SPEECH BANDWIDTH ENHANCEMENT USING STATE SPACE SPEECH DYNAMICS

Sheng Yao and Cheung-Fat Chan

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

ABSTRACT

Extending narrowband speech (0-4 kHz) to wideband speech (0-8 kHz) has applications in telephone systems and speech recognition systems where wideband training speech data may not be available. A couple of methods have been proposed to retrieve the missing high-band information (4-8 kHz) from narrowband speech. Memoryless systems are likely to produce large hissing artifacts since mutual information between low-band (0-4 kHz) and high-band (4-8 kHz) spectra are actually quite low. Generally speaking, bandwidth extension cannot recover original high-band information but good approximates with less over-estimation of the high-band energy, which usually refers to hissing artifact, can be obtained by considering the neighboring speech frames. In this paper, we propose a new bandwidth extension system with memory by using a state-space model to capture the long-term speech dynamics. The model parameters can be trained in the sense of maximum likelihood (ML) and the enhancement is obtained via wideband state vector estimation and Kalman filtering. The performance in terms of spectral distortion is shown to be much better than other memoryless systems and is comparable with early Continuous Density Hidden Markov Model (CDHMM) memory system. The new state-space method is inherent sequential and has advantages of less processing delays and robustness against block detection errors.

1. INTRODUCTION

In telephone systems, the transmission effect is similar to a low-pass filter with cutoff frequency at 3.4 kHz. Due to the loss of high frequency components, speech sounds such as fricatives (/s/ /t/ /f/) and stops (/p/ /k/) need extra listening effort. In worse communication environment, alternative lengthy description such as 'peter' may be helpful of articulating English letter 'p', but this adds to the overhead of conversation. Bandwidth extension system can handle this muffing effect but in the same time unavoidably bring about hissing artifacts [7]. Hissing level is much higher in early memoryless systems [1,3,4,5] since there is no inter-frame relationship between wideband feature estimates. Disregarding inter-frame relationship leads to spectral jump in high-band and causes serious hissing effect. Recent approaches tend to look back into frame history and generally have the concept of 'state' [6,7]. History-related state usually helps to determine meaningful choice of pre-trained model parameters whose correctness is judged by state estimation rules.

In this paper, we apply linear state-space model to the problem by assuming input narrowband speech feature vectors as observations and targeted wideband speech feature vectors as states to be estimated. In linear state-space model, state dynamics is modeled to be linearly evolving and observations are also linear mapping of the states.

The rest of paper is organized as follows. Section 2 briefly describes the formulation of linear state-space model. Section 3 deals with the details of system structure, followed by simulation and performance comparison in Section 4. Finally Section 5 is conclusion.

2. LINEAR STATE-SPACE MODEL

The formulation of the model has the following mathematical representation:

$$x(k+1) = Ax(k) + u + w(k) \quad (1)$$

$$o(k) = Cx(k) + v(k) \quad (2)$$

initialized by

$$E[x(0)] = \mu(0), \quad E[x(0)x(0)^T] = \Sigma(0)$$

In the above, $x(k) \in R^p$ is a hidden state vector at time k, $o(k) \in R^m$ is an observation vector, $w(k)$ and $v(k)$ are uncorrelated zero-mean Gaussian noise vectors with covariances $E[w(k)w(l)^T] = Q\delta_{kl}$ and $E[v(k)v(l)^T] = R\delta_{kl}$, respectively. $u(k) \in R^p$ is a deterministic input vector or control vector. The parameter set of state space model is

$$\theta = \{A, u, Q, C, R\}$$

Equation (1) is also called state equation. The equation represents linear dynamics of the state variable using linear prediction (or auto-regressive). The noise term $w(k)$ represents the degree of inaccuracy in using the linear state equation to describe the true state dynamics. The greater the determinant of Q is, the greater degree of inaccuracy there is.

Observation equation (2) is static in nature since the time indices of x and o are the same. It represents the noisy relationship between the state vector and the observation vector. The noise term v represents the degree of inaccuracy in using

the linear mapping $o(k) = Cx(k)$ to describe the true relationship between the state and observation vectors. Due to the presence of noise term and possible non-invertiblity of matrix C in (2), the state vector $x(k)$ cannot be uniquely determined given the observation vector $o(k)$, which reflects the one-to-many mapping between narrowband and wideband speech features [7]. Thus the state dynamics described in (1) is hidden dynamics like hidden state in Hidden Markov Model (HMM).

3. SYSTEM STRUCTURE

3.1. Feature Vector Definition

Narrowband speech feature vector has elements of 10 narrowband line spectral frequencies (LSF) and wideband speech feature vector consists of envelope energy ratio R and 18 wideband LSF. LSF feature is widely used in speech coding scheme, whose parameter structure has quantization merit. Besides, linear mapping system in [4] also uses this feature. R measures the energy level ratio between low-band and high-band spectral envelope.

$$R = (L - H)/(L + H - 2\lambda)$$

in dB scale, where L and H are average energy of low-band and high-band portions of spectral envelope respectively. λ is a fixed and pre-defined low-bound to guarantee $(L-\lambda)$ and $(H-\lambda)$ are both positive. Therefore R is normalized and ranges in (-1,1). Then we give the following vector definition for linear state-space model:

$$o = \{nLSF_1^{10}\}$$
$$x = \{R, wLSF_1^{18}\}$$

As time index or frame index k increases, $x(k)$ evolves according to equation (1) and $o(k)$ changes linearly with $x(k)$ according to equation (2). However the model parameter set θ cannot be stationary since a single state space model is far from sufficient to model all kinds of speech dynamics. Therefore, we make θ piecewise stationary, which means the parameter set change mode for every L frames. L is fixed frame block length and experimentally determined (to be discussed in Section 4).

3.2. Model Training

In state space literature, maximum likelihood learning criterion gives the following formulation:

$$[\hat{A} \quad \hat{u}] = [\sum_{n=1}^{N}\sum_{k=1}^{L_n} E(x_{kn}x_{(k-1)n}^T \mid o) \quad \sum_{n=1}^{N}\sum_{k=1}^{L_n} E(x_{kn} \mid o)] \cdot \begin{bmatrix} \sum_{n=1}^{N}\sum_{k=1}^{L_n} E(x_{(k-1)n}x_{(k-1)n}^T \mid o) & \sum_{n=1}^{N}\sum_{k=1}^{L_n} E(x_{(k-1)n} \mid o) \\ \sum_{n=1}^{N}\sum_{k=1}^{L_n} E(x_{(k-1)n}^T \mid o) & \sum_{n=1}^{N} L_n \end{bmatrix}^{-1}$$

$$\hat{C} = [\sum_{n=1}^{N}\sum_{k=1}^{L_n} E(o_{kn}x_{kn}^T \mid o)][\sum_{n=1}^{N}\sum_{k=0}^{L_n} E(x_{kn}x_{kn}^T \mid o)]^{-1}$$

$$\hat{Q} = \frac{1}{\sum_{n=1}^{N} L_n}\sum_{n=1}^{N}\sum_{k=1}^{L_n}\{E(x_{kn}x_{kn}^T \mid o) - [\hat{A} \quad \hat{u}][E(x_{kn}x_{(k-1)n}^T \mid o) \quad E(x_{kn} \mid o)]^T\}$$

$$\hat{R} = \frac{1}{\sum_{n=1}^{N}(L_n + 1)}\sum_{n=1}^{N}[\sum_{k=0}^{L_n}(o_{kn}o_{kn}^T \mid o) - \hat{C}\sum_{k=0}^{L_n} E(o_{kn}x_{kn}^T \mid o)^T]$$

where N is the number of wideband training sequence and L_n is the length of *n*'th sequence, which is now fixed to be L.

Note that state vector x_k is actually observable during training phase. Therefore we get:

$$E_L(x_k \mid o) = x_k$$
$$E_L(x_k x_k^T \mid o) = x_k x_k^T$$
$$E_L(o_k x_k^T \mid o) = o_k x_k^T$$

Generally speaking, when x_k is hidden state, the above estimates have to be calculated and $\hat{\theta}$ is estimated iteratively according to Estimation-Maximization (EM) algorithm. But now we can directly obtain θ from the first calculation, and therefore off-line training time is substantially reduced in comparison with [8], where EM iteration is unavoidable.

3.3. State Estimation

State estimation is actually wideband feature recovery, which is formally solved by Kalman filter algorithm. Its formulation can be summarized as:

For $k = 1,2,...,L,$

Kalman Prediction

$$\hat{x}_{k|k-1} = A\hat{x}_{k-1|k-1} + u$$
$$\Sigma_{k|k-1} = A\Sigma_{k-1|k-1}A^T + Q$$

Kalman Gain

$$\mathrm{K}_k = \Sigma_{k|k-1}C^T(C\Sigma_{k|k-1}C^T + R)^{-1}$$

Kalman Correction

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + \mathrm{K}_k(o(k) - C\hat{x}_{k|k-1}),$$
$$\Sigma_{k|k} = \Sigma_{k|k-1} - \mathrm{K}_k(C\Sigma_{k|k-1}C^T + R)\mathrm{K}_k^T.$$

Kalman filter algorithm is a sequential algorithm, initialized usually by

$$\hat{x}_{0|0} = E[x(0)] = \mu(0)$$
$$\Sigma_{0|0} = E[x(0)x(0)^T] = \Sigma(0)$$

Hence the above algorithm is also called Kalman forward recursion. The sequential property greatly reduces the overall delay of bandwidth extension system compared with the approach in [8], where Viterbi algorithm has to check the whole observation trellis diagram in the detected speech block and trace back for optimal state sequence.

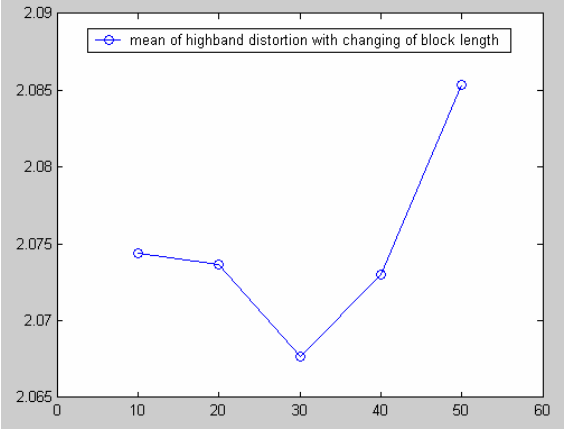


Figure 1. Average high-band distortion versus block length

4. SIMULATION

There are several practical issues during implementation. The first one is the choice of sequence length L. Another one is how many modes the state-space model is supposed to have, which has great relation with L. The longer the sequence length is, the more dynamics the sequence would have and the more modes are required. In simulation, we try different L values from 10 to 50 frames with frame step 10. For each experiment with fixed L, all the training speech (both wideband and its narrowband version) is actually chopped into L-frame long blocks. We apply KMEANS algorithm several times to narrowband blocks for different cluster number settings. As known to all, the average cluster variance decreases as the number of clusters increases. The final number of clusters (modes) is determined to be the one which leads to largest average cluster variance decrease in percentage from its previous cluster number trial as we make cluster number gradually increasing.

The feature vector for KMEANS should summarize the general spectral evolution and maintain low dimension. So along radius frequency, difference LSF is calculated:

$$nLSF'(i,k) = nLSF(i,k) - nLSF(i-1,k)$$

$i = 2,3,...,10$

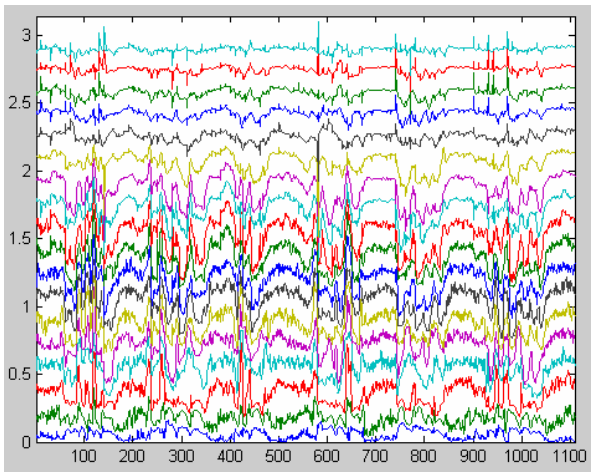
$$nLSF'(1,k) = nLSF(1,k)$$

Along time index k, mean and standard deviation is calculated for $nLSF'(i,k), i = 1,2,...,10$ until half of the block is reached, $k = L/2$. The same calculation is then repeated for the other half, $L/2 < k \le L$. Finally the block feature vector is 40-dimensional (10 means and 10 standard deviations for each half of the block). Figure 1 shows system performances with L value increasing from 10 to 50. The high-band distortion is calculated in dB as:

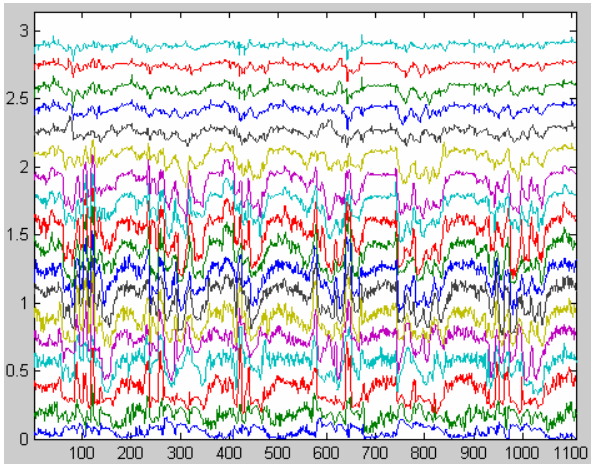
$$D = \left(\frac{2}{\pi} \int_{\pi/2}^{\pi} \left(10 \log_{10} S_{org}(\omega) - 10 \log_{10} S_{ext}(\omega)\right)^2 d\omega \right)^{\frac{1}{2}} \quad (3)$$

We set $L = 30$ in the following experiments.

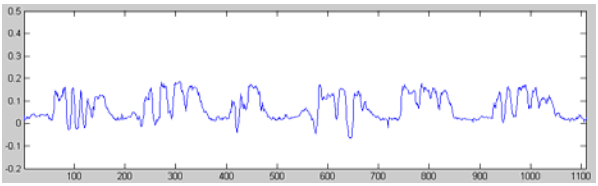
Another important issue is the blocking effect. On the boundary of blocks, state-space model changes its mode but the estimated x_k should not change instantly. Therefore we initiate $\hat{x}_{0|0}$ of block m as $\hat{x}_{L|L}$ of block m-1. $\Sigma_{0|0}$ also changes accordingly. Then Kalman filter algorithm iterates with previous estimate as starting point. For the very first block, we assume $\hat{x}_{0|0}$ to be typical noise-like $\{R, wLSF_1^{18}\}_{noise}$. Figure 2 illustrates feature vector estimation for 1100-frame speech in comparison with its original version.



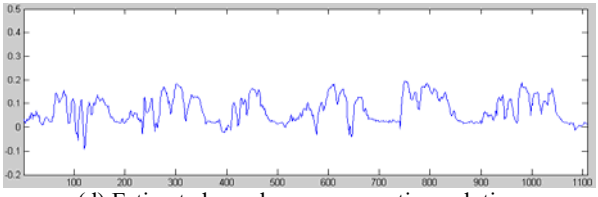
(a) Original 18 wideband LSF evolution



(b) Estimated 18 wideband LSF evolution

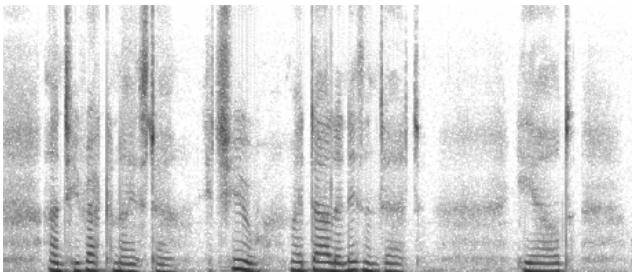


(c) Original envelope energy ratio evolution

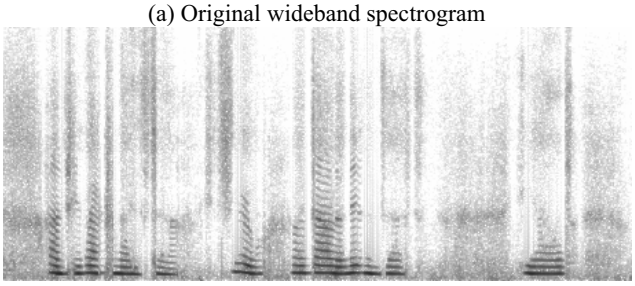


(d) Estimated envelope energy ratio evolution

Figure. 2 Feature vector estimation compares with the original version



(a) Original wideband spectrogram



(b) Estimated wideband spectrogram

Figure 3. Speech spectrogram comparison

Note that in figure 2 (b), the estimation in low-order LSF is quite accurate while in high-order LSF, the estimated curve is somehow smoothed. Smooth evolution in high-band LSF means loss in high-band spectral envelope detail, which is a common problem in all bandwidth systems. However, Kalman filter algorithm can maintain the general shape of the high-band evolution via linear modeling of state dynamics. Moreover, unlike memoryless method, there are few sparkles or sudden jump in high-band curve, which means hissing artifact is reduced. Figure 3 illustrates the slight smoothness effect in spectrogram. As shown, the general shape of high-band spectrum is maintained and speech sound such as fricatives and stops can be enhanced.

Both training and test speech data is chosen from of 16-kHz sampling IViE corpus consisting of three 6-minute paragraph reading speeches from three different speakers (male, female and child). Another three 1-minute speeches of the same three speakers is taken else where from corpus for evaluation. Our speech codec is based on harmonic-plus-noise model [2]. Wideband speech is low-pass filtered (cutoff frequency 4 kHz) and re-synthesized by multi-band excitation (MBE) synthesizer to form its narrowband version. Vector Quantization (VQ) codebook searching, linear mapping (LM), GMM and CDHMM methods are implemented for comparison. Speech feature vector definitions for these methods are the same (10 nLSF for narrowband feature, composite energy ratio plus 18 wLSF for wideband feature). Objective criterion (equation 3) is used to measure system performance. Table 1 shows the result of state-space model for speaker-dependency test. Table 2 shows the performance comparison with other methods.

5. CONCLUSION

Informal listening test shows that speech quality of proposed method is much better than memoryless method with less hissing artifact. Enhanced wideband speech sounds clearer than narrowband speech with crispy high-band components.

	D_{MEAN} (DB)	D_{SD} (DB)	OUTLIER(>5.0DB)	OUTLIER(>7.5DB)
male	2.07230	1.09978	1.11157%	0.08910%
female	2.37647	1.12854	1.61438%	0.12322%
child	2.35546	1.09809	1.88808%	0.06734%
Sp-ind	2.38083	1.36110	2.11157%	0.18910%

Table 1. Speaker-dependency test

	D_{MEAN} (DB)	D_{SD} (DB)	OUTLIER(>5.0DB)	OUTLIER(>7.5DB)
VQ	3.07432	3.63478	4.91432%	1.49311%
LM	2.99903	3.37214	4.57610%	1.11672%
GMM	2.73877	3.03571	3.71240%	0.99712%
CDHMM	2.43788	1.52771	2.22861%	0.20507%
Proposed	2.38083	1.36110	2.11157%	0.18910%

Table 2. Performance comparison with other methods

Objective measure also proves the advantage in terms of both mean and standard deviation of high-band spectral distortion. However, computational complexity is higher than memoryless systems (VQ, LM, GMM), which is usually the case for advanced memory system. Also the performance is comparable to CDHMM approach with the advantage of sequential processing to save system delay time and robustness against block detection error, since we fix the speech block length and effectively handle boundary discontinuity. Future work may focus on changing equation (1)(2) into non-linear form to possibly increase the capability of modeling fine details, making high-band spectrum a little richer.

6. REFERENCES

[1] 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.

[2] 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.

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

[4] S.Chennoukh, A.Gerrits, G.Miet and R.Sluijter, "Speech Enhancement via Frequency Bandwidth Extension using Line Spectral Frequencies", Proc. ICASSP, pp. 665-668, 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] P. Jax, and P. Vary, "On artificial Bandwidth Extension of Telephone Speech", Signal Processing, pp. 1707-1719, 2003.

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

[8] S.Yao and C.F.Chan, "Block-based Bandwidth Extension of Narrowband Speech Signal by using CDHMM", Proc. ICASSP, pp. I793-I796, 2005

[9] Li Deng and Douglas O'Shaughnessy, "Speech Processing – A Dynamic and Otimization-Oriented Approach"