formants and smoothing of the noisy linear prediction spectra and synthesis of the excitation signal. Trade-offs between computational complexity, delay, and accuracy of different methods are discussed.

Keywords: Whispered speech; voice parameter extraction; voice parameter modification; voice prosthesis

1 Introduction

In normally phonated speech, air from the lungs causes the vocal folds of the larynx to vibrate, exciting the resonances of the vocal tract. In whispered speech, the glottis is opened and turbulent flow created by exhaled air passing through this glottal constriction provides a source of sound. Typically, this sound source is distributed through the lower portion of the vocal tract [1] resulting in speech that is completely noise excited, with 20 dB lower power than its equivalent phonated speech [2]. The spectrum of whispers also rolls off under 500 Hz [3] due to an introduced spectral zero [1] and is typically flatter than the voiced spectrum between 500 and 2000 Hz [4]. Although the whispered speech signal is aperiodic, the sensation of pitch still exists [5]. These changes in perceived pitch are related to the speech intensity level as well as the formant frequencies and bandwidths [6].

Another significant change is caused by the increased coupling between the trachea and the vocal tract created by the open glottis. During aspiration, additional poles and zeros, as well as dampening of the first formant, have been observed [7]. Formant shifts, especially increases in the first formant frequency, occur in whispered speech [3,8] as well as in mechanically excited speech with an open glottis [9].

There are many different parametric models of speech production. The Mixed Excitation Linear Prediction (MELP) model, which has been used successfully for low bit-rate coding, uses a source-filter decomposition of the speech signal with a fully parametric excitation model [10] as shown in Figure 1. The parameters derived from the speech waveform are denoted by ovals. In this paper, a method for determining the parameters of voiced speech from observed whispers is proposed. Section 2 contains methods for compensating for short-time spectral differences in whispering, while the estimation of the excitation parameters is discussed in Section 3. The effectiveness of these methods are tested in Section 4.

2 Spectral Modification

The linear prediction spectrum used by the MELP coder is distorted in three ways. The first is a constant bias due to spectral excitation differences during normally voiced portions of speech. The second is the bias in the formant locations created by coupling with subglottal structures. The final difference is the variance of the linear prediction spectrum estimate caused by the noise-like excitation signal in whispering. In the MELP model, as well as in many other low
Fig. 1. Simplified block diagram of MELP model

Fig. 2. Block diagram of enhancement scheme

bit-rate coders, the linear prediction coefficients are transformed into line spectrum pairs (LSPs). The two LSP polynomials are given by

\[ P(z) = A(z) + z^{-(p+1)}A(z^{-1}), \]
\[ Q(z) = A(z) - z^{-(p+1)}A(z^{-1}), \]

where \( A(z) \) is the linear prediction polynomial of order \( p \). These polynomials are factored as

\[ P(z) = (z^{-1} + 1) \prod_{i=1}^{p/2} \left( z^{-2} - 2z^{-1} \cos \theta_i + 1 \right), \]
\[ Q(z) = (z^{-1} - 1) \prod_{i=1}^{p/2} \left( z^{-2} - 2z^{-1} \cos \phi_i + 1 \right), \]

where \( \theta_i \) and \( \phi_i \) are the interleaved line spectral pair frequencies. These frequencies have a one-to-one correspondence to the linear prediction polynomial, and are often used by speech coders because of their quantization properties. The proposed techniques for spectral enhancement all operate on these frequencies.

2.1 Pre-filter

The modification of the long-term source spectrum differences is accomplished by a static linear filter. The magnitude response was determined from the spectra of isolated voiced and whispered vowels. The experimentally derived average log-spectral difference between voiced and whispered sounds is plotted in

Figure 3. This spectrum was smoothed using a Mel-weighted filterbank. The truncated minimum-phase realization as described in [11] for this spectrum is shown as a solid line. As expected, the spectral region under 500 Hz is boosted higher than the rest of the spectrum. In the actual implementation, this magnitude response is reduced because noise of higher energy than the whispering is often present in the lower frequencies.

2.2 Spectral Smoothing

The MELP coder models the short term spectrum with the LSP frequencies and a power estimate. When the excitation is Gaussian white noise, the covariance matrix of the LSP vector \( \mathbf{f} \), for a single block of speech is

\[ \text{Cov}(\mathbf{f}) = \frac{\sigma^2}{N} \left[ \frac{\partial \mathbf{a}^T}{\partial \mathbf{f}} \mathbf{R} \frac{\partial \mathbf{a}}{\partial \mathbf{f}} \right]^{-1}, \] (1)

where \( N \) is the estimation window length, \( \mathbf{R} \) is the covariance matrix of the waveform, \( \sigma^2 \) is the variance of the residual signal, \( \mathbf{a} \) contains the linear prediction coefficients, and \( \frac{\partial \mathbf{a}}{\partial \mathbf{f}} \) is the Jacobian of \( \mathbf{a} \) with respect to \( \mathbf{f} \). The Jacobian matrix can be calculated analytically from the linear prediction polynomial [17]. The variance of the gain parameter in dB is equal to 37.9/\( N \), which corresponds to a standard deviation of 1.8 dB for the standard MELP block size of 180 samples. This noise leads to a rapidly varying spectrum during steady vowels. This variation, which is much higher for whispers than voiced speech, creates unnatural spectral dynamics in the synthesized speech. Similar variations have been shown to significantly reduce perceptual quality [12].

The goal of this section is to smooth the spectrum during the vowels without destroying the rapidly varying spectral content of the consonants. It has been shown that the estimation error covariance matrix \( \text{Cov}(\mathbf{f}) \) is diagonal [13], so the estimation noise is uncorrelated between the coefficients. This property of LSPs simplifies the enhancement process by allowing for inde-
pendent filtering of the LSP frequencies. Three methods are discussed, each with their advantages and disadvantages. To determine the statistics of the LSPs for normal speech, a long series of normal speech utterances were analyzed to produce their LSPs. The mean frequencies were subtracted, and these values were found to be well modeled as a first order autoregressive system. Under these assumptions, the LSPs are modeled by

\[ x_n = Ax_{n-1} + w_n \]

\[ f_n = x_n + v_n + m \]

where \( x_n \) is the vector containing the true LSP vector for block \( n \), \( f_n \) is the estimated LSP vector, and \( m \) is the constant LSP mean vector. The matrices \( A \) and \( \text{Cov}(w_n) \) are both constant and diagonal and are chosen to fit the normal speech data. The covariance matrix of \( v_n \) is set at each time step as described in Equation 1. These diagonal matrices allow for independent processing of the pairs, which reduces computational complexity.

### 2.2.1 Linear Filtering

The minimum mean squared error (MMSE) estimator for a dynamic linear model driven by Gaussian noise is the Kalman filter. For the variances involved in this problem, this provides very little smoothing. Upon further inspection of the normal speech, the residual plant noise \( w_n \) has heavy tails that are not characteristic of a Gaussian distribution. A histogram of the residual noise of the first LSP is contained in Figure 4. From this plot, it is apparent that the Gaussian model does not fit the data well. These tails result in a fairly high estimation of the system noise, which results in very little smoothing. However, the data is fit very well by a two Gaussian mixture. The Kalman filter tested assumes that the plant noise has a variance equal to the narrow mixture. This choice introduces some smoothing, but results in distortion during rapid transitions.

### 2.2.2 Nonlinear Filtering

Two nonlinear filters for LSPs are proposed in this paper. The most simple is the median filter of order \( K \):

\[ \hat{x}_n^{(m)} = \text{median} \left( f_n^{(m)}, f_{n-1}^{(m)}, ..., f_{n-K+1}^{(m)} \right) \]

where the superscript \( m \) denotes the element of the vector. This technique has the advantages of being extremely simple to implement, and preserves many rapid changes in the signal. The filter’s weakness is that it introduces a \((K - 1)/2\) block delay in the spectrum and is less efficient at eliminating Gaussian observation noise.

The second approach is to model \( w_n \) as a Gaussian mixture. Optimal estimation of this model involves a bank of Kalman filters [14]. This method is optimal when the observations come from the distribution

\[ w_n \sim p_1 N(0, R_1) + p_2 N(0, R_2), \]

where \( p_i \) is the probability of mixture \( i \), \( R_1 \) and \( R_2 \) are the diagonal covariance matrices of the two mixtures, and \( N(\mu, R) \) is the Gaussian distribution function. The exact estimator using this method requires \( 2^N \) filters on the \( n \)th iteration, with each filter representing the best estimate given the mixture choices. However, the algorithm can be efficiently implemented by merging Kalman filters with similar states after each iteration [14].

### 2.3 Formant Shifting

Several studies have shown that the formant locations and bandwidths of whispered speech differ from normally phonated speech [3,15,16]. The results from Kallail and Emanuel’s study on isolated whispered vowels from male subjects are summarized in Figure 5. This plot contains several ARPAbet characters that represent the formants of the vowels used in their experiment. The character location represents the relationship between the average whispered formant location and the frequency difference to the equivalent formant in a phonated vowel. In order to compensate for this difference, an algorithm for shifting the formants in the line spectrum domain has been developed [17]. For this system, the formants are shifted downward with the shift dependent on the formant location. This relationship is shown by the line in Figure 5.
The shifting algorithm algorithm uses the approximately linear relationship between the LSP frequencies and the formant locations and bandwidths. The block diagram of this algorithm is shown in Figure 6. In this technique, the Jacobian matrices $\partial \mathbf{F} / \partial \mathbf{f}$ and $\partial \mathbf{S} / \partial \mathbf{f}$ are computed, where $\mathbf{F}$ is the vector of formant locations, and $\mathbf{S}$ is a vector containing both the formant locations and bandwidths. Given desired formant and bandwidth shifts, $\Delta \mathbf{F}$ and $\Delta \mathbf{B}$ respectively, the corresponding changes in the LSPs, $\Delta \mathbf{f}_F$ and $\Delta \mathbf{f}_B$, are as follows:

$$\min \| \mathbf{D} \Delta \mathbf{f}_F \|_2, \text{ s.t. } \frac{\partial \mathbf{F}}{\partial \mathbf{f}} \Delta \mathbf{f}_F = \Delta \mathbf{F},$$

$$\min \| \Delta \mathbf{f}_B \|_2, \text{ s.t. } \frac{\partial \mathbf{S}}{\partial \mathbf{f}} \Delta \mathbf{f}_B = \begin{bmatrix} 0 \\ \Delta \mathbf{B} \end{bmatrix},$$

where $\mathbf{D}$ is a scaling matrix designed to help preserve the bandwidths in the formant shift. The final LSPs, $\mathbf{f}_M$, are found by summing the shifts,

$$\mathbf{f}_M = \mathbf{f} + \Delta \mathbf{f}_F + \Delta \mathbf{f}_B,$$

followed by enforcement of LSP ordering to ensure a stable prediction filter.

### 3 Excitation Estimation

To generate speech, the parameters of the excitation signal must be determined. The MELP excitation voicing is controlled in five frequency bands. The proposed method fixes the lower four frequency bands as voiced, while fixing the upper band as unvoiced, since this is a reasonable approximation for the majority of speech sounds. The gain is smoothed to account for the higher estimation variance. Using the observed correlation between intensity and perceived pitch of whispers [5], the pitch parameter is estimated by filtering the gain parameter. This method also allows for some user feedback control of the pitch during real-time synthesis.

### 4 Results

In order to show the spectral differences between the whispered and normal spectrum, the phrase “sway” was recorded using both modes. The linear prediction spectrum for this phrase is shown in Figures 7 and 9. As expected, the unvoiced portions of the spectra are very similar. However, in the normally voiced portion of the waveform, one can observe the differences between the two utterances. In the voiced speech, there is a smooth transition from the /w/ glide through the /ei/ diphthong, while in the whispered utterance, the formants are not as stable. In addition, the first formant in the whisper is higher.

Figure 9 contains the spectral estimates created by three different methods for modifying the spectrum. In all three, the pre-filter and formant shifting algorithm is implemented, while the smoothing method differs. The median filter example utilizes a five-point frame. This method performs very well at preserving the formant bandwidths and rapid transitions, however, it creates a 2 frame, or 45 ms, extra delay to the signal that may be unacceptable for real-time applications. The Kalman filter provides some smoothing, although not as much as the median filter. The non-linear filter based on Gaussian mixture priors achieves similar performance to the Kalman filter during the rapid transition, but is able to smooth more during the more slowly evolving diphthong.

In Figure 8, the estimated probability of the plant noise coming from the higher variance mixture is plotted. This plot can be interpreted as follows: when this probability is high, the signal is believed to be changing rapidly, so the weight of the previous samples for filtering is diminished; when this probability is low, the changes in the LSP frequencies are attributed to noise, and they are averaged out. In this phrase, the algorithm correctly assumes that the spectrum varies rapidly during the /w/ phone. However, there are several false alarms in the other parts of the phrase such as the spike at 0.6 seconds.
Conclusions

In this paper, several methods for reconstructing the LPC spectra of voiced speech from whispered speech have been tested. These algorithms compensate for the long-term spectral differences, the formant shifts, and the estimation noise inherent in whispered speech. The choices of different smoothing algorithms depend on trade-offs between delay, computational complexity, and performance. The median filter was found to be an effective solution with low computational cost, but it introduces a significant delay, while the Gaussian mixture based filter requires more computation, but provides a nearly delay free estimate. These spectral estimates can be used to synthesize normal speech using the MELP model vocoder. A real-time demonstration of this system has been created using a MELP synthesizer that allows for interactive experimentation with the proposed estimation techniques.

References


Fig. 9. Estimated and smoothed spectra of whispered “sway”