Dual-Mode Wideband Speech Recovery from Narrowband Speech

Yasheng Qian *Peter Kabal*

Department of Electrical and Computer Engineering
McGill University, Montreal, Canada

Abstract

The present public telephone networks trim off the lowband (50–300 Hz) and the highband (3400–7000 Hz) components of sounds. As a result, telephone speech is characterized by thin and muffled sounds, and degraded speaker identification. The lowband components are deterministically recoverable, while the missing highband can be recovered statistically. We develop an equalizer to restore the lowband parts. The highband parts are filled in using a linear prediction approach. The highband excitation is generated using a bandpass envelope modulated Gaussian signal and the spectral envelope is generated using a Gaussian Mixture Model. The mean log-spectrum distortion decreases by 0.96 dB, comparing to a previous method using wideband reconstruction with a VQ codebook mapping algorithm. Informal subjective tests show that the reconstructed wideband speech enhances lowband sounds and regenerates realistic highband components.

1 Introduction

The telephone speech transmitted in current public telephone networks is bandpass-filtered to 300–3400 Hz. The filter used is characterized by a response template in the ITU-T G.712 standard. The lowband boundary is set to suppress the power line longitudinal interference. Typically, there is more than 22 dB attenuation at 50–60Hz. The upper band boundary is specified to reduce the bandwidth requirements while retaining high intelligibility, though sacrificing naturalness. The loss of the lowband components makes the speech sound thin. The missing highband (3400–7000 Hz) leads to muffled sounds. In addition, it is difficult to distinguish between unvoiced phonemes such as /s/ and /f/, because their difference is essentially manifested in the highband range.

With the rapid evolution of telecommunication technology, hands-free telephony, teleconferencing, and future 3G wireless communications systems, wideband speech coding technology will deliver bandwidths up to 7 kHz with high subjective quality. A cost-effective way to provide wideband quality at the interface between newer wideband systems and conventional narrowband systems is to generate wideband speech from transmitted narrowband speech.

Based on a speech production model, the challenge is the regeneration of the excitation signal and spectrum envelope for the highband signal and restoration of the attenuated lowband components. Several papers have addressed these topics. The excitation signal can be modelled as a pulse train for voiced frames or a Gaussian noise for unvoiced frames in a manner similar to low bit-rate LPC vocoders [1]. An enhanced version employs harmonic-noise modelling, in which the pulse train sequence is replaced by a sum of pitch harmonics [2]. Another alternative utilizes spectral folding by downsampling and upsampling the available bandpass residual [3], as in early RELP speech coders. The spectrum envelope of the missing highband components can be reconstructed by a VQ codebook mapping or a statistical modelling approach [4, 5].

Our new approach considers two different models for a wideband recovery system. A deterministic model (an equalizer) is used to restore the attenuated lowband components. For the highband excitation, a 2–3 kHz bandpass-envelope modulated Gaussian noise (BP-MGN) is used for the highband excitation. A modulation gain is introduced to provide appropriate highband amplitude. A statistical Gaussian Mixture Model (GMM) of the wideband speech spectrum envelope parameters, the narrow band pitch gain and the modulation gain are the features used to estimate the missing highband spectrum envelope and the modulation gain.

2 The BP-MGN Excitation

We observed that the Linear Prediction (LP) residual of voiced phonemes contains weak pitch harmonics and noise-like components over 4 kHz, while the residual below 3.5 kHz shows strong pitch harmonics, as shown in Fig. 1. The unvoiced residuals are noisy in the highband as well as in the lowband as shown in the same figure. We use BP-MGN as a substitute for the highband part of the excitation.

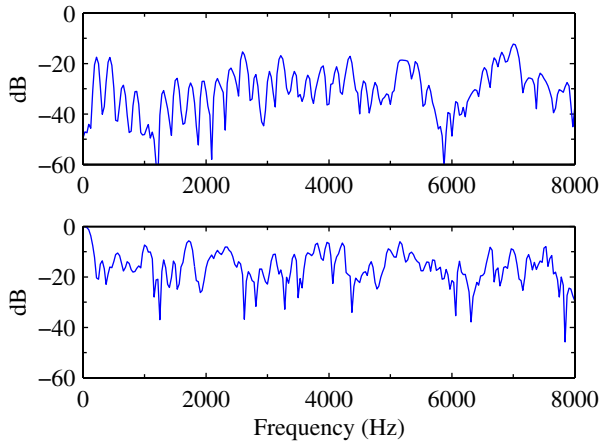

Fig. 1 The LP residual spectrum of a voiced phoneme (upper trace) and the LP residual spectrum of an unvoiced phoneme (lower trace).

The block diagram for generating BP-MGN is shown

in Fig. 2. The upsampled narrowband speech passes through a 2–3 kHz bandpass filter. The bandpass signal is

$$s_{bp}(n) = s_{bb}(n)\cos(2\pi f_o n). \quad (1)$$

where $f_o = 2.5$ kHz and $s_{bb}(n)$ is a baseband signal. The envelope of the bandpass signal is $|s_{bp}(n)|$. The spectrum of the envelope is $S_{bpe}(\omega)$,

$$S_{bpe}(\omega) = \sqrt{S_{bp}(\omega)S_{bp}(\omega)}. \quad (2)$$

The BP-MGN excitation, $e(n)$ is a bandpass-envelope modulated by a Gaussian noise. The spectrum of the BP-MGN excitation $E(\omega)$ is the convolution of the Gaussian noise spectrum $G_n(\omega)$ and S_{bpe} in frequency domain,

$$E(\omega) = G_n(\omega) * S_{bpe}(\omega). \quad (3)$$

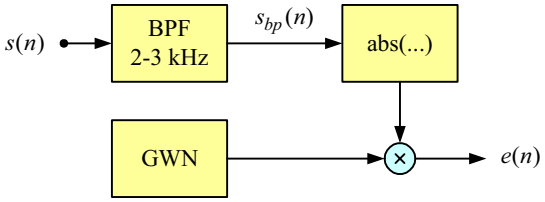

Fig. 2 BP-MGN excitation generation.

Figure 3 (top) shows the spectrum of the bandpass signal, $S_{bp}(\omega)$ of a voiced phoneme. It has strong pitch harmonics. Figure 3 (middle) gives the spectrum of the bandpass-envelope, showing the presence of pitch harmonics. Figure 3 (bottom) shows the BP-MGN spectrum, which will be used in the high frequency region.

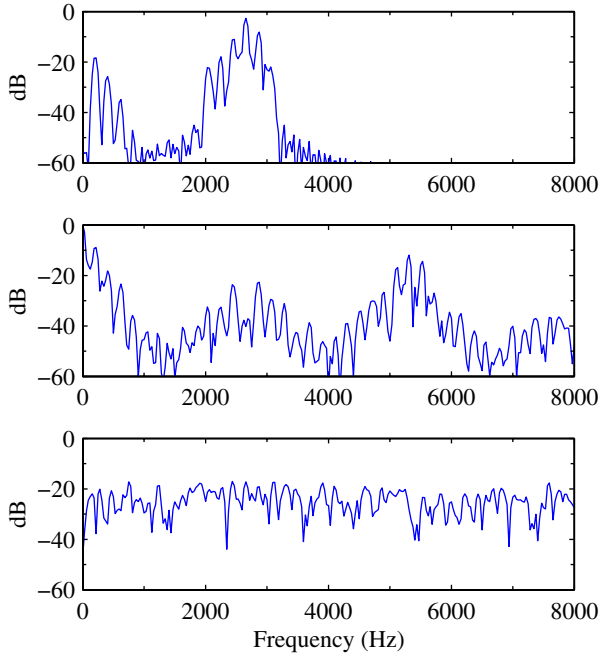

Fig. 3 The spectrum of the bandpass signal (top); the spectrum of the bandpass-envelope (middle); BP-MGN spectrum.

Figure 4 (top) shows the spectrum of the highband components of a voiced phoneme. Figure 4 (bottom) shows the recovered highband components with BP-MGN excitation under perfect highband spectrum regeneration. This latter spectrum maintains the perceptual cues for the original spectrum. Our informal subjective listening tests for several sentences verify that the BP-MGN excitation works well as a substitute for the missing highband excitation.

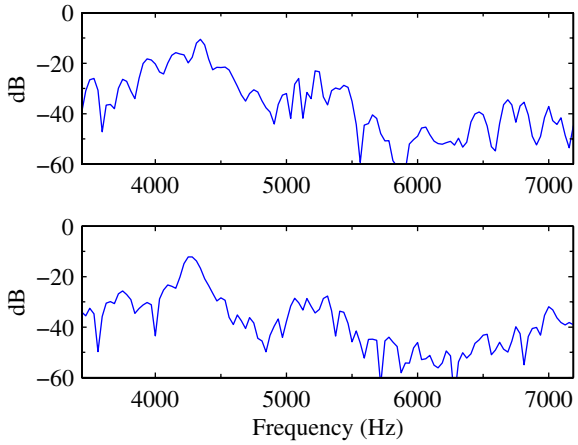

Fig. 4 The highband spectrum of a voiced phoneme (top); the reconstructed highband spectrum with BP-MGN.

A modulation gain, g, is used to set the energy of the resynthesized highband components. The reconstructed highband signal is

$$y(n) = g \cdot [e(n) * h(n)]. \quad (4)$$

The modulation gain g is the square root of the energy ratio of the original highband to the resynthesized one of each frame. The gain - a random variable, is represented by a GMM with the narrowband spectrum parameters and the pitch gain. We can calculate the GMM parameters with a large set of training data. Therefore, we are able to estimate the gain from the narrowband spectrum, pitch gain and GMM parameters as in highband spectrum reconstruction. The details are explained in the next section.

3 Reconstruction of the Highband Spectrum

The missing highband spectrum is reproduced by a statistical GMM of narrowband and highband spectrum parameters. Because of the well-known properties (ordering and quantization error resilience) of Line-Spectrum-Frequencies (LSF) for representing the speech spectrum, 14 and 10 LSFs are utilized for the narrowband and highband spectrum, respectively. The highband LSFs are LP spectrum-expanded by 60 Hz. In our previous work using a VQ codebook [6], we found that an acoustic-phonetic classification based on the pitch gain - β was beneficial. For the GMM approach we take pitch gain as an extra parameter.

The LSFs, β and the modulation gain, g, belong to a class of random vectors, whose probability density func-

tion (pdf) can be approximated by a GM pdf. Figure 5 (top) shows a histogram of the 3rd LSF of the highband components of a 529 second utterance by 12 female and 12 male speakers. Figure 5 (bottom) represents the GMM fit. The GM pdf is a weighted sum of M D-dimensional joint Gaussian density distributions.

$$p_Z(\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{i=1}^{M} \alpha_i b_i(\boldsymbol{z}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i). \quad (5)$$

where M is the number of individual Gaussian components, α_i, $i = 1,\ldots,M$ are the (positive) mixture weights, and Z is a D-dimensional random vector. Each density is a D-variate Gaussian PDF of the form,

$$b_i(\boldsymbol{z}|\mu_i,\boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp(\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_i)). \quad (6)$$

with mean vector $\boldsymbol{\mu}_i$, and covariance matrix $\boldsymbol{\Sigma}_i$. The GM pdf is defined by the mean vectors, the covariance matrices and the mixture weights for the Gaussian components.

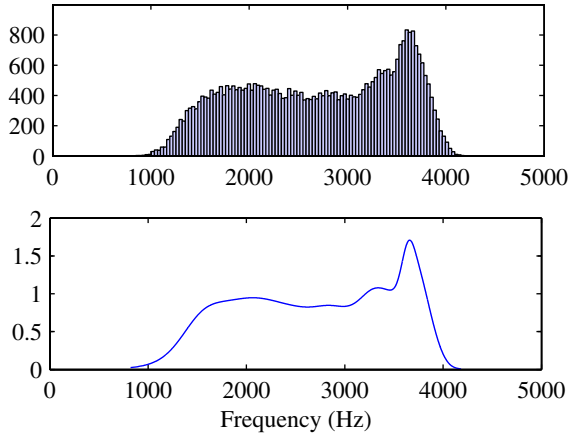

Fig. 5 The histogram of the 3rd high frequency LSF (top); GM pdf (bottom).

The parameter set, $\{\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ can be estimated by the maximum likelihood (ML) method. The ML algorithm finds the GMM parameters with maximum probability density for the given training data. We employ the popular expectation-maximization (EM) algorithm [7] to determine the set of GM density parameters iteratively.

Figure 6 shows the GM pdf of the 2nd and the 3rd highband LSFs of the training data. The training data of wideband speech are taken from Speech Database with a total of 39 479 frames each of 20 ms. The LSF GMM with the pitch gain, β, has a total of 25 dimensions. The number of mixtures, M, is 128. The covariance matrices, $\boldsymbol{\Sigma}_i$, are diagonal. The regeneration of the highband spectrum is based on the GMM joint density distribution of Eq. (5).

Let the random vector $\boldsymbol{x}$ be the combination vector of the narrowband LSF vector and the pitch gain β. The vector $\boldsymbol{y}$ is the highband LSF vector. For a given estimate, $\hat{\boldsymbol{y}}$, the mean-square error is

$$\varepsilon^2 = \int_{\Omega_y} ||\boldsymbol{y}-\hat{\boldsymbol{y}}||^2 p_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\,d\boldsymbol{y}. \quad (7)$$

The estimate which minimizes the error is found from

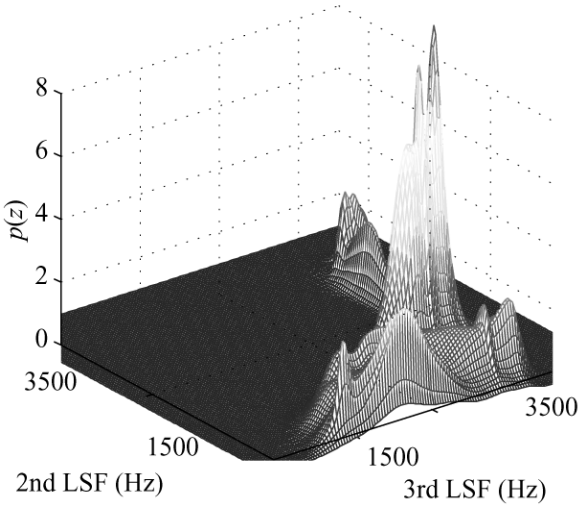

Fig. 6 The 2nd and 3rd LSF GM pdf.

$\partial\varepsilon^2/\partial\hat{\boldsymbol{y}} = 0$.

$$\hat{\boldsymbol{y}} = \frac{\int_{\Omega_y} \boldsymbol{y} p_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\,d\boldsymbol{y}}{\int_{\Omega_y} p_{Y|X}(\boldsymbol{y}|\boldsymbol{x})\,d\boldsymbol{y}} = \frac{\sum_{i=1}^{M}\alpha_i b_i(\boldsymbol{x})\boldsymbol{\mu}_{iy}}{\sum_{j=1}^{M}\alpha_j b_j(\boldsymbol{x})}. \quad (8)$$

where $\boldsymbol{\mu}_{iy}$ is the mean vector of the highband LSFs of the i-th Gaussian component. The estimate of the highband LSF vector is the expectation of the highband mixture mean vectors, given the narrowband LSF vector and the pitch gain. Similarly, we have established a GMM for the narrowband LSF vector, the pitch gain and the BP-MGN modulation gain. The modulation gain can be estimated with an equation similar to Eq. (8).

4 The Lowband Equalizer

The ITU-G.712 standard frequency response template for analogue to analogue channels between 2-wire ports specifies the limits of the response, as shown in Table 1. However, current telephone networks provide at least 22 dB attenuation at 50–60 Hz range to suppress the power line coupling interference with a highpass filter at the transmission side. A typical frequency response of a highpass filter is shown in Table 2.

Table 1 ITU-T G.712 frequency template for analogue-to-analogue channels between 2-wire ports.

f (Hz)	[0,200)	[200,300)	300
$\lvert H(f)\rvert$ (dB)	$[-\infty, 0]$	$[-\infty, 0.6]$	$[-2, 0.6]$

Table 2 Typical frequency response of a highpass filter.

f (Hz)	50	100	150	200	300
$\lvert H(f)\rvert$ (dB)	−25	−10	−5	−0.8	−0.2

We have designed an equalizer to recover the loss in the lowband. The equalizer has a boost of 10 dB at 100 Hz. The frequency response of the signal after the equalizer is almost flat from 100–300 Hz.

5 The Recovery System and Results

We have implemented the dual-model wideband recovery system from telephony speech, as shown in Fig. 7. The narrowband speech is interpolated to a 16 kHz sampling frequency. The upsampled speech is then passed to four branches: the first one goes to the lowband equalizer; the second one is a telephone speech direct path. The third branch passes to an LP analysis stage to extract the narrowband LSFs and the pitch gain. This is followed by highband LSF and modulation gain estimation using the GMM approach. The fourth branch generates the BP-MGN excitation. The LP synthesis filter reconstructs the missing highband components. Finally, we combine those three band outputs, the equalized lowband, the narrowband and the recovered highband to form a wideband speech signal.

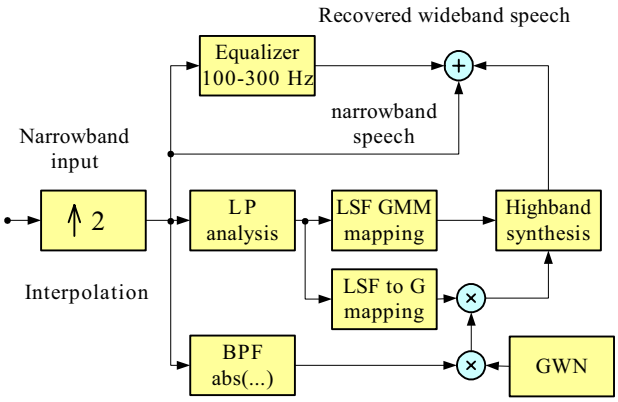

Fig. 7 The dual-mode wideband recovery system.

We have measured the mean log spectrum distortion (SD) in the missing highband (3.5–7 kHz). The definition of SD is as follows:

$$\mathrm{SD}^2 = \frac{1}{\pi}\int_{\omega_l}^{\omega_h} 20\log_{10}\left(\frac{\dfrac{g}{|A_{\mathrm{hb}}(e^{j\omega})|}}{\dfrac{g_{\mathrm{gmm}}}{|A_{\mathrm{gmm}}(e^{j\omega})|}}\right)^2 d\omega. \qquad (9)$$

where ω_l and ω_h are the cut-off frequencies of the missing band; g and g_{gmm} are the real modulation gain and the GMM-estimated modulation gain; $|A_{\mathrm{hb}}(e^{j\omega})|$ is the magnitude of the inverse filters of the highband signals of the wideband speech; $|A_{\mathrm{gmm}}(e^{j\omega})|$ is the estimated highband magnitude of response using the GMM parameters. The GMM method has brought down the SD by 0.96 dB compared with VQ codebook mapping method. The spectrograms (Fig. 8) show the reconstructed wideband signal and the boosted lowband components. Informal listening shows that the proposed wideband recovery algorithm generates substantially better and more natural speech than conventional telephone speech.

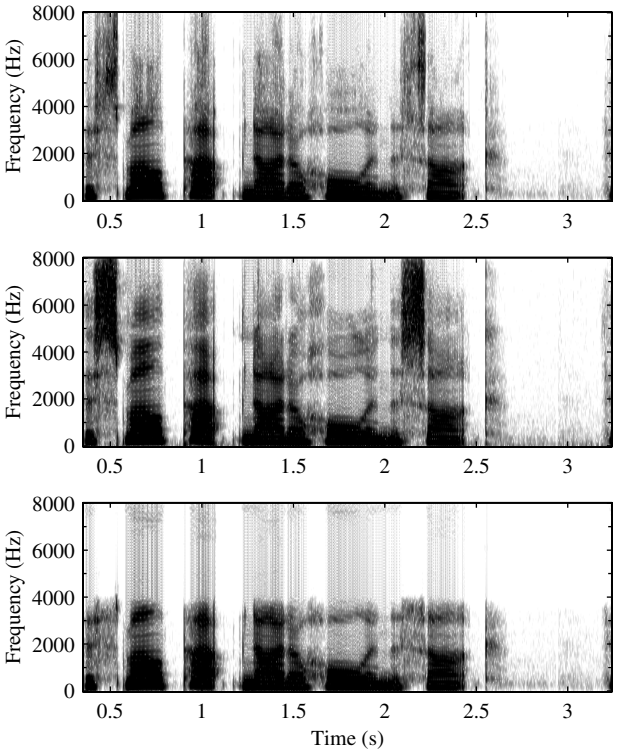

Fig. 8 The recovered spectrogram (top) compared with the original wideband (middle) and the narrowband input (bottom).

References

[1] K. Y. Park and H. S. Kim, "Wideband Conversion of Speech using GMM Based Transformation", *Proc Int. Conf. Acoustics Speech and Signal Processing,* pp. 1843–1846, 2000.

[2] D. G. Raza and C.-F. Chan, "Enhancing Quality of CELP Coded Speech via Wideband Extension by Using Voicing GMM Interpolation and HNM Re-Synthesis", *Proc.Int. Conf. Acoustics Speech and Signal Processing,* pp. I-241–I-244, 2002.

[3] M. Nilsson and W. B. Kleijn, "Avoiding Over-Estimation in Bandwidth Extension of Telephony Speech", *Proc. Int. Conf. Acoustics Speech and Signal Processing,* pp. 869–872, 2001.

[4] J. Epps and W. Holmes, "Speech Enhancement Using STC-based Bandwidth Expansion", *Proc. Int. Conf. on Speech Lang. Processing*, pp. 519–522, 1998.

[5] P. Jax and P. Vary, "An Upper Bound on the Quality of Artificial Bandwidth Extension of Narrowband Speech Signals", *Proc. Int. Conf. Acoustics Speech and Signal Processing*, pp. 237–240, 2002.

[6] Y. Qian and P. Kabal, "Wideband Speech Recovery from Narrowband Speech Using Classified Codebook Mapping", *The 9-th Australian Int. Conf. Speech Science, Technology*, pp. 106–111, 2002.

[7] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *J. Royal Statistical Soc.*, Series B, vol. 39, pp. 1–38, 1977.