BLOCK-BASED BANDWIDTH EXTENSION OF NARROWBAND SPEECH SIGNAL BY USING CDHMM

Sheng Yao and Cheung-Fat Chan

Department of Computer Engineering and Information Technology
City University of Hong Kong, Kowloon, Hong Kong
Sheng.Yao@student.cityu.edu.hk and itcfchan@cityu.edu.hk

ABSTRACT

In this paper we present a block-based bandwidth extension system to enhance the quality of narrowband speech signal (0-4kHz). In memoryless bandwidth extension systems, the missing high-band components are estimated from narrowband speech using the current frame only. As the narrowband-to-wideband mapping is a one-to-many problem, this memoryless system is likely to cause hissing and whistling artifacts in the reproduced speech. Our method estimates high-band components via narrowband-to-wideband state sequence mapping using continuous density hidden Markov model (CDHMM) on a block basis. The speech block is either one word or a sequence of words in narrowband utterance. CDHMM estimation method avoids the one-to-many property of low-band and high-band dependency. Both subjective and objective evaluations show that hissing and whistling artifacts are reduced and the spectrally extended wideband speech (0-8kHz) is pleasant to listen.

1. INTRODUCTION

Most of current speech transmission systems have bandwidth limit from 0.3kHz to 3.4kHz. The major degradation of narrowband speech, compared with wideband speech (0-8kHz), is its muffing effect. Speech sounds with important energy distribution beyond 3kHz, such as fricatives (/s/ /t/ /f/) and stops (/p/ /t/ /k/) are seriously degraded. Extra listening effort is required to distinguish those sounds. Although there is gradual growth of wideband voice terminal in industry, during the long transitional period, bandwidth extension system (BWE) would still exist because it is able to enhance speech quality without any modification of current infrastructure.

For all the BWE in literature, the focus is on the estimation of spectral envelope in high-band region. Various methods on high-band envelope recovery are reported, namely VQ codebook searching [3,4], linear mapping [5], GMM transformation [6,7], HMM-based VQ [8,9] and other statistical models. Regardless of the method used, hissing and whistling artifacts are a common problem. In [1,2], it is shown that the low-band and high-band relationship is a one-to-many mapping. It implies that a method obtaining high-band envelope from narrowband would probably result in perceivable spectral error in high-band, and successive frames of errors will result in hissing and whistling artifacts. Such effect is even more severe for memoryless estimation methods.

In this paper, we propose an extension of GMM method by using CDHMM estimation on a block-by-block basis. Harmonic plus noise model (HNM) [10] is employed to reproduce speech. Compared with LPC-synthesizer, which is commonly used by other BWE systems, MBE-synthesizer of HNM can produce more natural sounds. Moreover, in LPC model of speech, it is necessary to estimate high-band portion of excitation signal as well as spectral envelope. Inaccurate estimation would cause some other artifacts.

This paper is organized as follows. In section 2, the proposed BWE system is described, including enhancement structure and model training. Simulation and performance comparison with VQ and GMM methods are shown in section 3. Section 4 is for conclusion.

2. BANDWIDTH EXTENSION SYSTEM

2.1. Enhancement structure

It is well known that human speech generation is a non-stationary random process, which can be modeled by hidden Markov process. When people speak, the underlying statistical property changes which is reflected by changes in Markov states. After speech is bandwidth-limited, it can be considered as another hidden Markov process with different statistical properties. However, if HMM is defined to be left-to-right model, Markov states of narrowband speech and wideband speech have correlation even though their output pdfs of the corresponding states are different. Our method explores this correlation by mapping narrowband and wideband optimal state sequences.

The proposed enhancement system is shown in Figure 1. Narrowband speech is sampled at 8kHz. Speech frame is fed to HNM analyzer and features such as pitch, gain, voiced/unvoiced decision and spectral envelope represented by 10-order LSF coefficients, are extracted. The first three features are passed directly to MBE-synthesizer while narrowband LSF is used for wideband envelope recovery. Speech frame stream is then labeled. Specifically speaking, wherever a word is detected, the corresponding frame sequence is labeled with starting and ending points marked. Therefore, frames outside blocks are either noise-like or silent frames. Frame periodicity, zero crossing rate and normalized frame energy are used for this end-

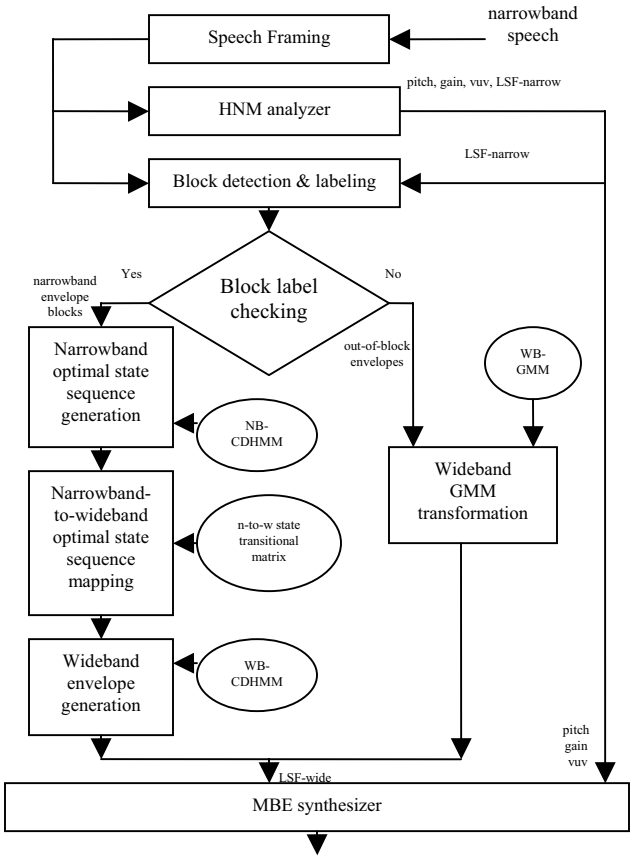

Figure 1. Flowchart of enhancement system

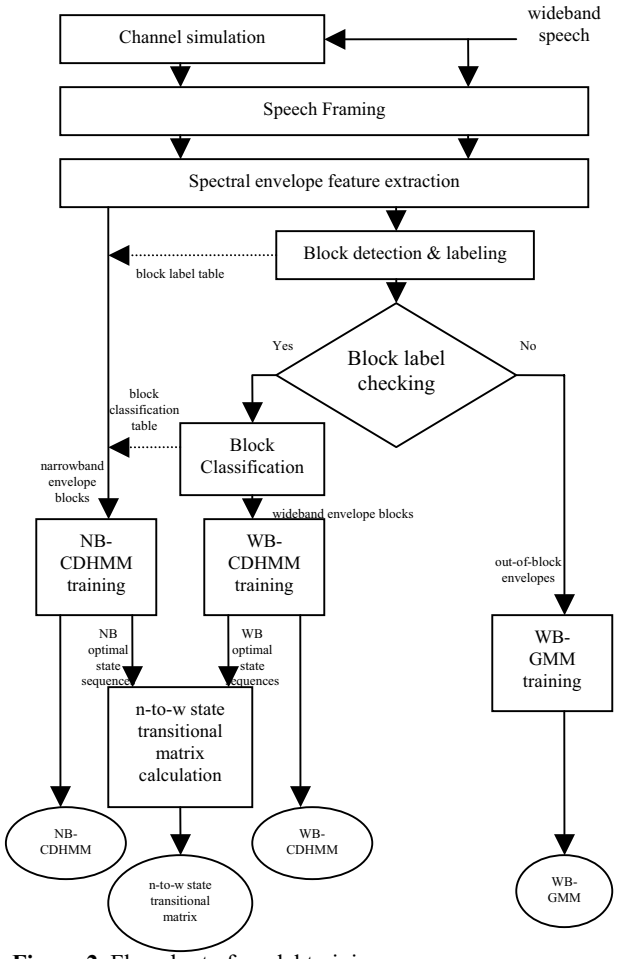

Figure 2. Flowchart of model training

point detection.

Let us consider an arbitrary block of narrowband envelopes to be an observation of CDHMM with Gaussian mixture density. Narrowband optimal state sequence is then obtained via Viterbi algorithm by using pre-trained narrowband CDHMM. It is then mapped to wideband optimal state sequence via dynamic programming by using pre-trained narrowband-to-wideband (n-to-w) state transitional matrix. Details are explained in section 2.3. Estimated wideband optimal state sequence, together with narrowband envelope block and pre-trained wideband CDHMM, generates wideband envelope block using the method described in section 2.4. Finally a MBE-synthesizer is used to reproduce wideband speech from the pitch, gain, vuv and estimated wideband envelope parameters.

2.2. Model training

Figure 2 is the flowchart of model training. As mentioned in previous section, two types of CDHMMs are employed, one for narrowband speech and the other for wideband speech. Wideband training speech (0-8kHz) firstly goes through transmission simulation process, which consists of a lowpass filter with cutoff frequency at 4kHz, HNM encoder and HNM decoder. LPC-MFCC and LSF are extracted from each speech frame to represent spectral envelope. LPC-MFCC is used as feature vector of CDHMM while the corresponding LSF parameters are used for speech synthesis. Wideband envelope frame stream is labeled as blocks using the process described in section 2.1. Note that the same labeling is applied to the corresponding narrowband envelope stream.

The detected envelope block reflects formant evolution and spectral energy distribution when certain words are pronounced. Since the smallest speech block unit is word, the number of envelope stream patterns is numerous and a single CDHMM is far from sufficient to model this diversity. Therefore, envelope blocks have to be classified into several clusters first. Clustering method used is binary splitting LBG algorithm for matrix quantization. Note that LBG clustering is applied to wideband envelope blocks only and narrowband blocks follow the same classification. A pair of CDHMMs is trained via EM algorithm; one for the wideband cluster and one for the narrowband clusters. The two CDHMMs are both left-to-right structured with the same number of states and Gaussian mixtures. Optimal state sequences for all wideband and narrowband blocks can be obtained via Viterbi algorithm from the trained wideband CDHMM and narrowband CDHMM, respectively. Since a wideband block is associated with a corresponding narrowband block (the same labeling and clustering guarantee the association), wideband and narrowband optimal state sequences also appear in pairs.

Under the constraint of left-to-right structure, wideband and narrowband states have correlation. It is well known that

successive reappearance of a state indicates that output pdf would remain stable during this period of time and state jump implies change in pdf. Although output pdfs of wideband CDHMM and narrowband CDHMM are generally different, it is not the case for state evolution. For example, when narrowband observation is believed to be within a steady state, the corresponding wideband state sequence probably resides in a certain state too. If there is a state jump in narrowband observation at some time instant, a state jump is most likely to happen in wideband observation at some neighboring time instant (including the same time instant). In order to estimate wideband optimal state sequence, we have to know which narrowband state is connected to which wideband state and how frequent they are connected. This correlation can be measured by counting the number of n-to-w state transitions in optimal state sequence pairs. The measurement is represented in square matrix form. If c_{ij} is the number of occurrence of frames when narrowband state is *i* and wideband state is *j*, matrix entry a_{ij} is the probability of mapping from narrowband state *i* to wideband state *j*, which is calculated as follows:

$$a_{ij} = \frac{c_{ij}}{\sum_k c_{ik}}$$

where *k* goes through all Markov states. The trained matrix usually has large probability values in diagonal and off-diagonal entries. This implies that a state in narrowband model is related more closely to nearby states in wideband model. If there are *M* clusters of CDHMM pairs, there are *M* n-to-w state transitional matrixes.

2.3. Narrowband-to-wideband optimal state sequence mapping

Given narrowband optimal state sequence and n-to-w state transitional matrix, most likely wideband optimal state sequence can be found through dynamic programming as indicated in figure 3. Specifically, since the cost of optimal path in DP grid is the product of n-to-w transitional probabilities of all nodes along the path, this DP problem is known as node-typed "eastbound salesman problem", which can be solved under Bellman optimality principle. Since wideband CDHMM is also left-to-right, additional constraint on wideband optimal state sequence is $s_i \le s_{i+1}$ for any frame index or time instant *i*.

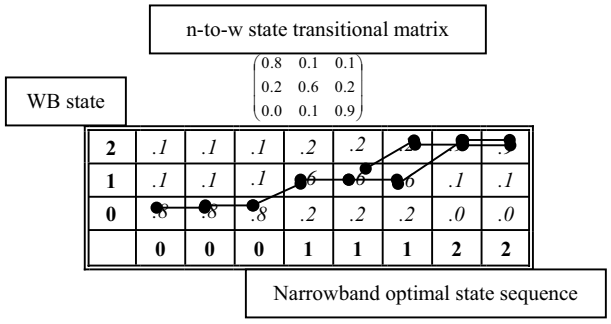

Figure 3. Illustration of a simple DP grid when narrowband optimal state sequence is "0 0 0 1 1 1 2 2". The marked paths are two most likely wideband optimal state sequences.

Conventional GMM transformation method uses either a single wideband GMM or two wideband GMMs (one for voiced frame and the other for un-voiced frame) to estimate wideband envelope. In our CDHMM approach, each wideband state has its own distinct GMM. The choice of appropriate GMM depends on which wideband state current narrowband frame is mapped to. Besides, there are many wideband CDHMMs for all kinds of block clusters, so the choice of GMM depends on the classification of current envelope block as well. In order to explore one-to-many low-high property, *N* most likely wideband optimal state sequences are found and weighted by normalizing their path costs. Multiple optimal paths can be solved by increasing the dimension of memory buffer of each node in DP grid, storing *N* most probable predecessors instead of a single one. Therefore our method is an extension of conventional GMM.

2.4. Wideband envelope generation

With input narrowband envelope block and an estimated wideband optimal state sequence, wideband envelope block for this wideband optimal state sequence can be calculated. Denote x_i to be the feature vector of envelope at *i*'th frame within a block, s_i to be the current wideband state for the wideband optimal state sequence, and $\{\pi_m, \mu_m, \Sigma_m\}_i$ to be the parameter set of output pdf for s_i. And y_i is feature vector of the current wideband envelope, which is calculated as follows:

$$y_i = \sum_m \omega_{i,m} \mu_{i,m}$$

$$\omega_{i,m} = \frac{\pi_{i,m} g(x_i, \mu_{i,m}, \Sigma_{i,m})}{\sum_m \pi_{i,m} g(x_i, \mu_{i,m}, \Sigma_{i,m})}$$

where *m* is mixture index, *g*() is Gaussian distribution density function. When calculation goes through all *i* within a block, we get the estimated wideband envelope block for that wideband optimal state sequence. If there are *N* wideband optimal state sequences picked out of DP grid, final wideband envelope block is interpolated by *N* estimated wideband envelope blocks with interpolating coefficient to be the normalized weight of each wideband optimal state sequence.

2.5. Wideband speech synthesis

Firstly, original narrowband envelope is copied to the low-band portion of estimated wideband envelope. At the same time, energy level of high-band portion is adjusted to match the low-band-to-high-band energy ratio of estimated wideband envelope. This procedure guarantees that the same low-band information is maintained. Then wideband speech is generated with the technique described in [10]. Magnitudes of all harmonic bands are derived from wideband envelope and frame gain. Voicing level of each band is calculated with V/UV mixture curve, which is obtained also from wideband envelope. Finally sinusoidal oscillator of each band generates harmonic signal with its harmonic magnitude scaled by voicing level. White noise is fed to a time-varying lattice filter whose coefficients are derived from V/UV-scaled wideband envelope to produce noise portion of speech. Synthesized speech is generated by adding these two signals together.

3. SIMULATION

Speech data is from phonetically balanced IViE corpus (Intonational Variation in English). The corpus consists of about 36 hours of 16 kHz-sampled speech spoken by 56 males and 56 females from 9 different cities in UK. Around 90% is used for training and the rest is for evaluation. Evaluation data is lowpassed first and then processed through narrowband codec (encoded and decoded) before use.

Detected wideband envelope blocks are partitioned into 768 clusters for CDHMM training. All 768 wideband and 768 narrowband CDHMMs have 8 states, 8 Gaussian mixtures and upper triangular transitional matrix. The n-to-w state transitional matrix is thus 8-by-8. Both wideband and narrowband output GMM pdfs take 32-order LPC-MFCC as feature vector and have diagonal co-variance matrix. During EM learning of wideband CDHMM, the centriods of the grouped 18-order LSF vectors are iteratively updated accordingly. The trained LSF vectors will be used for wideband envelope generation and speech synthesis. 10 most likely wideband optimal state sequences are picked out from DP grid for generating the wideband spectral envelope. As for speech frames outside blocks, a wideband GMM with 256 mixtures is trained to expand envelope in a memoryless manner.

The VQ [3] and conventional GMM [6] methods for bandwidth expansion are implemented for comparison purposes. The feature vector parameters of the two methods are also 32-order LPC-MFCC calculated from 18-order LSF. VQ codebook size for voiced frame is 1024 and 512 for unvoiced frame. Mixture number for voiced GMM and unvoiced GMM are both 256. Objective measurement is high-band spectral distortion *D*:

$$D = \left(\frac{2}{\pi} \int_{\pi/2}^{\pi} \left(10 \log_{10} S_{org}(\omega) - 10 \log_{10} S_{ext}(\omega)\right)^2 d\omega \right)^{\frac{1}{2}}$$

Table 1 shows the objective performance of three methods. Smaller percentage of outliers indicates less hissing and whistling artifacts. Our proposed method is superior to the other two methods in terms of lower average spectral distortion and smaller percentage of outliers.

	D_{mean}	D_{σ}	OUTLIERS (>10DB)	OUTLIERS (>15DB)
VQ	3.19217	3.50798	5.085%	1.583%
GMM	3.16272	3.47197	4.700%	1.073%
CDHMM	2.92003	2.82911	2.174%	0.245%

Table 1. Objective performance comparison

Result of subjective tests indicates that the enhanced wideband speech sounds more natural with crispy high-frequency components. The hissing and whistling artifacts as often present in other bandwidth expansion systems are reduced significantly. However, some artifacts are still perceivable occasionally. Future research will try to further reduce these artifacts with better block classification and state mapping techniques.

4. CONCLUSION

In this paper, a block-based CDHMM method is proposed to estimate the missing high-band spectral envelope from narrowband signals. The method explores inter-frame relationship within the detected block via mapping between narrowband and wideband optimal state sequences. Both objective and subjective tests show that the proposed method can produce good quality wideband speech from a narrowband input. The hissing and whistling artifacts commonly present in most bandwidth expansion systems are significantly reduced.

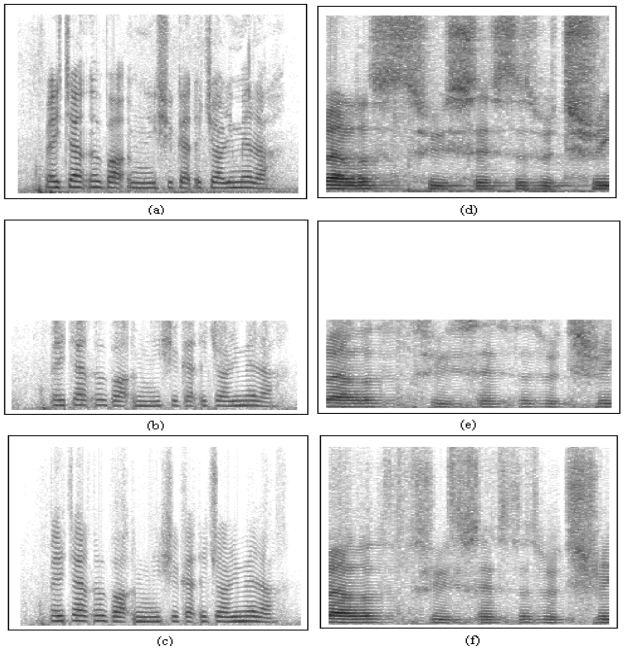

Figure 4. Spectrogram comparison: (a) original wideband female sound (b) narrowband female sound (c) extended wideband female sound (d) original wideband male sound (e) narrowband male sound (f) extended wideband male sound

5. REFERENCES

[1] Y. Agiomyrgiannakis, and Y. Stylianou, "Combined Estimation/coding of Highband Spectral Envelopes for Speech Spectrum Expansion", Proc. ICASSP, pp. 469-472, 2004.

[2] M. Nilsson, H. Gustafsson, S.V. Andersen, and W.B. Kleijn, "Gaussian Mixture Model Based Mutual Information Estimation between Frequency Bands in Speech", Proc. ICASSP, pp. 525-528, 2002.

[3] N. Enbom, and W.B. Kleijn, "Bandwidth Expansion of Speech Based on Vector Quantization of the Mel Frequency Cepstral Coefficients", Proc. Speech Coding, pp. 171-173, 1999.

[4] M. Nilsson, and W.B. Kleijn, "Avoiding Over-estimation in Bandwidth Extension of Telephony Speech", Proc. ICASSP, pp. 869-872, 2001.

[5] Y. Nakatoh, M. Tsushima, and T. Norimatsu, "Generation of Broadband Speech from Narrowband Speech Based on Linear Mapping", Electronics and Communications in Japan, Part 2, Vol 85, No. 8, pp. 44-53, 2002.

[6] K.Y. Park, and H.S. Kim, "Narrowband to Wideband Conversion of Speech Using GMM Based Transformation", Proc. ICASSP, pp. 1843-1846, 2000.

[7] 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. ICASSP, pp. 241-244, 2002.

[8] P. Jax, and P. Vary, "On artificial Bandwidth Extension of Telephone Speech", Signal Processing, pp. 1707-1719, 2003.

[9] G. Chen, and V. Parsa, "HMM-based Frequency Bandwidth Extension for Speech Enhancement Using Line Spectral Frequencies", Proc. ICASSP, pp. 709-712, 2004.

[10] W.M. Yu, and C.F. Chan, "Harmonic+noise Coding Using Improved V/UV Mixing and Efficient Spectral Quantization", Proc. ICASSP, pp. 477-480, 1999.