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

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
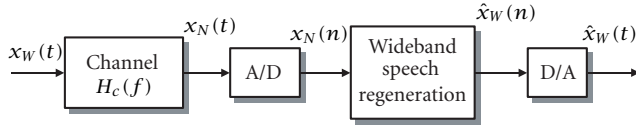
FIGURE 1: Context of the wideband speech regeneration system.

FIGURE 2: 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
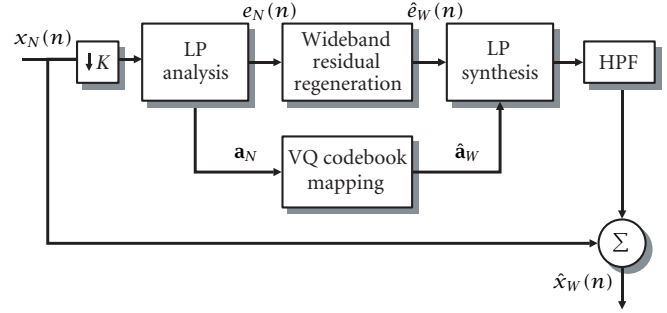
signal. For each speech frame, this produces a narrowband residual or error signal, $e_N(n)$, and a set of narrowband LPCs, denoted $\mathbf{a}_N = [a_1, a_2, \ldots, a_p]^T$. To regenerate the wideband speech, both of these components must be converted into their wideband counterparts.

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

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

and discrete-time frequency response

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

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

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

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

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

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
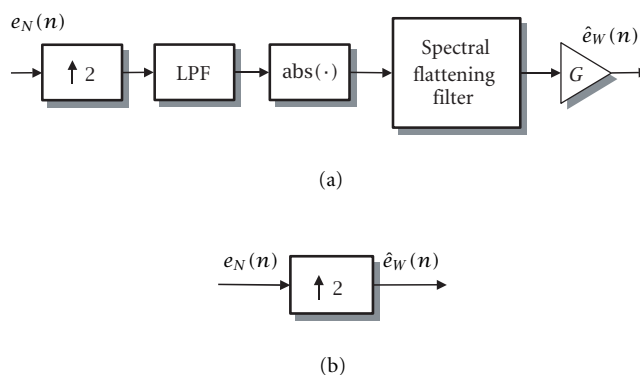

(a)

(b)

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.
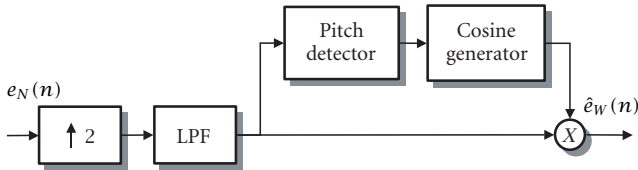
FIGURE 4: Proposed spectral shifting high frequency regeneration algorithm.

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
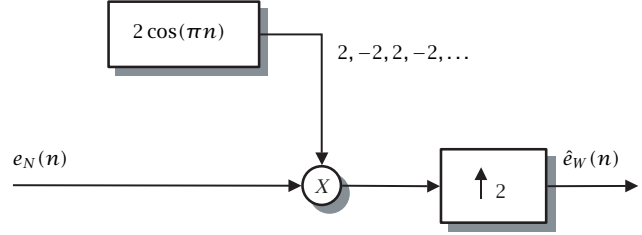
fundamental frequency and the results added with one another (after appropriate filtering to prevent overlap). This will further increase the signal bandwidth.

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

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

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

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

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

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

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

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.

fastcase®
Smarter legal research.