EURASIP Journal on Applied Signal Processing 2001:0, 1–9

Techniques for the Regeneration of Wideband Speech from Narrowband Speech

Jason A. Fuemmeler

University of Dayton, Department of Electrical and Computer Engineering, 300 College Park, Dayton, OH 45469-0226, USA
Email: fuemmeja@hotmail.com

Russell C. Hardie

University of Dayton, Department of Electrical and Computer Engineering, 300 College Park, Dayton, OH 45469-0226, USA
Email: rhardie@udayton.edu

William R. Gardner

Qualcomm, Inc., 5775 Morehouse Drive, San Diego, CA 92121, USA
Email: wgardner@qualcomm.com

Received 31 July 2001 and in revised form 30 September 2001

This paper addresses the problem of reconstructing wideband speech signals from observed narrowband speech signals. The goal of this work is to improve the perceived quality of speech signals which have been transmitted through narrowband channels or degraded during acquisition. We describe a system, based on linear predictive coding, for estimating wideband speech from narrowband. This system employs both previously identified and novel techniques. Experimental results are provided in order to illustrate the system's ability to improve speech quality. Both objective and subjective criteria are used to evaluate the quality of the processed speech signals.

Keywords and phrases: wideband speech regeneration, narrowband speech, linear predictive coding, speech processing, speech coding.

1. INTRODUCTION

In voice communications, the quality of received speech signals is highly dependent on the received signal bandwidth. If the communications channel restricts the bandwidth of the received signal, the perceived quality of the output speech is degraded. In many voice transmission or storage applications the high-frequency portions of the input speech signal are eliminated due to the physical properties of the channel or to reduce the bandwidth required. The resulting lowpass speech often sounds muffled or "far away" compared to the original, due to the lack of high frequency content.

One way to compensate for these effects is to efficiently encode the speech at the transmitter so that less channel bandwidth is required to transmit the same amount of information. Of course, this requires that the receiver have an appropriate decoder to recover the original signal. Because of this burden on both the receiver and transmitter, the use of wideband vocoding techniques is difficult to apply to systems that have already been standardized (e.g., analog telephone communications). It would be more convenient to devise a system at the receiver that could regenerate the lost high-frequency content.

Some work has already been done in the area of wideband speech regeneration [1, 2, 3, 4, 5, 6, 7, 8]. These works have primarily used linear predictive (LP) techniques. By using these techniques, the reconstruction problem is divided into two separate tasks. The first task is forming a wideband residual error signal, while the second is recreating a set of wideband linear predictive coefficients (LPCs). Once these two components have been generated, the wideband residual is fed into the wideband LP synthesis filter resulting in a regenerated wideband speech signal.

This paper provides a brief review of LP techniques used in wideband speech regeneration and proposes a new LP technique with several novel aspects. In particular, a new method for the creation of a wideband residual is proposed which overcomes difficulties encountered using many methods previously employed. Furthermore, a relatively new distortion measure is investigated for use in codebook generation and in mapping narrowband LPCs to wideband LPCs. To the best

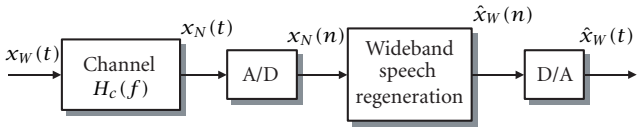

Figure 1: Context of the wideband speech regeneration system.

of the authors' knowledge, this new metric has not previously been applied to the speech regeneration problem. Finally, a new method for calculating optimal gain coefficients for application to the wideband LPCs is described.

The organization of the rest of this paper is as follows. In Section 2, an overview of the wideband speech regeneration system is provided. Since the proposed technique is based on LP coding, Section 3 provides a brief review of LP analysis and synthesis. Section 4 describes the wideband residual error regeneration and Section 5 describes the codebook mapping of the LPCs. Experimental results are presented in Section 6, and conclusions and areas for future work are presented in Section 7.

2. OVERVIEW OF THE WIDEBAND SPEECH REGENERATION SYSTEM

The context in which the wideband speech regeneration system may be employed is shown in Figure 1. Here, a narrowband continuous-time speech signal, $x_N(t)$, is formed by passing its wideband counterpart, $x_W(t)$, through a bandlimited channel. The channel, with frequency response $H_c(f)$, is assumed to be a continuous-time lowpass filter with cutoff frequency f_c. This lowpass filter need not be ideal. However, if the sampling frequency of the analog-to-digital (A/D) and digital-to-analog (D/A) converters is denoted as f_s, the following restriction is assumed:

$$|H_c(f)| \approx 0 \quad \text{for } |f| > \frac{f_s}{2K}, \tag{1}$$

where K is a positive integer greater than one. In other words, the channel transfer function is such that the digital signal entering the wideband speech regeneration system is oversampled by a factor of at least K. Furthermore, this oversampling will in turn allow the wideband speech regeneration system to increase the bandwidth of the received signal by a factor of at least K. Examples of system parameters for application to analog telephone speech might be $f_c = 3.3$ kHz, $f_s = 16$ kHz, and $K = 2$.

In many speech applications, the input speech waveform is already filtered and sampled at a lower frequency (e.g., 8 kHz typically in many wireline and wireless communication systems). In this case, the speech signal can be upsampled and filtered to provide a higher frequency sampling rate (e.g., 16 kHz as in this work) with no high frequency content above f_c.

The structure of the wideband speech recovery system is shown in Figure 2. The system begins by performing a standard LP analysis of the downsampled, narrowband speech

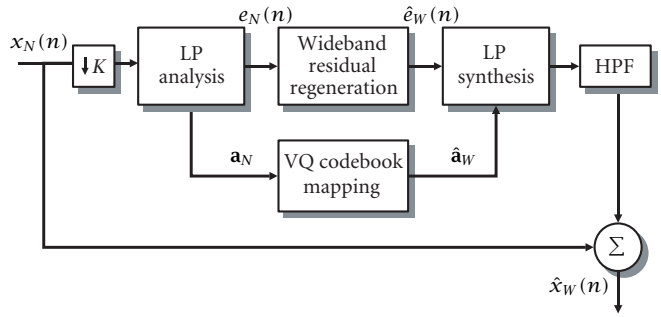

Figure 2: The wideband speech regeneration system.

signal. For each speech frame, this produces a narrowband residual or error signal, $e_N(n)$, and a set of narrowband LPCs, denoted $\mathbf{a}_N = [a_1, a_2, \ldots, a_p]^T$. To regenerate the wideband speech, both of these components must be converted into their wideband counterparts.

The process for creating a wideband residual, $\hat{e}_W(n)$, from a narrowband residual is referred to herein as high frequency regeneration (HFR). The process for generating wideband LPCs, $\hat{\mathbf{a}}_W$, from the narrowband LPCs is referred to as codebook mapping. Having regenerated these two critical speech components, it is then possible to construct wideband speech through LP synthesis (the inverse process of LP analysis). It is important to realize that, although no upsampling is explicitly shown in Figure 2, the HFR block outputs a signal at K times the sampling rate of its input. Finally, the regenerated wideband speech is passed through a highpass filter and added to the input narrowband speech signal to create the final speech waveform. This is done in order to preserve the accurate low frequency information contained in the original input speech signal. Thus, the proposed processing will not alter the low frequency content of the input signal but will simply add high frequency components. The following three sections describe in detail the LP, HFR, and codebook mapping blocks, respectively.

3. LINEAR PREDICTIVE TECHNIQUES

LP analysis determines the coefficients of a linear prediction filter designed to predict each speech sample as a weighted sum of previous samples. The prediction filter output can be written as

$$\hat{x}(n) = \sum_{k=1}^{P} a_k x(n-k), \tag{2}$$

where $a_1, a_2, \ldots, a_P$ are the LPCs. The prediction error signal (also known as a residual or residual error signal) is defined as the difference between the actual and predicted signals as follows:

$$e(n) = x(n) - \sum_{k=1}^{P} a_k x(n-k). \tag{3}$$

This expression simply defines a finite impulse response

filter with impulse response

$$a(n) = \delta(n) - \sum_{k=1}^{P} a_k \delta(n-k), \tag{4}$$

and discrete-time frequency response

$$A(\omega) = 1 - \sum_{k=1}^{P} a_k e^{jk\omega}. \tag{5}$$

This filter is known as the LP analysis filter and is used to generate the residual from the original discrete-time signal.

To perform optimal linear prediction, the LPCs are chosen such that the power in the residual is minimized. The LPCs are typically also chosen to be gain normalized and such that the LP analysis filter is minimum phase. Such LPCs can be found efficiently through the application of the Levinson-Durbin algorithm. Because the LP analysis filter is minimum phase, it has a stable inverse known as the LP synthesis filter. This synthesis filter is an all-pole, infinite impulse response filter. While the LP analysis filter is used to generate the residual, the LP synthesis filter creates the original signal from the residual. The system difference equation for the LP synthesis filter can be found by rewriting (3) as

$$x(n) = e(n) + \sum_{k=1}^{P} a_k x(n-k). \tag{6}$$

LP techniques are especially useful in speech processing. Because of the error minimization used to find the LPCs, the residual error signal tends to be spectrally flat. This means that the shape of the speech signal's spectral envelope is represented in the LPCs. The residual then contains the amplitude, voicing, and pitch information. The speech signal's spectral envelope can be approximately written in the frequency domain as

$$S(\omega) = \frac{\sigma}{|A(\omega)|}, \tag{7}$$

where σ is the square root of the residual power. In speech processing systems, LP analysis is typically performed on frames of about 10–20 ms, since speech characteristics are relatively constant over this time interval.

4. HIGH FREQUENCY REGENERATION

In this section, methods for regenerating wideband residual errors from narrow band errors are described. Previously defined techniques are described, motivation for exploring alternative methods is presented, and a novel method, referred to here as "spectral shifting," is developed.

4.1. Previous techniques

The HFR methods employed in previous systems can basically be divided into two classes. These two approaches are illustrated in Figure 3. For simplicity, the discussion is restricted to the case where $K = 2$. The first class of HFR uses

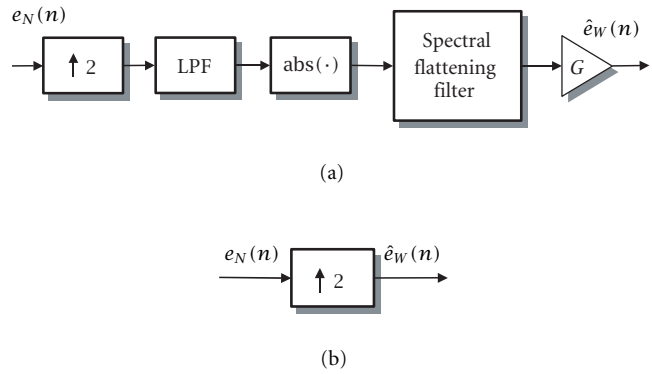

Figure 3: (a) HFR using rectification and filtering, (b) HFR using spectral folding.

rectification of the upsampled narrowband residual to generate high-frequency spectral content. The signal is then filtered with an LP analysis filter to generate a spectrally flat residual. An appropriate variable gain factor must also be applied to this new wideband residual so that its signal power will not be too large or small when compared with the original narrowband version. This approach is illustrated in Figure 3a. The main drawback to this method is that the spectral components generated by the rectification (a nonlinear operation) are largely unpredictable. As a result, it often generates noisy or rough high frequency components, especially when the speech is voiced.

The second class of HFR techniques, shown in Figure 3b, is termed spectral folding and involves expansion of the narrowband residual through insertion of zeros between adjacent samples. Although simple, this method has several potential problems when applied to voiced speech.

(1) First, it is unlikely that the new high frequency harmonics will reside at integer multiples of the voiced speech's fundamental frequency. Often, this does not result in a large perceptual effect as long as the low frequency content has the harmonics spaced correctly and as long as the energy in the low frequency components is significantly greater than that in the higher frequencies.

(2) Second, as the pitch of the narrowband residual moves higher or lower in frequency, the high-frequency portions of the new wideband residual move in the opposite direction. This will be seen later in Section 6. Because of this effect, the resultant speech can sound somewhat garbled—especially if there are wide variations in fundamental frequency.

(3) Finally, a greater problem occurs when the cutoff frequency of the bandlimiting process is lower than half the narrowband sampling frequency. Although LP analysis tends to produce spectrally flat residuals, this is generally not possible when a portion of the input spectrum has been eliminated. In these regions, the narrowband residual therefore exhibits little spectral energy. When spectral folding is applied to such a residual, the resultant wideband speech exhibits a band gap in the middle of the spectrum. This will also be seen in Section 6. This partial lack of spectral content can degrade perceptual speech quality.

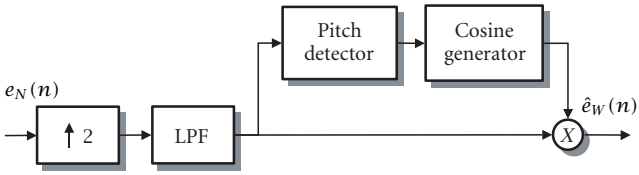

FIGURE 4: Proposed spectral shifting high frequency regeneration algorithm.

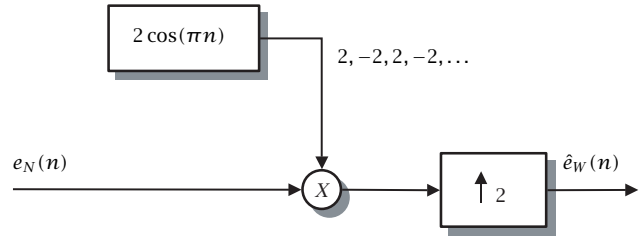

FIGURE 5: Simplest form of the spectral shifting method.

4.2. The spectral shifting method

A new method of HFR is proposed, which is illustrated in Figure 4. This method relies on spectral shifting rather than on spectral folding. If used to its full extent, it is capable of overcoming the three problems associated with the spectral folding method listed above.

The first step in the spectral shifting method involves upsampling the input narrowband residual. The lowpass filter used has cutoff frequency of $\omega_c = 2\pi f_c/f_s$ radians/sample and a gain of $K = 2$. A pitch detector then assumes that the speech is voiced and finds the fundamental pitch period, T_f, of the resultant signal. If the speech is in fact unvoiced, the pitch period computed is of little importance and the output of the pitch detector can still be used. The upsampled residual is then multiplied (mixed) with a cosine of amplitude 2 at a radian frequency ω_g, where ω_g is a multiple of the fundamental frequency that is close to (but does not exceed) the cutoff frequency ω_c. One expression that calculates such a radian frequency is

$$\omega_g = \frac{2\pi}{T_f}\,\text{floor}\left(\frac{T_f\omega_c}{2\pi}\right), \tag{8}$$

where floor$(\cdot)$ computes the maximum integer less than or equal to its argument.

The multiplication by a cosine results in a shift of the original spectrum. If the discrete-time Fourier transform of the upsampled and filtered narrowband residual is denoted as $E_N(\omega)$, it is easily shown that the resultant wideband spectrum is given by

$$\begin{aligned} E_W(\omega) &= E_N(\omega - \omega_g) + E_N(\omega + \omega_g) \\ &\approx E_N(\omega - \omega_c) + E_N(\omega + \omega_c). \end{aligned} \tag{9}$$

This expression clearly reveals the reason this method is referred to as a "spectral shifting" method. Unlike the spectral folding method, this method does not preserve the original narrowband residual information. However, this is not a problem because a highpass filter will subsequently be applied to the output of the LP synthesis filter, eliminating the narrowband portion of the regenerated signal anyway. Since $\omega_g \approx \omega_c$, the bandwidth of the wideband residual is approximately twice that of the narrowband residual. However, if more bandwidth is desired, the narrowband residual can be multiplied with cosines at higher multiples of the fundamental frequency and the results added with one another (after appropriate filtering to prevent overlap). This will further increase the signal bandwidth.

In practice, the pitch detection algorithm is applied to each frame (10–20 ms) of speech. However, instantaneously updating the frequency of the cosine at each frame boundary results in undesired frequency components and noisy speech. Applying linear interpolation to the values of ω_g between adjacent frames has been found to sufficiently eliminate this effect.

If reduction in computational complexity is desired, certain modifications can be made to the spectral shifting method. One such modification is to eliminate the pitch detector and always use a cosine of frequency ω_c. If this is done, the spectral shifting method will eliminate only problems (2) and (3) described in Section 4.1. However, as noted in Section 4.1, problem (1) is generally not significant, and thus, this simplification may have little impact on speech quality.

An additional simplification can be made by using a cosine at frequency $\pi/2$ rather than at ω_c (if $\omega_c = \pi/2$, system quality is not additionally compromised). In this case, no lowpass filter is necessary for interpolation because every other cosine value will be zero—eliminating all interpolated values. In this case, the spectral shifting method reduces to the system shown in Figure 5. Note that this system will solve only problem (2) (assuming $\omega_c \neq \pi/2$). Problem (3) is only somewhat alleviated since the size of the band gap in the spectrum will be cut in half. However, the performance is still an improvement over the spectral folding method with the only additional computational complexity being a sign change of every other sample.

5. CODEBOOK MAPPING

In this section, the mapping of the narrowband LPCs to wideband LPCs is addressed. The use of a dual codebook in a vector quantization scheme is discussed, and solutions to the problem of applying an appropriate gain to the wideband LPCs are presented.

5.1. Narrowband to wideband LPC conversion

In a vector quantization system, codebook generation is most commonly performed through training using the Linde,

Buzo, Gray (LBG) algorithm [9]. In the case of a narrowband to wideband mapping, it is necessary to generate a *dual* codebook. Part of the codebook contains narrowband codewords and the other part contains the corresponding wideband codewords. These codewords contain representative LPCs in some form.

Generation of the dual codebook requires training data sampled at the desired higher rate of f_s with the full low and high frequency content intact. These data are artificially degraded and downsampled to match the bandwidth of the actual signals to be processed. The LBG algorithm is then applied to the speech frames in the narrowband version of the training data. While the LBG algorithm operates on the narrowband data, each operation is mimicked on the wideband version of the training data to form the wideband portion of the dual codebook. In this way, the dual codebook will contain a set of representative narrowband codewords and the corresponding codewords based on the wideband data. This dual codebook now contains the a priori information needed to allow the wideband speech regeneration algorithm to extend the bandwidth of a speech signal.

During wideband speech regeneration, narrowband codewords are computed from the input speech frames and the best match in the narrowband portion of the codebook is found. The corresponding wideband codeword from the wideband portion of the codebook is then used to generate the output speech. The assumption underlying the use of the dual codebook is that there is correlation between the low frequency spectral envelope of a speech frame and its high frequency envelope for a given speaker or class of speakers. That is, when the algorithm recognizes the narrowband spectral envelope of a speech frame (provided by the narrowband LPCs), the training data will allow us to predict what the spectral envelope should be for the full broadband version (contained in the wideband LPCs). Performance will clearly depend on the level of correlation between the low and high frequencies, and on how representative the training data is of the actual data. It is worth mentioning that improved performance was reported by Epps and Holmes [8] when separate codebooks were used for voiced and unvoiced speech frames. However, this method was not employed here.

The operation of the LBG algorithm and the codebook mapping requires some quantitative measure of closeness for sets of LPCs. The only requirement is a distance (or distortion) measure for which a centroid calculation exists. The centroid of a bin or group is defined as the codeword that minimizes the total distortion over that bin. The quality of the resultant codebook is greatly affected by the correlation between the quantitative distance metric and human perception of difference in the reconstructed speech frames.

One commonly used distortion metric is the Itakura-Saito measure defined as

$$d_{\mathrm{IS}}\left(\frac{\sigma}{A(\omega)},\frac{\hat{\sigma}}{\hat{A}(\omega)}\right) = \int_{-\pi}^{\pi}\left(\frac{\sigma^2\left|\hat{A}(\omega)\right|^2}{\hat{\sigma}^2\left|\hat{A}(\omega)\right|^2}-\ln\frac{\sigma^2\left|\hat{A}(\omega)\right|^2}{\hat{\sigma}^2\left|\hat{A}(\omega)\right|^2}-1\right)d\omega. \tag{10}$$

An efficient method for calculating this exists which uses the estimated autocorrelation for each speech frame. The use of this distortion measure for codebook generation was first described in [10]. Another common metric is log spectral distortion (LSD), given by

$$d_{\mathrm{LSD}}\left(\frac{\sigma}{A(\omega)},\frac{\hat{\sigma}}{\hat{A}(\omega)}\right) = \int_{-\pi}^{\pi}\left(\ln\frac{\sigma^2}{\left|A(\omega)\right|^2}-\ln\frac{\hat{\sigma}^2}{\left|\hat{A}(\omega)\right|^2}\right)^2 d\omega. \tag{11}$$

With respect to wideband speech regeneration, this distance measure has been previously applied by sampling the logarithm of the Fourier transform of the LPCs [1, 8].

It has been shown by Gardner and Rao [11] that a distortion measure, using a weighted mean squared error of the line spectral pair frequencies, is equivalent to LSD and to the Itakura-Saito measure for high rate vector quantizers. It is thought that this measure may offer the performance of the LSD metric in the current application, yet be more computationally efficient. The computational savings comes primarily from the fact that Fourier transforms and logarithms need not be computed.

5.2. Optimal gain constant calculation

In addition to the generation of the wideband LPCs discussed above, the gain applied to each wideband LP synthesis filter must also be determined such that the new wideband information has the appropriate energy. The optimal gain constant is defined here as that which minimizes the distance between the reconstructed and original wideband spectral envelopes in the narrowband region.

To derive the optimum gain, we first trace a wideband spectral envelope through the system. Represent the spectral envelope of the original wideband speech signal as a real, positive, symmetric function in the frequency domain. After the bandlimiting filter and subsequent downsampling, the narrowband spectral envelope can be represented as

$$S_N(\omega) = S_W\left(\frac{\omega}{K}\right)\left|H\left(\frac{\omega}{K}\right)\right|, \tag{12}$$

where $H(\omega)$ is the impulse-invariant discrete-time system frequency response used to model the continuous channel, $H_c(f)$. After LP analysis, the spectral envelope of the narrowband residual is

$$\begin{aligned} S_E(\omega) &= S_N(\omega)\left|A_N(\omega)\right| \\ &= S_W\left(\frac{\omega}{K}\right)\left|H\left(\frac{\omega}{K}\right)\right|\left|A_N(\omega)\right|. \end{aligned} \tag{13}$$

The high-frequency regeneration technique does not alter the basic shape of the residual spectral envelope. However, it does create a signal sampled at a higher rate. Thus, the wideband residual spectral envelope can be approximated as $S_E(K\omega)$. The wideband LP synthesis filter transforms the wideband residual spectral envelope into the reconstructed wideband

speech signal spectral envelope as

$$\begin{aligned}\hat{S}_W(\omega) &= S_E(K\omega)\frac{\sigma}{|\hat{A}_W(\omega)|}\\ &= S_W(\omega)|H(\omega)||A_N(K\omega)|\frac{\sigma}{|\hat{A}_W(\omega)|}.\end{aligned} \tag{14}$$

Note the presence of the gain constant σ. Using (14), we are able to relate the original and reconstructed spectral envelopes in the narrowband region.

One approach to finding the optimal gain constant is to store a calculated gain constant for each training frame in a given bin [1]. This constant can be computed using the relative powers of the narrowband and wideband residuals for each frame. At the end of training, a centroid of these gains is found for each bin using the gain values for each frame, the narrowband training codewords for each frame, and the newly computed representative wideband codeword for the bin.

However, it is assumed that the representative narrowband codeword is indeed representative of all the narrowband training codewords in that bin. Thus, an alternative approach is to use only the representative narrowband codeword, the representative wideband codeword, and an estimate of the bandlimiting transfer function to compute the optimal gain for each bin. This eliminates the need to store the gain constants during training and also the need to use multiple narrowband codewords in the optimal gain calculation. With this approach, the gain constant can be computed by minimizing

$$\begin{aligned}&d(S_W(\omega),\hat{S}_W(\omega))\\ &\quad= \left(S_W(\omega), S_W(\omega)|H(\omega)||\hat{A}_N(K\omega)|\frac{\sigma}{|\hat{A}_W(\omega)|}\right),\end{aligned} \tag{15}$$

with respect to σ over a selected narrowband frequency range, where $d(\cdot,\cdot)$ is the relevant distance metric. If the LSD distance measure is being used, the optimal gain constant is given by

$$\begin{aligned}\sigma = \exp\Bigg(&\frac{1}{4\omega_n}\int_{-\omega_n}^{\omega_n}\ln\left(|\hat{A}_W(\omega)|^2\right)\\ &-\ln\left(|H(\omega)|^2\right)-\ln\left(|\hat{A}_N(K\omega)|^2\right)d\omega\Bigg).\end{aligned} \tag{16}$$

If the Itakura-Saito distance measure is being used, the optimal gain constant is given by

$$\sigma = \sqrt{\frac{1}{2\omega_n}\int_{-\omega_n}^{\omega_n}\frac{|\hat{A}_W(\omega)|^2}{|H(\omega)|^2|\hat{A}_N(K\omega)|^2}\,d\omega}. \tag{17}$$

In both expressions, ω_n, is a radian frequency in the range $(0,\pi/K)$. This frequency should be selected to include only those portions of the narrowband spectrum in which $H(\omega)$ is invertible. In the simplest case, where $H(\omega)$ is assumed to be an ideal lowpass filter, it is most appropriate to use $\omega_n = \omega_c$. In practice, the various spectra are generated and numerically integrated by summing fast Fourier transform (FFT) results.

Table 1: Parameters used in system testing.

Bandlimiting filter type	High-order Butterworth
Bandlimiting filter cutoff	3.3 kHz
Narrowband sampling rate	8 kHz
Wideband sampling rate	16 kHz
Frame size	30 ms
Frame rate	100 Hz
Codebook size	512
Narrowband LPCs (Itakura-Saito)	7
Wideband LPCs (Itakura-Saito)	15
Narrowband LPCs (Gardner-Rao)	8
Wideband LPCs (Gardner-Rao)	14

This calculation does not slow system performance because it is performed only once—after training and before the system actually operates.

6. EXPERIMENTAL RESULTS

Testing of the wideband speech recovery system has been performed in MATLAB using speech samples obtained from the TIMIT speech corpus.[1] For simplicity, codebook training was performed using 30 randomly selected female utterances from dialect region 3 (North Midland region). These utterances comprised approximately 90 seconds of speech data. Testing of the resultant system was performed using 2 utterances not in the training data, but from the same gender and dialect region.

The parameters used in both the training and testing phases are shown in Table 1. As stated earlier, speech waveform characteristics are relatively stationary over a 10 ms interval. Thus, computations are performed for frames taken at a 100 Hz rate. The frames, however, are 30 ms soft-windowed overlapping frames. This allows for smooth transitions between adjacent frames. The numbers of LPCs used for the Itakura-Saito distortion measure have been selected to make FFT computations more convenient. However, the Gardner-Rao method requires that the numbers of LPCs be even, explaining why slightly different numbers were used for this distortion measure.

Additionally, it should be noted that frames with energies below an empirically determined threshold were considered silence, and were thus excluded from training and testing. The spectral shifting method was implemented by multiplying the narrowband residual by two cosine functions: one at 3.3 kHz and one at 4.7 kHz. These results were added together after appropriate filtering to prevent overlap. Note that pitch-detection was not employed in these tests.

Sample results from these tests are shown in Figure 6. Figure 6a shows an original wideband speech signal spectro-

[1] Texas Instruments/Massachusetts Institute of Technology Acoustic-Phonetic Countinuous Speech Corpus October 1990 (www.ntis.gov/fcpc/cpn4129.htm).

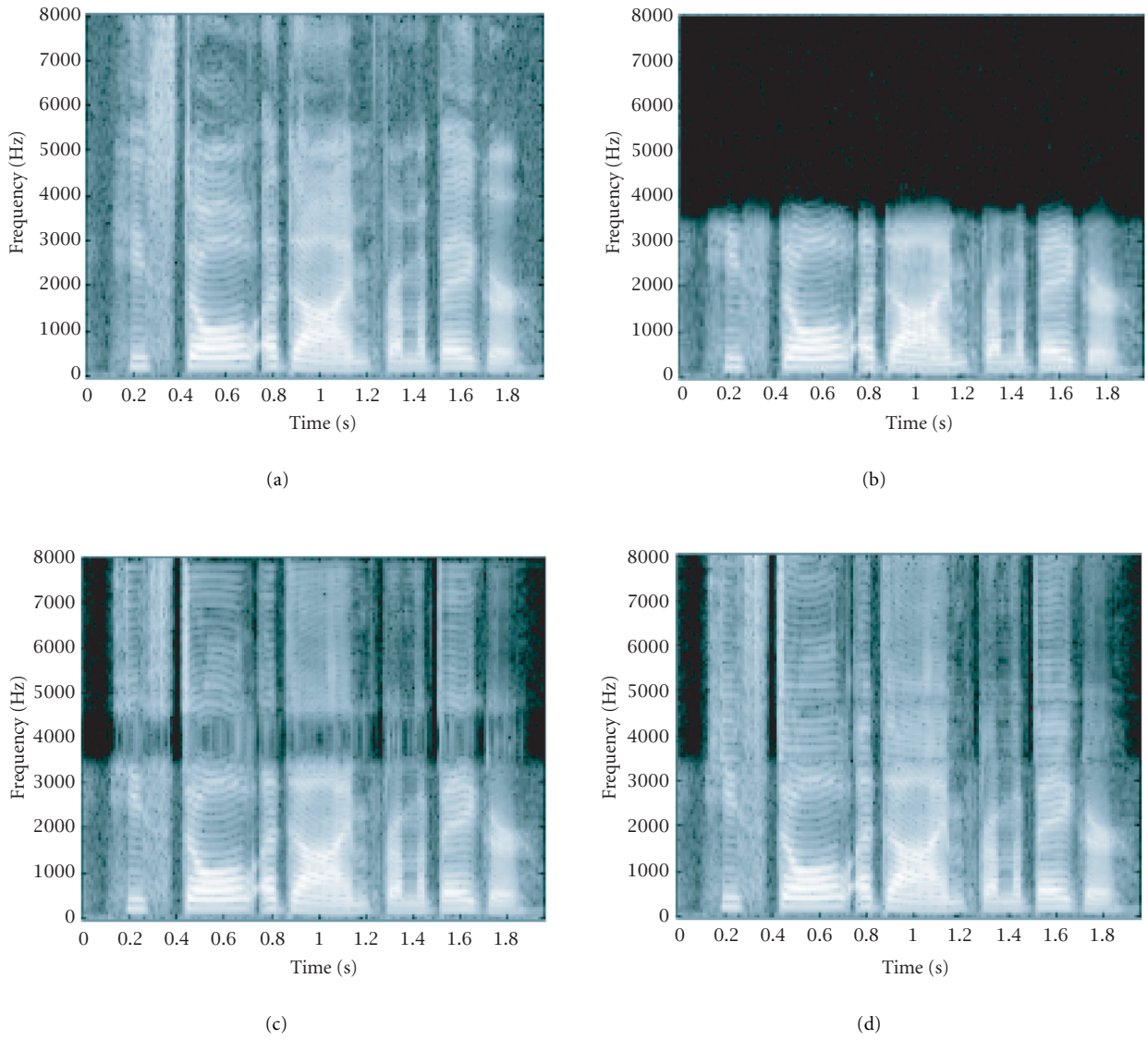

FIGURE 6: Spectrograms for the (a) original signal, (b) bandlimited/narrowband signal, (c) reconstructed signal using spectral folding, and (d) reconstructed signal using spectral shifting.

gram, while Figure 6b shows this same signal bandlimited to approximately 3.3 kHz. The spectrogram of the reconstructed signal using the Itakura-Saito distortion measure and spectral folding is shown in Figure 6c. The spectrogram of the reconstructed signal using the Itakura-Saito distortion measure and spectral shifting is shown in Figure 6d. The result using the spectral folding method contains a gap in the middle of the frequency spectrum. Also, the pitch contours in the high-frequency portion of the spectrum run counter to those in the true wideband signal and in the low-frequency portion. This is especially evident in the 0.5–0.7 sec time range. In contrast, the spectral shifting method eliminates both of these problems. A comparison of Figures 6a and 6d reveals that the high-frequency portion of the original spectrum is approximated reasonably well in the high-frequency portion of the reconstructed spectrum.

Subjective evaluations have also been obtained using a survey, to which 18 participants responded. These participants were asked to compare 5 pairs of speech files and select one from each pair that they thought had the best overall sound quality. The choices and the results are summarized in Table 2. A p-value is also given to indicate a level of confidence in the results. For example, a p-value of 0.05 indicates that there is a 95% chance that the choice preferred in the survey would still be preferred if an infinite number of responses had been received.

It can be concluded from the survey results that it is indeed possible to enhance the perceived quality of speech signals. This is seen most clearly in Choice 3. It is believed that artifacts created by the reconstruction process—especially around unvoiced consonants—adversely affected the results for Choice 2. Techniques for eliminating these artifacts need

Table 2: Survey results for the A/B testing.

Choice 1		
(A) Bandlimited signal vs.		
(B) Original wideband signal		
Preferred: B	Proportion: 100%	p-value: ~ 0
Choice 2		
(A) Bandlimited signal vs.		
(B) Rebuilt signal (Itakura-Saito and spectral shifting)		
Preferred: B	Proportion: 61%	p-value: 0.173
Choice 3		
(A) Rebuilt signal (Gardner-Rao and spectral shifting) vs.		
(B) Bandlimited signal		
Preferred: A	Proportion: 78%	p-value: 0.009
Choice 4		
(A) Rebuilt signal (Itakura-Saito and spectral folding) vs.		
(B) Rebuilt signal (Itakura-Saito and spectral shifting)		
Preferred: B	Proportion: 100%	p-value: 0
Choice 5		
(A) Rebuilt signal (Itakura-Saito and spectral shifting) vs.		
(B) Rebuilt signal (Gardner-Rao and spectral shifting)		
Preferred: B	Proportion: 56%	p-value: 0.319

to be developed in order to ensure consistent perceptual quality.

The results for Choice 4 clearly indicate that the spectral shifting method outperforms the spectral folding method as a method for HFR for these data. Choice 5 appears to indicate that the Gardner-Rao measure produces results comparable to those of the more often used Itakura-Saito measure on these data. For very high rate quantizers, these two measures should perform comparably to the LSD measure although there may be some advantage to using the Gardner-Rao measure over that of the Itakura-Saito measure for lower rate quantizers.

7. CONCLUSIONS

Wideband speech regeneration techniques enable one to improve perceptual speech quality at the receiving end of voice communication systems, without using specialized encoding techniques. Because they involve processing only at the receiver, these systems may be of use in voice communications systems already in place. However, further work is necessary before such systems achieve optimal performance.

We have demonstrated that the regeneration of the wideband residual from the narrow band can have a dramatic impact on perceived speech quality. The spectral shifting method appears very promising, based on the subjective evaluation results and examination of the spectrograms. The other key aspect of the wideband regeneration system is the codebook generation. The techniques presented here rely on the fact that correlation exists between the low frequency spectral envelope of a speech frame and its high frequency envelope, and that this correlation can be captured in the dual codebook. The relationship between the low and high frequency envelopes may vary widely from speaker to speaker. Thus, speaker-dependent training may be a way to more fully capture and exploit this.

Future work will involve further testing of the techniques presented in this paper. Such testing will be applied to a wide range of speakers and dialects to study the sensitivity of the algorithm to speaker variations. More subjective testing using larger numbers of participants, greater amounts of testing data, and more controlled listening environments will also be helpful in evaluation of the proposed technique. The robustness of these techniques in the presence of noise, music, multiple speakers, or other audio distortions is another important aspect to study.

We have identified the problem of occasional "hissing" artifacts heard around unvoiced consonants. It is thought that this arises because such consonants have very similar low-frequency characteristics and are difficult to distinguish from one another. Using separate codebooks for voiced and unvoiced sounds (as in [8]) may aid in eliminating this problem.

REFERENCES

[1] H. Carl and U. Heute, "Bandwidth enhancement of narrowband speech signals," in *Proceedings of EUSIPCO'94*, vol. 2, pp. 1178–1181. 1994.

[2] M. Abe and Y. Yoshida, "More natural sounding voice quality over the telephone," *NTT Review*, vol. 7, no. 3, pp. 104–109, 1995.

[3] C. Avendano, H. Hermansky, and E. A. Wan, "Beyond Nyquist: towards the recovery of broad-bandwidth speech from narrow-bandwidth speech," in *Proceedings of the Fourth European Conference on Speech Communication and Technology*, pp. 165–168. Madrid, Spain, 1995.

[4] C.-F. Chan and W.-K. Hui, "Wideband enhancement of narrowband coded speech using MBE re-synthesis," in *Proceedings of the International Conference on Signal Processing, ICSP*, vol. 1, pp. 667–670. 1996.

[5] H. Yasukawa, "A simple method of broadband speech recovery from narrowband speech for quality enhancement," in *Proceedings of the IEEE Digital Signal Processing Workshop*, pp. 173–175. 1996.

[6] H. Yasukawa, "Restoration of wideband signal from telephone speech using linear prediction error processing," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, pp. 901–904. 1996.

[7] H. Yasukawa, "Wideband speech recovery from bandlimited speech in telephone communications," in *Proceedings of IEEE ISCAS'98*, vol. 4, pp. 202–205. Monterey, CA, USA, 1998.

[8] J. Epps and W. H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," in *IEEE Workshop on Speech Coding*, pp. 371–374. Porvoo, Finland, 1999.

[9] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[10] A. Buzo, A. H. Gray, R. M. Gray, and J. D. Markel, "Speech

coding based upon vector quantization," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 562–574, 1980.

[11] W. R. Gardner and B. D. Rao, "Optimal distortion measures for the high rate vector quantization of LPC parameters," in *IEEE Trans. on Speech and Audio Processing*. 1995.

Jason A. Fuemmeler graduated Summa Cum Laude from the University of Dayton in December, 2000 with a Bachelor of Electrical Engineering degree from the Department of Electrical and Computer Engineering. As a member of the University of Dayton Honors program, his scholastic work included the development of an Honors Thesis in wideband speech regeneration. Mr. Fuemmeler is currently employed as a signal processing engineer at Rockwell Collins in Cedar Rapids, Iowa.

Russell C. Hardie graduated Magna Cum Laude from Loyola College in Baltimore Maryland in 1988 with a B.S. degree in Engineering Science. He obtained an M.S. and Ph.D. degree in Electrical Engineering from the University of Delaware in 1990 and 1992, respectively. Dr Hardie served as a Senior Scientist at Earth Satellite Corporation in Maryland prior to his appointment at the University of Dayton in 1993. He is currently an Associate Professor in the Department of Electrical and Computer Engineering and holds a joint appointment with the Electro-Optics Program. Along with several collaborators, Dr Hardie received the Rudolf Kingslake Medal and Prize from SPIE in 1998 for work on multi-frame image resolution enhancement algorithms. In 1999, he received the School of Engineering Award of Excellence in Teaching at the University of Dayton. His research interests include a wide variety of topics in the area of digital signal and image processing, with emphasis on signal and image restoration.

William R. Gardner received his BSEE in 1989 from the University of Delaware, his MSEE in 1990 from Stanford University, and his Ph.D. in 1994 from the University of California, San Diego. He is currently Vice President of Engineering at Qualcomm, Inc., a leading provider of wireless communications technology. He has published numerous papers related to speech coding, and he holds 14 issued patents related to speech processing and communications technology.