Signal Processing, Vol. 82, 2002, pp. 205-227

Nonstationary-State Hidden Markov Model Representation of Speech Signals for Speech Enhancement

*Hossein Sameti and Li Deng**

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

Current Address: Li Deng, Microsoft Research, Redmond, WA; E-mail: deng@microsoft.com

Running title: Nonstationary-State HMM for Speech Enhancement

Keywords: speech enhancement, noise removal, nonstationary-state hidden Markov model, minimum mean square error estimate

Abstract

A novel formulation of the nonstationary-state hidden Markov model (NS-HMM), employed as the speech model and serving as the theoretical basis for the construction of a speech enhancement system, is presented in this paper. The NS-HMM is used as a compact and parametric model, generalized from the stationary-state HMM, for describing the clean speech statistics in deriving the minimum mean square error (MMSE) estimator. The feature selection problem associated with the use of the NS-HMM in designing the speech enhancement system is addressed. The MMSE formulation is established where the NS-HMM is used as the clean speech model and Gaussian-mixture, stationary-state HMM as the additive noise model. Speech enhancement experiments are conducted, demonstrating superiority of the NS-HMM over the stationary-state HMM in the speech enhancement performance for low SNRs. Detailed diagnostic analysis on the speech enhancement system's operation shows that the superiority arises from the ability of the NS-HMM to fit the spectral trajectory of the signal embedded in noise more closely than the stationary-state HMM.

==========

* Current address for correspondence: Dr. Li Deng, Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

1 Introduction

Current speech enhancement techniques can be categorized into two major classes: the model-free methods and the model-based methods. The model-free techniques are deficient in several aspects compared to the model-based methods. Some model-free techniques need to use two microphones for both noise and speech recordings[1]. This is usually not possible, especially in on-line enhancement applications (e.g. in hearing-aid applications). Some others use the assumed periodicity of speech for filter design[2, 3, 4]. This causes difficulties in removing the noise from nonperiodic utterances such as fricatives. The source of the problems associated with model-free methods is the unreasonable assumption for the noise being relatively stationary. The spectral subtraction technique[5, 6, 7] is considered as the most efficient model-free speech enhancement method. This technique updates the noise estimate using the data occurring during speech-silence periods or nonspeech segments, enabling noise nonstationarity to be handled to some extent. However, the results of the spectral subtraction are usually unsatisfactory when noise characteristics change relatively fast or when the nonspeech segments do not occur often enough. Further, in applying the spectral subtraction method, the process of dynamic reduction of spectral energy necessarily introduces the audible "musical"-like artifact acting as signal dependent interference[8]. Several attempts have been made to deal with the problem of musical noise and to handle noise nonstationarity more efficiently [9], with some limited success. Unlike model-free methods where musical noise is an inherent problem requiring special techniques to overcome, it is desirable to avoid the musical-noise problem from the very beginning and model-based methods are inherently free from the problem.

In our previous works [8, 10], we focused on hidden Markov model (HMM) based speech enhancement algorithms[11, 12]. We improved the minimum mean square error (MMSE) enhancement method presented by Ephraim[13, 14] by three major modifications: 1) incorporating of mixture components in the noise HMM in order to better handle noise nonstationarity; 2) applying two efficient methods in the enhancement algorithm implementation that make the system real-time implementable; and 3) devising an adaptation method to the noise type in order to accommodate a wide variety of noise types expected under the enhancement system's operating environment. Our experimental results on comparisons between the performance of the spectral subtraction and HMM-based speech enhancement methods were reported in [10].

The study described in this paper represents a significant step in improving the speech enhancement system from that based on the conventional, stationary-state HMM[10] to that constructed using the nonstationary-state HMM (NS-HMM) as the model for clean speech. (The latter HMM includes the former HMM as a special, degenerate case). Although the stationary-state HMM is

itself a generally nonstationary model (via its multiple states whose transition probabilities do not happen to trap into a stationary distribution), it assumes that within a given state the speech data is stationary. No mechanism is provided in the stationary-state HMM setup to handle detailed variations of intrinsically dynamic speech signals and their temporal relationships. In the earlier work of Deng *et al.*, [15, 16, 17] the NS-HMM was developed and evaluated for speech recognition applications. In this model, the dynamic nature of the acoustic signals in the speech is described in terms of statistical nonstationarity which is hierarchically organized at two distinct levels. At the global level, the nonstationarity is modeled when phonetic contents change over time in a relatively slow fashion. A Markov chain is used to describe this change. The local nonstationarity, on the other hand, is described by state-conditioned regression functions on time explicitly.

The experimental results published in [16, 17] show evidence that the NS-HMM fits the detailed temporal variations of the speech data much better than the stationary-state HMM. Since an accurate estimate of the clean speech spectra from their noisy version is essential for speech enhancement purposes, we expect that the superior data-fitting capability provided by the NS-HMM will translate to superior performance in speech enhancement as well. Of course, compared with the speech recognition experiments where the main concern is to solve the mismatch problem in fitting models to speech data under disparate training and testing conditions (noise-free), the speech enhancement task involves an additional complexity of estimating the clean speech spectrum from its noisy version. In this paper, we formulate a general framework for MMSE speech enhancement incorporating the NS-HMM as the model representing statistical characteristics of the clean speech spectra. It is an extension of the framework with use of stationary-state HMMs published in [13]. The results of the diagnostic experiments and some limited speech enhancement experiments (described in Sections 4.2 and 4.3) have corroborated the results of the previous studies on the superiority of the NS-HMM over the the stationary-state HMM as a more appropriate model for speech signals. The scope of this paper, however, is mainly centered on the formulation of the NS-HMM speech enhancement system, as well as on the design and implementation of the formulated system. Comprehensive evaluation of the system, especially with use of subjective criteria and with use of real-world noisy speech materials, will be our future work.

The organization of this paper is as follows. In Section 2 we first address the issue of speech feature selection, which is tightly associated with use of the NS-HMM as the speech model, and then provide a detailed description of the MMSE formulation for speech enhancement in which NS-HMMs are employed. Section 3 contains implementation details of the speech enhancement algorithm developed based on the above theoretical formulation. The speech enhancement system and its evaluation are reported in Section 4, where two sets of diagnostic experiments and some limited speech enhancement experiments are described. The experimental results show superior

performance of the new system over the earlier version of the system with use of only the conventional, stationary-state HMM (i.e., degenerate case of the NS-HMM) as the speech model. Finally, in Section 5, a summary of the study is provided and conclusions are drawn.

2 Nonstationary-State HMM as the Speech Model in the MMSE Enhancement System

The standard stationary-state HMM, which is widely used for speech modeling both in speech recognition and enhancement applications, accommodates the inherent nonstationarity of speech by introducing different states within the model with the state transitions governed by the Markovian property. In this model, a signal is assumed to be stationary given a state. Nonstationarity is modeled by transition among different states, thereby changing the statistical properties over time. There are two identifiable problems with the standard HMM for speech enhancement applications. First, in order to accommodate continuously and rapidly varying characteristics of the speech signal, it is necessary to increase the number of states in the model. This requires an undesirably large size of the model while accepting the risk of impairing the performance when the signal variation is slower than anticipated. Second, even with an increased number of states, the continuity of the speech features is subject to deterioration with the standard HMM, since within each state a constant mean is assumed for the signal and this value can be noticeably different for consecutive states.

2.1 Formulation of the Nonstationary-State HMM

The NS-HMM introduced in [17] and further investigated in [15] has been intended to overcome the weaknesses of the standard HMM by permitting nonstationarity within the states of the HMM. In the NS-HMM, the observation vector, $\mathbf{y}_t$, is composed of a deterministic trend function plus the residual

$$\mathbf{y}_t = \mathbf{f}_t + N_t, \tag{1}$$

where $\mathbf{f}_t$ is the deterministic trend function at time frame t, and N_t is the stationary residual. $\mathbf{y}_t$, $\mathbf{f}_t$, and N_t are all $K \times 1$ vectors. N_t is taken to be an IID, zero-mean Gaussian source. A NS-HMM is completely characterized by the following parameter set: 1) Transition probabilities, $a \triangleq \{a_{ij}\}$, $i, j = 1, 2, \ldots, M$ of the Markov chain; 2) Mixture weights, $c \triangleq \left\{c_{m|i}\right\}$, $i = 1, 2, \ldots, M, m = 1, 2, \ldots, L$ of the Markov chain with a total of M states and L mixture components; 3) Parameters $\Theta \triangleq \{\Theta_{i,m}\}, i = 1, 2, \ldots, M, m = 1, 2, \ldots, L$ in the deterministic trend

function $\mathbf{f}_t(\Theta_{i,m})$, dependent on state i and mixture component m in the Markov chain; and 4) Covariance matrices, $\Sigma \stackrel{\triangle}{=} \{\Sigma_{i,m}\}, i = 1, 2, \ldots, M, m = 1, 2, \ldots, L$, of the zero-mean Gaussian IID residual $N_t(0, \Sigma_{i,m})$ which are also state and mixture-component dependent.

Given the above model parameters, the observation vector sequence, $\mathbf{y}_t$ (indexed by (i, m)), $t = 0, 1, \ldots, T-1$ is generated from the model according to

$$\mathbf{y}_t^{(i,m)} = \mathbf{f}_t(\Theta_{i,m}) + N_t(0, \Sigma_{i,m}), \tag{2}$$

where state i and mixture component m at a given time frame t is determined by the evolution of the Markov chain characterized by a_{ij} and $c_{m|i}$.

The mixture version of the NS-HMM has the same underlying Markov chain as in the conventional, stationary-state HMM. In the NS-HMM used in this study, the time-varying means are expressed explicitly as polynomials of state-sojourn time. We thus have

$$\mathbf{y}_t^{(i,m)} = \sum_{r=0}^{R} B_{i,m}(r)\mathbf{h}_r(t - \tau_i) + N_t(0, \Sigma_{i,m}), \tag{3}$$

where the first term is the state(i) and mixture component (m) dependent polynomial regression function (order R) with $B_{i,m}(r)$ being the polynomial coefficients[1] with τ_i registering the time when state i in the HMM is just entered before regression on time takes place. $\mathbf{h}_r(.)$ is an r-th order polynomial function. In our model implementation, we choose orthogonal polynomials for their superior stability properties in parameter estimation. The final term in Eqn.(3) is the residual noise assumed to be the output of an IID, zero-mean Gaussian source with state-dependent, mixture-component-dependent, but time-invariant covariance matrix $\Sigma_{i,m}$. Note that in Eqn.(3) only the covariance matrix $\Sigma_{i,m}$, and polynomial coefficients $B_{i,m}(r)$ (for state i, mixture component m, and polynomial order r) are considered as *true* model parameters; τ_i is merely an auxiliary parameter for the purpose of obtaining maximal accuracy in estimating $B_{i,m}(r)$ (over all possible τ_i values).

In order to automatically train the NS-HMM parameters from the clean speech data, an efficient algorithm has been developed. The algorithm is motivated by and is an extension of the segmental k-means algorithm developed in the past for training conventional HMMs [18]. Like the segmental k-means algorithm, the algorithm developed here also involves two iterative steps — the segmentation step and the regression step. For speech enhancement, we use a fully-connected (ergodic) HMM for speech. Therefore, for the segmentation step we used a modified version of the segmentation algorithm described in [19] to take into account the changes for fully-connected HMM. The reestimation (regression) step here is identical to that of [19].

[1]To take account of all components of vector $\mathbf{y}_t$ of dimension K, we have a matrix with dimension $K \times (R+1)$ for all the polynomial coefficients.

2.2 Speech Feature Selection

For every task in speech processing, including speech recognition and enhancement, selection of suitable features to represent speech data is crucial. The features we use to represent speech should preserve the information of the speech data that the task needs. We discuss some issues on feature selection specific to the speech enhancement task using the NS-HMM approach pertaining to this work.

Reversibility of the features is a major issue. While removing speech redundancy, some useful information in the data may be lost and this complicates the reversibility issue. For reconstructing the speech data the lost information may have to be estimated. Generally, for speech recognition, reversibility is less of a consideration than for speech enhancement because the speech waveforms are not to be reconstructed once the recognizer has selected the matching speech units. Thus for speech recognition, it is desirable to reduce the variability of the data resulting from various speakers' characteristics, accents, contexts, etc.. In contrast, for speech enhancement, we would like to preserve these variabilities and are ultimately interested in reconstructing the speech waveform as faithfully as possible. In the speech enhancement system described in this paper (see Section 3 for details), the selected speech features are used to construct weighted, time-varying Wiener filters. The filtering is then performed in frequency domain and the frequency-domain filtered speech is transformed to the time-domain signal. The feature reversibility issue is therefore translated to the issue of transformability to spectral – or frequency – domain that enables the construction of the Wiener filters.

Aside from the reversibility issue, another constraint in selecting a suitable feature for a specific speech processing task is the additivity of the features as required by the formulation of the noisy speech model. The assumption of speech-noise additivity and lack of correlation between the speech and the noise degrading the speech impose strong constraints on the speech features being fed to the model. The selected feature has to preserve such additivity properties in order to be consistent with the problem formulation.

The discrete Fourier transform (DFT) represents speech characteristics in frequency-domain in a direct manner. Using the short time Fourier transform (STFT) the variations in time can be captured directly. For further compressing the STFT data, mel-frequency representations may seem to be appropriate. By passing the squared spectral magnitude of the signal through the mel-frequency triangular filter-banks and calculating the log energy output of each filter, the mel-frequency spectral coefficients (MFSC)[20] are obtained. Carrying out a cosine transform on the MFSC yields the mel-frequency cepstral coefficients (MFCC)[20]. The MFSC, MFCC, and delta MFCC have resulted in good performance and have become virtually standard features in speech

recognition. For enhancement, unfortunately, the reversibility issue becomes a serious problem for either MFSC or MFCC/delta-MFCC since the averaging effect of the triangular filters makes the reversibility problematic. Moreover, MFSC and MFCC/delta-MFCC are not consistent with speech-noise additivity assumption, because the log operation makes the transformation nonlinear and consequently the noisy speech cannot be a simple sum of speech and noise in the MFSC or the MFCC domain even if it is so in the time domain.

Another common representation of speech is the linear predictive coefficients (LPC)[21] where it is assumed that speech is an auto-regressive (AR) process produced by feeding a Gaussian noise process as input to an all-pole transfer function. The magnitude of the speech spectrum can be easily found from the AR coefficients and the time-domain data can be reconstructed using the original phase of the signal (noisy signal phase for the enhancement case), but the LPC coefficients do not have the additivity property. Since autocorrelation coefficients (R_{xx}) are transformable to AR coefficients and vice-versa, the autocorrelation coefficients are considered faithfully reversible, too. Experiments we have performed to test the reversibility of autocorrelation coefficients have confirmed this reversibility.

Finally, the smoothness of the selected feature versus time is also an important issue when that feature is to be modeled using the NS-HMM. Nonstationary-state hidden Markov modeling is based on the assumption of smooth variation of the mean of the speech data trajectories in a way that can be represented by a deterministic trend function. Since the AR coefficients do not have the smoothness property assumed in speech features for NS-HMM, the concept of the deterministic trend function added to a Gaussian random process does not match the AR coefficients, whereas for autocorrelation coefficients this continuity is a strong inherent property. Due to their nonlinear transformation, AR coefficients do not preserve the time-domain additivity. Line spectrum pair (LSP) parameters are a linear function of the LPC polynomial and thus have the same properties as AR coefficients; i.e., they are non-smooth and non-additive. For the features discussed earlier, their smoothness property is briefly described here. The DFT magnitude is a smooth function of time (frame). Similarly, MFSC and MFCC are also smooth enough to be modeled by the NS-HMM. The reflection coefficients k's are smooth and reversible like autocorrelation coefficients, but they are not additive due to the nonlinear functions involved in their calculation. However, this non-additivity is not a serious problem because the reversibility allows the power spectrum to be calculated uniquely from the reflection coefficients. Since the spectrum domain features are additive, Wiener filter computations can be performed in the spectrum domain. This means that the k coefficients in principle can also be used as a feature in our NS-HMM speech enhancement

system due to their smoothness and reversibility. [2]

A summary of the properties of the different features discussed above for the representation of speech data is provided in Table 1. The above discussion and the properties shown in Table 1 suggest that a number of features are suitable for the NS-HMM representation in use for speech enhancement; such features include autocorrelation coefficients. In this paper, we will describe the speech enhancement system employing only DFT magnitudes as the feature, with the results to be presented in Section 4.

2.3 MMSE Formulation with Nonstationary-State HMM as the Model for Clean Speech

Assume the clean speech signal in a chosen feature domain $\mathbf{y} \stackrel{\triangle}{=} \{\mathbf{y}_t, t = 0, \ldots, T-1\}$ is corrupted with independent additive noise $\mathbf{v} \stackrel{\triangle}{=} \{\mathbf{v}_t, t = 0, \ldots, T-1\}$, resulting in the noisy speech signal $\mathbf{z} \stackrel{\triangle}{=} \{\mathbf{z}_t = \mathbf{y}_t + \mathbf{v}_t, t = 0, \ldots, T-1\}$. Let $\mathbf{y}_0^t, \mathbf{v}_0^t$, and $\mathbf{z}_0^t$ denote the vectors of clean speech, noise, and noisy speech processes, respectively, from time frame 0 to t. Let λ_y, λ_v, and λ_z denote the model parameters for the clean speech, noise and noisy speech, respectively. Further, let $\hat{\mathbf{y}} \stackrel{\triangle}{=} \{\hat{\mathbf{y}}_t, t = 0, \ldots, T-1\}$ denote the MMSE estimate of the clean speech sequence. For the mean square error

$$\text{Error} = E[(\hat{\mathbf{y}}_t - \mathbf{y}_t)^{Tr}(\hat{\mathbf{y}}_t - \mathbf{y}_t)] \tag{4}$$

to be minimized, $\hat{\mathbf{y}}_t$ should be:

$$\hat{\mathbf{y}}_t = E\{\mathbf{y}_t \mid \mathbf{z}_0^t\}, \tag{5}$$

according to the well-established estimation theory, where $E\{\cdot \mid \cdot\}$ denotes conditional expectation. We are interested in estimating a function g of the clean speech. The optimal estimate[3] is

$$\begin{aligned} \hat{g(\mathbf{y}_t)} &= E\{g(\mathbf{y}_t) \mid \mathbf{z}_0^t\} \\ &= \int_{-\infty}^{\infty} g(\mathbf{y}_t) f_{\lambda_y|\lambda_z}(\mathbf{y}_t \mid \mathbf{z}_0^t) d\mathbf{y}_t, \end{aligned} \tag{6}$$

where $f_{\lambda_y|\lambda_z}$ denotes the conditional probability density function defined according to the model parameters λ_y and λ_z. Using Bayes' rule we can write:

$$f_{\lambda_y|\lambda_z}(\mathbf{y}_t \mid \mathbf{z}_0^t) = \frac{f_{\lambda_z|\lambda_y}(\mathbf{z}_0^t \mid \mathbf{y}_t) f_{\lambda_y}(\mathbf{y}_t)}{f_{\lambda_z}(\mathbf{z}_0^t)}. \tag{7}$$

[2] We have not explored such a possibility in the current study.

[3] Note that $\hat{g(\mathbf{y}_t)}$ is an estimate of function $g(\mathbf{y_t})$, and is dependent on noisy speech $\mathbf{z}_0^t$.

Substituton of Eqn.(7) into Eqn.(6) gives:

$$E\{g(\mathbf{y}_t) \mid \mathbf{z}_0^t\} = \frac{\int_{-\infty}^{\infty} g(\mathbf{y}_t) f_{\lambda_z|\lambda_y}(\mathbf{z}_0^t \mid \mathbf{y}_t) f_{\lambda_y}(\mathbf{y}_t) d\mathbf{y}_t}{f_{\lambda_z}(\mathbf{z}_0^t)}. \tag{8}$$

For two independent random variables $\mathbf{y}$ and $\mathbf{v}$, the pdf of their sum $\mathbf{z}$ equals the convolution of their respective pdf's.[4] Hence,

$$\begin{aligned} f_{\lambda_z|\lambda_y}(\mathbf{z}_0^t \mid \mathbf{y}_t) &= \int_{-\infty}^{\infty} f_{\lambda_y}(\mathbf{y}_0^t \mid \mathbf{y}_t) f_{\lambda_v|\lambda_y}(\mathbf{z}_0^t - \mathbf{y}_0^t \mid \mathbf{y}_t) d\mathbf{y}_0^t \\ &= \int_{-\infty}^{\infty} f_{\lambda_y}(\mathbf{y}_0^{t-1}) f_{\lambda_v}(\mathbf{v}_0^t) d\mathbf{y}_0^t \\ &= f_{\lambda_v}(\mathbf{v}_t) \int_{-\infty}^{\infty} f_{\lambda_y}(\mathbf{y}_0^{t-1}) f_{\lambda_v}(\mathbf{v}_0^{t-1}) d\mathbf{y}_0^t \\ &= f_{\lambda_v}(\mathbf{v}_t) f_{\lambda_z}(\mathbf{z}_0^{t-1}), \end{aligned} \tag{9}$$

Note that in the above we assume $f_{\lambda_y}(\mathbf{y}_0^t \mid \mathbf{y}_t) = f_{\lambda_y}(\mathbf{y}_0^{t-1})$. This implies independence[5] of vectors $\mathbf{y}_t$ for different time frame t (given the model parameters λ_y), as assumed in all types of HMMs including our NS-HMM.[6] However, the key difference between the NS-HMM and the conventional HMM is the different parameter sets λ_y, one defining smoothly-varying trajectories and the other does not. Given such parameterization, the $\mathbf{y}_t$ vectors necessarily exhibit the dynamic behavior in the NS-HMM despite the independence assumption on the residual noise sequence (and hence independence assumption on $\mathbf{y}_t$ given λ_y).

By substituting Eqn.(9) into Eqn.(8), we obtain:

$$E\{g(\mathbf{y}_t) \mid \mathbf{z}_0^t\} = \frac{\int_{-\infty}^{\infty} g(\mathbf{y}_t) f_{\lambda_v}(\mathbf{v}_t) f_{\lambda_z}(\mathbf{z}_0^{t-1}) f_{\lambda_y}(\mathbf{y}_t) d\mathbf{y}_t}{f_{\lambda_z}(\mathbf{z}_0^t)}. \tag{10}$$

Note that Eqn.(10) above is a general result with no assumptions on any particular models used to describe the statistics of the speech and noise signals. In order to devise a practical speech enhancement algorithm, some modeling assumptions are needed. In this work, the NS-HMM is assumed for the short-time DFT (magnitude) sequences of clean speech, and the stationary-state

[4]While in the DFT magnitude domain the additivity and statistical independence do not hold mathematically for general signals $\mathbf{y}$ and $\mathbf{v}$, as a good approximation they are practically valid for the signals consisting of speech in noise. The evidence for this validity comes from the reasonable success of the popular spectral subtraction method which works in the DFT magnitude domain.

[5]Note that $f_{\lambda_y}(\mathbf{y}_0^t \mid \mathbf{y}_t) = f_{\lambda_y}(\mathbf{y}_0^{t-1}, \mathbf{y}_t \mid \mathbf{y}_t) = f_{\lambda_y}(\mathbf{y}_0^{t-1}, \mathbf{y}_t)/f_{\lambda_y}(\mathbf{y}_t)$, and hence independence.

[6]The reason for such independence comes from the assumption of i.i.d. residual noise in the definition of the trajectory as in Eqn.(2). This i.i.d. assumption does not affect the essence of the model which uses deterministic trajectories to describe the speech dynamics.

HMM used for the short-time DFT sequences of noise. Given such modeling assumptions, the pdf's in Eqn.(10) can be written out in detail as follows. $f_{\lambda_y}(\mathbf{y}_t)$ and $f_{\lambda_v}(\mathbf{v}_t)$ can be written, according the law of total probability, as

$$\begin{aligned} f_{\lambda_y}(\mathbf{y}_t) &= \sum_{\beta=1}^{M}\sum_{\gamma=1}^{L}\sum_{d=1}^{t} f_{\lambda_y}(\mathbf{y}_t \mid s_t=\beta, m_t=\gamma, d_t=d)\cdot \\ &\quad f_{\lambda_y}(s_t=\beta, m_t=\gamma, d_t=d) \qquad (11) \\ f_{\lambda_v}(\mathbf{v}_t) &= \sum_{\xi=1}^{N}\sum_{\delta=1}^{P} f_{\lambda_v}(\mathbf{v}_t \mid n_t=\xi, p_t=\delta)\cdot f_{\lambda_v}(n_t=\xi, p_t=\delta), \qquad (12) \end{aligned}$$

where s_t and m_t denote the speech state and mixture component at time t, respectively, d_t is the duration of state s_t from the time of entry into that state to time t, and n_t and p_t denote the noise state and mixture component at time t, respectively. Similarly, $f_{\lambda_z}(\mathbf{z}_0^t)$ can be written as:

$$\begin{aligned} f_{\lambda_z}(\mathbf{z}_0^t) &= \sum_{\mathcal{S}_0^t}\sum_{\mathcal{M}_0^t}\sum_{\mathcal{N}_0^t}\sum_{\mathcal{P}_0^t} f(\mathcal{S}_0^t, \mathcal{M}_0^t, \mathcal{N}_0^t, \mathcal{P}_0^t, \mathbf{z}_0^t) \\ &= \sum_{\mathcal{S}_0^t}\sum_{\mathcal{M}_0^t}\sum_{\mathcal{N}_0^t}\sum_{\mathcal{P}_0^t} f_s(\mathcal{S}_0^t) f_m(\mathcal{M}_0^t \mid \mathcal{S}_0^t) f_n(\mathcal{N}_0^t \mid \mathcal{M}_0^t, \mathcal{S}_0^t) \\ &\quad f_p(\mathcal{P}_0^t \mid \mathcal{N}_0^t, \mathcal{M}_0^t, \mathcal{S}_0^t) f_{\lambda_z}(\mathbf{z}_0^t \mid \mathcal{P}_0^t, \mathcal{N}_0^t, \mathcal{M}_0^t, \mathcal{S}_0^t), \qquad (13) \end{aligned}$$

where $f_s(\mathcal{S}_0^t)$ is the probability of the clean speech state sequence $\mathcal{S}_0^t$, $f_m(\mathcal{M}_0^t \mid \mathcal{S}_0^t)$ is the probability of the sequence of clean speech mixture-component sequence $\mathcal{M}_0^t$ given the state sequence of clean speech $\mathcal{S}_0^t$, $f_n(\mathcal{N}_0^t \mid \mathcal{M}_0^t, \mathcal{S}_0^t)$ (which, due to independence of the clean speech and noise sequences, equals $f_n(\mathcal{N}_0^t)$) is the probability of the noise state sequence, $f_p(\mathcal{P}_0^t \mid \mathcal{N}_0^t, \mathcal{M}_0^t, \mathcal{S}_0^t) = f_p(\mathcal{P}_0^t \mid \mathcal{N}_0^t)$ is the probability of the noise mixture-component sequence given noise state sequence, and $f_{\lambda_z}(\mathbf{z}_0^t \mid \mathcal{P}_0^t, \mathcal{N}_0^t, \mathcal{M}_0^t, \mathcal{S}_0^t)$ is the pdf of the noisy speech output sequence $\mathbf{z}_0^t$ given $\{\mathcal{P}_0^t, \mathcal{N}_0^t, \mathcal{M}_0^t, \mathcal{S}_0^t\}$.

The components of the product on the right-hand side of Eqn.(13) are found according to

$$\begin{aligned} f_s(\mathcal{S}_0^t) &= \prod_{\tau=0}^{t} a_{s_{\tau-1}s_\tau} \qquad (14) \\ f_m(\mathcal{M}_0^t \mid \mathcal{S}_0^t) &= \prod_{\tau=0}^{t} f_m(m_\tau \mid s_\tau) = \prod_{\tau=0}^{t} c_{m_\tau|s_\tau} \qquad (15) \\ f_n(\mathcal{N}_0^t \mid \mathcal{M}_0^t, \mathcal{S}_0^t) &= f_n(\mathcal{N}_0^t) = \prod_{\tau=0}^{t} a'_{n_{\tau-1}n_\tau} \qquad (16) \\ f_p(\mathcal{P}_0^t \mid \mathcal{N}_0^t, \mathcal{M}_0^t, \mathcal{S}_0^t) &= f_p(\mathcal{P}_0^t \mid \mathcal{N}_0^t) \\ &= \prod_{\tau=0}^{t} f_p(p_\tau \mid n_\tau) = \prod_{\tau=0}^{t} c'_{p_\tau|n_\tau} \qquad (17) \end{aligned}$$

$$f_{\lambda_z}(\mathbf{z}_0^t \mid \mathcal{P}_0^t, \mathcal{N}_0^t, \mathcal{M}_0^t, \mathcal{S}_0^t) = \prod_{\tau=0}^{t} f_{\lambda_z}(\mathbf{z}_\tau \mid s_\tau, m_\tau, n_\tau, p_\tau, d_\tau)$$
$$= \prod_{\tau=0}^{t} b(\mathbf{z}_\tau \mid s_\tau, m_\tau, n_\tau, p_\tau, d_\tau), \quad (18)$$

where a_{ij} denotes the speech state transition probability from state i to state j and $c_{m|j}$ is the probability of choosing speech mixture component m given speech state j. a'_{ij} and $c'_{m|j}$ are similarly defined for the noise model. $b(\mathbf{z}_t \mid s_t, m_t, n_t, p_t, d_t)$ denotes the pdf of noisy observation $\mathbf{z}_t$ at time t given the quadruple of speech state, speech mixture component, noise state, and noise mixture component (s_t, m_t, n_t, p_t) and the duration of the speech state up to time t, d_t. Given (s_t, m_t, d_t), the clean speech pdf at time t is Gaussian and so is the noise pdf at time t given (n_t, p_t). Due to the independence of the clean speech and noise processes, the noisy speech pdf (Gaussian) can be written as:

$$b(\mathbf{z}_t \mid s_t, m_t, n_t, p_t, d_t) = N_t[\Lambda(s_t, m_t, n_t, p_t, d_t), \Sigma_{s_t,m_t} + \Sigma_{n_t,p_t}]. \quad (19)$$

where Σ_{s_t,m_t} denotes the covariance matrix of the clean speech for (s_t, m_t) and Σ'_{n_t,p_t} denotes the noise covariance matrix for (n_t, p_t). In Eqn.(19), the mean of the Gaussian, $\Lambda(s_t, m_t, n_t, p_t, d_t)$, is defined as

$$\Lambda(s_t, m_t, n_t, p_t, d_t) = \left[\mathbf{z}_t - \sum_{r=0}^{R} \mathcal{B}_r(s_t, m_t, n_t, p_t)\mathbf{h}_r(d_t)\right], \quad (20)$$

where $\mathbf{h}_r(d_t)$ is the value of the Legendre orthogonal polynomial of order r for the duration d_t, and where $\mathcal{B}_r(s_t, m_t, n_t, p_t)$ is the r-th order trend polynomial coefficient of the noisy speech model. These polynomial coefficients are found from:

$$\mathcal{B}_0(s_t, m_t, n_t, p_t) = B_0(s_t, m_t) + \mu(n_t, p_t)$$
$$\mathcal{B}_r(s_t, m_t, n_t, p_t) = B_r(s_t, m_t) \qquad \text{for} \quad r > 0. \quad (21)$$

In Eqn.(21), $B_r(s_t, m_t)$ denotes the r-th order polynomial coefficients of the clean speech NS-HMM for the state and mixture-component pair (s_t, m_t), and $\mu(n_t, p_t)$ is the mean of the noise standard HMM for the state and mixture-component pair (n_t, p_t).

Now returning to the computation of the MMSE estimate specified in Eqn.(10). Use of Eqns.(13)-(18) results in

$$f_{\lambda_z}(\mathbf{z}_0^{t-1}) =$$
$$\sum_{\mathcal{S}_0^{t-1}} \sum_{\mathcal{M}_0^{t-1}} \sum_{\mathcal{N}_0^{t-1}} \sum_{\mathcal{P}_0^{t-1}} \prod_{\tau=0}^{t-1} a_{s_{\tau-1}s_\tau} c_{m_\tau|s_\tau} a'_{n_{\tau-1}n_\tau} c'_{p_\tau|n_\tau} b(\mathbf{z}_\tau \mid s_\tau, m_\tau, n_\tau, p_\tau, d_\tau). \quad (22)$$

Eqn. (22) can now be used to calculate $f_{\lambda_z}(\mathbf{z}_0^t)$ in the denominator of Eqn.(10) by replacing $t-1$ with t.

In order to write the computation of the MMSE estimate in a clean form, we first define

$$G_t(\beta,\gamma,\xi,\delta,d,\mathbf{z}_0^t) \triangleq f_{\lambda_z}(\mathbf{z}_0^t \mid s_t=\beta, m_t=\gamma, n_t=\xi, p_t=\delta, d_t=d) \tag{23}$$
$$= \sum_{\{\mathcal{S}_0^t:s_t=\beta\}} \sum_{\{\mathcal{M}_0^t:m_t=\gamma\}} \sum_{\{\mathcal{N}_0^t:n_t=\xi\}} \sum_{\{\mathcal{P}_0^t:p_t=\delta\}} \prod_{\tau=0}^{t} a_{s_{\tau-1}s_\tau} \cdot c_{m_\tau|s_\tau} \cdot a'_{n_{\tau-1}n_\tau} \cdot c'_{p_\tau|n_\tau} \cdot b(\mathbf{z}_\tau \mid s_\tau, m_\tau, n_\tau, p_\tau, d_\tau).$$

The denominator of Eqn.(10) can then be written as

$$f_{\lambda_z}(\mathbf{z}_0^t) = \sum_{\beta=1}^{M}\sum_{\gamma=1}^{L}\sum_{\xi=1}^{N}\sum_{\delta=1}^{P}\sum_{d=1}^{t} G_t(\beta,\gamma,\xi,\delta,d,\mathbf{z}_0^t). \tag{24}$$

Further, the product of the three pdf's in the integrand of numerator in Eqn.(10) can be written in a compact form derived from Eqn.(11), Eqn.(12), Eqn.(22), and Eqn.(24):

$$\begin{aligned} &f_{\lambda_z}(\mathbf{z}_0^{t-1}) f_{\lambda_y}(\mathbf{y}_t) f_{\lambda_v}(\mathbf{v}_t) = \\ &\quad \sum_{\beta=1}^{M}\sum_{\gamma=1}^{L}\sum_{\xi=1}^{N}\sum_{\delta=1}^{P}\sum_{d=1}^{t} G_t(\beta,\gamma,\xi,\delta,d,\mathbf{z}_0^t) \\ &\quad f_{\lambda_y}(\mathbf{y}_t \mid s_t=\beta, m_t=\gamma, n_t=\xi, p_t=\delta, d_t=d). \end{aligned} \tag{25}$$

On substitution of Eqn.(22) and Eqn.(25) in Eqn.(10), we finally obtain the MMSE estimate

$$\begin{aligned} &E\{g(\mathbf{y}_t) \mid \mathbf{z}_0^t\} = \\ &\quad \sum_{\beta=1}^{M}\sum_{\gamma=1}^{L}\sum_{\xi=1}^{N}\sum_{\delta=1}^{P}\sum_{d=1}^{t} \mathcal{W}_t(\beta,\gamma,\xi,\delta,d,\mathbf{z}_0^t) \cdot \\ &\quad \int_{-\infty}^{\infty} g(\mathbf{y}_t) f_{\lambda_y}(\mathbf{y}_t \mid s_t=\beta, m_t=\gamma, n_t=\xi, p_t=\delta, d_t=d) d\mathbf{y}_t \\ &\quad = \sum_{\beta=1}^{M}\sum_{\gamma=1}^{L}\sum_{\xi=1}^{N}\sum_{\delta=1}^{P}\sum_{d=1}^{t} \mathcal{W}_t(\beta,\gamma,\xi,\delta,d,\mathbf{z}_0^t) \cdot \\ &\quad\quad E\{g(\mathbf{y}_t) \mid s_t=\beta, m_t=\gamma, n_t=\xi, p_t=\delta, d_t=d\} \end{aligned} \tag{26}$$

where the weights $\mathcal{W}_t(\beta,\gamma,\xi,\delta,d,\mathbf{z}_0^t)$ are defined as

$$\mathcal{W}_t(\beta,\gamma,\xi,\delta,d,\mathbf{z}_0^t) \triangleq \frac{G_t(\beta,\gamma,\xi,\delta,d,\mathbf{z}_0^t)}{\sum_{\beta=1}^{M}\sum_{\gamma=1}^{L}\sum_{\xi=1}^{N}\sum_{\delta=1}^{P}\sum_{d=1}^{t} G_t(\beta,\gamma,\xi,\delta,d,\mathbf{z}_0^t)} \tag{27}$$
$$\text{for} \qquad 1\le\beta\le M,\ 1\le\gamma\le L,\ 1\le\xi\le N,\ 1\le\delta\le P,\ 1\le d\le t.$$

A most interesting interpretation emerges from an examination of Eqn.(26): the MMSE estimate of the speech signal can be expressed as a weighted average of the state- and mixture-component-conditioned signal expectations over all possible combinations of speech and noise

HMM states and mixture-components. The main burden for computing the MMSE estimate is now reduced to the calculation of these state- and mixture-component-conditioned expectations:

$$E\{g(\mathbf{y}_t) \mid s_t = \beta, m_t = \gamma, n_t = \xi, p_t = \delta, d_t = d\} = \int_{-\infty}^{\infty} g(\mathbf{y}_t) f_{\lambda_y}(\mathbf{y}_t \mid s_t = \beta, m_t = \gamma, n_t = \xi, p_t = \delta, d_t = d) d\mathbf{y}_t, \tag{28}$$

which we address below.

To compute Eqn.(28), function $g(\cdot)$ has to be specified first. For the stationary-state HMM (a special case of the NS-HMM presented in this paper), the conditional expectation Eqn.(28) has been evaluated by Ephraim [13]. Different choices of the $g(\cdot)$ function incur varying computational costs. The least amount of computation happens with the choice of

$$g(\mathbf{y}_t) = \{Y_t(k), k = 0, \ldots, K-1\}, \tag{29}$$

where $Y_t(k)$ is the k-th DFT (magnitude) component of $\mathbf{y}_t$. In addition to the motivations discussed in detail in Section 2.2, this computational consideration gives another motivation for use of DFT as the speech feature in this work; that is, we choose $g(\cdot)$ according to Eqn.(29) in the computation of the MMSE estimate.

In determining the integral in Eqn.(28), we generalize a result of [13] from stationary-state HMMs to NS-HMMs. The generalized result is that the linear estimate using the MMSE criterion for the k-th component of g (denoted by $g(k)$) given by

$$E\{g(k) \mid \mathbf{z}_t, s_t, m_t, n_t, d_t\} = \int Y_t(k) f_{\lambda_y \lambda_v}(Y_t(k) \mid \mathbf{z}_t, s_t, m_t, n_t, d_t) dY_t(k) \tag{30}$$

is Gaussian distributed with mean $\widetilde{H}_{s_t,m_t,n_t,p_t,d_t}(k) Z_t(k)$. Here, $Z_t(k)$ is the k-th component of the DFT of $\mathbf{z}_t$, and $\widetilde{H}_{s_t,m_t,n_t,p_t,d_t}(k)$ is the k-th component of the DFT of the frequency-domain Wiener filter output (given state s_t, mixture component m_t, and duration d_t in the clean speech model, and given state n_t and mixture component p_t in the noise model). Note that this Wiener filter changes its transfer function over every single time frame because the transfer function is determined by the polynomial trend function in the NS-HMM which changes smoothly over time frames by design.

Returning to the computation of the sum in Eqn.(24) which determines the weights in Eqn.(27) contributing to the aggregate Wiener filter, we are able to write down an efficient recursive form. The recursive formula for the calculation is

$$G_t(\beta, \gamma, \xi, \delta, 0, \mathbf{z}_0^t) = \left[\sum_{\beta'=1}^{M} \sum_{\gamma'=1}^{L} \sum_{\xi'=1}^{N} \sum_{\delta'=1}^{P} \sum_{d'=0}^{t} G_{t-1}(\beta', \gamma', \xi', \delta', d', \mathbf{z}_0^{t-1}) \cdot a_{\beta'\beta} \cdot a'_{\xi'\xi} \right] \cdot$$

$$
\begin{aligned}
& c_{\gamma|\beta} \cdot c'_{\delta|\xi} \cdot b(\mathbf{z}_t \mid \beta, \gamma, \xi, \delta, 0) \\
G_t(\beta, \gamma, \xi, \delta, d, \mathbf{z}_0^t) \;\; = \;\; & \left[\sum_{\xi'=1}^{N} \sum_{\delta'=1}^{P} G_{t-1}(\beta, \gamma, \xi', \delta', d-1, \mathbf{z}_0^{t-1}) \cdot a_{\beta\beta} \cdot a'_{\xi'\xi} \right] \cdot \\
& c_{\gamma|\beta} \cdot c'_{\delta|\xi} \cdot b(\mathbf{z}_t \mid \beta, \gamma, \xi, \delta, 0), \qquad \text{for } \; 0 < d \leq t. \qquad (31)
\end{aligned}
$$

Note that the above recursion has an additional dimension of state-sojourn time (d') which results from the time-varying means (i.e. polynomial trend functions) in the NS-HMM output distributions.

The computation in Eq.26 and that in the related terms by recursive updating formula Eq.31 form the core computation in the MMSE algorithm. Given the time-varying Wiener filters, the multiple nested summations in Eq.26 gives a total of $M \times L \times N \times P \times t$ summations for each frame (at time t). To compute all T frames, the total number of summations is $M \times L \times N \times P \times T^2/2$. Similarly, the total number of summations required in computing the quantities in Eq.31 is $M \times L \times N \times P \times T^2/2$.

2.4 Training Speech and Noise Models Based on Nonstationary-State HMM

As described earlier, in our development of the speech enhancement system reported in this work, we used DFT magnitudes as the speech feature. Since our enhancement system is based on use of the NS-HMM as the speech model, training is needed to determine the parameters of the model. This training procedure is independent of the selected feature for the model. A separate preprocessor produces the required speech feature and training is then performed on the preprocessed speech data.

Consider the NS-HMM as the speech model with a K-dimensional acoustic feature, a total of M states and a total of L mixture components, and with the polynomial order R in the trend functions. The parameters of this model have been described in Section 2.3 in detail. In particular, the parameters for the trend functions are polynomial coefficients: $\mathbf{B} \triangleq \{B_{j,m}(r)\}$, $j = 1, 2, \ldots, M$, $m = 1, 2, \ldots, L$, $r = 0, \ldots, R$.

During our training of the model parameters, we first perform vector quantization on the training data to initialize the model parameters. We then use an extended Viterbi algorithm similar to that described in [15] to iteratively update the model parameters. The training for the noise HMM is essentially the same as that for conventional, stationary-state HMM. For more details on the training procedure see [22].

3 Implementation of the Speech Enhancement Algorithms

Our earlier work [8, 10] has shown significant advantages of the MMSE algorithm over other types of enhancement algorithms. The advantages arise not only from its superior theoretical motivation according to estimation theory, but also from practical considerations such as its elimination of the need for iteration. In this section, we will describe implementation of the MMSE algorithm in our speech enhancement system. We will also describe an approximate method to the MMSE algorithm, which has been used in this work in analyzing the enhancement system behavior.

3.1 Implementation of the MMSE Method

A simplified block diagram of the NS-HMM-based enhancement system using the DFT magnitude as speech features is shown in Figure 1. Theoretically, in the enhancement procedure, all possible Wiener filters have to be calculated. The number of possible Wiener filters for the enhancement system based on the NS-HMM as the clean speech model is far more than that in the standard HMM case. Since the mean of the clean speech model for a given state and mixture-component pair is still a function of the additional parameter — state sojourn time, another dimension is added to the calculation of the MMSE forward algorithm for finding the filter weights. For each time frame t, a number of $t \times M \times L \times N \times P$ Wiener filters and their corresponding weights have to be calculated. The weights as shown in Eqn. (27) are functions of the speech and mixture-component pairs of both clean speech and noise models and the clean speech duration up to time t. In theory, the possible number of duration values which a speech state may have at time t is t itself. So the factor t is multiplied by the number of possible Wiener filters for the standard HMM-based speech enhancement system.

In parallel with calculation of the Wiener filters, each frame of time-domain noisy input speech is preprocessed and transferred to the model feature domain (DFT magnitude). Then for each frame of the noisy speech the NS-MMSE forward algorithm (explained in Section 2.3) is performed. For this, the pdf of the noisy speech has to be calculated. The noisy speech pdf can be specified with the trend polynomial coefficients ($\mathbf{B}$ matrix) and the covariance matrix. Since speech and noise are additive and independent, the polynomial coefficients of the noisy speech ($\mathbf{B}$ matrix) is found by adding up the corresponding elements of the $\mathbf{B}$ matrix of the clean speech model and the means of the noise model. For each time frame t, a total of $M \times L \times N \times P$ matrices of polynomial coefficients have to be calculated. These calculations are extremely simple compared to the calculations necessary for the construction of the noisy pdf in the AR-HMM structure even after the approximations employed to reduce the computation cost [10]. By using diagonal

covariance matrices for the NS-HMMs, the problem of high computation cost for the inversion of the covariance matrix for calculation of the output likelihoods is also avoided. [7] It should be noted that the AR-HMM is not capable of accommodating nonstationary states due to the lack of smoothness in AR coefficients versus time.

Despite all the simplicity for finding the pdf of the noisy speech in the NS-HMM framework, the encoding space is still significantly larger than that in the standard HMM case due to the additional dimension of the speech state duration. This problem has been solved using the double pruning algorithm explained in [10]. For the NS-HMM the pruning algorithm is of utmost importance, because in addition to pruning the calculations for finding the filter weights, it is employed to reduce the number of Wiener filter transfer functions to be calculated. In the standard HMM enhancement framework, it is practical to calculate all possible Wiener filters and save them before the actual enhancement procedure starts. However, in the NS-HMM framework, due to the extensive number of possible Wiener filters, it is better to avoid such calculation. Therefore, in the NS-HMM enhancement algorithm, the necessary Wiener filters for a specific time frame of noisy speech are calculated after the calculation of the filter weights for that specific time frame.

After calculating the pdf of the noisy speech, the extended MMSE forward algorithm is used to find the Wiener filter weights for each time frame. The weights are employed to generate a single weighted Wiener filter from the inventory of the previously calculated Wiener filters. The weighted filter is applied to the noisy frame of speech in frequency-domain. The filtered frequency-domain speech signal is transformed to time-domain using the original noisy signal phase and the output enhanced signal is synthesized with the overlap-add method.

3.2 Approximate MMSE Enhancement Method Using the Dominant Wiener Filter

The MMSE algorithm which employs a soft-decision method for building a weighted-sum filter can be approximated by using a single Wiener filter for each time frame. In the approximate MMSE (AMMSE) method, the most likely Wiener filter for each time frame is selected using a modified version of the extended Viterbi algorithm for the NS-HMM explained in [15]. A block diagram of this enhancement method for the case of the DFT magnitude as the speech feature is shown in Figure 2.

In this system, similar to the MMSE method, the noisy signal is preprocessed and transformed to the autocorrelation or DFT magnitude. Using the selected noise model and the clean model

[7] In our earlier experiments, we had implemented an expensive version of the system which used full covariance matrices. Empirically, we found comparable results to the current system with the diagonal covariance matrices.

NS-HMM, an NS-HMM is built for the noisy speech. With the noisy speech model the extended Viterbi algorithm is performed on the preprocessed noisy speech data to provide the segmentation information for the noisy observations. The extended Viterbi algorithm determines the most likely speech state, speech mixture-component, noise state, and noise mixture-component quadruple for each frame of the noisy speech signal. For every time-frame, having the most likely state, mixture component, and state duration for the clean speech and the most likely state and mixture-component pair for the noise, a single Wiener filter is built from the model parameters. The noisy speech data belonging to the frame is filtered with the constructed Wiener filter. The resultant frequency-domain enhanced frame of speech is transformed to time-domain using the noisy speech phase and the overlap-add method.

In the AMMSE enhancement method, the problem is to fit the clean speech data given the noisy speech. More accurate data-fitting is equivalent to a more appropriate Wiener filter and better noise cancellation as a result. Therefore, this method can be used to analyze the enhancement system behavior and diagnose probable malfunctioning of the system. The AMMSE enhancement system is simpler than the MMSE method in that it uses only one Wiener filter for each frame and avoids the costly calculation of the filter weights and the weighted-sum filter. However, the AMMSE algorithm relies on the segmentation information which is extracted by the extended Viterbi algorithm. The extended Viterbi algorithm for the multiple state and mixture noise model would itself be computationally costly. The algorithm confronts a five-dimensional search space comprising the noise and speech states and mixture components in addition to the speech state duration. For a single-state noise, however, the AMMSE system is more efficient than the MMSE enhancement method.

The computation of the AMMSE algorithm mainly lies in that of the extended Viterbi algorithm, whose computational complexity for the NS-HMM has been analyzed in detail in [15]. To summarize, the computation is quadratic in T (as opposed to the conventional Viterbi algorithm which has the computation linear in T). However, if the durational properties of speech are taken into account, the computation can be reduced to that linear in T with the large constant equal to the maximum duration of the speech unit. This computation reduction has been analyzed in [15] also.

4 Speech Enhancement System and Experimental Results

4.1 Speech Enhancement System Overview and Experimentation Environment

The speech data used in the speech enhancement experiments reported in this section were selected from the sentences in the TIMIT database. One hundred sentences spoken by 13 different speakers (with sampling rate of 16 kHz) were used for training the clean speech model. One frame of speech covers 256 speech samples (equivalent to 16 ms). No interframe overlap was used in training the speech model. The sentences used for enhancement tests were selected such that there were no common sentences or speakers between the enhancement and training sets. A 50% overlap between adjacent frames was used in the enhancement procedure. For the standard HMM-based enhancement systems, the models used for clean speech and noise are both AR-HMMs of different AR orders. In this work, an AR order of 14 for clean speech and an AR order of 6 for noise are used. Therefore, in contrast to the stationary-state HMM-based enhancement system where the preprocessor extracts the AR coefficients of the noisy speech, in the NS-HMM based enhancement systems the preprocessor extracts the DFT magnitude components.

The implemented MMSE enhancement system follows the procedures displayed in Figure 1, with the block diagram of the noise-model selection component shown in Figure 3. The noise-model selection method has been described in detail in [10]. Briefly, the noisy signal during long periods of non-speech activity is first fed into a Viterbi forward algorithm. Then, the likelihood for each pre-trained noise HMM is calculated and compared with likelihoods for the other noise HMMs and the model associated with the highest likelihood determines the selected noise model. Using the selected noise HMM parameters and the clean speech model, the noisy speech model is artificially generated through calculating the pdf of the noisy speech.[8]

In the meantime, all Wiener filters for all combinations of the state and mixture-component pairs in the speech and noise models are calculated. A single weighted filter is constructed for each frame of noisy speech using the calculated filter weights and the pre-trained Wiener filters. The filtering of the noisy signal is carried out using the weighted filter. This generates the spectral magnitude of the enhanced speech signal. Using this magnitude and the phase of the noisy speech, an inverse FFT is performed to obtain the time-domain enhanced speech via the standard overlap-

[8]The calculation procedures for the pdf of noisy speech are rather different between the standard-HMM and NS-HMM based systems. The respective procedures have been described in in [10] and in Section 2.3 of this paper. Briefly, In the standard-HMM based system, the preprocessed noisy speech is input to the MMSE forward algorithm which generates the weights for the Wiener filters. For the NS-HMM based system, the procedure the weights for the Wiener filters are calculated according to Eqn.(27).

add method. For the MMSE enhancement, the noise HMMs we have implemented contained 3 states and 3 mixture components.[9]

A global measure of signal-to-noise ratio (SNR) was used as the objective evaluation criterion, which is calculated by

$$SNR = 10 \log \frac{\sum_{n=1}^{K} y^2(n)}{\sum_{n=1}^{K} [y(n) - \hat{y}(n)]^2}, \tag{32}$$

where K is the frame-length, and $y(n)$ and $\hat{y}(n)$ are the n-th components of the time-domain clean speech and of the time-domain enhanced speech signals, respectively.

For the experimentation with the new speech enhancement system incorporating the NS-HMM, we carried out the experiments in two phases, the diagnostic experimentation phase and the speech enhancement experimentation phase, to be described below.

4.2 Diagnostic Experiments

In the diagnostic experiments reported in this subsection, some individual components of the enhancement system were analyzed. The AMMSE method was used as the enhancement algorithm since it enables us to analyze detailed data-fitting behaviors in the enhancement process and to investigate the relation between the observed goodness of fit and the overall performance of the enhancement system.

In our AMMSE based enhancement system, the extended Viterbi algorithm extracts the possible segmentation information for the desired *clean* speech signal using the available *noisy* speech signal. To achieve this goal, the noisy speech NS-HMM is artificially built from the pre-trained clean speech model and the noise model. As explained in Section 3, the AMMSE system requires the extended Viterbi algorithm to search over a five-dimensional space. This high cost in computation has limited us to perform the diagnostic experiments only with white noise and with a single-state noise model.

An utterance comprising a portion (1.1 second) of a sentence from the TIMIT speech database [10] was used for both training and testing. DFT magnitude vectors containing 129 components with a resolution of 62.5 Hz were used as the speech and noise features. Each frame of the noisy speech was taken to be 256 samples long corresponding to 16 ms of data. The overlap between adjacent frames was set to 50%. NS-HMMs with 4 states and 4 mixture components and of orders 0, 1, and 2 were trained with the utterance. Noisy speech was generated by adding Gaussian

[9]Noise models containing 5 states and 5 mixture components have also been used in a few tests and we found that they did not result in notable improvements over the noise model comprising 3 states and 3 mixture components.

[10]The complete sentence was: "Woe betide the interviewee if he answered vaguely."

white noise to the test utterance with zero-dB SNR. The AMMSE enhancement procedure was performed on the test utterance and the output SNRs and total likelihoods for varying polynomial orders of the speech NS-HMM were then calculated. Data-fitting results were obtained during the segmentation stage where the polynomial functions of time concatenated sequentially according to the selected HMM state/mixture sequence were used to approximate the data trajectories. To establish the highest possible performance of the speech enhancement system, the clean speech spectra were used in the system to construct Wiener filters instead of using the polynomials obtained from the output of the extended Viterbi algorithm. The SNR obtained via use of the clean speech information sets the upper limit of the system performance. The output SNRs and the associated likelihoods signaling goodness of data-fitting to the models are presented in Table 2. A direct correlation between the goodness of data-fitting and the speech enhancement performance measured by the output SNRs is clearly demonstrated by these results. The results in Table 2 also show that as the polynomial order in the NS-HMM is increased from zero to two, the performance of enhancement is approaching the theoretical limit determined by using the clean speech spectra to derive the Wiener filters.

To examine the detailed behavior in the comparative system performance exhibited in Table 2, we show data-fitting results for modeled DFT feature vectors (polynomials) fitting to real DFT feature vectors. For illustration purposes, only one representative DFT component (f=187.5 Hz) is shown in Figure 4 and another representative DFT component (f=1687.5 Hz) shown in Figure 5. The results in Figures 4 and 5 and our observations on a number of other DFT components show that the DFT components which are less affected by noise have been estimated more accurately than those affected more by noise. This implies that the method of calculating the noisy speech pdf and that of building a good noisy speech model are crucial in finding a good estimate of the clean speech signal. Note that the likelihood obtained from the extended Viterbi algorithm is the likelihood of the noisy speech observations given the artificially generated noisy speech model, and it is desirable that this likelihood be a reliable indication of the likelihood for the clean speech data given the clean speech model. The data-fitting observations shown in Figures 4 and 5 are supportive of a positive correlation between the two likelihoods.

The results in Table 2 discussed earlier showed that use of higher polynomial orders in the NS-HMM results in both better likelihoods and higher SNRs than use of lower orders (in particular, order zero). The data-fitting results shown in Figures 4 and 5 have provided some underlying reasons for this improvement. That is, the NS-HMM is doing its job of smoothly (except when HMM state transitions occur) following speech spectral variations over time, with a better job done using higher orders of polynomial trend functions. The consistency of the clean speech data-fitting results with the associated likelihoods indicates that the parameters of the noisy speech model are

reasonably accurately estimated. Further, the consistency between goodness of data-fitting and the output SNRs indicates that the construction of the dominant Wiener filter and the procedure of frequency-domain filtering and reconstruction of time-domain speech signals have been done correctly.

It may be argued that the smooth trajectories across frames inherent in the NS-HMM could also be realized by the Gaussian-mixture model if different mixture components could be automatically chosen to fit the smooth trajectories in the speech data. To examine such a possibility, we conduct speech enhancement experiments by constraining the number of HMM states to be one and obtaining a Gaussian-mixture model for clean speech. Keeping the same number of Gaussian components (16 in total), we obtained the output SNR which is worse than the HMM — 9.05 dB for the Gaussian-mixture model versus 9.45 dB for the conventional HMM and 11.01 for the NS-HMM. This suggests that the Gaussian-mixture model appears unable to automatically choose the "correct" mixture components given only the noisy speech data available. Since the NS-HMM forces smooth transitions across frames in the model structure itself, it guarantees a natural fit between the data trajectory and the model component, thereby outperforming both the Gaussian-mixture model and the conventional HMM.

The next diagnostic experiment has been motivated by the following reasonings. As explained in Section 3, the AMMSE speech enhancement method employs the extended Viterbi algorithm to obtain the segmentation information for the clean speech signal. Given a valid segmentation of the clean speech, proper estimation of the signal is possible using the clean speech model parameters. However, this estimation is not sufficient to generate high-quality enhanced speech. This is due to the fact that the estimated clean speech is the mean of a random process, and only a single realization of the random process gives rise to the enhanced speech. Since the noisy version of this realization is available, use of Wiener filtering is one way of obtaining the clean speech signal. In other words, a Wiener filter is constructed using the acquired estimation of the clean speech and the noise spectrum, and frequency-domain filtering of the noisy speech spectrum is performed to obtain the enhanced speech. Now, to illustrate the role of the Wiener filtering in speech enhancement, we in this diagnostic experiment deliberately eliminated the Wiener filtering stage. In the implementation, we first estimated the means of the polynomial trajectories in the NS-HMM (for the clean speech spectral sequence) using the extended Viterbi algorithm, and then straightforwardly transformed the estimated means into time domain. (This transformation does not lose information because the DFT used is a fully reversible feature.) The SNRs of the reconstructed time-domain speech utterance were calculated, with the results shown in Table 3. Here, we observe that the qualities of such reconstructed speech signals are significantly lower than their counterparts obtained with use of Wiener filters (cf. Table 2), losing SNRs about 3 dB

or greater. Nevertheless, we also observe from Table 3 that the absolute SNRs still improve as the polynomial order in the NS-HMM used to fit the speech data increases from zero to two. This observation is consistent with that seen earlier in Table 2.

The main conclusion drawn from the diagnostic experiments presented in this subsection is that the superior speech enhancement performance (measured by the output SNRs) achieved by the NS-HMM, compared to the conventional stationary-state HMM, roots in the superior ability of the NS-HMM in fitting the speech data. Therefore, the problem of speech enhancement is to a large extent equivalent to that of correct segmentation of and accurate spectral-mean approximation (fitting) to the "hidden" clean speech signal given the noisy speech data.

4.3 Speech Enhancement Results

In this section the results of speech enhancement experiments using the NS-HMM are presented. The MMSE enhancement algorithm described in detail in Section 3 was used throughout the experiments. The training and enhancement details have been described in Section 4.1. We have tested polynomial orders of 0, 1, 2, 3, and 4 in the NS-HMM, and have used two types of noise, white noise and simulated helicopter noise,[11] as the additive noise to generate noisy speech data in all our experiments reported here. All experiments have been carried out using 129 components of the DFT magnitude as the features for speech and noise data.

We have run experiments on sevarl arbitrarily chosen TIMIT sentences, each consisting of about two to four seconds of an utterance. The two types of additive noise, each having input SNRs ranging from 0 dB to 15 dB with an increment of 5 dB, are used in the experiments. The formal measure of the enhancement performance is the output SNR. The results of the experiments are shown in Tables 4 and 5, for the use of simulated white noise and simulated helicopter noise, respectively, in a typical sentence. In the tables, the output SNR as the performance measure is shown as a function of the input SNR and of the polynomial order in the NS-HMM used in the MMSE enhancement algorithm.

In analyzing the results for the sentences we have run, we first observe general consistencies of the results across difference speech utterances and across the noise type, the latter indicating that our enhancement algorithm is equally effective for stationary and nonstationary noise. Second, the enhancement algorithm is most effective for low input SNRs (0 dB and 5 dB), and for higher input SNRs, output SNRs are slightly lower than input SNRs indicating some undesirable distortions. Third, importantly, advantages of the NS-HMM (order greater than zero) over the stationary-state HMM (i.e. order-zero NS-HMM) have been observed across all sentences, all input SNRs, and

[11]The helicopter noise is generated by amplitude modulation of white noise with a fixed modulation frequency.

across both noise types. Fourth, the SNR improvements gained using the NS-HMM with orders higher than one is marginal but the superiority of the order-one NS-HMM over the order-zero NS-HMM is considerable. This, again, has been observed to be consistent across all sentences, all input SNRs, and across both noise types. Some informal listening we ourselves have experienced has shown a high degree of consistency with the output SNR results shown in Tables 4 and 5. In particular, informal listening results indicate that the improvement from use of order-zero NS-HMM to use of order-one NS-HMM is perceptually detectable by listeners, but no perceptual differences are found with use of other varying polynomial orders in the NS-HMM.

5 Summary and Conclusions

The focus of this study has been on the incorporation of the NS-HMM, a more accurate model for dynamic speech spectral patterns than the benchmark stationary-state HMM, in the speech enhancement system. The NS-HMM represents "local" speech nonstationarity within a state in the HMM, and models detailed and relatively smooth variations in the intrinsically dynamic speech signal. This model is formally related to but different from the model described in [23, 24] which has also been used in speech enhancement. The NS-HMM used in the current work directly describes the dynamics for the feature vectors as polynomial functions of time frame. No recursion (autoregression) is used in defining the dynamics. In the model of [23, 24], the dynamics is modeled at the time-sample level, which requires autoregression in the description of the dynamics. The coefficients in the autoregressive model, rather than the feature vectors, are modeled as polynomial functions at time-frame level.

In order to incorporate the NS-HMM in the enhancement system, significant modifications have been made to the conventional HMM-based system in this work. One conceptually most important modification arises from the use of new speech features since the auto-regressive filter coefficients used in conventional HMM-based systems do not possess the smoothness characteristics required by the NS-HMM. A set of conventional speech features commonly in use for speech recognition are reviewed in terms of several key properties (smoothness, additivity, and reversibility) required by the NS-HMM based speech enhancement algorithm. The DFT magnitudes and the autocorrelation functions are selected as appropriate features because of their fulfillment of all of the smoothness, additivity, and reversibility requirements. (In this work, we only implemented and evaluated the DFT-magnitude features.) Based on use of these desirable speech features, the MMSE formulation has been derived where the NS-HMM is used as the speech model and the Gaussian-mixture, stationary-state HMM as the noise model. An approximate MMSE method is also devised and implemented which makes the enhancement system analysis (reported in Section 4.2 on diagnostic

experiments) possible.

In the experimental evaluation part of this work, two types of noise — white noise, and simulated helicopter noise — are used to corrupt the speech signals. The experiments are carried out in two phases, the diagnostic experimentation phase and the speech enhancement experimentation phase. In the diagnostic experimentation phase, individual components of the enhancement system are analyzed. The training and the testing utterances are made identical in the experiments so that a detailed study of the system behavior becomes possible. According to the results from the diagnostic experiments, we identify the problem of speech enhancement as one equivalent to accurate fitting of the modeled data trajectory to the "hidden" (unobservable) clean speech data trajectory. Such fitting uses only separate pre-trained speech and noise models, as well as noisy speech data sequences as the observable information. We find in the diagnostic experiments that reasonably accurate fitting to the "hidden" clean speech can be achieved given noisy speech data within the framework of the NS-HMM, but not within the framework of the conventional HMM. Correlated superiority of non-zero polynomial orders for the NS-HMM in data-fitting and in enhancement performances is demonstrated in the diagnostic experiments. Such results have provided considerable insights to the functionality of the MMSE algorithm based on the NS-HMM representation of spectral trajectories of the speech signal.

In the speech enhancement experimentation phase, separate sets of training and testing (enhancement) data are used. Experiments are performed for the enhancement system with the polynomial orders varying from zero to four. The experimental results show consistent superiority of higher polynomial orders in the NS-HMM over the order-zero counterpart as the benchmark, consistent with the observations made in the diagnostic experiments.

The contributions of this work lie more in the insights gained to the nature of the speech enhancement problem than in the practical implementation of our particular system. The most significant lesson we learned from this work is the importance of a good, analyzable speech model in speech enhancement applications. Striving for such a model is indeed also the enterprise of other speech technology areas, notably speech recognition. In fact, the NS-HMM used as the backbone of the speech enhancement system reported in this paper was originally developed for speech recognition. Its success motivated us to port this model to the current speech enhancement application. However, different applications (enhancement versus recognition) do require considerable efforts in understanding several key issues such as feature selection, formulation of different optimization criteria, and development of different optimization procedures. All these issues for the speech enhancement application have been dealt with carefully in this work and been discussed in this paper.

Finally, we would like to point out the high complexity in implementing the system described

in this paper. The pruning method we devised has substantially reduced the computational cost in executing the MMSE algorithm but its tradeoff with the possible performance degradation has not been studied thoroughly. This and a number of other implementation issues are currently under investigation. Also, as future work, different noise types with varying input SNRs will need to be tested on a greater number of testing sentences. Tests on real world noisy data (not made noisy in the lab) will also need to be carried out with not only objective evaluation (such as SNR improvement) but also with subjective evaluation. These more rigorous and comprehensive tests will further refine the algorithm developed in the current study and eventually validate the model-based approach proposed in this paper.

Nevertheless, the results reported in this paper showing superiority of better speech models in enhancement performance are already encouraging. This is particularly so because we have developed diagnostic tools, as reported in the diagnostic experiments, which permit us to analyze the underlying reasons for the superiority of the enhancement performance and thus give us a sure way to not only avoid pitfall implementations but also monitor performance improvement with theoretical guidances. This should open a wide door in the future for incorporating even better speech models, some of which are currently under active development in our laboratory, into speech enhancement systems as well as into other areas of speech technology applications.

Acknowledgements

We thank Dr. Chin Lee who provided constructive comments on this work. We also thank the anonymous reviewers, whose comments have significantly improved the quality of the paper presentation and the paper content. This work was supported by Unitron Industries Ltd., Canada, Ontario URIF fund, and by NSERC, Canada.

References

[1] S. F. Boll and D. C. Pulsipher. "Suppression of Acoustic Noise in Speech Using Two Microphone Adaptive Noise Cancellation". *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28(6):752–753, Dec. 1980.

[2] R. H. Frazier, S. Samsam, L. D. Braida, and A. V. Oppenheim. "Enhancement of Speech by Adaptive Filtering". *Proceedings of the IEEE ICASSP*, pages 251–253, 1976.

[3] J. S. Lim, A. V. Oppenheim, and L. D. Braida. "Evaluation of an Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-26(5):354–358, Aug. 1978.

[4] T. W. Parsons. "Separation of Speech from Interfering of Speech by Means of Harmonic Selection". *Journal of the Acoustical Society of America*, 60(4):911–918, Oct. 1976.

[5] J. S. Lim. "Evaluation of a Correlation Subtraction Method for Enhancing Speech Degraded by Additive White Noise". *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-26(5):471–472, Oct. 1978.

[6] S. F. Boll. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction". *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-27(2):113–120, Apr. 1979.

[7] J. S. Lim and A. V. Oppenheim. "Enhancement and Bandwidth Compression of Noisy Speech". *Proceedings of the IEEE*, 67(12):1586–1604, Dec. 1979.

[8] H. Sheikhzadeh, H. Sameti, L. Deng, and R. L. Brennan. "Comparative Performance of Spectral Subtraction and HMM-Based Speech Enhancement Strategies with Application to Hearing Aid Design". *Proceedings of the IEEE ICASSP*, 1:13–16, 1994.

[9] R. J. McAulay and M. L. Malpass. "Speech Enhancement Using a Soft-Decision Noise Suppression Filter". *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28:137–145, 1980.

[10] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan. "HMM-Based Strategies for Enhancement of Speech Embedded in Non-Stationary Noise". *IEEE Transactions on Speech and Audio Processing*, 6:445–455, 1998.

[11] Y. Ephraim, D. Malah, and B. H. Juang. "On the Application of Hidden Markov Models for Enhancing Noisy Speech". *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-37(12):1846–1856, Dec. 1989.

[12] Y. Ephraim. "Statistical-Model-Based Speech Enhancement Systems". *Proceedings of the IEEE*, 80(10):1526–1555, Oct. 1992.

[13] Y. Ephraim. "A Minimum Mean Square Error Approach for Speech Enhancement". *Proceedings of the IEEE ICASSP*, pages 829–832, 1990.

[14] Y. Ephraim. "On Minimum Mean Square Error Speech Enhancement". *Proceedings of the IEEE ICASSP*, pages 997–1000, 1991.

[15] L. Deng, M. Aksmanovic, D. Sun, and J. Wu. "Speech Recognition Using hidden Markov Models with Polynomial Regression Functions as Nonstationary States. *IEEE Transactions on Speech and Audio Processing*, 2(4):507–520, Oct. 1994.

[16] L. Deng. "A Stochastic Model of Speech Incorporating Hierarchical Nonstationarity". *IEEE Transactions on Speech and Audio Processing*, 1(4):471–475, Oct. 1993.

[17] L. Deng. "A Generalized Hidden Markov Model with State-Conditioned Trend Functions of Time for the Speech Signal". *Signal Processing*, 27(1):65–78, Apr. 1992.

[18] B. H. Juang and L. R. Rabiner. "The Segmental K-Means Algorithm for Estimating Parameters of Hidden Markov Models". *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-38(9):1639–1641, Sep. 1990.

[19] L. Deng and H. Sameti. "Transitional Speech Units and Their Representation by the Regressive Markov States: Applications to Speech Recognition". *IEEE Transactions on Speech and Audio Processing*, 4(4):301–306, 1996.

[20] S. D. Davis and P. Mermelstein. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28(4):357–366, Aug. 1980.

[21] J. Makhoul. "Linear Prediction: A Tutorial Review". *Proceedings of the IEEE*, 63(4):561–580, Apr. 1975.

[22] H. Sameti. *"Model-Based Approaches to Speech Enhancement: Stationary-State and Nonstationary-State HMMs"*. PhD thesis, University of Waterloo, Waterloo, Ontario, Canada, 1994.

[23] J. Rheem K. Lee and K. Shirai. "Recursive estimation based on the trended HMM in speech enhancement". *Proceedings of IEEE Asia Pacific Conf. on Circuits and Systems*, 1:239–242, 1996.

[24] G. Ruske K. Lee, J. Kee and K. Y. Kee. "Speech Recognition and Enhancement by a Nonstationary AR HMM with Gain Adaptation under Unknown Noise". *IEEE Transactions on Speech and Audio Processing*, 1999.

Feature	Smoothness	Additivity	Reversibility
Waveform	$\surd$	$\surd$	$\surd$
DFT (STFT)	$\surd$	$\surd$	$\surd$
MFSC	$\surd$	–	–
MFCC	$\surd$	–	–
LPC (AR)	–	–	$\surd$
R_{xx}	$\surd$	$\surd$	$\surd$
k	$\surd$	–	$\surd$
LSP	–	–	$\surd$

Table 1: Properties of various speech features.

	Order 0	Order 1	Order 2	Clean Data
Output SNR (dB)	9.45	10.53	11.01	11.43
Log-Likelihood	-1.6196e+05	-1.5883e+05	-1.5769e+05	–

Table 2: The output SNRs and log-likelihoods from the diagnostic tests with different orders of the NS-HMM trend polynomial and the output SNR using clean speech information.

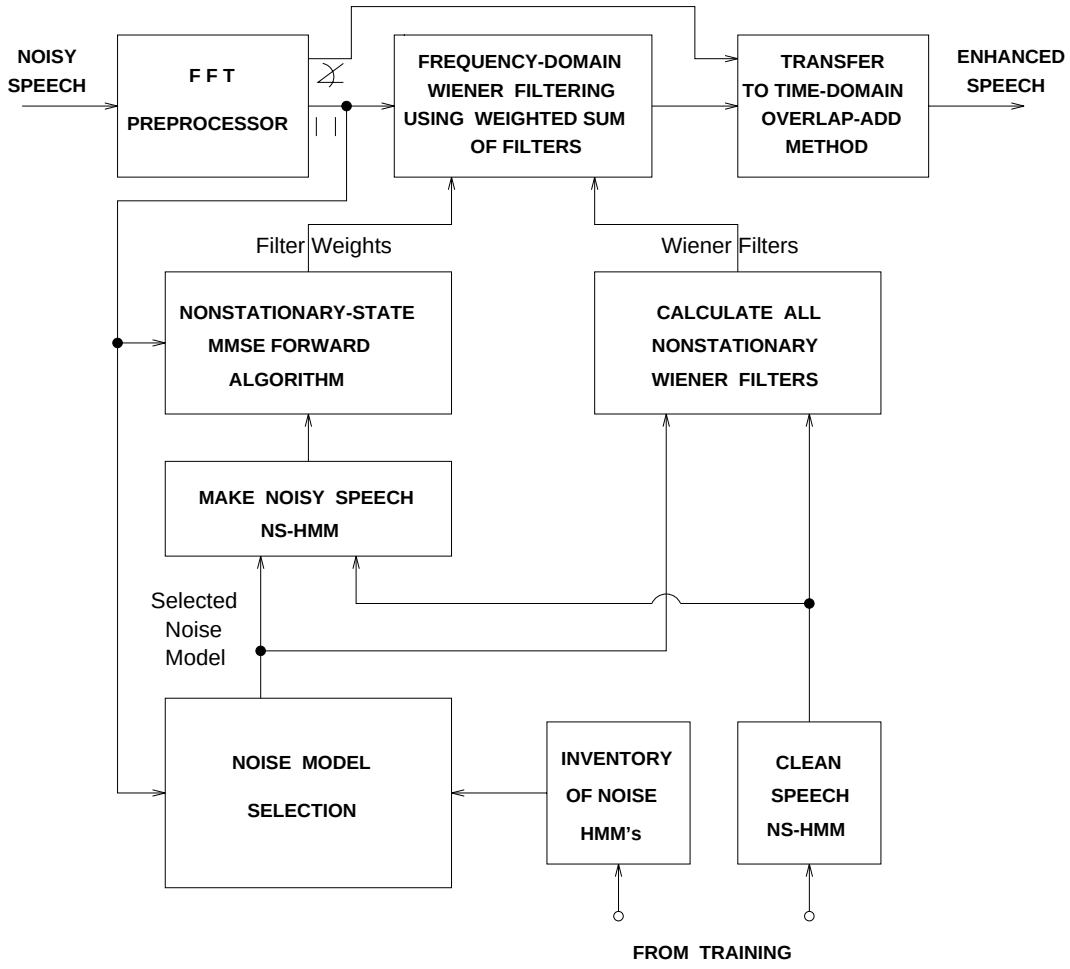


Figure 1: Block diagram of the MMSE enhancement system using the NS-HMM for clean speech.

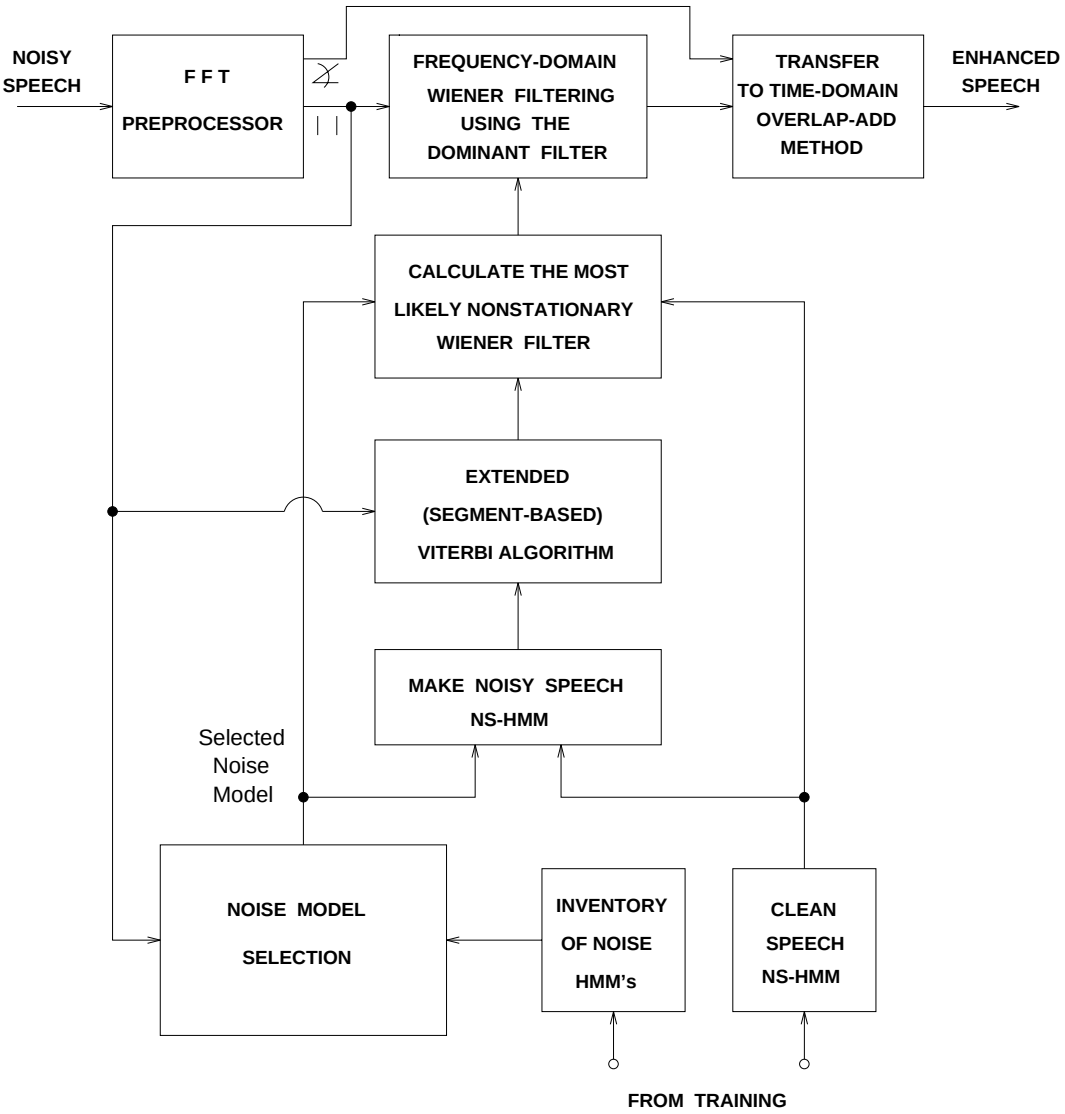


Figure 2: Block diagram of the approximate MMSE enhancement system using the NS-HMM for clean speech.

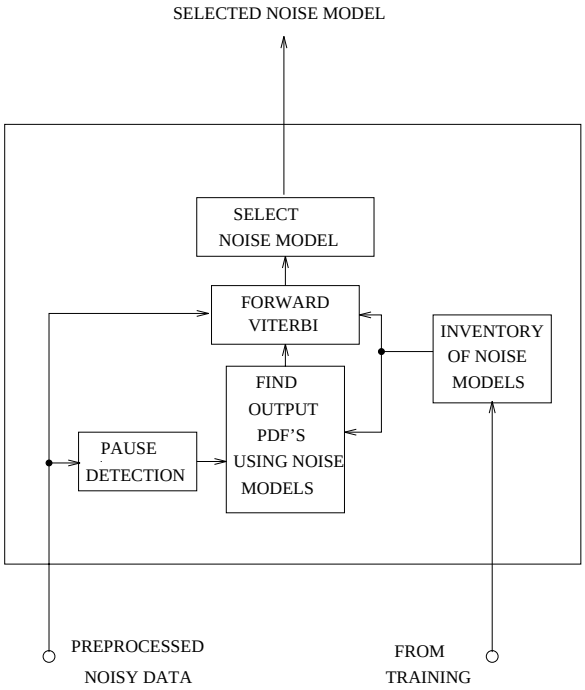


Figure 3: Block diagram of the noise-model-selection module in the speech enhancement system

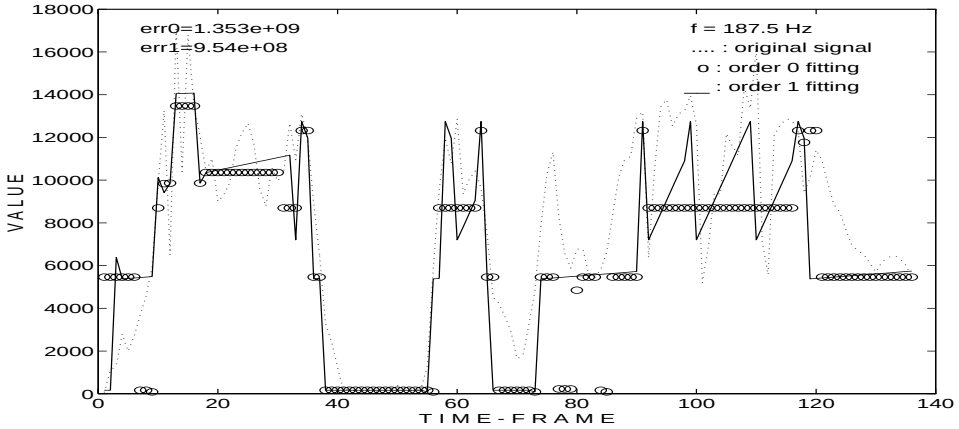
	Order 0	Order 1	Order 2
Output SNR (dB)	6.60	7.50	7.81

Table 3: The output SNRs from the diagnostic tests using only the estimated mean trajectories of the clean speech, with different orders of the NS-HMM trend polynomial.

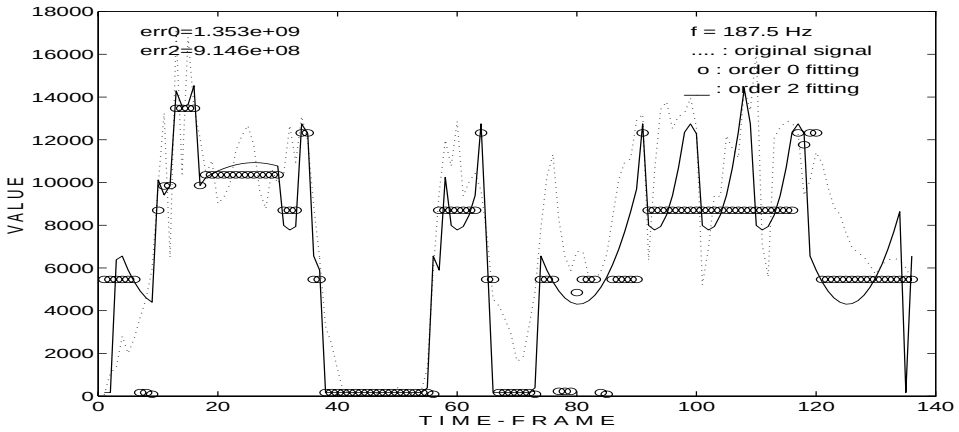
Input SNR (dB)	Order 0	Order 1	Order 2	Order 3	Order 4
0	7.51	7.60	7.62	7.65	7.67
5	8.22	9.10	9.12	9.13	9.15
10	8.97	10.95	10.95	11.02	11.15
15	9.91	12.15	12.33	12.38	12.50

Table 4: The output SNRs for different input SNRs and different orders of NS-HMM. White additive noise is used.

a)



b)



c)

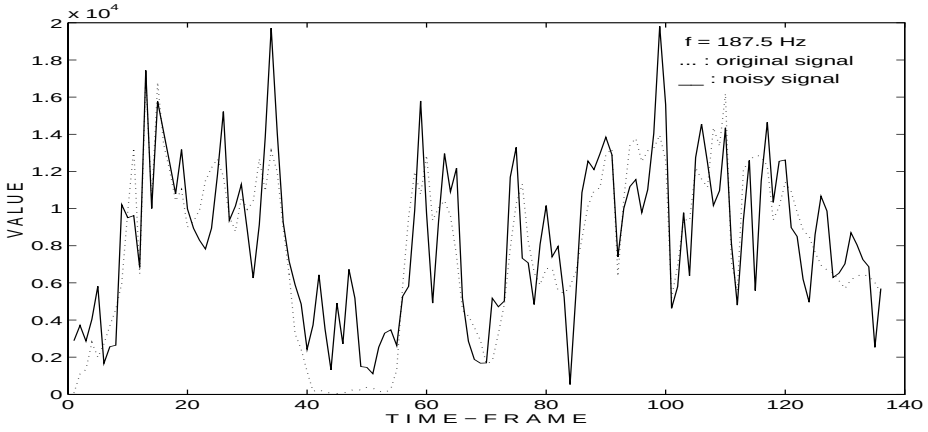
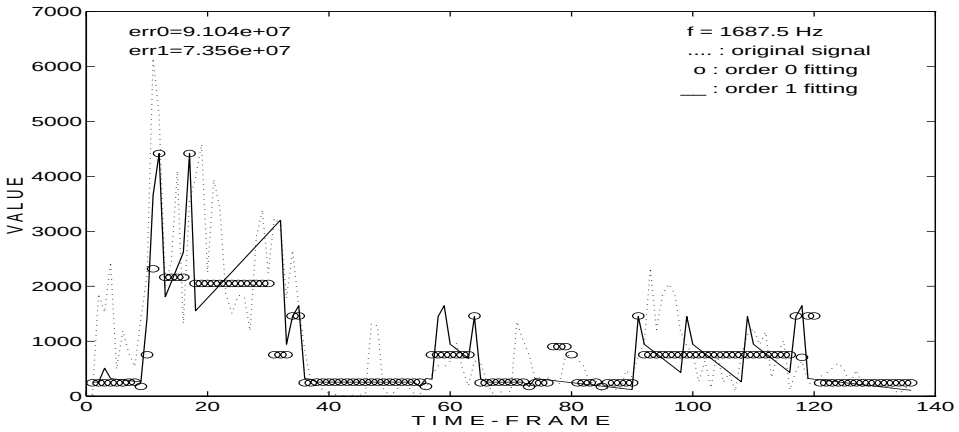
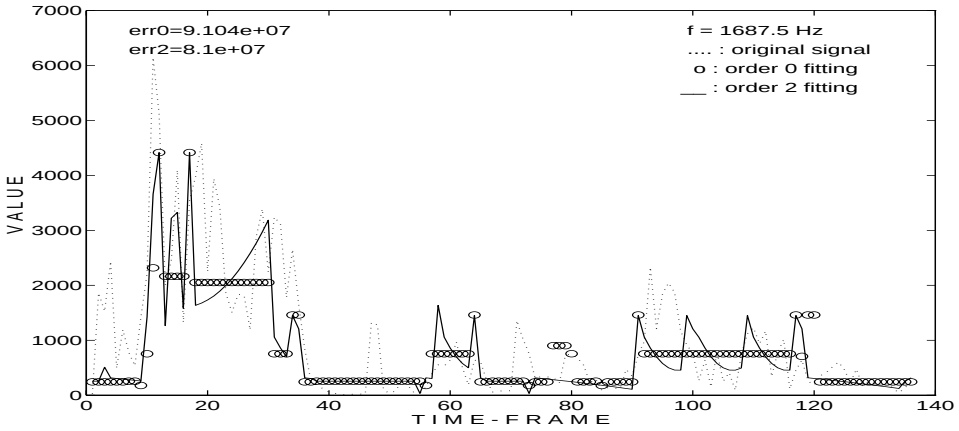


Figure 4: HMM fitting with the clean signal using noisy input with varying polynomial orders. DFT magnitude component 3, frequency=187.5 Hz. a) order 0 and order 1 fitting b) order 0 and order 2 fitting c) noisy input and clean signal. err0: total fitting error between the original signal and the order-0 model; err1: error between the original signal and the order-1 model; err2: error between the original signal and the order-2 model

a)



b)



c)

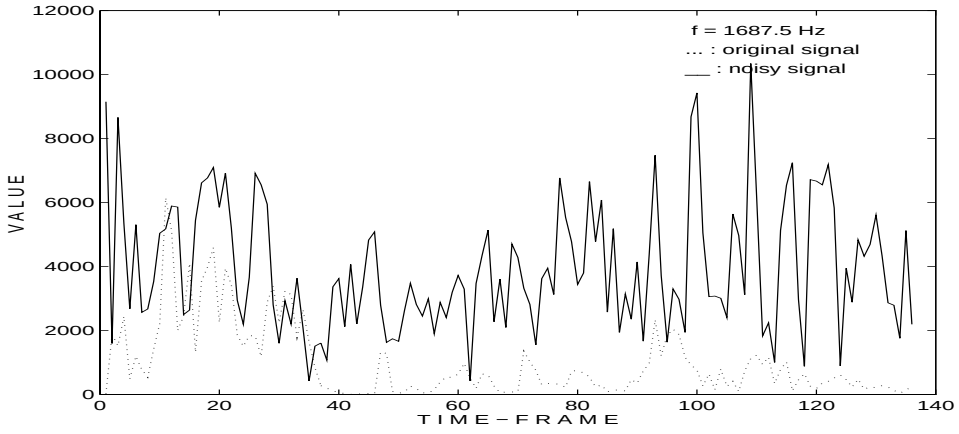


Figure 5: HMM fitting with the clean signal using noisy input with varying polynomial orders. DFT magnitude component 27, frequency=1687.5 Hz. a) order 0 and order 1 fitting b) order 0 and order 2 fitting c) noisy input and clean signal. err0: total fitting error between the original signal and the order-0 model; err1: error between the original signal and the order-1 model; err2: error between the original signal and the order-2 model

Input SNR (dB)	Order 0	Order 1	Order 2	Order 3	Order 4
0	7.98	8.15	8.21	8.22	8.30
5	8.70	9.62	9.71	9.81	9.88
10	9.15	10.92	10.95	11.12	11.20
15	9.81	12.20	12.32	12.38	12.55

Table 5: The output SNRs for different input SNRs and different orders of Simulated helicopter noise is used.