BLOCK-BASED SPEECH BANDWIDTH EXTENSION SYSTEM WITH SEPERATED ENVELOPE ENERGY RATIO ESTIMATION

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

The major issue in extending bandwidth of narrowband speech signal (0-4kHz) is the estimation of high-band portion (4-8 kHz) of spectral envelope. It is found that, apart from the shape of high-band spectral envelope, the relative energy level of the missing high band to the observable low band is also crucial to the system performance. In this paper, the two-fold problem is solved by two different estimation rules. In memoryless bandwidth extension systems, the missing high-band information is estimated from narrowband speech using the current frame only. As the narrowband-to-wideband mapping is a one-to-many problem ([1]), memoryless system is likely to cause hissing and whistling artifacts. Our method treats envelope shape estimation on a block basis. Detected narrowband speech block is either one word or a sequence of words, which is modeled by CDHMM (continuous density hidden Markov model) and mapped to a wideband CDHMM pre-trained by original version of the speech block. High-band energy level, present as normalized energy ratio to observable low-band energy, is estimated on an MMSE rule. 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/ /z/ /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 would still exist due to its capability of enhancing speech quality without any modification of current infrastructure.

In some reported enhancement systems ([2,3,4,5]), how to reduce hissing and whistling artifact is a common problem. In [1], it is shown that the low-band and high-band relationship is a one-to-many mapping. It implies an arbitrary solution can somehow find out high-band estimate that optimally meets pre-defined criterion, but can hardly tell how close the estimate is to the original. This kind of uncertainty makes hissing artifact unavoidable. In memoryless system, where the estimate is independent of context, hissing is even more severe. However, if additional envelope energy ratio can be estimated and original ratio trajectory be generally tracked, the outcome would have less hissing artifact even though the fine detail in high-band portion (spectral shape) is different from the original.

In this paper, envelope shape and energy ratio are both estimated with memory. Shape estimation is realized by block-based CDHMM state mapping (Section 2.1). Energy ratio trajectory is separately tracked in an MMSE manner (Section 2.2). H+N model (Harmonic plus noise model) [7] is employed to synthesize speech. Simulation and performance comparison with VQ (Vector Quantization codebook mapping) and GMM (Gaussian Mixture Model conversion) methods are shown in section 3.

2. BANDWIDTH EXTENSION SYSTEM

2.1. Envelope shape estimation

2.1.1 Enhancement structure

It is well known that human speech generation is a non-stationary random process, which can be approximated by hidden Markov process (piecewise stationary). When people is speaking, the hidden Markov state is also changing, whose evolution can be readily calculated via Viterbi algorithm. If speech is bandwidth-limited, it can be considered as the output of another new hidden Markov process with slightly different statistical properties. However, once HMM is firstly defined as left-to-right, experiments show that Markov states of narrowband speech and wideband speech have correlation even though their output pdfs (probability density functions) of the corresponding states are different. We measure this correlation with a so-called state-transitional matrix and use it to map narrowband optimal state sequence to wideband one.

The proposed enhancement system is shown in Figure 1. Narrowband speech is sampled at 8 kHz. Speech frame is fed to H+N analyzer and features such as pitch, gain, voiced/unvoiced decision and spectral envelope represented by 10-order LSF (Line Spectrum Frequencies) coefficients, are extracted. The first three features are passed directly to MBE (Multi-Band Excitation) synthesizer while narrowband LSF is used for wideband envelope shape recovery. Frame periodicity, zero crossing rate and frame energy are used for block detection. The

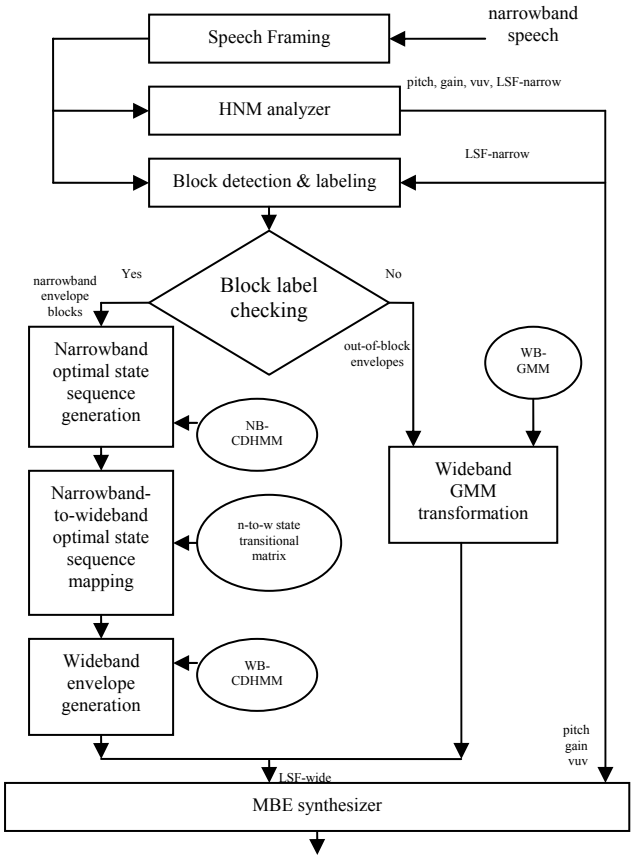

Figure 1. Flowchart of enhancement system

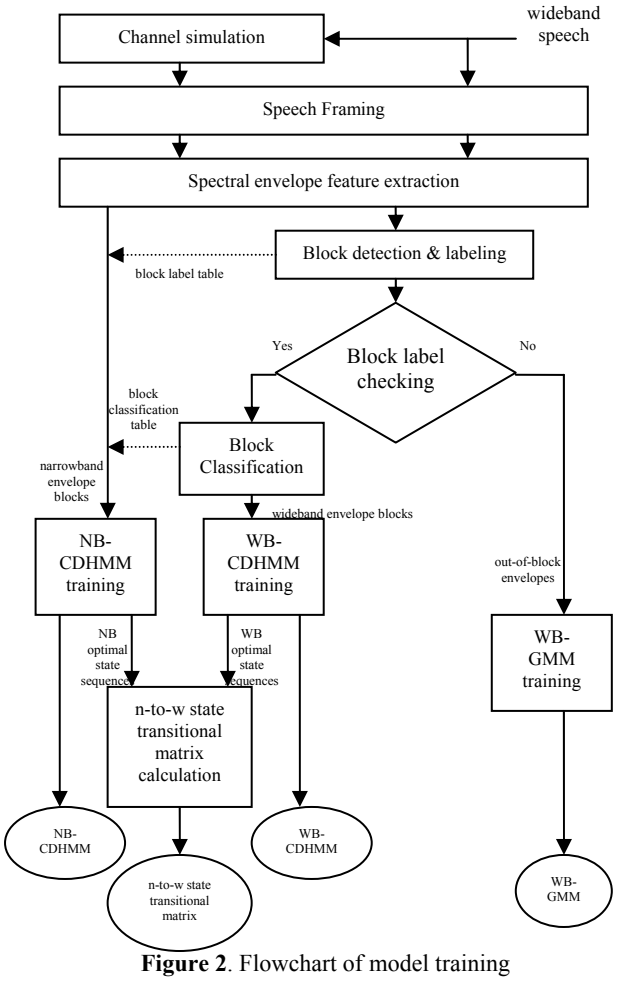

Figure 2. Flowchart of model training

detector simply makes sure any frame outside the block is either noise-like or silent frame. With endpoints of speech block marked, speech frame stream is then labeled.

Let us consider an arbitrary block of narrowband envelopes as the observation of pre-trained left-to-right CDHMM with Gaussian mixture pdf. Narrowband optimal state sequence is then obtained via Viterbi algorithm. It is then mapped to wideband optimal state sequence via dynamic programming using pre-trained n-to-w (narrowband-to-wideband) state transitional matrix mentioned above and to be discussed in Section 2.1.3 at some length. The estimated wideband optimal state sequence together with narrowband observation is fed to wideband CDHMM to get wideband observation using the method described in section 2.1.4. Finally MBE synthesizer collects all necessary parameters such as pitch, gain, vuv and estimated wideband envelope in LSF format to produce wideband speech.

2.1.2. Model training

Figure 2 shows how models in Figure 1 are trained. As mentioned in previous section, two types of CDHMMs are employed, one for narrowband speech block and the other for wideband. Wideband training speech (0-8kHz) firstly goes through a simulation of speech transmission, which consists of a low-pass filter with cutoff frequency at 4 kHz, H+N encoder and decoder. LPC-MFCC (Mel Frequency Cepstrum Coefficients of LPC coded spectral envelope) and LSFs are extracted frame-by-frame. LPC-MFCC is used as feature vector of CDHMM while LSF parameters are grouped with LPC-MFCC for speech synthesis. Wideband envelope frame stream is labeled as blocks. Note that the same labeling is applied to the corresponding narrowband envelope frame stream.

Since the smallest speech block unit is word, the number of envelope block patterns is numerous and a single CDHMM is far from sufficient to model this diversity. Therefore, all envelope blocks have to be clustered first. Clustering method for the matrix quantization is binary splitting LBG algorithm. Note that LBG clustering is applied to wideband envelope blocks only and narrowband blocks follow the same classification. All CDHMM pairs are trained via EM (Expectation-Maximization) algorithm. CDHMM pair has the same number of states and Gaussian mixtures per state.

As for training data, since each wideband speech 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. Optimal state sequence pairs are used to measure n-to-w state transitional matrix.

Under the constraint of left-to-right structure, wideband and narrowband states have large correlation. It is well known that successive reappearance of a certain state indicates that output pdf would remain stable during this period of time and state jump implies sudden change in pdf. Although output pdfs of

wideband CDHMM and narrowband CDHMM are generally different, it is not the case for state evolution. Our experiments show that, when narrowband state evolution is within a steady state, the corresponding wideband state evolution probably resides in a certain wideband state too. If there is a state jump in narrowband state at some time instant, a state jump is most likely to happen in wideband state evolution at some nearby time instant. In order to estimate wideband optimal state sequence, we have to measure how probable an arbitrary narrowband state is related to each wideband state. The whole measurement presents 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 wideband states. The trained matrix usually has large probability values in diagonal and off-diagonal entries, which implies most narrowband state is related more closely to its nearby wideband states in left-to-right structure. Note that if there are *M* clusters of CDHMM pairs, there are *M* n-to-w state transitional matrixes.

2.1.3. 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 searched from the DP (dynamic programming) grid as indicated in figure 3. Specifically, since the cost of optimal path 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*.

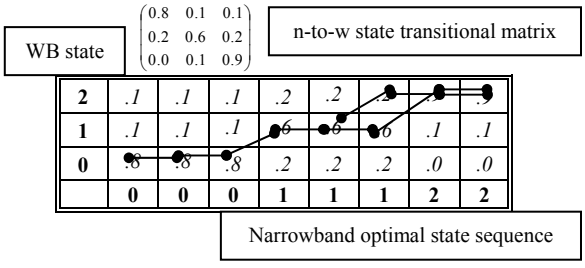

Figure 3. Illustration of a simple DP grid when narrowband optimal state sequence is "0 0 0 1 1 1 2 2". The marked paths are two most likely wideband optimal state sequences.

To handle the situation where several DP pathes are competing, *N* most likely wideband optimal state sequences are searched and weighted by normalizing their path costs. Multiple optimal path searching 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.

Conventional GMM conversion method uses either a single large-mixture wideband GMM or two large-mixture wideband GMMs (one for voiced frame and the other for un-voiced frame) to estimate wideband envelope from current narrowband input. In our approach, the desired Gaussian pdf of current narrowband frame is searched out by firstly determining the wideband speech block cluster, current belonging narrowband block is associated to, and then picking out the particular wideband state, to which current narrowband frame is mapped. Therefore our method is an extension of conventional GMM method.

2.1.4. Wideband envelope shape 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:

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

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

where *m* is mixture index, *g*() is Gaussian distribution density function. When *i* goes through all frames within a block, we get the estimated wideband envelope block for that particular 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 coefficients to be the normalized weights of each wideband optimal state sequence.

2.2. Envelope energy ratio estimation

2.2.1. Estimation structure

Energy ratio between low-band portion and high-band portion is another important factor that would affect the whole performance of bandwidth extension system, because energy ratio trajectory gives the general look of output spectrogram. We borrow the estimation structure in [5], which implies that, if high-band feature can be well clustered in its vector space, estimate of the feature can be a function of those cluster centroids. And the function's parameters are determined from narrowband feature input in a statistically optimal manner. In [5], its HMM-plus-VQ structure and MMSE estimator turn this function to be linear combination of those centroids. In [6], the authors list other two alternatives. Since their approach works well for multi-dimensional high-band feature vector, in our case, we define the HMM state as energy ratio centroid (one-dimensional) and keep other procedures similar to [5]. The centroids are computed via LBG algorithm and stored in a SQ (scalar quantization) codebook. We obtain state sequences by calculating energy ratio and searching through SQ codebook. State sequences are then used to calculate initial probability π and transitional probabilities a_{ij} for the HMM. Output pdf $p(x|s)$, the pdf of narrowband feature vector, on condition that energy ratio belongs to state S, is modeled by Gaussian mixture.

2.2.2. Estimation rule

Let us denote $X_1^m = \{x(1), x(2), ..., x(m)\}$, where m is frame index. Referring to [5], once π, a_{ij} and $p(x|s)$ are available, the quantity $P(s_i(m)|X_1^m)$, which is the degree of association in current frame with i'th state when all previous frames are observed, can be readily calculated. The MMSE estimate about energy ratio of current frame is computed as $y(m)_{MMSE} = E\{y(m)|X_1^m\} = \sum_{i=1}^{N} e_i P(s_i(m)|X_1^m)$, where N is the number of HMM states or SQ codebook entries and e_i is i'th entry in the codebook. The following formulas are for calculation of $P(s_i(m)|X_1^m)$.

$$P(s_i(m)|X_1^m) = p(s_i(m), X_1^m) \Big/ \sum_{j=1}^{N} p(s_j(m), X_1^m),$$

where the hardly tractable pdf $p(X_1^m)$ is replaced by the marginal density of the joint pdf $p(s_i(m), X_1^m)$.

$$p(s_i(m), X_1^m) = p(x(m)|s_i(m), x(1)...x(m-1)) p(s_i(m), x(1)...x(m-1))$$
$$= p(x(m)|s_i(m)) p(s_i(m), X_1^{m-1}) = p(x(m)|s_i(m)) \alpha_i(m)$$

, where $\alpha_i(m) = p(s_i(m), X_1^{m-1})$ is successively calculated:

$$\alpha_i(1) = P(s_i) = \pi_i$$
$$\alpha_i(m+1) = \sum_{j=1}^{N} \alpha_j(m) p(x(m)|s_j(m)) P(s_i(m)|s_j(m))$$
$$= \sum_{j=1}^{N} \alpha_j(m) p(x(m)|s_j(m)) a_{ji}$$

The energy ratio $R = (L-H)/(L+H-2\theta)$ is in dB scale, where L and H are average energy of low-band and high-band portions of spectral envelope respectively. θ is a fixed and pre-defined low-bound to guarantee $(L-\theta)$ and $(H-\theta)$ are both positive. Therefore R is normalized and ranges in (-1,1).

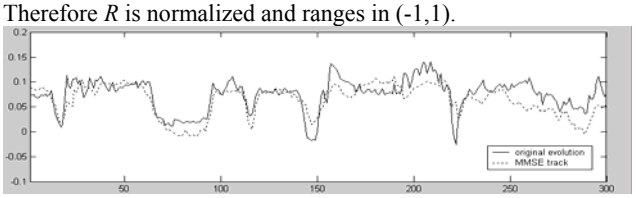

Figure 4. Energy ratio retrieval

In Figure 4, original energy ratio evolution for 300 frames long is marked solid while MMSE estimate is dotted. Configuration of the test is 64-entry SQ codebook (64 HMM states) and 8 Gaussian mixtures per state for $p(x|s)$. The MMSE ratio estimation follows the original trajectory, thus giving output spectrogram a similar general look.

3. SIMULATION

Speech data is from phonetically balanced IViE corpus (Intonational Variation in English) sampled at 16 kHz. Around 90% is used for training and the rest is for evaluation. We also have a degraded version (8 kHz) of all data for both training and evaluation.

For shape estimation, detected wideband envelope blocks are partitioned into 768 clusters and narrowband blocks follow the same classification. Each cluster is fit with an 8-state left-to-right CDHMM with output pdf having 8 Gaussian mixtures per state and diagonal covariance matrix. Therefore, there are totally 768 CDHMM pairs. Output pdf of both wideband and narrowband CDHMM takes 32-order LPC-MFCC as feature vector. Since there are 8 states for all CDHMMs, n-to-w state transitional matrix is therefore 8-by-8 dimensional. During EM training of wideband CDHMM, the centriods of the grouped 18-order LSF vectors, which is for speech synthesis, are iteratively updated accordingly. 10 most likely wideband optimal state sequences are picked out from DP grid to handle the case of competing paths. As for speech frames outside blocks, a wideband GMM with 256 mixtures is trained to map narrowband envelope in a memoryless sense. In energy ratio estimation, the size of SQ codebook is 64 and $p(x|s)$ is 8-Gaussian mixture density function. Feature vector of narrowband speech is also 32-order LPC-MFCC. Wideband speech is synthesized according to [7].

The VQ mapping [2] and conventional GMM conversion [4] are implemented for comparison purposes. The feature vector parameters of the two methods are also 32-order LPC-MFCC. VQ mapping has two kinds of codebooks. 1024-entry codebook is for voiced frame and 512-entry codebook for unvoiced frame. GMM conversion also has two different model configurations for voiced and un-voiced frames. Mixture number for voiced GMM and unvoiced GMM are both 256. Objective measurement D is spectral envelope distortion in high band:

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

Table 1 shows the objective performance comparison. Note that smaller percentage of outliers indicates less hissing and whistling artifacts. Our proposed method is superior to others with lower average spectral distortion and smaller variance and percentage of outliers.

	D_{mean}	D_{σ}	outliers(>5dB)	outliers(>7.5dB)
VQ	3.19217	3.50798	5.085%	1.583%
GMM	3.16272	3.47197	4.700%	1.073%
PROPOSED	3.00050	2.24138	2.025%	0.207%

Table 1. Objective performance comparison

Subjective tests show that the enhanced wideband speech sounds more natural with crispy high-frequency components. Particularly, the unavoidable hissing and whistling artifacts are greatly reduced in our proposed system.

4. 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] 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.

[3] 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.

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

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

[6] P. Jax and P. Vary, "Artificial Bandwidth Extension of Speech Signals using MMSE Estimation Based on A Hidden Markov Model", Proc. ICASSP, pp. I680-I683, 2003

[7] 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.