# A MAXIMUM A POSTERIOR-BASED RECONSTRUCTION APPROACH TO SPEECH BANDWIDTH EXPANSION IN NOISE

*Hyunson Seo*[*]*, Hong-Goo Kang*

Yonsei University
DSP Lab., Department of E.E.
Seoul, South Korea

*Frank Soong*

Microsoft Research Asia
Beijing, China

**ABSTRACT**

We propose a novel bandwidth expansion algorithm for extending narrowband speech signal to wideband by exploiting segment examples pre-stored in a speaker independent database. Both narrowband and wideband representation of speech signals are pre-stored in the corpus and they are dynamically chopped into variable length segments. Narrowband segments are used dynamically to explain a given narrowband input sentence while the wideband expanded version of the input sentence is constructed correspondingly. The matching process in the narrowband favors a longer segment patch by the chosen Maximum A Posterior (MAP) criterion. As a result, the multiple choices in matching process are significantly reduced with the MAP criterion in decoding. The approach is further generalized to deal with noise corrupted narrowband input signals and the well-known Vector Taylor Series (VTS) noise adaptation algorithm is incorporated into the matching and bandwidth expansion process. A series of experiments is performed to validate the approach on both clean and noise corrupted narrowband speech where both car noise and babble noise corrupted samples are tested.

***Index Terms—*** speech bandwidth expansion, maximum a posterior, corpus-model, noise reduction, vector Taylor series

## 1. INTRODUCTION

Quality and intelligibility of narrowband telephone speech can be improved by artificial bandwidth expansion (BWE), which extends the speech bandwidth using information available in the narrowband speech signal. The derivation of high-frequency components from the lower-band frequencies of speech signals is a non-trivial problem because the mutual information between the lower-band and upper-band frequencies is relatively low within any frame of speech features [1]. In [2], it was shown that the low-band and high-band relationship is an one-to-many mapping problem.

Various methods have been proposed to tackle the problem and extend the narrowband speech to wideband one. Most of BWE algorithms in literature, theoretically, are based on the source-filter speech production model, and the techniques to estimate the wideband spectral envelopes have been developed in many literatures using various methods [3–6]. Since statistical algorithms, such as Gaussian mixture model (GMM) and hidden Markov model (HMM), are flexible in modeling the statistics of speech signal and identify the high-band parameters with a soft decision scheme, they showed better performance than other simpler statistical methods, e.g., codebook or linear mapping [7, 8]. However, phonetic models used in those systems were only capable of representing the neighboring phonetic contexts, thus, hissing and whistling artifacts still remained as a common problem.

Several approaches to explicitly utilize cross-frame correlations in longer units have been proposed. In [9], an extension of GMM method by using HMM on a block basis was proposed, where the speech block was either one word or a sequence of words in narrowband utterance. Bansal exploited cross-frame contextual dependencies rather than spectral vectors within a frame using non-negative matrix factorization, where sequences of wideband magnitude spectral vectors are represented as linear combinations of non-negative bases [10]. The bases were several frames wide and learnt from a target speaker's brief recording of both wideband and narrowband speech. In [11], an memory-based approximation of GMM was formulated for BWE.

In this paper, we propose a new bandwidth expansion method which exploits a Maximum A Posterior (MAP)-based reconstruction approach. The estimate of wideband spectrum is formed by retrieving the *longest* matching segments from a pre-stored wideband speech corpus. Both narrowband and wideband representation of speech signals are pre-stored in the corpus and they are used for matching and reconstruction, respectively, in a synchronous way. Given any new narrowband signal, the narrowband *longest* segments that best explain the input signal are estimated. Accordingly, wideband extended version of the input sentence is derived by combining the corresponding segment examples stored in the wideband corpus. Identifying the *longest* matching segments is realized by the chosen MAP criterion, which necessarily favors longer matching segments [12]. With the MAP criterion, multiple choices in matching process are significantly reduced and the defect of little amount of information to identify high-frequency components within a frame is compensated with more contextual dependencies. This new approach to modeling and segment matching is based on the work in [12]. We generalize the idea to estimating missing part of speech, i.e., high-frequency components, given the narrowband speech signal. Throughout the paper, we use the term *corpus-model* to describe a set of models trained on the parallelized wideband and narrowband speech databases on which the segment matching and bandwidth expansion are performed with the MAP criterion.

In the latter part of this study, the approach is further generalized to deal with noise corrupted narrowband input signals, where the well-known Vector Taylor Series (VTS) [13] noise adaptation algorithm is incorporated into the matching process. The work presented here tightly integrates the noise reduction and bandwidth expansion process using a single shared *corpus-model*. In Section 2, we describe overall process of the proposed BWE framework. Section 3 generalizes the proposed technique to the noise corrupted speech. Then, a series of experiments on clean and noise corrupted narrowband speech is followed in Section 4 to validate our approach.

[*]Part of this work was done while the first author worked as an intern under supervision of Speech Group at Microsoft Research Asia.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

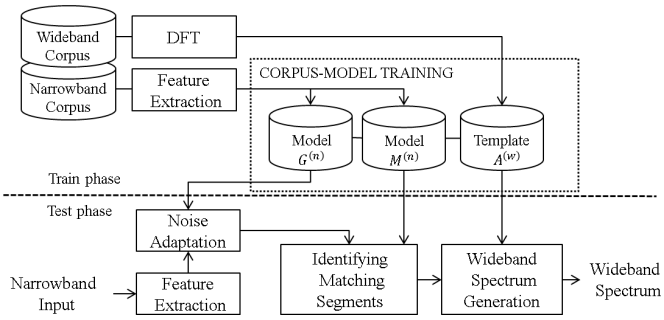


**Fig. 1**. Proposed speech bandwidth expansion framework

## 2. THE PROPOSED SPEECH BANDWIDTH EXPANSION FRAMEWORK

In this section, the overall BWE process is presented mainly with the following three steps: (i) *corpus-model* training (ii) matching segment identification (iii) wideband speech reconstruction. Fig. 1 shows the block diagram of the proposed BWE framework. The noise adaptation block is presented in Section 3.

### 2.1. Modeling spectral and temporal dynamics of narrowband speech

The first step of the proposed BWE method starts from constructing a *corpus-model* [12] that can represent exact and fine structure of spectral and temporal dynamics of speech using a parallel narrowband and wideband training data. The model consists of three components, $\{G^{(n)}, M^{(n)}, A^{(w)}\}$, where the notations $(n)$ and $(w)$ represents the bandwidth of speech signals on which each model is trained. In this work, for simplicity, the wideband bandwidth is defined as meaning the 8kHz full bandwidth (i.e.,0-8kHz) and narrowband the original 4kHz bandwidth (i.e., 0-4kHz).

Let $\mathbf{s} = \{\mathbf{s}_i : i = 1, 2, ..., I_s\}$ be the complete set of narrowband training features with $I_s$ frames and $\mathbf{s}_i$ be the feature vector at time frame $i$. The first model denoted by $G^{(n)}$ is a Gaussian mixture model (GMM) trained on $\mathbf{s}$ such as

$$G^{(n)} = \sum_{k=1}^{K} \lambda_k N(\mathbf{s}; \mu_k, \mathbf{\Sigma}_k), \tag{1}$$

where $N()$ is a Gaussian, $\lambda_k$ the mixture weights, and $\mu_k$ and $\mathbf{\Sigma}_k$ the mean and covariance of the Gaussians, respectively. Then, based upon $G^{(n)}$, it is possible to represent the complete training set $\mathbf{s}$ by identifying the Gaussian component in $G^{(n)}$ that maximizes likelihood of the frame. This results in a time sequence of maximum-likelihood Gaussian components such as:

$$M^{(n)} = \{m_i : i = 1, 2, ..., I_s\}, \tag{2}$$

where $m_i$ is a kernel index of a Gaussian in $G^{(n)}$ that produces the maximum likelihood for the segment $\mathbf{s}_i$. This sequence model $M^{(n)}$ captures long range of temporal dynamics of speech signals. Technically, any segment of any length in the training sentences can be modeled. Lastly, we generate a template model, or dictionary, $A^{(w)} = \{\mathbf{a}_i : i = 1, 2, ..., I_s\}$, where $\mathbf{a}_i$ is the clean wideband magnitude spectrum of $i$th frame in corpus. These wideband version segments in $A^{(w)}$ are later fetched in reconstruction phase in a way that the paralleled narrowband speech segments best explain the given narrowband input sentence.

### 2.2. Finding the best-matching segment examples in the narrowband

Given the model $G^{(n)}$ and $M^{(n)}$ trained in the manner described above, in the test phase, we identify segment examples from the narrowband corpus that most likely matches the input narrowband speech. Then, the wideband extended version of the input sentence is derived by retrieving the corresponding segment examples from the wideband corpus. In this subsection, we describe the method to find the segments of the narrowband training data which best explains the input speech sentence. Notice that the segment matching process is executed in the narrowband.

Let $y_{t:t+\tau}$ be a test segment taken from time frame $t$ to $t+\tau$ of the sentence $\mathbf{y} = \{y_t : t = 1, 2, ..., T\}$ and $M^{(n)}_{u:u+\tau} = \{m_i : i = u, u+1, ..., u+\tau\}$ the sequence of Gaussian component modeling consecutive frames from $u$ to $u+\tau$ in the training dataset $\mathbf{s}$. We measure the similarity between the two segments by using the posterior probability of the corpus segment $M^{(n)}_{u:u+\tau}$ given the test segment $y_{t:t+\tau}$ such as:

$$M^{(n)}_{t,u:u+\tau} = \arg\max_{\tau} p(M_{u:u+\tau}|y_{t:t+\tau}), \tag{3}$$

$$p(M^{(n)}_{u:u+\tau}|y_{t:t+\tau}) = \frac{p(y_{t:t+\tau}|M^{(n)}_{u:u+\tau})p(M^{(n)}_{u:u+\tau})}{p(y_{t:t+\tau})}. \tag{4}$$

Assuming that all the training patterns seen in the training dataset occur with an equal probability in the testing condition - an equal prior probability of $p(M^{(n)}_{u:u+\tau})$, the numerator term in (4) becomes the likelihood of the test segment $y_{t:t+\tau}$ associated with the segment of training dataset modeled with $M^{(n)}_{u:u+\tau}$. The likelihood is calculated as:

$$p(y_{t:t+\tau}|M^{(n)}_{u:u+\tau}) = \prod_{\varepsilon=0}^{\tau} g(y_{t+\varepsilon}|m^{(n)}_{u+\varepsilon}), \tag{5}$$

where the independence between adjacent frames is assumed. The denominator can be calculated as the summation of $p(y_{t:t+\tau}|M^{(n)}_{u:u+\tau})$ over all the possible pattern of $M^{(n)}_{u:u+\tau}$ stored in the model $M^{(n)}$. As shown in [12], an important characteristics of the posterior probability $p(M^{(n)}_{u:u+\tau}|y_{t:t+\tau})$ is that it favors the continuity of the matching segments, in terms of giving larger values for longer matching between $y_{t:t+\tau}$ and $M_{u:u+\tau}$. Thus, with the matching process based on the chosen MAP criterion, the multiple choices in matching process are significantly reduced. The identified longest segment as a whole matched unit includes rich contextual dependencies across frames. This alleviates one-to-many mapping problem in typical BWE problem so that the hissing or whistling artifacts are reduced.

### 2.3. Forming the estimate of the wideband spectrum

Suppose now we found the *longest* matching segments $M^{(n)}_{t,u:u+\tau_{max}}$ to the given narrowband speech $y_{t:t+\tau_{max}}$ at all $t$. Then, we form an estimate of the wideband expanded version of input speech spectra as follows:

$$\widehat{S}^{(w)}_{\varepsilon} = \frac{\sum_t A^{(w)}(u^t_{\varepsilon}) p(M^{(n)}_{t,u:u+\tau_{max}}|y_{t:t+\tau_{max}})}{\sum_t M^{(n)}_{t,u:u+\tau_{max}}|y_{t:t+\tau_{max}}}, \tag{6}$$

where $A^{(w)}(u^t_{\varepsilon})$ represents a prototype magnitude spectrum associated with the frame of wideband training dataset corresponding to $M^{(n)}_{t,u:u+\tau_{max}}$, where $u^t_{\varepsilon}$ indicates the most-likely time path $u = \{u, u+1, ..., u+\tau_{max}\}$ at time $t$. The estimate of the wideband mag-

nitude spectrum for time frame $\varepsilon$ is obtained by taking all the adjacent matched segments that contain $\varepsilon$ and, then, averaging $A^{(w)}(u_\varepsilon^t)$ over $t$. In the averaging, frames within the same segments share a common weight, which is the posterior probability of the segment. Notice that each frame is estimated through identification of the longest matching segment, then each wideband magnitude estimate is smoothed over successive longest matching segments. This improves robustness to imperfect segment matching, i.e., between observation segments and narrowband corpus segments.

## 3. SPEECH BANDWIDTH EXPANSION IN NOISY ENVIRONMENTS

In this section, we describe how the BWE framework described in the previous section can be generalized to the noise corrupted observation. To compensate the statistical mismatch between the *corpus-model* which is trained on a clean speech corpus and the input noisy observation, the well-known Vector Taylor Series (VTS) [13] noise adaptation algorithm is incorporated into the matching process. The work presented here tightly integrates the noise reduction and bandwidth expansion process using the single shared *corpus-model*.

### 3.1. An overview of environment model adaptation using vector Taylor series

Let us assume that in the time domain the clean speech $x[m]$ is corrupted by additive noise $n[m]$ and channel distortion $h[m]$:

$$y[m] = x[m] * h[m] + n[m], \tag{7}$$

where $y[m]$ is the corrupted speech signal. In the mel-frequency cepstral coefficient (MFCC) domain, this is equivalent to

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \mathbf{C}\log\left(1 + \exp\left(\mathbf{C}^{-1}\left(\mathbf{n} - \mathbf{x} - \mathbf{h}\right)\right)\right), \tag{8}$$

where $\mathbf{C}$ represents the discrete consign transform (DCT) matrix and $\mathbf{y}$, $\mathbf{x}$, $\mathbf{h}$, $\mathbf{n}$ the MFCC vectors corresponding to distorted speech, clean speech, channel, and noise, respectively. It is assumed that $\mathbf{x}$, $\mathbf{h}$, $\mathbf{n}$ are statistically independent and Gaussian with means $\mu_x$, $\mu_h$ and $\mu_n$ and covariance matrices $\mathbf{\Sigma}_x$, $\mathbf{\Sigma}_h$ and $\mathbf{\Sigma}_n$, respectively. The Jacobian of $\mathbf{y}$ in (7) with respect to $\mathbf{x}$, $\mathbf{h}$ and $\mathbf{n}$ evaluated at $\mu = \mu_n - \mu_x - \mu_h$ can be expressed as

$$\begin{aligned} \left.\frac{\partial y}{\partial x}\right|_{(\mu_x,\mu_h,\mu_n)} &= \left.\frac{\partial y}{\partial h}\right|_{(\mu_x,\mu_h,\mu_n)} = \mathbf{A}, \\ \left.\frac{\partial y}{\partial n}\right|_{(\mu_x,\mu_h,\mu_n)} &= \mathbf{I} - \mathbf{A} = \mathbf{F}, \end{aligned} \tag{9}$$

where

$$\mathbf{A} = \mathbf{C}\cdot\mathrm{diag}\left(\frac{1}{1+\exp\left(C^{-1}\left(\mu_x - \mu_h - \mu_n\right)\right)}\right)\cdot\mathbf{C}^{-1}, \tag{10}$$

and diag$(\cdot)$ in (10) represents the diagonal matrix whose elements equal to the value of the vector in the argument. Then, the nonlinear equation in (8) can be approximated by using a first order VTS expansion around $(\mu_x, \mu_h, \mu_n)$ as follows:

$$\mathbf{y} \approx \mu_x - \mu_h - \mathbf{g} + \mathbf{A}\left(\mathbf{x} - \mu_x\right) + \mathbf{A}\left(\mathbf{h} - \mu_h\right) + \mathbf{F}\left(\mathbf{n} - \mu_n\right), \tag{11}$$

where

$$\mathbf{g} = \mathbf{C}\log\left(1 + \exp\left(\mathbf{C}^{-1}\left(\mu_x - \mu_h - \mu_n\right)\right)\right). \tag{12}$$

By taking the expectation of (11), the mean of $\mathbf{y}$, $\mu_y$, and its covariance matrix $\mathbf{\Sigma}_y$ can be obtained as

$$\mu_y \approx \mu_x + \mu_y + \mathbf{g}\left(\mu_n - \mu_x - \mu_h\right), \tag{13}$$

$$\mathbf{\Sigma}_y \approx \mathbf{A}\Sigma_x\mathbf{A}^T + \mathbf{A}\Sigma_h\mathbf{A}^T + \mathbf{F}\Sigma_n\mathbf{F}^T. \tag{14}$$

### 3.2. Integrating BWE with environment adaptation

Given the input narrowband noisy speech, the model $G^{(n)}$ estimated from clean training data cannot provide the proper a prior to the matching process. Thus, to adapt the model to the target noisy environment, we introduce the VTS model adaptation scheme described in the previous section to the matching process. The updated model $\hat{G}^{(n)}$ to the target environment is estimated as follows:

$$\hat{G}^{(n)} = \sum_{k=1}^{K} \lambda_k N(\mathbf{s}; \hat{\mu}_k, \hat{\mathbf{\Sigma}}_k), \tag{15}$$

where $\hat{\mu}_{y_k} \approx \mu_{x_k} + \mu_{y_k} + \mathbf{g}_k\left(\mu_n - \mu_{x_k} - \mu_{h_k}\right)$. A single Gaussian is assumed for the noise model. Assuming $\mathbf{x}$ and $\mathbf{n}$ are independent, and given the noise covariance $\mathbf{\Sigma}_k$ on clean training data, the covariance matrix of $k$th component adapted to the target environment is computed as $\hat{\mathbf{\Sigma}}_{y_k} \approx \mathbf{A}_k\Sigma_{x_k}\mathbf{A}_k^T + \mathbf{A}_k\Sigma_{h_k}\mathbf{A}_k^T + \mathbf{F}_k\Sigma_n\mathbf{F}_k^T$. The sequence model $M^{(n)}$, which defines the temporal dynamics in training corpus, is not changed so that it enables to find the corresponding speech segments in clean training corpus given the noise corrupted speech. Notice that the posterior probability $p(M_{u:u+\tau}^{(n)}|y_{t:t+\tau})$ given in (4) is now calculated based upon $\hat{G}^{(n)}$, which is adapted by the noise statistics estimated from each test utterance. Finally, $A^{(w)}$ is referred to fetch the corresponding spectrum patches which are stored in forms of prototype clean and wideband magnitude spectrum.

## 4. EXPERIMENTS AND ANALYSIS

To test the effectiveness of the proposed BWE algorithm, we performed a series of experiments on clean and noise-corrupted narrowband speech. To train the *corpus-model*, we utilized the training set from the TIMIT database [14]. To create a parallel narrowband corpus, the speech was downsampled to 8 kHz and filtered according to the G.712 telephony channel specification. The training data for the GMM $G^{(n)}$ consists of 1088 sentences produced by 136 speakers. Narrowband features were created by extracting 42-dim log mel spectral vectors from the power spectrum, and then converting these to 40-dim cepstra, including c0. We used feature vectors with higher dimensionality compared to the one typically used for speech recognition in order to retain more detail of both the spectral envelope and harmonic components. The narrowband cepstra were used to train a GMM $G^{(n)}$ with 4096 Gaussian components using conventional expectation and maximization scheme. The sequence model $M^{(n)}$ was trained with the training data size $Is = 338,546$.

To evaluate the proposed BWE algorithm on noisy speech, the test narrowband speech was mixed with samples of noise from NoiseX92 corpus [15] using FaNT [16], which filters speech data with a frequency characteristic as defined by ITU for telephone equipment, e.g., G.712 and/or adds noise to speech recordings at a desired SNR. Car and babble noise were used for evaluation with SNRs between 0 and 25 dB with 5 dB intervals. When VTS adaptation was performed, the noise model parameters were estimated using the first and last 10 frames, 20 frames in total, of each utterance.

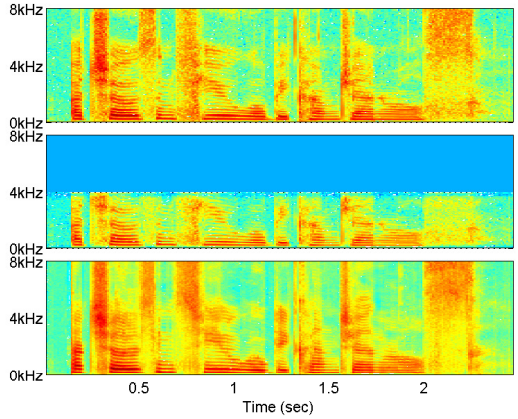


**Fig. 2**. Spectrum comparison on clean test data: (a) original wideband speech, (b) clean narrowband speech, (c) extended wideband speech

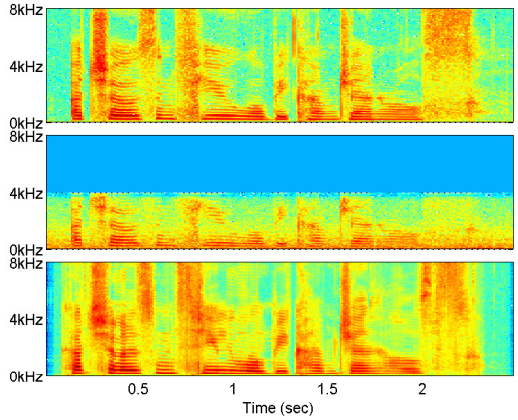


**Fig. 3**. Spectrum comparison on noisy test data: (a) original wideband speech, (b) noisy narrowband speech (car noise with 5dB SNR), (c) extended wideband speech.

We first evaluated the proposed BWE algorithm on clean narrowband speech. A test corpus of telephone speech was created from TIMIT female core test set, which consists of 64 sentences produced by 8 speakers. These utterances were converted to narrowband speech in the manner describe above. Fig. 2 shows an example of reconstructed wideband speech compared to the observed narrowband speech and original wideband speech. Though differences in test data make direct comparison difficult, the performance of the proposed BWE algorithm on clean speech is comparable to that of other recently proposed BWE algorithms in the literature [9, 11]. In particular, compared to the previous studies in which the largest errors were found in sibilant sounds or other fricatives [17], the proposed approach had little degradation on those classes. We hypothesize that identifying the matching criterion based on long-range speech dynamics involves more contextual dependencies, thus, compensates the defect of little amount of information available in lower-band frequencies. Fig. 3 shows an example of BWE given the noisy narrowband observation. The noise statistics estimated in each test utterance were used to update the pre-trained model $G^{(n)}$ in test phase. Notice that the system tightly integrates the noise reduction and bandwidth extension process in a unified framework, where the output spectrum is reproduced using the clean and wideband magnitude spectrum segments stored in a speech corpus.

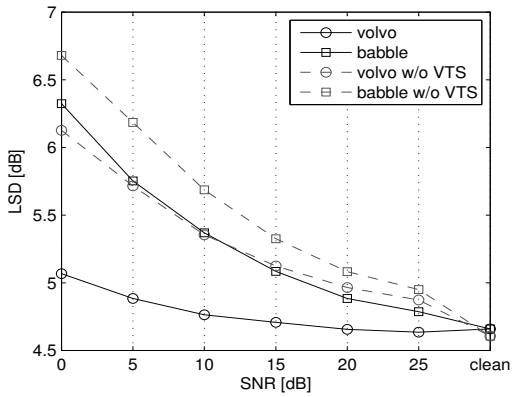


**Fig. 4**. Spectral distortion of the extended high frequency spectrum vs. SNR obtained from noise-corrupted narrowband speech.

To further validate the system on noisy observation, log-spectral distance (LSD) between original and estimated wideband speech was measured against various SNRs. The LSD was calculated from FFT power spectra using the formula

$$LSD_i = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left[10\log_{10}\frac{P_i(k)}{\widehat{P}_i(k)}\right]^2}, \qquad (16)$$

where $LSD_i$ is the spectral distortion of frame $i$, $P_i$ is the power spectrum of the ground-truth clean wideband at $i$-th frame, and $\widehat{P}_i(k)$ is the power spectrum of reconstructed wideband speech using the proposed BWE algorithm. $K$ is the number of FFT bins corresponding to wideband frequency range. Fig. 4 shows the results using car and babble noise, which has relatively stationary and non-stationary characteristics, respectively. When we apply the corpus-model based BWE without integrating VTS scheme, the noise corrupted narrowband speech induced large errors in matching process. This resulted in large LSD between the estimated wideband spectrum and the original one. The additional noise compensation process, i.e., VTS model adaptation, mitigated the degradation on both car and babble noise corrupted speech signals. Since the car noise is relatively more stationary than the babble noise, it took more advantage from VTS compensation. On the clean narrowband speech, applying the noise adaptation scheme resulted in a negligible degradation in performance.

## 5. CONCLUSION

In this study, a new algorithm of speech bandwidth expansion was proposed based upon a recently proposed corpus-model based speech reconstruction algorithm. In contrast to memoryless decoding or HMM-based estimation with additional information from adjacent frames, the proposed method tried to exploit the long-range, up to a sentence length, speech dynamics to synthesize the wideband speech. It was realized by maximizing the posterior probability in the matching process with segment examples pre-stored in a speaker independent database. The approach was generalized to noise corrupted narrowband speech input and Vector Taylor Series (VTS) was incorporated to combine bandwidth expansion with noise reduction in a unified framework. Both clean and noise corrupted narrowband speech signals were tested and the feasibility of the proposed algorithm was verified.

## 6. REFERENCES


[1] Mattias Nilsson, Soren Vang Andersen, and W Bastiaan Kleijn, "On the mutual information between frequency bands in speech," in *Proc. ICASSP*, 2000, vol. 3, pp. 1327–1330.

[2] Yannis Agiomyrgiannakis and Yannis Stylianou, "Combined estimation/coding of highband spectral envelopes for speech spectrum expansion," in *Proc. ICASSP*, 2004, vol. 1, pp. I–469.

[3] Yan Ming Cheng, Douglas O'Shaughnessy, and Paul Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 544–548, 1994.

[4] Kun-Youl Park and Hyung Soon Kim, "Narrowband to wideband conversion of speech using gmm based transformation," in *Proc. ICASSP*, 2000, vol. 3, pp. 1843–1846.

[5] Peter Jax and Peter Vary, "Artificial bandwidth extension of speech signals using mmse estimation based on a hidden markov model," in *Proc. ICASSP*, 2003, vol. 1, pp. I–680.

[6] Geun-Bae Song and Pavel Martynovich, "A study of hmm-based bandwidth extension of speech signals," *Signal Processing*, vol. 89, no. 10, pp. 2036–2044, 2009.

[7] Niklas Enbom and W Bastiaan Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," in *Proc. IEEE Workshop on Speech Coding*, 1999, pp. 171–173.

[8] Yoshihisa Nakatoh, Mineo Tsushima, and Takeshi Norimatsu, "Generation of broadband speech from narrowband speech based on linear mapping," *Electronics and Communications in Japan (Part II: Electronics)*, vol. 85, no. 8, pp. 44–53, 2002.

[9] Sheng Yao and Cheung-Fat Chan, "Block-based bandwidth extension of narrowband speech signal by using cdhmm," in *Proc. ICASSP*, 2005, pp. 1793–1796.

[10] Dhananjay Bansal, Bhiksha Raj, and Paris Smaragdis, "Bandwidth expansion of narrowband speech using non-negative matrix factorization.," in *Proc. Interspeech*, 2005, pp. 1505–1508.

[11] Amr H Nour-Eldin and Peter Kabal, "Memory-based approximation of the gaussian mixture model framework for bandwidth extension of narrowband speech.," in *Proc. Interspeech*, 2011, pp. 1185–1188.

[12] Ji Ming, "Maximizing the continuity in segmentation-a new approach to model, segment and recognize speech," in *Proc. ICASSP*, 2009, pp. 3849–3852.

[13] Alex Acero, Li Deng, Trausti T Kristjansson, and Jerry Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition.," in *Proc. Interspeech*, 2000, pp. 869–872.

[14] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The darpa speech recognition research database: specifications and status," in *Proc. DARPA Workshop on speech recognition*, 1986, pp. 93–99.

[15] Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[16] H-Guenter Hirsch, "Fant-filtering and noise adding tool," 2005.

[17] Hannu Pulakka, Laura Laaksonen, Martti Vainio, Jouni Pohjalainen, and Paavo Alku, "Evaluation of an artificial speech bandwidth extension method in three languages," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1124–1137, 2008.