Proc. Interspeech 2004 (Jeju Island, Korea), pp. 2717-2720, Oct. 2004

Highband Spectrum Envelope Estimation of Telephone Speech Using Hard/Soft-Classification

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

The bandwidth for telephony is generally defined to be from 300–3400 Hz. This bandwidth restriction has a noticeable effect on speech quality. We present an algorithm which recovers the missing highband parts from telephone speech. We describe an MMSE estimator using hard/soft-classification to create the missing highband spectrum envelope. The classification is motivated by acoustic phonetics: voiced vowels and consonants, and unvoiced phonemes demonstrate different characteristic spectra. The classification also captures gender differences. A hard classification on phoneme characteristic parameters, such as a voicing degree and a pitch lag, reduces the MMSE of the highband spectrum envelope estimates. An estimator using HMM-based soft-classification can further bring down the estimated highband spectrum distortion by taking the time evolution of the spectra into consideration. Objective measures (mean log-spectrum distortion) and spectrograms confirm the improvement noted in informal subjective tests.

1 Introduction

The International Telecommunication Union (ITU) in G.712 specifies the standard frequency mask for public telephone networks to include frequencies from 300–3400 Hz. The low frequencies are cut out to reduce interference from power line induction and the high frequencies are cut to avoid aliasing in sampled systems. The naturalness, the intelligibility of some syllables and the speaker identity is compromised by the bandwidth restriction of telephone speech. To substantially improve the voice quality of the present networks, wideband speech can be, approximately, reconstructed by a wideband recovery system at the receiver side without the need for an overlay wideband (with frequencies up to 7 kHz) network.

The principle of the wideband recovery system is shown in schematic form in Fig. 1. Narrowband speech is input to the wideband speech recovery system. For our experimental work, we actually start off with wideband speech, from which we create the narrowband speech for the recovery system. This allows us to compare the actual wideband speech with the synthetic wideband speech. The narrowband speech is passed to three branches: the right one goes to spectrum envelope estimators, which generate both the highband spectrum envelope and highband gain estimates; the middle branch produces the missing highband excitation; the left branch is the narrowband speech. The excitation signal multiplied by the gain excites an LP synthesis filter (estimated spectrum envelope) to reconstruct the missing highband components. Finally, the narrowband and the recovered highband signals are combined to form a wideband speech signal.

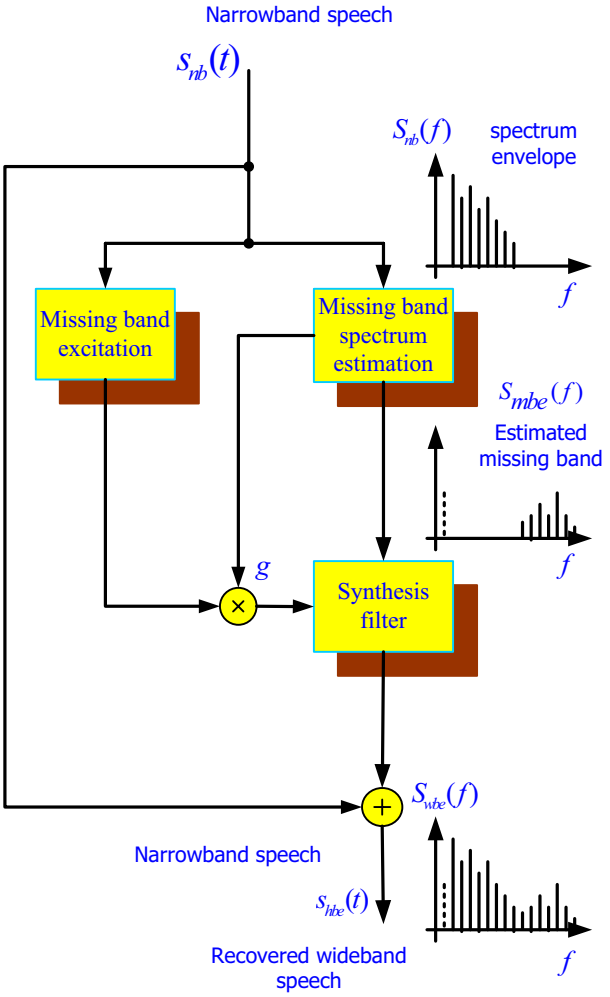

Fig. 1 Wideband speech recovery from telephone speech

In our work we generate the highband excitation using a bandpass modulated Gaussian noise (BP-MGN) [3]. We have found this excitation when combined with actual highband spectra generates a very high quality wideband signal. The main challenge is how to estimate the highband spectrum envelope just from the telephone speech.

In this work, we do not explicitly consider means to regenerate the low frequency components (but see [3]).

The spectrum envelope of the missing highband components can be reconstructed by a VQ codebook mapping or a statistical modelling approach [4],[7]. A statistical Gaussian Mixture Model (GMM) of the wideband speech spectrum parameters can also be used to estimate the missing highband spectrum envelope [1]. The feasibility of the wideband recovery scheme is based on the assumption that the missing highband components are statistically correlated, to a certain degree, with the narrowband speech. The higher the correlation, the better the highband reconstruction. We note that the highband reconstruction needs to be realistic but not necessarily the same as in the original wideband signal — the listener does not have the original wideband signal for comparison. We explore further statistical dependencies using acoustic-phonetic hard/soft classification. The goal is to reduce the spectrum distortion of the estimated highband components. To this end, we develop an MMSE estimator based on hard/soft-classification. to restore the missing highband spectrum envelope. The acoustic-phonetic hard classification reduces the MMSE of the highband spectrum envelope estimation. An estimator using HMM-based soft-classification further reduces the highband spectrum distortion.

2 Acoustic-Phonetics for Classification

Acoustic-phonetics describes distinctive waveforms, spectrum envelopes, pitch (fundamental frequency F0), pitch gains and power properties of speech sounds, or phonemes. There are 42 phonemes of American English, including 5 unvoiced fricative phonemes, such as /s/, /f/, the whisper, /h/, 2 affricatives, 4-voiced fricatives, 6 stop phonemes, 11 vowels, 6 diphthongs, 4 semivowels, and 3 nasal consonant phonemes.

Typical spectrum envelopes of a voiced vowel /o/ (solid line), a voiced explosive consonant /k/ (dotted line) and an unvoiced fricative phoneme /s/ (dashed line) are illustrated in Fig. F:phoneme.

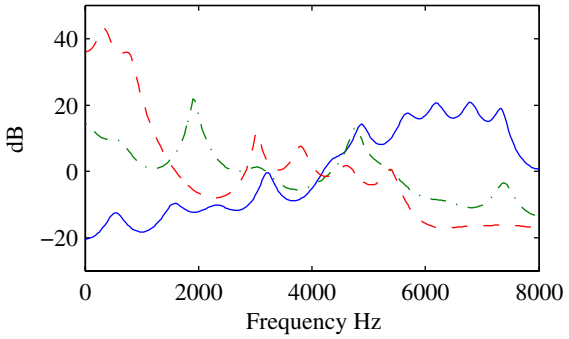

Fig. 2 Spectrum envelopes: voiced phoneme 'o' (solid line); unvoiced phoneme 's' (dashed line); voiced phoneme 'k' (dotted line)

We have classified the phonemes into 3 groups: unvoiced, voiced and mixed phoneme groups, based on their highband vs. lowband energy ratio and voicing degree β, which represents the pitch prediction filter gain [5]. In this paper we divide phonemes into three other groups by the voicing degree β and the pitch frequency. This groups correspond (roughly) to voiced female speech, voiced male speech, and unvoiced male or female speech. The pitch F0 differentiates female speakers from male speakers. An average value F0 for males is about 132 Hz while F0 for females is about 233 Hz. We have found that the statistical features of the spectrum envelopes of male and female voiced phonemes are quite different, although there is little discrepancy for unvoiced phonemes. We use the Line-Spectrum-Frequency (LSF) representation for the speech spectrum. Ten LSFs are used to represent the highband spectrum envelope. As an example, the histograms of the 8th highband LSF for 680 utterances by males (11 speakers) and 718 utterances by females (12 speakers) are depicted in Fig 3. The use of classified statistical characteristics can be beneficial in estimating the missing highband spectrum envelope, as we will be seen in the next section.

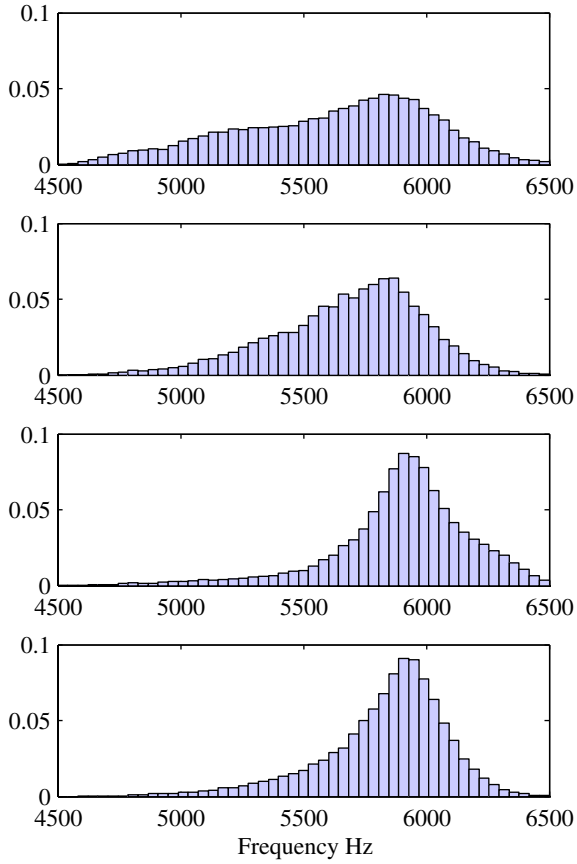

Fig. 3 The histograms of the 8th highband LSF: voiced phonemes, female talkers (top); voiced phonemes, male talkers (second); unvoiced phonemes, female talkers (third), unvoiced phonemes, male talkers (bottom).

3 The Highband Estimation Using Classified GMMs

The probability density function pdf of the LSF random vectors can be modelled by a mixture of Gaussian pdfs (a Gaussian Mixture Model, GMM). The GMM pdf is a

weighted sum of M D-dimensional joint Gaussian density distributions.

$$p_Z(\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{i=1}^{L} \alpha_i b_i(\boldsymbol{z}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i), \tag{1}$$

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

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

with mean vector $\boldsymbol{\mu}_i$, and covariance matrix $\boldsymbol{\Sigma}_i$. The GMM 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 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 [6] to determine the set of GMM parameters iteratively.

The training data of wideband speech are divided into three subsets: voiced male (25,407 frames); voiced female (35,808 frames); unvoiced male and female (82,113 frames). They come from a speech database with a total of 143,328 frames each of 20 ms. Since male pitch varies from 80–200 Hz and female pitch ranges from 133–390 Hz, we set the threshold at 160 Hz. The error probability ε_{pit} for the pitch classification, is 0.097. This is the probability of mis-classifying the gender. This value of error does not substantially affect the estimation error. Thereby, we find three sets of GMM parameters representing three pdfs: $p_{f_Z}(\boldsymbol{\alpha}_f,\boldsymbol{\mu}_f,\boldsymbol{\Sigma}_f)$ for voiced female frames; $p_{m_Z}(\boldsymbol{\alpha}_m,\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)$ for voiced male frames; $p_{uv_Z}(\boldsymbol{\alpha}_{uv},\boldsymbol{\mu}_{uv},\boldsymbol{\Sigma}_{uv})$ for unvoiced frames.

Let the random vector $\boldsymbol{x}_f$ be the vector of the narrowband LSF vector of a voiced female frame. The vector $\boldsymbol{y}_f$ is the highband LSF vector of a voiced female frame. For a given estimate, $\hat{\boldsymbol{y}}_f$, the mean-square error is

$$\varepsilon^2 = \int_{\Omega_{yf}} ||\boldsymbol{y}_f - \hat{\boldsymbol{y}}_f||^2 p f_{Y|X}(\boldsymbol{y}_f|\boldsymbol{x}_f)\, d\boldsymbol{y}_f. \tag{3}$$

The estimate which minimizes the error is found from $\partial\varepsilon^2/\partial\hat{\boldsymbol{y}}_f = 0$,

$$\hat{\boldsymbol{y}}_f = \frac{\int_{\Omega_{yf}} \boldsymbol{y}_f p f_{Y|X}(\boldsymbol{y}_f|\boldsymbol{x}_f)\, d\boldsymbol{y}_f}{\int_{\Omega_{yf}} p f_{Y|X}(\boldsymbol{y}_f|\boldsymbol{x}_f)\, d\boldsymbol{y}_f} = \frac{\sum_{i=1}^{L} \alpha_i b_i(\boldsymbol{x}_f)\boldsymbol{\mu}_{iy_f}}{\sum_{j=1}^{L} \alpha_j b_j(\boldsymbol{x}_f)}, \tag{4}$$

where $\boldsymbol{\mu}_{iy_f}$ is the mean vector of the female 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.

The estimates for the highband male and unvoiced phoneme spectrum envelopes are similar in form.

We have measured the RMS-log spectrum distortion (SD) in the missing highband (3.5–7 kHz) to evaluate our scheme. The SD is defined as

$$\mathrm{SD}^2 = \frac{1}{\pi}\int_{\omega_l}^{\omega_h} 20\log_{10}\left(\frac{\dfrac{g}{|A_{\mathrm{hb}}(e^{j\omega})|}}{\dfrac{g_{\mathrm{gmm}}}{|A_{\mathrm{gmm}}(e^{j\omega})|}}\right)^2 d\omega, \tag{5}$$

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 filter of the highband signal of the wideband speech; $|A_{\mathrm{gmm}}(e^{j\omega})|$ is the estimated highband magnitude of response using the GMM parameters.

Table 1 shows the measured mean SD and the number of outliers. We have also compared to the mean SD of highband estimation using a non-classified GMM. These are listed in the last column of the table. The mean SD improvement is about 0.6 dB for voiced frames. The 10 dB outliers for voiced speech are reduced. There is no improvement for unvoiced frames. Although the RMS-LogSD values in Table 1 are much larger than what is considered to be the threshold of transparency for narrowband spectra (1 dB), our listening tests show that a RMS-LogSD of about 6.0 dB still can deliver high quality of reconstructed wideband speech.

Table 1 RMS-LogSD in (dB)

	Voiced	Unvoiced	V/UV	Not classified
Mean	5.58	6.22	5.96	6.20
Outliers 10 dB	6.5%	13.4%	10.6%	13.1%
Outliers 15 dB	0.0%	0.0%	0.0%	1.2%

4 Soft-Classification Using HMM

We consider above-mentioned classification using the pitch and voicing degree as a hard classification. In order to further push down the estimation error, we apply soft-classification using a Hidden Markov Model (HMM) of the missing highband. An HMM is an embedded stochastic process with an underlying state Markov stochastic process that is not observable (hidden), but which can only be observed through another set of stationary process that produces the sequence of observations. We interpret the state as a phoneme group and the observable sequence as spectrum envelope parameters, LSFs, voicing degree and pitch values, etc. In our application, the true state, or the group, is not directly observable. We can only resolve the probability of the state, given an observed parameters and HMM parameters. Thus, we consider it as a soft-classification to distinguish from the hard-classification. The HMM also tracks the time-evolution of the spectral envelopes.

An HMM is determined by the five parameters: the state transition probability matrix $A = \{a_{jk}\}$, where a_{jk} stands for the state transition probability from the state $S_j(m)$ at time instant m to the state $S_k(m+1)$ at time instant $m+1$, $j,k = 1,\ldots,M$; the initial state probability vector $\pi = \pi_j$ at $m = 1$; the GMM parameter set of L matrices of each state $B = \{\boldsymbol{\alpha}_{ij},\boldsymbol{\mu}_{ij},\boldsymbol{\Sigma}_{ij}\}$, where $i = 1,\ldots,L, j = 1,\ldots,M$; the total number of states is M.

For convenience, a compact notation is introduced to indicate the complete parameter set of the HMM.

$$\lambda = \{A, B, \pi\}. \tag{6}$$

A joint prior probability or pdf of the observed parameter sequence up to time m, $\boldsymbol{X}(m) = \boldsymbol{x}(1),\ldots,\boldsymbol{x}(m)$, which ends up at j-state, $S_j(m)$, $\gamma_j(m) = p(S_j(m),\boldsymbol{X}(m))$, is introduced to calculate the

conditional pdf $p(S_j(m)|\boldsymbol{X}(m))$. The joint prior pdf, $\gamma_j(m)$ at time m, a classical problem of HMM, can be recursively solved as

$$p(S_j(m), \boldsymbol{X}(m)) = \sum_{k=1}^{M} \gamma_j(m-1) a_{kj} p(\boldsymbol{x}(m)|S_j(m)). \quad (7)$$

Eq. (7) can be verified by the definition of the conditional probability and the fundamental assumption of independence between the successive observations. We can recursively calculate the joint prior pdf $\gamma_j(m)$ with the initial condition $\gamma_j(1) = \pi_j p(\boldsymbol{x}(1)|S_j(1)$ by Eq. (7). The estimated highband spectrum envelope $\hat{\boldsymbol{y}}$ can be determined by a similar equation as Eq. (4) based on the MMSE estimation criterion. Notice that the conditional vector $\boldsymbol{x}$ is replaced by the past observation sequence $\boldsymbol{X}$. We drop the time index m for simplicity,

$$\hat{\boldsymbol{y}} = \int_{\Omega_y} \boldsymbol{y} p_{Y|X}(\boldsymbol{y}|\boldsymbol{X})\, d\boldsymbol{y} \quad (8)$$

$$p_{Y|X}(\boldsymbol{y}|\boldsymbol{X}) = \sum_{j=1}^{M} p_{Y|X}(\boldsymbol{y}|S_j) p(S_j|\boldsymbol{X}), \quad (9)$$

where $p(S_j|\boldsymbol{X}) = p(S_j, \boldsymbol{X}) / \sum_{j=1}^{M} p(S_j, \boldsymbol{X})$. Since $\int_{\Omega_y} \boldsymbol{y} p(\boldsymbol{y}|Sj) d\boldsymbol{y} = \boldsymbol{C}_j$ is the VQ codevector of the j-state, the estimated highband envelope

$$\hat{\boldsymbol{y}}(m) = \frac{\sum_{j=1}^{M} \boldsymbol{C}_j \gamma_j(m)}{\sum_{j=1}^{M} \gamma_j(m)}. \quad (10)$$

The estimate in Eq. (10) depends on the joint prior pdf $\gamma_j(m)$ and VQ code vectors of all states, given an observation sequence and HMM.

We have trained the HMM parameters λ using the Baum-Welch ML iterative algorithm [9] and the VQ codebook by the well-known General Max-LLoyd method [10]. A small VQ codebook was used in the experiments, $M = 8$.

We used the Eq. (10) to estimate highband spectrum envelope and compared the results to hard-classified estimation. The spectrograms (Fig. 4) show that hard- and soft-classification give reasonbable highband components. The soft-classification gives a slightly richer highband. Listening shows that the highband spectrum estimation using both the hard and soft classification gives wideband speech which is much preferred to the original narrowband speech.

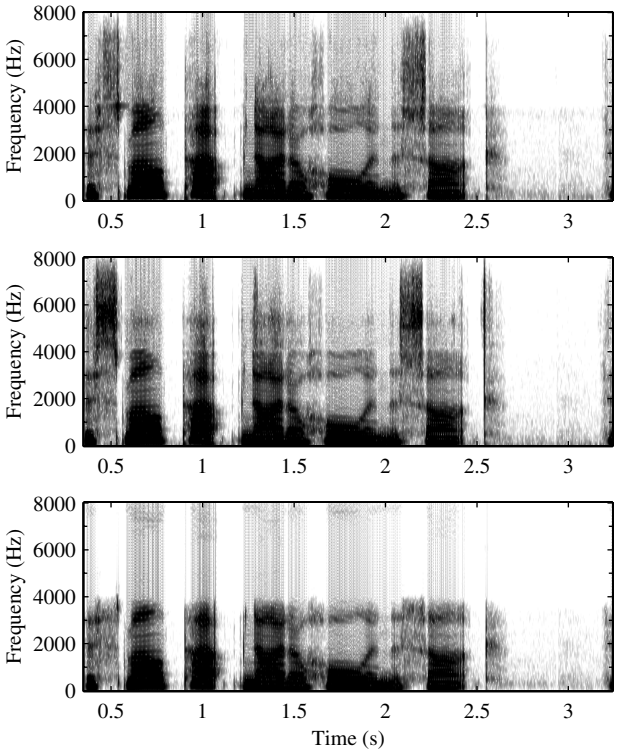

Fig. 4 The recovered hard-classified spectrogram (top) compared with the soft-classified spectrogram (middle) and the narrowband input (bottom).

References

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

[2] 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.

[3] Y. Qian and P. Kabal, "Dual-Mode Wideband Speech Recovery from Narrowband Speech", *Proc. European Conf. Speech, Commun. Tech.,* pp. 1433–1437, Sept. 2003.

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

[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] 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.

[7] 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.

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

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

[10] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for Vector Quantizer Design", *IEEE Trans. Commun.*, vol. 28, pp. 84-95, Jan. 1980.