ELectronics EXpress

IEICE Electronics Express, Vol.8, No.14, 1143–1148

Linear-scale perceptual feature extraction for Speech Bandwidth Extensions

Kuekjae Lee[a)], **Sang Bae Chon, Mingu Lee, and Koeng-Mo Sung**

Applied Acoustics Lab., Institute of New Media and Communications,
Department of Electrical Engineering, Seoul National University
San 56–1, Sillim-dong, Kwanak-gu, Seoul 151–742, Republic of Korea
a) *soulphoenix@acoustics.snu.ac.kr*

Abstract: This paper presents a new method to extract linear-scale perceptual feature as a subsitute of MFCCs for highband (3.4 kHz∼) in Speech Bandwidth Extensions(BWE). The feature extraction method is based on the mel-scale constrained Nonnegative Matrix Factorization(NMF), which decompose linear-scale log spectrum into a linear combination of mel-scale latent variables. While MFCCs parametrization contains non-invertible procedures, suggested feature is represented in linear-scale and proper to recover the highband time-domain speech. Experiment results report that suggested feature shows better instrumental performance with narrowband MFCCs than real cepstrum without additional computation.

Keywords: BWE, NMF, MFCCs
Classification: Science and engineering for electronics

References

[1] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, Aug. 2003.
[2] D. D. Lee and S. H. Seung, "Algorithms for non-negative matrix factorization," *NIPS*, vol. 13, pp. 556–562, 2001.
[3] Nour-Eldin, Amr H and Kabal and Peter "Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech," *Interspeech-2008*, pp. 53–56.
[4] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski, "Speech reconstruction from mel frequency cepstral coefficients and pitch frequency," *Proc. Int. Conf. Acoust. Speech and Signal Processing (ICASSP)*, pp. 1299–1302, 2000.
[5] P. Jax, "*Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds*," phD thesis, Aachen University (RWTH), Aachen, Germany, 2002.
[6] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM-based transformation," *Proc. ICASSP*, Istanbul, Turkey, vol. 3, pp. 1847–1850, June 2000.

DOI: 10.1587/elex.8.1143
Received February 06, 2010
Accepted March 23, 2011
Published July 25, 2011

1 Introduction

Speech signal over the telephone system has own characteristic sound due to the limited bandwidth (0.3∼3.4 kHz) of communication system. Bandwidth Extension(BWE) is to enhance the speech quality by recovering highband signal (3.4 kHz∼). Commonly BWE algorithms are based on source-filter model, where narrowband speech signal is separated into a source signal and filter coefficients. And their counterparts of highband are estimated from them respectively. Since human hearing is insensitive to phase information, the reconstruction method of source signal is insignificant. So most researchs on BWE are focused on estimation of filter coefficients which contains envelope information.

To reconstruct highband envelope, it is important to find a set of features which shows high relevance between narrowband envelope and highband envelope. Cepstral coefficients are common choices for features because they represent log spectral information and pitch information is easily removed by truncating higher order coefficients. Mel-Frequency Cepstral Coefficients(MFCCs),which approximately convert linear-scale log spectra into mel-scale log spectra, are more frequently used features in speech processing. MFCCs are reported to show more relevance between narrowband feature and highband feature [1].

In the following section (Section 2), we describe one of the conventional MFCCs parametrization and their limited uses for highband. And some trials using MFCCs for highband are briefly introduced. In Section 3, we suggested new perceptual highband feature extraction method to overcome the limitation of MFCCs. Section 4 presents experiments and results to validate our suggested feature extraction method.

2 Conventional MFCCs parametrization and its limitation

One of MFCCs parametrizations [3] is summarized as follows

1. *Pre-emphasis* : A single-pole (at $z = -0.97$) high-pass filter is used to emphasize the highband formants
2. *Windowing* : 50 % overlapped hamming window
3. *Linear-scale power spectra* : FFT is applied followed by a magnitude operation.
4. *Mel-scale filterbank binning* : Mel-scale triangular filters are applied to the magnitude spectrum with normalized mel-scaled filter energies.
5. *Log operation* : logarithm is taken to the enegies.
6. *IFFT* : IFFT of the log-energies is applied

After MFCCs parametrization, higher order cepstral coefficients are truncated to remove pitch information. 0th coefficient, which represents the gain

DOI: 10.1587/elex.8.1143
Received February 06, 2010
Accepted March 23, 2011
Published July 25, 2011

of log spectrum, is typically removed since it is environmentally sensitive variable.

Mel-scale filterbank binning in procedure 4 is conducted by non-invertible matrix. Therefore it is difficult to synthesize highband speech from MFCCs. There are some trials to use MFCCs for highband by using high-resolution IDCT or nonnegative least square method. But high-resolution IDCT do not recover linear-scale log spectra but interpolates finer cepstral coefficients from mel-scale log spectra [3]. And nonnegative least square method requires pre-designed basis functions and iterative computation [4].

3 Linear-scale perceptual highband feature extraction

Superiority of MFCCs is mainly originated from redundant information in linear-scale log spectra being reduced by finding smaller dimensional mel-scale latent variables. While mel-scale filterbank binning in Fig. 1 (a) is obtained in a perceptual view, it is possible to get perceptual features in a generative view. Its advantage over Mel-scale filterbank binning is that obtained power spectra is not mel-scale but linear-scale which is appropriate to synthesize highband speech signal. It is described in Fig. 1 (b).

Linear-scale power spectra S_y, which is computed by FFT and magnitude operation, is first decomposed into a linear combination of latent variables.

$$S_y(m,n) = \sum_l W(m,l)H(l,n) + N(m,n) \tag{1}$$

where m, n and l represent linear-scale frequency index, mel-scale frequency index and time index.

In the assumption that noise term *N(m,n)* contains useless information, decomposed factors are re-producted to obtain noise-removed linear-scale power spectra S'_y.

$$S'_y(m,n) = \sum_l W(m,l)H(l,n) \tag{2}$$

Since values of power spectra are non-negative, this procedure is realized by non-negative matrix factorization. With selecting the Kullback-Leibler divergence as a cost function [2],

$$D(S_y|WH) = \sum_{mn} \left([S_y]_{mn} \log \frac{[S_y]_{mn}}{[WH]_{mn}} - [S_y]_{mn} + [WH]_{mn} \right) \tag{3}$$

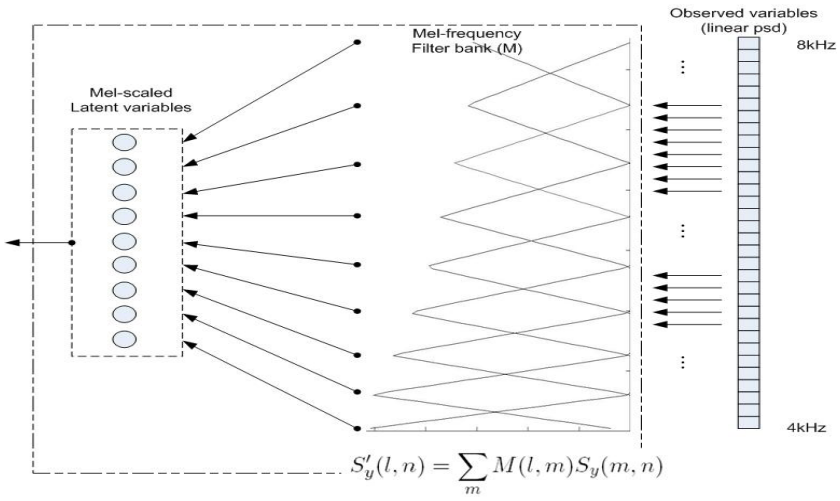
Multiplicative update rules to train basis matrix W and sets of coefficients H from the set of observed variable S_y are followed as

$$[H]_{ln} \longleftarrow [H]_{ln} \cdot \frac{\sum_m [W]_{ml} \left[[S_y]_{mn} / [WH]_{mn} \right]}{[W^T U]_{ln}} \tag{4}$$

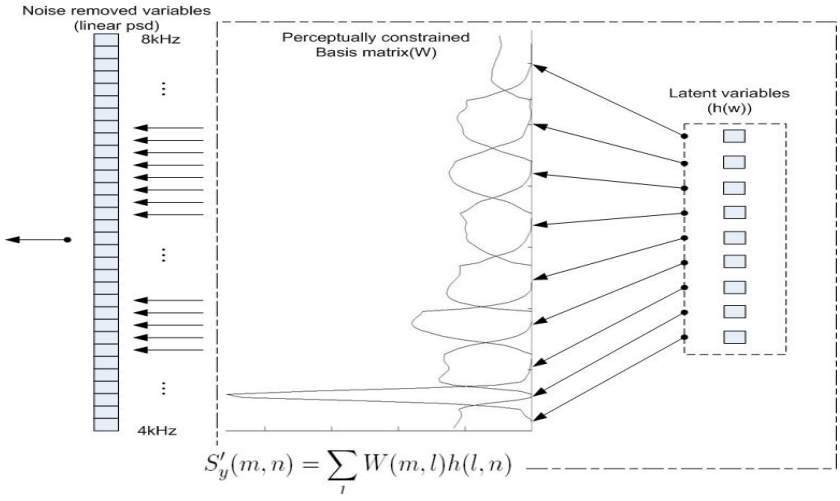
$$[W]_{ml} \longleftarrow [W]_{ml} \cdot \frac{\sum_n \left[[S_y]_{mn} / [WH]_{mn} \right] [H]_{ln}}{[UH^T]_{ml}} \tag{5}$$

$$\sum_m [W]_{ml} = 1 \tag{6}$$

DOI: 10.1587/elex.8.1143
Received February 06, 2010
Accepted March 23, 2011
Published July 25, 2011



(a)



(b)

Fig. 1. Diagrams of (a) mel-scale filterbank binning and (b) linear-scale perceptual feature extraction

To find perceptually meaningful latent variables, mel-scale constraint of basis matrix W is imposed. Since zero values cannot be updated to positive values and vise versa in multiplicative update rules, the constraint is easily imposed by initializing W as

$$W_{initial}(m,l) = \begin{cases} \text{positive value} & \text{for } f_c(l-1) \le f(m) < f_c(l+1) \\ 0 & \text{elsewhere.} \end{cases} \quad (7)$$

where f_c is a center frequency of mel-scale filterbank.

With above initial condition of W, H and W are alternately updated by Eq. (4) and Eq. (5) until it converges. Since obtained basis matrix W is data-dependent, their shapes are irregular as shown in Fig. 1 (b). Once noise-removed power spectra is obtained from Eq. (2), We can get cepstral coefficients, which enables us to sythesize highband speech.

DOI: 10.1587/elex.8.1143
Received February 06, 2010
Accepted March 23, 2011
Published July 25, 2011

4 Experiments and results

For convienience of phase reconstruction, narrowband speech signals (0∼4 kHz) are extended to wideband speech signals (∼8 kHz). MFCCs are extracted by MFCCs parametrization in Section 2. Mel-scale filterbank is designed with 31 outputs for 0∼8 kHz wideband. For narrowband, 23 mel-scale power spectra are converted into truncated 9th order MFCCs. In our experiments, 0th coefficient is replaced by prediction error.

$$\rho = \frac{1}{2}\log\frac{\sigma_{s_n}^2}{\sigma_e^2} \tag{8}$$

where $\sigma_{s_n}^2$ and σ_e^2 are variances of narrowband speech and linear prediction error signal respectively. It is added as a complementary clue for voicedness/unvoicedness classfication and related to spectral flatness on the Gaussian assumption. 6th order real cepstrum is used for highband feature. One is obtained from NMF with initialization of 9 dimensional mel-scale filterbank and other is obtained by conventional method is extracted for comparison.

Because the purpose of experiments is to check the performance of linear-scale perceptual feature extraction method, we follows conventional BWE system which is designed by Peter Jax et al. [1]. 128 samples with 50% overlapped narrowband signal is processed by algorithm described in Fig. 2. Source signal is extended by spectral translation of narrowband source [1]. And envelope information is estimated by Gaussian Mixture Model(GMM) [6].

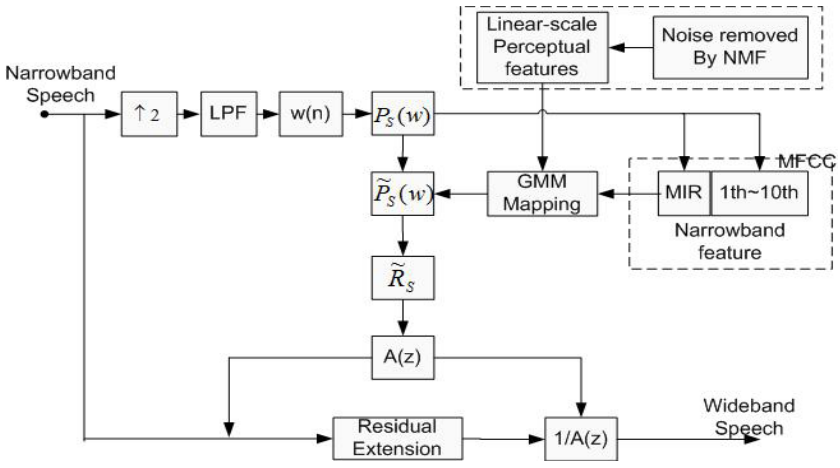


Fig. 2. BWE system

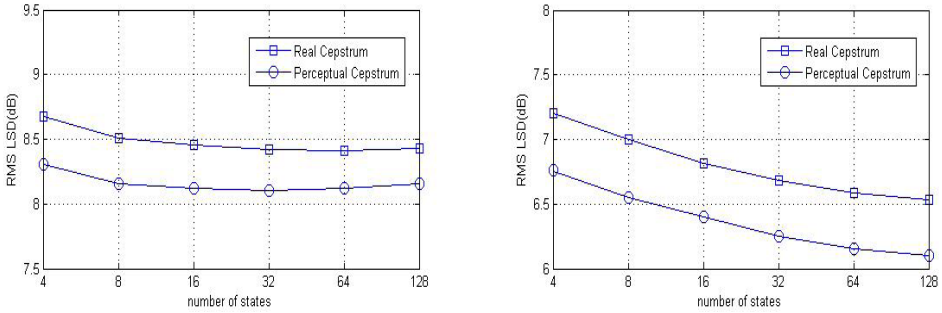
11 sentences(about 45 seconds each) spoken by 25 different speakers are used for training and experiment. The speaker-dependent models were trained individually using only training data from a single speaker. And the speaker-independent models were trained using all training data excluding a single speaker. The instrumental performance evaluation was performed in terms of the *root mean square*(RMS) *log spectral distortion*(LSD) of the estimated spectral envelope with in the highband. This sub-band spectral distortion measure can be determined by calculating the mean square esti-

DOI: 10.1587/elex.8.1143
Received February 06, 2010
Accepted March 23, 2011
Published July 25, 2011

mation error.

$$d_{LSD} \approx \frac{\sqrt{2}10}{\ln 10}\sqrt{\sum_{i=0}^{n-1}(y_i - \tilde{y}_i)^2} \tag{9}$$

where y_i is ith cepstral coefficient of highband.



(a) (b)

Fig. 3. Mean log spectral distortion of the estimated spectral envelope of (a) speaker-independent model and (b) speaker-dependent model

The results are illustrated in Fig. 3. For both speaker-independent and speaker-dependent modeling, new perceptual feature extracted by suggested method yield a consistent imporvement in comparision to real cepstrum. Since there exists loosely theoretical lower bound of achievable RMS log spectral distortion in dependence of the mutual information and differential entropy [5], 0.4∼0.5 dB improvement in performance without additional computation is meaninful.

5 Conclusion

In this paper, a new linear-scale perceptual highband feature extraction method is suggested. Unlike MFCCs, it is based on a generative view so that invertible linear-scale power spectra is obtained. It shows more relevance to narrowband MFCCs than real cepstrum, whose performance is confirmed by instrumental measurements. Since it requires no additional computation, it can be used in highband feature extraction to improve performance of BWE.

DOI: 10.1587/elex.8.1143
Received February 06, 2010
Accepted March 23, 2011
Published July 25, 2011