Classified Highband Excitation for Bandwidth Extension of Telephony Signals

Yasheng Qian *Peter Kabal*

Department of Electrical and Computer Engineering
McGill University, Montreal, Canada H3A 2A7
yasheng@tsp.ece.mcgill.ca, kabal@ece.mcgill.ca

Abstract

Current telephone networks compromise bandwidth for efficiency. The impairment of the audio quality in telephony has become a problem for the rapidly emerging sophisticated wideband telecommunications systems. We present a classified bandwidth extension algorithm which recovers the missing highband portion of telephony signals. We describe a new highband excitation generator, a Pitch-Synchronized-BandPass-Shifted-Sum excitation for strongly harmonic signals such as some voiced phonemes or some music audio signals. For other signals, a BandPass Envelope Modulated Gaussian Noise is used as the highband excitation. The highband spectrum envelope and the excitation gain are estimated using classified Gaussian Mixture Models. Objective measurements of spectrum sections and informal subjective tests of both reconstructed telephony speech and audio signals show more highband harmonic textures for strongly harmonics signals than previous bandwidth extension methods.

1 Introduction

Current telephone networks employ a bandwidth of 300–3400 Hz. Meanwhile, the international telecommunication community has foreseen the fast deployment of wideband telecommunications networks, such as for third generation wireless systems (3GPP, 3GPP2 and MPEG) and has specified wideband speech codec standards, as SMV and AMR-WB. The wideband systems deliver signals with bandwidth of 50–7000 Hz which preserve perceptually better naturalness — "presence", better intelligibility and better speaker identity. Bandwidth extension is an alternative way to substantially improve the quality of legacy networks. Wideband signals can be, approximately, reconstructed by bandwidth extension at the network terminal receiver side.

Based on a linear prediction (LP) synthesis model, the great challenge of the bandwidth extension system is how to recreate an excitation and a spectrum envelope of the missing band (3400–7000 Hz) from a telephony signal of 300–3400 Hz bandwidth. The basic procedure is shown in schematic form in Fig. 1. The narrowband signal is first upsampled to 16 kHz to match the digital wideband signal requirement. The narrowband signal then is passed to two branches: the right one is used to generate the missing highband spectrum envelope and to estimate the highband gain. The left branch produces the missing highband excitation. The excitation signal multiplied by the estimated gain g is input to an LP synthesis filter (modelling the estimated spectrum envelope) to reconstruct the missing highband components. The reconstructed highband signals is combined with the narrowband to form a wideband signal.

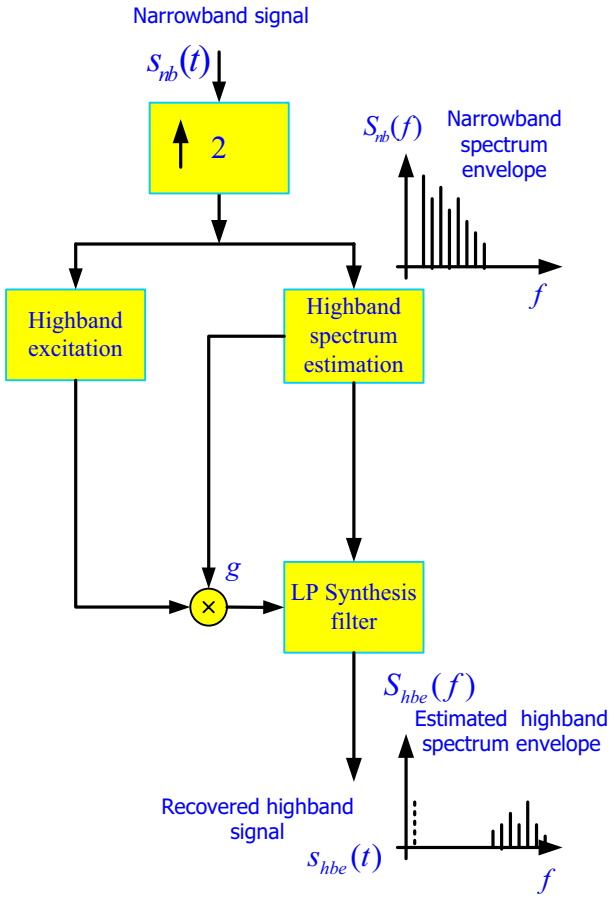

Fig. 1 The LP synthesis model of highband regeneration

Because only the narrowband signal is available at the receiver side, recovery of the lost highband components strongly depends on the correlation between the missing highband components and the narrowband signal. The higher the correlation, the better the highband reconstruction. The degree of the statistical correlation sets the boundary of the bandwidth extension methods.

A number of researchers have addressed the key issue for evaluating the spectrum envelope of the missing band (3400–7000 Hz)[1], [2], [3], [4]. We have tried to employ several methods, such as VQ codebook mapping [5], an MMSE Gaussian Mixture Model (GMM) estimator, or a Hidden Markov Model (HMM) estimator to estimate the spectrum

envelope of the missing highband components using the parameters of a lowband spectrum envelope, a pitch prediction gain and the pitch frequency F0 [6], [7], [8]. Our objective measurements of RMS-Spectral Distortion (RMS-SD) are close to previously published values, about 6 dB. An RMS-SD of 5–6 dB, probably is the bound of the state-of-the-art highband spectrum estimation algorithms. Although that value of RMS-SD is much larger than the well-known transparent criterion of 1 dB for low bit-rate narrowband speech coding, it proves to be adequate for highband estimation in bandwidth extension.

Another important issue is the generation of the missing highband excitation signal. Basically, a deterministic approach is used to create an approximate substitutes for the highband excitation. Either spectral folding of a lowband signal or its residual, as in RELP (Residual-Excitation LP) speech coders, is applied in many papers. This approach results in too high a level of high frequency components and introduces phase distortion. The excitation can also be constructed by periodic pulses for voiced phonemes or noise for unvoiced ones. That brings about more distortion as artificial sounds in reconstruction signals as in early low bit-rate vocoders.

In this paper, we focus on a new highband excitation generator, a Pitch-Synchronized-BandPass-Shifted-Sum excitation for strongly harmonic signals, such as some voiced phonemes or some music signals. For other signals, a Band-Pass Envelope Modulated Gaussian Noise is used to be the missing band excitation. The highband spectrum envelope and excitation gain are estimated using classified Gaussian Mixture Models. The classification is based on a pitch predication gain and zero-crossings. In addition, we employ two post-filters (one for frequencies below 4 kHz the other for frequencies above 4 kHz) to further enhance the quality of bandwidth extended signals.

2 The Characteristics of Highband Spectra

We have observed the characteristics of highband spectra of phonemes of female and male speech, and audio signals. We note that some voiced phonemes, particularly for female speakers, show strong harmonics in the highband portion, for instance in the 'a' in the word 'small', as shown in the top of Fig. 2. The similar features have been found in spectra of other voiced phonemes, ('en' in 'bent', 'aw' in 'gnaw' and 'i' in 'fish', etc.). Some voiced phonemes display noisy-like characteristics in the highband, as 'u' in 'pup' for a female speaker and 'i' in 'wide' of a male speaker, although they have strong harmonics in the lowband, as depicted in the second and third parts of Fig. 2. Most of male voiced phonemes do not have harmonics in the highband. Noise-like spectra of unvoiced phonemes are well known, as shown in the bottom of Fig. 2.

Music signals exhibit similar strong harmonics of some notes in the highband while the highband is noise-like for other notes, as shown in Fig. 3.

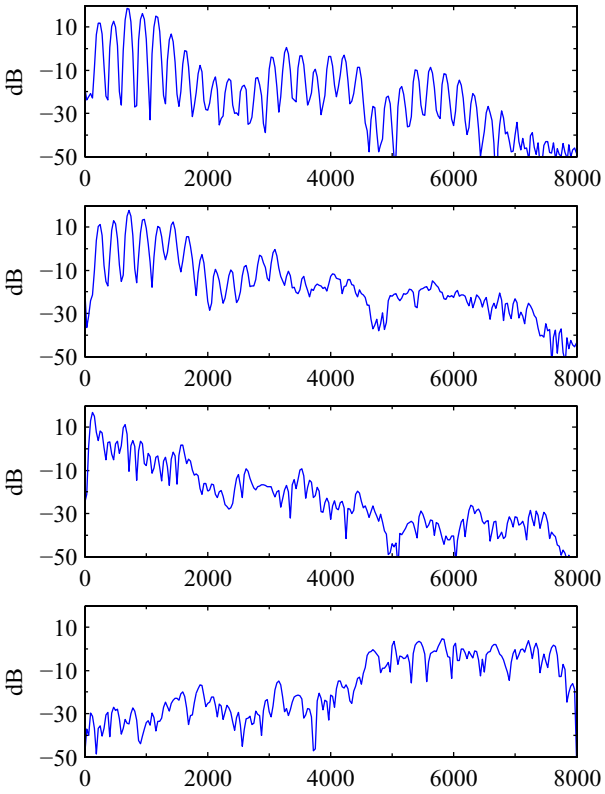

Fig. 2 The highband spectra of typical phonemes: female speaker, phoneme 'a' (top); female speaker, phoneme 'u' (second from top); male speaker, phoneme 'i' (third from top); female speaker, unvoiced 's' (bottom)

The spectrum of the reconstructed highband signal, $S_{hbe}(f)$, can be expressed as

$$S_{hbe}(f) = g\ e_{hb}(f)\, H_{hb}(f). \tag{1}$$

where $e_{hb}(f)$ is the highband excitation spectrum, $H_{hb}(f)$ is the estimated LP highband spectrum envelope, and g is the estimated gain. The harmonics structure depends on the $e_{hb}(f)$ in the highband excitation spectrum. It is not possible to generate harmonics in the highband, if there are no harmonics in $e_{hb}(f)$.

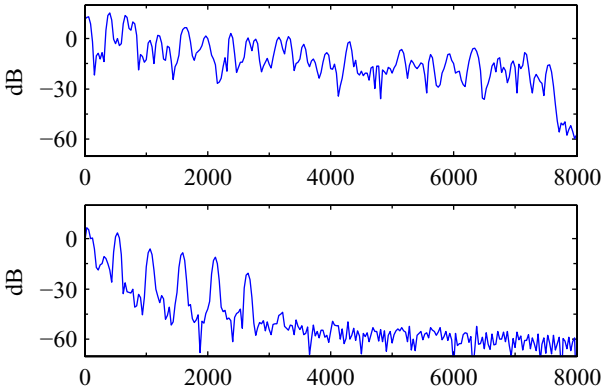

Fig. 3 The highband spectra of orchestral music: Notes with strong harmonics (top); notes without harmonics in highband. (bottom)

That motivates us to develop a Pitch-Synchronized-BandPass-Shifted-Sum (PSBPSS) excitation for creating

a strong-harmonic texture in those corresponding voiced phonemes or music notes.

3 Pitch-Synchronized-BandPass-Shifted-Sum (PSBPSS) excitation

The principle of the PSBPSS excitation is illustrated in Fig. 4. The narrowband signal is first passed into a bandpass filter of a bandwidth of 1 000 Hz. The centre frequency of the bandpass filter is 3 500 Hz. The motivation for choosing a bandpass signal (3 000–4 000 Hz) originates from the observations that the highband harmonics are close to the texture of this bandpass signal and are quite different from its lowband counterparts (300–3 000 Hz) as shown in Fig. 2. Then, three modulators shift the bandpass signal to three upper bands (4 000–5 000 Hz, 5 000–6 000 Hz and 6 000–7 000 Hz), pitch-synchronously. Each modulator is an upper band Single-SideBand modulator using a phase descrimination method [9].

$$S_{USSB}(n) = 0.5\, S_{bp}(n)\cos(\Delta\omega\, n) \;\; - 0.5\, \hat{S}_{bp}(n)\sin(\Delta\omega\, n). \tag{2}$$

where $S_{bp}(n)$ is the bandpass signal centered at 3 500 Hz, $\hat{S}_{bp}(n)$ is the Hilbert Transform of $S_{bp}(n)$. $\Delta\omega$ is the shifted frequency which is an integer multiples of the pitch frequency F0. All the consecutive bandpass signals are spaced by a pitch F0. Finally, the three bandpass signals are summed up to form a PSBPSS excitation signal in the upper band. Fig. 4

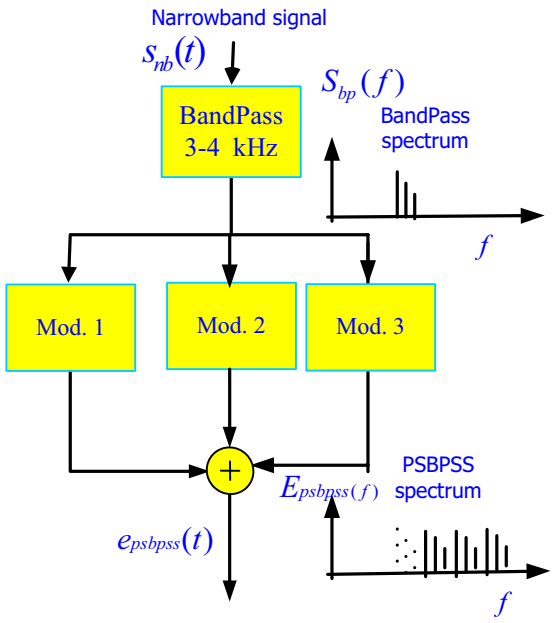

Fig. 4 The highband PSBPSS generation

also shows the schematic spectra of those bands and the PSBPSS in the highband. As an example, we have plotted the PSBPSS spectra, $E_{psbpss}(f)$ of a frame of a phoneme 'a' (female speaker) in the word of 'small' in Fig. 5.

We have employed an Enhanced BandPass Envelope Modulated Gaussian noise (EBP-MGN) in our early bandwidth extension system. Although the EBP-MGN has been proven effective for those noise-like or weak harmonic signals, there is no apparent harmonics in the excitation of the highband.

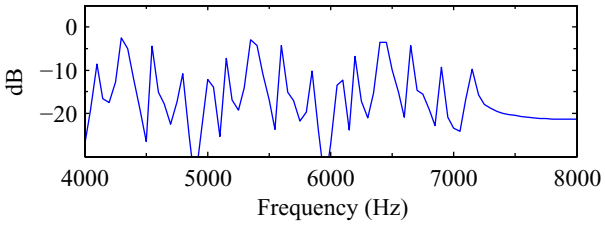

Fig. 5 The highband PSBPSS spectrum of a phoneme 'a'

Fig. 6 illustrates the highband EBP-MGN spectrum of the

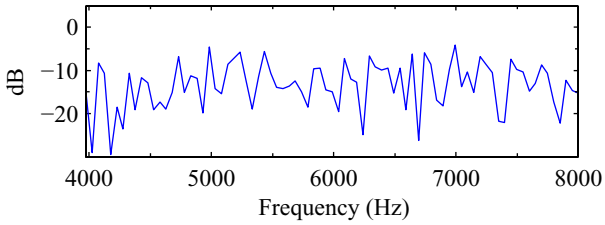

Fig. 6 The highband EBP-MGN spectrum of a female phoneme 'a'

same phoneme 'a' as in Fig. 5. Obviously, EBP-MGN is not able to reconstruct the strong-harmonic texture in the highband. Therefore, it's better to classify the signal into two modes: a strong-harmonic mode and a noise-like mode to match their desirable distinct features of excitation. The strong-harmonic mode employs PSBPSS as its excitation while the other modes use EBP-MGN for excitation.

4 The Classified gain estimation using GMMs

An excitation gain, g, is an important parameter to be estimated in the LP synthesis model in Fig. 1. The g is introduced to scale the synthesized highband components to an appropriate energy. The energy of the reconstructed highband components should be equal to the energy of the corresponding frequency band in wideband signal. The excitation gain g is calculated as the square root of the energy ratio of the original highband signal, $S_{hb}(f)$, to the synthesized one, $S_{res}(f) = E_{hb}(f) \cdot H_{hb}(f)$ in Eq. 1, of each frame.

$$g = \sqrt{\frac{||S_{hb}(f)||^2}{||S_{res}(f)||^2}}. \tag{3}$$

Since the excitation gain g depends on the excitation $E_{hb}(f)$, we train two classified GM modes of the estimated values : g_{shm} with PSBPSS excitation for a strong-harmonic signal, g_{nm} with EBP-MGN for other signals.

We derive the statistical parameters of two Gaussian mixture pdfs. Each of them is a joint pdf of the narrowband spectrum and the excitation gain parameters from the training program. We employ probabilistic estimation to get an estimated g_{shm} or g_{nm} on Minimum Mean Square Error criterion.

Because of the well-known properties (ordering and quantization error resilience) of the Line-Spectrum-Frequencies (LSF) representation, we use 14 and 10 LSFs to represent the narrowband and highband spectrum, respectively. The LSFs and the excitation gain, g_{shm} or g_{nm} , constitute a random vector, whose probability density function (pdf) can be approximated by a GM pdf.

The GM pdf is a weighted sum of M D-dimensional joint Gaussian density distributions.

$$p_Z(\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{i=1}^{M} \alpha_i b_i(\boldsymbol{z}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i). \quad (4)$$

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

$$b_i(\boldsymbol{z}|\mu_i,\boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp(\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_i)). \quad (5)$$

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.

The parameter set $\{\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ can be, iteratively, determined by the popular expectation-maximization (EM) algorithm using maximum likelihood (ML) [10] for the given training data. The training data of wideband speech are taken from Speech Database with a total of 150 000 frames each of 20 ms with 1 320 utterances, spoken by 24 speakers (half male and half female).

Z is a 15-dimensional random vector, representing the 14 narrowband LSFs, the excitation gain g_{shm} or g_{nm}. The number of mixtures, M, is 128. The covariance matrices, $\boldsymbol{\Sigma}_i$, are diagonal. The $\hat{g}_{shm}$ or $\hat{g}_{nm}$ estimate is based on the GM joint density distribution of Eq. (4). Let the random vector $\boldsymbol{x}$ be the combination vector of the narrowband LSF vector. For the sake of simplicity, we drop the the subscript $_{shm}$ and $_{nm}$ in the expression of g_{shm} or g_{nm}. The optimal estimate which minimizes the error is found from $\partial\varepsilon^2/\partial\hat{g} = 0$.

$$\hat{g}_{opt} = \frac{\int_g g\, p_{g|X}(g|\boldsymbol{x})\, dg}{\int_g p_{g|X}(g|\boldsymbol{x})\, dg} = \frac{\sum_{i=1}^{M} \alpha_i b_i(\boldsymbol{x})\mu_{ig}}{\sum_{j=1}^{M} \alpha_j b_j(\boldsymbol{x})}. \quad (6)$$

where μ_{ig} is the mean of g_{shm} or $g_{nm}$of the i-th Gaussian component. The $\hat{g}_{opt}$ estimate is the conditional expectation of the μ_{ig} mixture mean, given a narrowband LSF vector. Similarly, we have established a GMM for the narrowband LSF vector and the highband LSF vector. The highband LSF vector can be estimated with an equation similar to Eq. (6).

5 Performance evaluation

We have accomplished objective and subjective evaluation of the reconstructed signal of the classified bandwidth extension system. The classification is based on the degree of periodicity, measured in a pitch prediction gain, zero-crossings and previous state. We observed that if the pitch prediction gain is in the range of 0.95 to 1.05 and zero-crossings are less than 85 in the lowband signal, the signal probably has strong harmonics in the highband. We have plotted out a spectrum section of a bandwidth extended signal of a female phoneme 'a' in word 'small' (top) and a frame in orchestra notes (bottom) in Fig. 7. The strong-harmonics textures are well preserved in bandwidth extended speech and audio signals.

we employ a lowband post-filter and a highband one with different parameters to further enhance the quality of bandwidth extended signals. Informal listening tests shows the classified bandwidth extension scheme has noticeably improve the quality of the telephony signals.

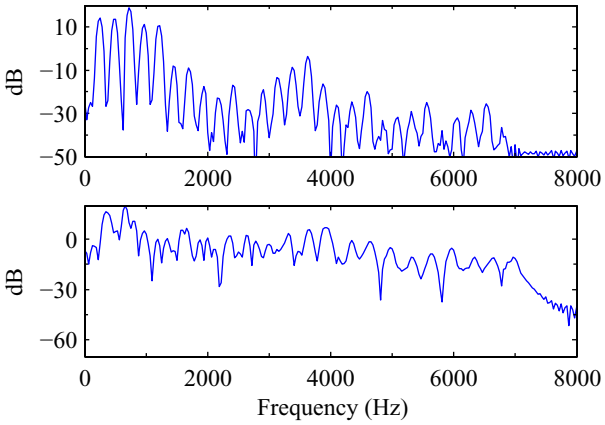

Fig. 7 The spectrum sections of a bandwidth extended signal of a female phoneme 'a' in word 'small' (top) and a frame in orchestra notes (bottom).

References

[1] P. Jax and P. Vary, "Feature Selection for Improved Bandwidth Extension of Speech Signal", *Proc Int. Conf. Acoustics, Speech, Signal Processing,* pp. I - 697–700, 2004.

[2] P. Jax and P. Vary, "On Artificial Bandwidth Extension of Telephone Speech", *Signal Processing*, vol. 83, pp. 1707–1719, Aug. 2003.

[3] B. Iser and G. Schmidt, "Neural Networks Versus Codebooks in an Application for Bandwidth Extension of Speech Signals", *Proc. European Conf. Speech, Commun. Tech.*, pp. 565–568, 2003.

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

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

[6] Y. Qian and P. Kabal, "Combining Equalization and Estimation for Bandwidth Extension of Narrowband Speech", *Proc.Int. Conf. Acoustics, Speech, Signal Processing,* pp. I - 713–716, May, 2004.

[7] Y. Qian and P. Kabal, "Highband Spectrum Envelope Estimation of Telephone Speech Using Hard/Soft-Classification", *Proc. Interspeech 2004*, pp. 2717-2720, Oct. 2004.

[8] L. Rabiner, "A Tutorial on HMM and Selected Applications in Speech recognition", *Proc. of the IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[9] S. Haykin, "Communication Sistems", *John Wiley & Sons Inc.*, Chapt.3, pp. 143-144, 1983.

[10] 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.