

# HMM-Based Reconstruction of Unreliable Spectrographic Data for Noise Robust Speech Recognition

Bengt J. Borgström, *Student Member, IEEE*, and Abeer Alwan, *IEEE Fellow*

**Abstract**—This paper presents a framework for efficient HMM-based estimation of unreliable spectrographic speech data. It discusses the role of Hidden Markov Models (HMMs) during minimum mean-square error (MMSE) spectral reconstruction. We develop novel HMM-based reconstruction algorithms which exploit intra-channel (across-time) correlation and/or inter-channel (across-frequency) correlation. For the sake of computational efficiency, this paper utilizes approximations to HMM-based decoding methods by developing models constructed from lower resolution quantizers. State configurations for lower resolution models are obtained through a tree-structured mapping of quantizer centroids, and model parameters are adapted accordingly. HMM downsampling avoids expensive re-training of models, and eliminates unnecessary memory requirements. Explicit general formulae are presented for the adaptation of steady-state and transitional statistics. Adaptation of observation statistics are derived from stochastic models of noise spectral magnitude estimation accuracies. The proposed estimation methods are applied in combination with oracle masks, which provide an upper performance bound, as well as masks derived from speech presence probability, which represent a more realistic scenario. Both methods are shown to boost noise robust recognition accuracies significantly relative to the Mel-Frequency Cepstral Coefficient (MFCC) baseline system. Furthermore, HMM downsampling greatly reduces the complexity of the HMM-based reconstruction method while negligibly affecting results.

**Index Terms**—Hidden Markov Models, Missing Features, Spectral Reconstruction, Noise Robust Speech Recognition, Mask Estimation.

## I. INTRODUCTION

In real world automatic speech recognition (ASR) applications, speech signals will generally suffer degradation by acoustic noise, resulting in decreased system performance [1]. Traditionally, the problem of noise robust ASR has been approached in front end feature extraction by reducing variability due to noise while retaining important discriminative information [2],[37]. As an alternative to the previous studies, the authors of [5] explore the missing features approach to speech recognition, in which unreliable spectral components are detected and compensated for accordingly.

Missing feature (MF) algorithms for robust speech recognition can be grouped into two main approaches: marginalization and data imputation. The marginalization approach utilizes information regarding the reliability of spectral features in the back-end by deemphasizing posterior observation probabilities corresponding to unreliable features during the recognition

process [6]-[8]. Similar MF-based techniques have been applied to channel mitigation for distributed speech recognition (DSR) [9]. The data imputation approach reconstructs unreliable spectral features in the front end so that recognition is based on estimated spectral information [10], [11]. Figure 1 illustrates the general overview for a noise robust ASR system based on the missing feature data imputation approach.

This paper presents a data imputation framework utilizing minimum mean-square error (MMSE) estimation of missing features for noise robust speech recognition. Hidden Markov Models (HMMs) have been extensively used in signal processing and communications due to their elegant framework which captures steady-state, transitional, and observation statistics [12]-[14]. Recently, HMMs have been successfully used to model feature trajectories during the estimation of missing data due to lost packets during the transmission of digital information [15]-[19].

In this paper, we propose an HMM-based approach for the reconstruction of spectral speech components degraded by acoustic noise. By implicitly quantizing spectrographic data, feature trajectories can be interpreted as transitioning through a HMM-defined trellis, with respect to time or with respect to frequency. The proposed method utilizes observed speech together with a local noise estimate to compute observation statistics. With the aforementioned transitional and observation information, the proposed HMM-based missing data algorithm uses the traditional forward-backward algorithm [12] to obtain optimal spectral estimates in the MMSE sense. This paper presents separate reconstruction methods which exploit intra-channel correlation, inter-channel correlation, or a combination of both.

A general downside to the use of HMMs is the induced computational load. In [25], the authors explore the use of multi-resolution models for DSR channel mitigation. We present efficient approximations to HMM-based estimation methods for the task of spectral reconstruction by means of HMM downsampling. Here, we utilize a tree-structured mapping of quantizer centroids, allowing HMMs to be downsampled, and corresponding statistical parameters to be adapted accordingly. By applying our downsampling framework instead of using predetermined multi-resolution models, we avoid expensive retraining of lower resolution models and unnecessary memory requirements. This paper provides explicit formulae for the adaptation of steady-state and transitional statistics. Adaptation of observation statistics are derived from stochastic models of noise spectral magnitude estimation accuracies. The proposed

Portions of this work were presented at ICASSP 2008 [19] and Interspeech 2008 [20].

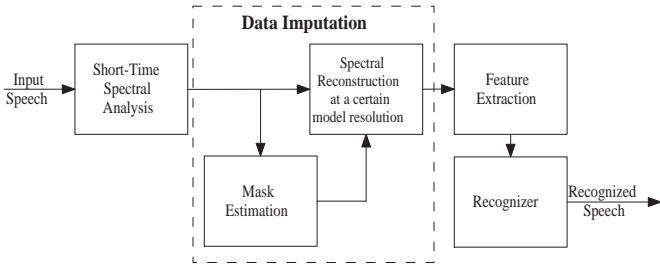


Fig. 1. Overview of Noise Robust Speech Recognition System, Featuring HMM-Based Spectral Reconstruction

model downsampling method is shown to greatly reduce the required complexity of missing feature estimation while negligibly affecting recognition results. We have previously applied the downsampled HMM framework with success to error concealment in remote speech recognition [19].

A major component of missing feature approaches for robust speech recognition systems is mask estimation, which detects the spectral location of reliable features. Many missing feature studies include results based upon *oracle masks*, for which exact knowledge of a clean version of the input speech signal is known. Knowledge of oracle masks provides an upper performance bound for data imputation techniques [5], [10], [11]. However, in order to convey realistic results, mask estimation must be performed. In [24] and [26], the authors present mask estimation techniques which train each frequency band separately as a Bayesian classifier, based on a variety of features. In this paper, we develop spectral masks based on speech presence probability (SPP) [27], in which the probability of active speech is based upon the statistical distribution of speech and noise spectral magnitudes. Note that optimization of spectral masks is beyond the scope of this paper: we provide SPP masks as an illustrative example.

This paper is structured as follows. In Section II, we present a framework for HMM-based estimation of missing spectrographic data. Specifically, we present separate techniques to exploit inter-channel correlation, intra-channel correlation, or a combination of both. In Section III, we present a method for HMM-downsampling for efficient estimation, and discuss the resulting reduction in complexity. Mask estimation is discussed in Section IV. Experimental results are presented in Section V, followed by a summary and conclusion in Section VI.

## II. RECONSTRUCTION OF MISSING SPECTRAL FEATURES

Missing feature approaches to robust speech recognition have been proven successful in difficult acoustic environments [5]. In this section, we present HMM-based estimation methods for missing spectrographic data.

### A. The Role of HMMs in Spectral Reconstruction

After short-time analysis, the observed magnitude spectro-temporal representation of a speech signal can be expressed as  $x_m(n)$ , where  $m$  denotes frequency channel index and  $n$  denotes discrete time. The underlying clean speech and noise magnitudes are denoted by  $s_m(n)$  and  $d_m(n)$ , respectively.

Note that the missing feature estimation algorithms presented in this study are robust to the exact spectro-temporal representation used. For example, common analysis methods used for automatic speech recognition (ASR) include the short-time Fourier Transform (STFT), the Mel-Filtered STFT, and short-time Linear Prediction Coefficients [14]. For illustrative purposes, this study will focus on the Mel-Filtered STFT.

A common method to model data trajectories is with an HMM [15]-[19]. Due to the discrete nature of HMM states, features must be quantized, at least implicitly, prior to the estimation process. Note that when used in this study, the term quantizer is used slightly differently than in the traditional sense, such as in source coding. The aim of quantization in our work is not data compression, but rather to designate to each decoded feature element a corresponding HMM state.

Let the quantizer of channel  $m$  be represented by the set of scalar-valued centroids  $\{c_m^1, c_m^2, \dots, c_m^N\}$ . Channel-specific quantizers can be designed based on empirical data using clustering techniques such as the K-means algorithm [32]. In this study, quantizers were trained on clean data from the Aurora-2 database.

Let the quantization of the clean spectro-temporal feature  $s_m(n)$  be performed by minimizing Euclidean distance:

$$q_m(s_m(n)) = c_m^i, \text{ where } i = \arg \min_j \|s_m(n) - c_m^j\|^2. \quad (1)$$

Furthermore, we define the set  $\{z_m^{0,1}, z_m^{1,2}, \dots, z_m^{N,N+1}\}$  as boundaries between centroids such that the region in feature space confined by  $z_m^{i-1,i}$  and  $z_m^{i,i+1}$  can be interpreted as the cell corresponding to  $c_m^i$ . Boundary values are determined as:

$$z_m^{i,i+1} = \begin{cases} \frac{1}{2}(c_m^i + c_m^{i+1}), & \text{if } 0 < i < N, \\ 0, & \text{if } i = 0, \\ \infty, & \text{if } i = N \end{cases}. \quad (2)$$

In order to apply estimation methods, separate HMMs are constructed for each channel-specific quantizer to model feature trajectories. Let the HMM applied to the output signal from channel  $m$  be referred to as  $\Lambda_m = (\mathbf{A}_m, \mathbf{B}_m, \boldsymbol{\pi}_m)$ , where  $\mathbf{A}_m$  provides transitional statistics,  $\mathbf{B}_m$  provides observation statistics, and  $\boldsymbol{\pi}_m$  provides steady-state statistics [12]. The steady-state probabilities of  $\Lambda_m$  can be determined empirically from training data as:

$$\pi_i^m = \frac{\text{no. of samples quantized to centroid } c_m^i(n)}{\text{total no. of samples}}. \quad (3)$$

The transitional probabilities of  $\Lambda_m$  can similarly be determined from training data as:

$$a_{ij}^m = \frac{\text{no. of samples transitioning from } c_m^i(n) \text{ to } c_m^j(n+1)}{\text{no. of samples quantized to } c_m^i(n)}. \quad (4)$$

In this study, steady-state and transitional statistics were obtained from Aurora-2 clean training data.

Figure 2 shows example graphical versions of transitional probability matrices obtained empirically from the clean training data in the Aurora-2 database [21]. For illustrative purposes, log-probabilities are shown. As can be observed, there

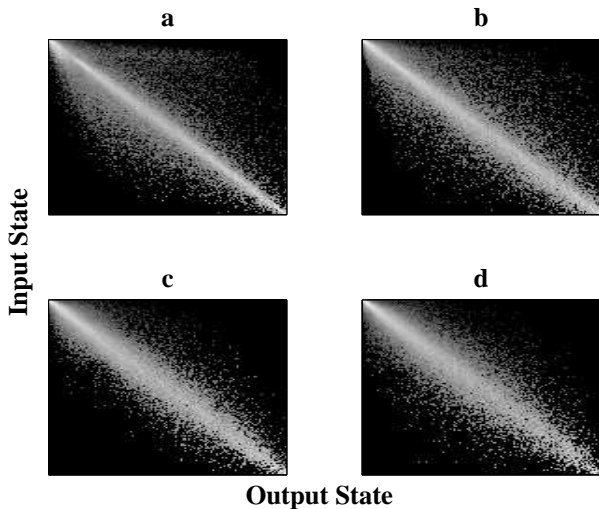


Fig. 2. Example Transitional Probability Matrices,  $\mathbf{A}_m$ : For illustrative purposes, the log-probabilities are shown. The y-axes correspond to the input state and the x-axes correspond to the output state. Panels **a** through **d** refer to Mel-Filtered Short-Time Fourier Transform (M-STFT) channels with center frequencies 313 Hz, 734 Hz, 1328 Hz, and 2188 Hz, respectively.

exist strong diagonal trends in the transitional probability matrices. These diagonal trends correspond to low-entropy signals, and such statistical patterns further motivate the use of transitional probabilities during the estimation process.

The observation statistic  $b_i^m(x_m(n))$  represents the probability that the underlying clean speech component  $s_m(n)$ , corresponding to centroid  $c_m^i$ , is observed as the noisy feature  $x_m(n)$ :

$$b_i^m(x_m(n)) = p(x_m(n) | c_m^i). \quad (5)$$

Steady-state and transitional statistics are described by discrete distributions due to the discrete configuration of underlying signal models. Observation statistics, on the other hand, are continuous distributions, since they incorporate the reliability of noise magnitude estimation.

The proposed algorithm builds upon an initial estimate of clean underlying speech spectrum. While more complex estimators can be utilized, we assume a simple initial spectral approximation which utilizes magnitude spectral subtraction (MSS) [33]:

$$\hat{s}_m(n) = x_m(n) - \hat{d}_m(x_m(n)), \quad (6)$$

where  $\hat{d}_m(x_m(n))$  represents a local estimate of the noise spectral magnitude. It is assumed that the estimate  $\hat{d}_m(x_m(n))$  is obtained with a certain level of accuracy, such that the actual hidden noise process,  $d_m(n)$ , can be expressed as:

$$d_m(n) = \hat{d}_m(x_m(n)) + \varepsilon(n), \quad (7)$$

where  $\varepsilon_m(n)$  is the estimation error. We model  $\varepsilon(n)$  as a stationary zero-mean Gaussian process with standard deviation  $\sigma_\varepsilon$ . By assuming  $\varepsilon(n)$  to be zero-mean, we imply that the

noise estimation technique utilized does not produce bias. However, if noise estimation is known to produce bias, this can be addressed by shifting the mean of  $\varepsilon(n)$ .

Note that  $\sigma_\varepsilon$  is related to the accuracy of the noise estimation process. For example, if  $\hat{d}_m(x_m(n))$  can be determined exactly, then  $\sigma_\varepsilon=0$ ; conversely as the noise estimation process becomes less accurate,  $\sigma_\varepsilon$  increases. It therefore follows intuitively that the standard deviation of the noise estimation process should be related to the magnitude of the estimated noise. In this study, we use the rough approximation:

$$\sigma_\varepsilon = \kappa \hat{d}_m(x_m(n)), \quad (8)$$

where  $\kappa$  was empirically set to 0.6.

Assuming a Gaussian distribution for the estimation error of the noise spectral magnitude, the observation probability distribution is given by:

$$b_i^m(x_m(n)) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \int_{z_m^{i-1,i}}^{z_m^{i,i+1}} e^{-\frac{(\tau - \hat{d}_m(x_m(n)))^2}{2\sigma_\varepsilon^2}} d\tau, \quad (9)$$

where  $\sigma_\varepsilon$  is approximated by Equation 8.

It is important to note that the methods described for obtaining the relationship in Equation 8 should not be considered the matched-case or stereo-training scenario, as in [35], which requires extensive corresponding clean and noisy speech databases. Instead, Equation 8 is a rough approximation involving a single parameter which is used to adapt the system to large variations in component-specific input SNR. More advanced relationships between  $\sigma_\varepsilon$  and  $\hat{d}_m(x_m(n))$  can be developed, and may be expected to result in more accurate observation probabilities.

## B. HMM-Based Estimation Methods

Once underlying signal sources have been modeled by the set  $\{\Lambda_1, \dots, \Lambda_M\}$ , unreliable spectrographic data can be approximated by means of HMM-based estimation. In this section, we present HMM-based spectral reconstruction methods which exploit intra-channel correlation, inter-channel correlation, or a combination of both.

1) *Utilizing Correlation Across Time*: The forward-backward algorithm provides an accurate algorithm for estimating missing or ambiguous data within an HMM framework [15],[16]. The FB algorithm determines the MMSE spectral estimate  $\hat{s}_m(n)$  given the first and last reliable features at temporal indices  $n - n_1$  and  $n + n_2$ , and given the series of observations  $x_m(n - n_1), \dots, x_m(n + n_2)$ . Figure 3 provides an illustrative example of the missing feature problem formulation.

The estimate of component  $s_m(n)$ , using the FB algorithm and exploiting intra-channel correlation, is determined as [12]:

$$\hat{s}_m(n) = \sum_{i=1}^N c_m^i \gamma_{T,m}^i(n), \quad (10)$$

where:

$$\gamma_{T,m}^i(n) = \frac{\alpha_m^i(n) \beta_m^i(n)}{\sum_{j=1}^N \alpha_m^j(n) \beta_m^j(n)}. \quad (11)$$

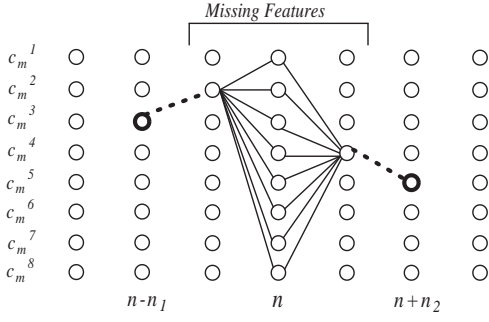


Fig. 3. An Illustrative Example of the Missing Feature Problem Formulation: In this case, the quantizer is comprised of  $N=8$  centroids. A series of 3 features are missing. In this figure, the feature at time index  $n$  is to be estimated, so that  $n_1=2$  and  $n_2=2$ .

The set of values  $\gamma_{T,m}^i$  represents the conditional distribution of  $q_m(s_m(n))$ , conditioned on the first and last reliable features, as well as on past and future noisy observations:

$$\gamma_{T,m}^i(n) = P(c_m^i | x_m(n-n_1), \dots, x_m(n+n_2)). \quad (12)$$

Note that the subscript  $T$  denotes the application of HMM-based decoding across time. The values  $\alpha_m^i(n)$  and  $\beta_m^i(n)$ , known as the forward and backward variables, respectively, can be determined recursively as:

$$\alpha_m^i(n) = \left[ \sum_{j=1}^N a_{ji}^m \alpha_m^j(n-1) \right] b_i^m(x_m(n)), \quad (13)$$

$$\beta_m^i(n) = \sum_{j=1}^N a_{ij}^m \beta_m^j(n+1) b_j^m(x_m(n+1)).$$

The forward variable conveys the probability of a given centroid conditioned on past and present observations, whereas the backward variable conveys the probability of a given centroid conditioned on future observations:

$$\alpha_m^i(n) = P(c_m^i | x_{n-n_1}, \dots, x_n) \quad (14)$$

$$\beta_m^i(n) = P(c_m^i | x_{n+1}, \dots, x_{n+n_2}).$$

The forward and backward variables at the last and first reliable feature vectors are known as:

$$\alpha_m^i(n-n_1) = \begin{cases} 1, & \text{if } q_m(\hat{s}_m(n-n_1)) = c_m^i \\ 0, & \text{else} \end{cases} \quad (15)$$

and:

$$\beta_m^i(n+n_2) = \begin{cases} 1, & \text{if } q_m(\hat{s}_m(n+n_2)) = c_m^i \\ 0, & \text{else} \end{cases} \quad (16)$$

The HMM-based spectral reconstruction method utilizing intra-channel (across-time) correlation is referred to as  $\mathbf{FB}_T$ .

2) *Utilizing Correlation Across Frequency Channels*: In many missing data scenarios, such as speech communication and remote speech recognition, unreliable data is a result of corrupt bits or dropped packets [15]-[19], and thus utilizing correlation along the frequency axis is not applicable. In the current study, however, this is not the case, and we wish to exploit the strong inter-channel correlation present within speech during spectrogram reconstruction. It is interesting to note that [11] previously utilized inter-channel correlation by employing a full covariance Gaussian mixture model during imputation of missing spectro-temporal components of speech. When performing estimation along the frequency axis, the stationarity assumed by the transitional statistics of  $a_{ij}^m$  does not hold. It is therefore necessary to introduce the probabilities:

$$P_{k:l}^{i:j} = \begin{cases} P(q_k(s_k(n)) = c_k^i | q_l(s_l(n)) = c_l^j), & \text{for } |k-l|=1, \\ 0, & \text{else} \end{cases} \quad (17)$$

which define the inter-channel transitional statistics of spectrographic data. That is,  $P_{k:l}^{i:j}$  represents the probability that the spectrographic feature in channel  $k$  corresponds to centroid  $i$ , given that the spectrographic feature in channel  $l$  is quantized to centroid  $j$ . Note that since HMM-based processing assumes a 1<sup>st</sup> order signal model, then  $|k-l|=1$  must hold. The probability  $P_{k:l}^{i:j}$  can be determined from training data as:

$$P_{k:l}^{i:j} = \frac{\text{no. of samples transitioning from } c_k^i(n) \text{ to } c_l^j(n)}{\text{no. of samples quantized to } c_k^i(n)}. \quad (18)$$

Modified forward and backward variables can be expressed for the estimation of missing data along the frequency axis:

$$\delta_m^i(n) = \left[ \sum_{j=1}^N P_{m:m-1}^{i:j} \delta_{m-1}^j(n) \right] b_i^m(x_m(n)), \quad (19)$$

$$\epsilon_m^i(n) = \sum_{j=1}^N P_{m:m+1}^{j:i} \epsilon_{m+1}^j(n) b_j^{m+1}(x_{m+1}(n)). \quad (20)$$

Note that  $\delta_m^i(n)$  and  $\epsilon_m^i(n)$  are determined similarly to the forward and backward variables expressed in Eq. (13), with the exception that the stationary transitional probability  $a_{ij}^m$  is replaced by  $P_{k:l}^{i:j}$ .  $\delta_m^i(n)$  and  $\epsilon_m^i(n)$  each represents a distribution of  $q_m(s_m)$  conditioned on different subsets of inter-channel observations:

$$\delta_m^i(n) = P(c_m^i | x_{m-m_1}(n), \dots, x_m(n)) \quad (21)$$

$$\epsilon_m^i(n) = P(c_m^i | x_{m+1}(n), \dots, x_{m+m_2}(n)).$$

Here,  $m-m_1$  and  $m+m_2$  denote the channel indices of the first and last reliable components within the current feature vector. In this application, referred to as  $\mathbf{FB}_F$ ,  $\gamma_{F,m}^i(n)$  represents the distribution of  $q_m(s_m(n))$  conditioned on reliable components and noisy observations from the feature vector at time index  $n$ :

$$\gamma_{F,m}^i(n) = P(c_m^i | x_{m-m_1}(n), \dots, x_{m+m_2}(n)). \quad (22)$$

The corresponding expression for determining  $\gamma_{F,m}^i(n)$  is given by:

$$\gamma_{F,m}^i(n) = \frac{\delta_m^i(n) \epsilon_m^i(n)}{\sum_{j=1}^N \delta_m^j(n) \epsilon_m^j(n)}. \quad (23)$$

The forward and backward variables for inter-channel estimation are known at the last and first reliable spectral components as:

$$\delta_{m-m_1}^i(n) = \begin{cases} 1, & \text{if } q_{m-m_1}(s_{m-m_1}(n)) = c_{m-m_1}^i \\ 0, & \text{else} \end{cases} \quad (24)$$

and:

$$\epsilon_{m+m_2}^i(n) = \begin{cases} 1, & \text{if } q_{m+m_2}(s_{m+m_2}(n)) = c_{m+m_2}^i \\ 0, & \text{else} \end{cases}. \quad (25)$$

3) *Utilizing Correlation Across Time and Across Frequency Channels:* We additionally propose the estimation of unreliable spectral components utilizing statistics along both the time and frequency axes, referred to as  $\mathbf{FB}_{2D}$ . In this scenario,  $\gamma_{2D,m}^i(n)$  is the distribution of  $x_m(n)$  conditioned on data contained in both channel  $i$  and feature vector  $\mathbf{f}(n)$ :

$$\gamma_{2D,m}^i(n) = P(c_m^i | x_m(n-n_1), \dots, x_m(n+n_2)) \times P(c_m^i | x_{m-m_1}(n), \dots, x_{m+m_2}(n)) \quad (26)$$

Equation 26 leads to the conditional probabilities  $\gamma_{2D,m}^i(n)$  being expressed as:

$$\begin{aligned} \gamma_{2D,m}^i(n) &= \gamma_{T,m}^i(n) \gamma_{F,m}^i(n) \\ &= \frac{\alpha_m^i(n) \beta_m^i(n) \delta_m^i(n) \epsilon_m^i(n)}{\sum_{j=1}^N \alpha_m^j(n) \beta_m^j(n) \sum_{k=1}^N \delta_m^k(n) \epsilon_m^k(n)}. \end{aligned} \quad (27)$$

For the conditional distributions given by Equations 23 and 27, the corresponding estimates  $\hat{s}_m(n)$  are determined similarly to Equation 10.

### III. EFFICIENT APPROXIMATION OF HMM-BASED ESTIMATION TECHNIQUES

#### A. Downsampling of Statistical Models

A well-known downside to HMM-based processing is the induced computational load. For large codebooks (large  $N$ ), the required complexity may cause HMM-based algorithms to be infeasible for resource constrained or delay-sensitive applications. We propose a framework for efficient HMM-based missing feature estimation based on downsampling of underlying statistical models, which we previously presented in [19] and [20]. Instead of using the original quantizer centroids to construct signal models, we use quantizers with less resolution to build statistical models. We implement a tree-structure mapping of centroids to allow downsampling of the discrete HMMs by factors of 2, with  $N=2^R$ . One could train a higher resolution quantizer, but test with various lower resolution systems without explicit retraining. By using the

model downsampling framework, the system has access to multiple resolutions of HMM-based processing while only retaining one model in memory. This is especially important for distributed and/or resource-constrained applications.

In general, let a  $R$ -bit quantizer for channel  $m$  be comprised of centroids  $\{c_m^{R,1}, c_m^{R,2}, \dots, c_m^{R,2^R}\}$ . (Note that typically the original quantizer is allocated up to 8 bits [15]-[17].) Similarly, the multi-resolution quantizations of clean speech are denoted by:

$$q_m^R(s_m(n)) = c_m^{R,i}, \text{ where } i = \arg \min_j \|s_m(n) - c_m^{R,j}\|^2. \quad (28)$$

Using tree-structure quantization, centroids can be mapped according to:

$$c_m^{8,i} \Rightarrow c_m^{R,j}, \text{ for } 1 \leq R < 8, \text{ where } j = \left\lfloor \frac{i}{2^{8-R}} \right\rfloor. \quad (29)$$

In Eq. 29, neighboring centroids are simply grouped in pairs as  $R$  decreases. In this manner, centroids can easily be regrouped as clusters without requiring expensive retraining of quantization codebook. Centroid boundaries,  $z^{(i,i+1),R}$ , are adapted similarly. Although the proposed framework is robust to the resolution chosen for  $\tau$ , we will set  $\tau=8$  for further derivations.

#### B. Adaptation of Statistical Parameters

The signal model, now referred to as  $\Lambda_m^R$ , has statistical parameters  $(\mathbf{A}_m^R, \mathbf{B}_m^R, \boldsymbol{\pi}_m^R)$ . The steady-state and transitional statistics can be approximated according to:

$$\pi_i^{m,R} = \sum_{k=0}^{2^{8-R}-1} \pi_{(2^{8-R}i+k)}^{m,8}, \quad (30)$$

and:

$$\begin{aligned} a_{ij}^{m,R} &= \\ \frac{1}{2^{8-R}} &\left[ \frac{\sum_{k=0}^{(2^{8-R}-1)} \pi_{(2^{8-R}i+k)}^{m,8} \sum_{l=0}^{(2^{8-R}-1)} a_{(2^{8-R}i+k, 2^{8-R}j+l)}^{m,8}}{\sum_{k=0}^{(2^{8-R}-1)} \pi_{(2^{8-R}i+k)}^{m,8}} \right]. \end{aligned} \quad (31)$$

However, when quantizers are designed to minimize expected distortion, as is done in the current algorithm, the steady-state probabilities of quantizer centroids can be expected to be close to uniform. In this case, the formula in Equation 31 can be efficiently approximated as:

$$a_{ij}^{m,R} = \frac{1}{2^{8-R}} \left[ \sum_{k=0}^{(2^{8-R}-1)} \sum_{l=0}^{(2^{8-R}-1)} a_{(2^{8-R}i+k, 2^{8-R}j+l)}^{m,8} \right]. \quad (32)$$

During estimation of spectral components degraded by acoustic noise, the observation statistics  $b_i^m(x_m(n))$  correspond to the probability that a clean spectral value belonging to centroid  $c_m^i$  is corrupted by additive noise and observed

as  $x_m(n)$ . Hence, observation probability distributions of downsampled models are:

$$b_i^{m,R}(x_m(n)) = p(x_m(n) | c_m^{R,i}) \quad (33)$$

### C. Approximated HMM-Based Estimation

The HMM-based estimation techniques discussed in Section II-B can be applied to the downsampled statistical framework developed in Section III. The estimate of component  $x_m(n)$  using the approximation of the  $\mathbf{FB}_T$  algorithm with  $R$  bits of resolution becomes:

$$\hat{s}_m^R(n) = \sum_{i=1}^{2^R} c_m^{R,i} \gamma_m^{R,i}(n), \quad (34)$$

where:

$$\gamma_{T,m}^{R,i}(n) = \frac{\alpha_m^{R,i}(n) \beta_m^{R,i}(n)}{\sum_{j=1}^{2^R} \alpha_m^{R,j}(n) \beta_m^{R,j}(n)}. \quad (35)$$

The values  $\alpha_m^{R,i}(n)$  and  $\beta_m^{R,i}(n)$  can be determined recursively as:

$$\alpha_m^{R,i}(n) = \left[ \sum_{j=1}^{2^R} a_{ji}^{m,R} \alpha_m^{R,j}(n-1) \right] b_i^{m,R}(x_m(n)), \quad (36)$$

$$\beta_m^{R,i}(n) = \sum_{j=1}^{2^R} a_{ij}^{m,R} \beta_m^{R,j}(n+1) b_j^{m,R}(x_m(n+1)). \quad (37)$$

Similarly,  $x_m(n)$  can be estimated using the  $\mathbf{FB}_F$  algorithm with  $R$  bits of resolution via Equation 34, where:

$$\gamma_{F,m}^{R,i}(n) = \frac{\delta_m^{R,i}(n) \epsilon_m^{R,i}(n)}{\sum_{j=1}^{2^R} \delta_m^{R,j}(n) \epsilon_m^{R,j}(n)}. \quad (38)$$

In this case, the forward and backward variables are determined as:

$$\delta_m^{R,i}(n) = \left[ \sum_{j=1}^N P_{m:m-1}^{i:j} \delta_{m-1}^{R,j}(n) \right] b_i^{m,R}(x_m(n)), \quad (39)$$

$$\epsilon_m^{R,i}(n) = \sum_{j=1}^N P_{m:m+1}^{j:i} \epsilon_{m+1}^{R,j}(n) b_j^{m+1,R}(x_{m+1}(n)). \quad (40)$$

Finally, the  $\mathbf{FB}_{2D}$  algorithm estimates  $x_m(n)$  via Equation 34, where:

$$\begin{aligned} \gamma_{2D,m}^{R,i}(n) &= \gamma_{T,m}^{R,i}(n) \gamma_{F,m}^{R,i}(n) \\ &= \frac{\alpha_m^{R,i}(n) \beta_m^{R,i}(n) \delta_m^{R,i}(n) \epsilon_m^{R,i}(n)}{\sum_{j=1}^{2^R} \alpha_m^{R,j}(n) \beta_m^{R,j}(n) \sum_{k=1}^{2^R} \delta_m^{R,k}(n) \epsilon_m^{R,k}(n)}. \end{aligned} \quad (41)$$

Thus, HMM-based spectral reconstruction can be performed efficiently utilizing lower resolution statistical models, without expensive retraining or unnecessary memory requirements.

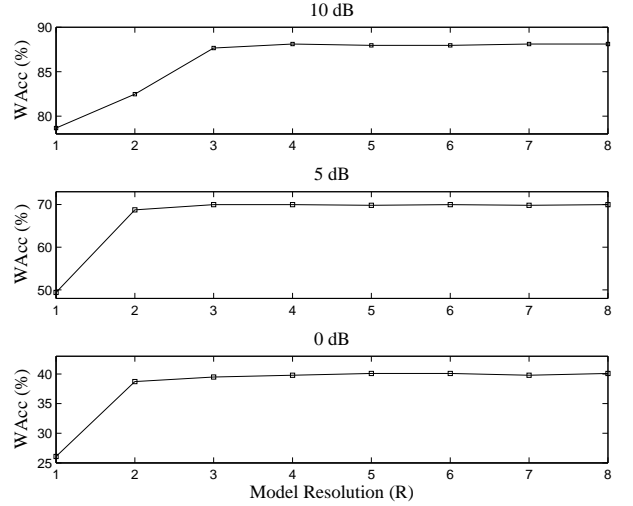


Fig. 4. The Effect of Model Downsampling on ASR Performance: Speech Presence Probability (SPP) Masks (see Section IV) were Used in Series with the  $\mathbf{FB}_T$  algorithm, at various signal SNRs, and for various model resolutions. The system was tested on the Aurora database [21], specifically the subset corrupted with vehicular noise.

### D. Performance and Complexity Analysis

The efficient HMM-based estimation techniques developed in Section III can intuitively be expected to degrade ASR performance as the resolution of underlying signal models is reduced. In this section, we examine the extent to which this performance degradation occurs, and the possible reductions in computational complexity. Figure 4 illustrates the effect of model downsampling on word accuracy. Here, SPP-based masks were used (see Section IV for details) in series with the  $\mathbf{FB}_T$  spectral reconstruction method, at various model resolutions. The system was tested on the Aurora database [21], specifically the subset corrupted by vehicular noise, at various SNR levels. As can be concluded from Figure 4, the proposed model downsampling method leads to minimal performance degradation for resolutions as low as  $R=3$ .

The proposed downsampling methods can greatly reduce the required complexity of the proposed estimation system. Note that the order of complexity for the forward-backward decoding algorithm applied to a fully connected model is  $O(2^{2N+1})$ , where  $N=2^R$  is the number of states comprising each underlying statistical model. The proposed estimation techniques were applied to a sample utterance, and the average number of operations per frame are provided in Table I. Note that Table I excludes operations required by calculation of observation statistics, which are independent of model resolution. The sample utterance from the Aurora-2 database was 2.2 seconds in duration, and was degraded by vehicular noise at 15 dB SNR. As can be concluded, the proposed HMM downsampling method greatly reduces the induced computational load of spectral reconstruction. For example, if the resolution of underlying models is decreased from the standard 8 bits to 3 bits, the required complexity is reduced by a factor of over 800.

Based on the performance analysis of Figure 4 and the

TABLE I

Average Operations Per Frame, as a Function of Model Resolution,  $R$ .  $\mathbf{FB}_T$  refers to intra-channel HMM-based estimation,  $\mathbf{FB}_F$  refers to inter-channel HMM-based estimation, and  $\mathbf{FB}_{2D}$  refers to a combination of both. These results were obtained on a single utterance from the Aurora-2 database, with the window update set to 100 Hz, and using 26 Mel channels.

Operation	multiplications	additions	
R=8	$\mathbf{FB}_T$	$4.342 \times 10^5$	$4.299 \times 10^5$
	$\mathbf{FB}_F$	$4.293 \times 10^5$	$4.253 \times 10^5$
	$\mathbf{FB}_{2D}$	$9.755 \times 10^5$	$9.664 \times 10^5$
R=5	$\mathbf{FB}_T$	7,132	6,591
	$\mathbf{FB}_F$	7,005	6,591
	$\mathbf{FB}_{2D}$	$1.594 \times 10^4$	$1.479 \times 10^4$
R=3	$\mathbf{FB}_T$	520	383
	$\mathbf{FB}_F$	501	375
	$\mathbf{FB}_{2D}$	1,145	854

complexity analysis of Table I, it can be concluded that the optimal tradeoff between recognition accuracy and induced computational load is obtained at a resolution of  $R=3$ . At this resolution, performance is affected negligibly relative to the standard 8-bit model system, at least for the Aurora-2 database. However, in terms of operations required, the complexity is reduced by a factor of over 800, and in terms of processing time, the complexity is reduced by a factor of over 130.

The analysis provided in this section regarding model down-sampling hints at the robustness of the recognizer to small changes in feature values. That is, the recognizer seems not to react to minor variability in features obtained from reconstructed spectrographic data at various resolutions. Instead, it appears that the general spectral shape of estimated data determines recognizer output.

#### IV. MASK ESTIMATION

A major component of missing feature approaches for robust speech recognition systems is mask estimation, which detects the spectral location of unreliable features. Optimization of spectral masks lies outside of the scope of this paper. However, in this section, we discuss oracle masks and present masks based on speech presence probability (SPP), which are used in combination with the proposed missing feature techniques to illustrate their effectiveness.

##### A. Oracle Masks

Previous studies involving missing feature approaches to speech recognition have included results based on oracle masks [5]-[11]. Oracle masks assume exact knowledge of the clean version of the input speech signal, and therefore provide an upper performance bound for missing feature techniques. Furthermore, they provide a method of grading the success of a missing feature algorithm that is independent of the mask estimation method used.

Let  $x_m(n)$  be spectrographic data from an observed noisy speech signal. Let  $s_m(n)$  be the spectrographic representation of the corresponding clean version, which represents oracle information. An oracle mask,  $M_{ORC}$ , can be determined as a simple SNR comparison:

$$M_{ORC}(m, n) = \begin{cases} 1, & \text{if } 20 \log \left( \frac{s_m(n)}{x_m(n) - s_m(n)} \right) > \zeta_{ORC} \\ 0, & \text{else} \end{cases}, \quad (42)$$

where  $\zeta_{ORC}$  is a predetermined cut-off. In this study,  $\zeta_{ORC}$  was optimized empirically to 0 dB. The oracle mask was post-processed by a 2-dimensional median filter [34] with the goal of retaining salient reliable segments, and of reducing false detection.

##### B. Spectral Masks Based on Speech Presence Probability

When implementing noise robust automatic recognition systems utilizing missing feature analysis, mask estimation is a key component, since in reality one cannot assume oracle information. Previous work has developed mask estimation techniques based on spectral subtraction and SNR criteria [29]. Other studies have derived mask estimation techniques which train frequency channels separately as individual binary classifiers, based on a variety of suitable features [24], [26]. In this section, we develop spectral masks based speech presence probability (SPP), which utilize statistical distributions of speech and noise spectral magnitudes. Similar approaches have been described in [27] and [30] for the task of speech enhancement, but to the extent of the authors' knowledge, have not been applied to missing feature-based ASR.

During the mask estimation process, we have access to the observed spectrographic data,  $x_m(n)$ , as well as an estimate of the spectral noise floor,  $\hat{d}_m(x_m(n))$ . It is reported in [22] and [31] that smoothing of observed spectrographic data with respect to time improves the accuracy of noise estimation by suppressing high frequency components of feature trajectories, which are characteristic of noise. We therefore smooth the observed data utilizing a 2-point running average.

In deriving SPP-based masks, we assume that in time, speech is observed from a two-state information source. The first state,  $H_0$ , corresponds to inactive speech, for which only noise is observed. The second state,  $H_1$ , corresponds to an active speech state, for which both speech and noise are observed.

Let  $\rho_m(n)$  represent the a posteriori SNR, and be defined as:

$$\rho_m(n) = \frac{x_m^2(n)}{\sigma_{N,m}^2(n)}, \quad (43)$$

where  $\sigma_{N,m}^2(n)$  is the channel-specific variance of the noise. Let  $\xi_m(n)$  denote the a priori SNR, which in this study is approximated by the ML estimate [38]:

$$\xi_m(n) = \frac{\max(x_m^2(n) - \sigma_{N,m}^2(n), 0)}{\sigma_{N,m}^2(n)} = \max(\rho_m(n) - 1, 0). \quad (44)$$

In this study, second-order noise statistics were approximated simply as:

$$\sigma_{N,m}^2(n) \approx \hat{d}_m^2(x_m(n)). \quad (45)$$

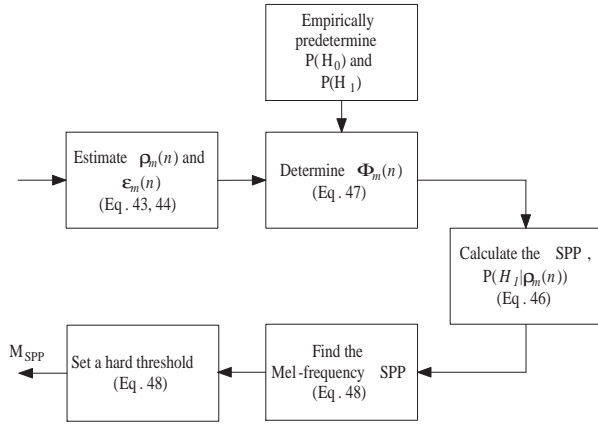


Fig. 5. Overview of the Mask Estimation Process for SPP-based Masks Assuming Knowledge of  $x_m(n)$  and  $\hat{d}_m(n)$

In determining a spectral mask, we are interested in calculating the value  $P(H_1|\rho_m(n))$ , since it represents the probability of the corresponding spectral location being in the active speech state given the current observed signal. If speech and noise spectral components are modeled as independent Gaussian processes, a Bayesian approach results in [38],[27]:

$$P(H_1|\rho_m(n)) = \frac{\Phi_m(n)}{1 + \Phi_m(n)}, \quad (46)$$

where the Generalized Likelihood Ratio (GLR) is given by:

$$\Phi_m(n) = \frac{P(H_1)}{P(H_0)} \frac{1}{1 + \xi_m(n)} e^{\xi_m(n)}. \quad (47)$$

In the above equation, the steady-state probabilities  $P(H_0)$  and  $P(H_1)$  can be trained from empirical data.

The previous derivation was performed on spectrographic data in the linear frequency scale. However, SPP-based masks can be transformed from the linear frequency scale to the Mel-filtered scale via multiplication with the matrix  $W(m, i)$ , a linearization of the Mel-filtering process [36].

According to Equation 46, the mask estimation method based on speech presence probability (SPP) returns a soft-decision value for each spectral location. However, the current task of missing feature reconstruction requires a binary mask. We therefore set a hard probability threshold,  $\zeta_{SPP}$ , and determine the Mel-scale binary SPP-based mask accordingly:

$$M_{SPP}(m, n) = \begin{cases} 1, & \text{if } \sum_i W(m, i) P(H_1|\rho_i(n)) > \zeta_{SPP} \\ 0, & \text{else} \end{cases} \quad (48)$$

In this study  $\zeta_{SPP}$  was empirically optimized to 0.4. Finally, the mask was post-processed by a 2-dimensional median filter. Figure 5 provides an overview of the mask estimation process for SPP-based masks.

Figure 6 illustrates examples of the SPP-based discussed mask estimation techniques. Panel **a** illustrates the Mel-Filtered spectrogram of a clean signal ("one two three seven seven four three"). Panel **b** represents the oracle mask of the corresponding signal at a noise level of 0 dB. Panel **c** represents the SSP-based mask for the corresponding noisy

signal. The given utterances were from the Aurora-2 database [21].

## V. EXPERIMENTAL RESULTS

The overall noise robust recognition system featuring spectral reconstruction was tested on the Aurora-2 database [21]. The proposed framework was trained using a model with resolution of  $R=8$ , but was tested with a downsampled model with  $R=3$ . This resolution was empirically shown in Section III-D to be the optimal tradeoff between performance and complexity for the given database. The baseline front-end extracted 39-dimensional MFCC vectors, including 1<sup>st</sup> and 2<sup>nd</sup> derivatives. 16-state, 3-mixture word models were used during recognition. The system was trained on clean data and tested on four environmental condition sets, including subway noise, babble, vehicular noise, and exhibition hall noise. In each case, the proposed spectral reconstruction algorithms were applied in the Mel-filtered spectral domain. Note that for efficiency, spectral reconstruction was only applied to those frames wherein active speech components were discovered during the mask estimation process, implicitly applying voice activity detection (VAD). Inactive frames were set to a spectral floor equal to  $0.15x_m(n)$ , similar to noise flooring described in [37].

### A. Recognition with Oracle Masks

Table II provides word-accuracy results for the proposed missing feature methods, when combined with oracle masks. The baseline, denoted as "none", performs recognition on unprocessed speech signals. Note that when oracle information is present, the derived noise robust ASR system suffers little degradation as noise levels increase.

It can be observed in Table II that exploiting inter-channel correlation provides better noise robust recognition than exploiting intra-channel correlation alone. On average, the  $\mathbf{FB}_F$  and  $\mathbf{FB}_{2D}$  algorithms provide the best results for the given database.

The use of oracle information in missing feature-based ASR provides an upper performance bound for a particular MF technique. If mask estimation methods used in this study are improved to better differentiate between reliable and unreliable spectrographic components, the overall recognition performance can be expected to approach the upper bound provided by Table II. The optimization of binary reliability masks, however, lies beyond the scope of this paper, and we offer SPP-based masks as an illustrative example.

Besides providing an upper performance bound, oracle masks allow decoupling of the mask estimation and data imputation components of missing feature systems. As in [8], we compare the proposed spectral reconstruction technique in the oracle mask case with data imputation based on spectral subtraction [33]. Spectral subtraction imputation is defined as:

$$\hat{s}_m(n) = \begin{cases} x_m(n), & \text{if } x_m(n) \text{ is deemed reliable} \\ \left(x_m^\tau(n) - \hat{d}_m^\tau(x_m(n))\right)^{1/\tau}, & \text{else} \end{cases} \quad (49)$$

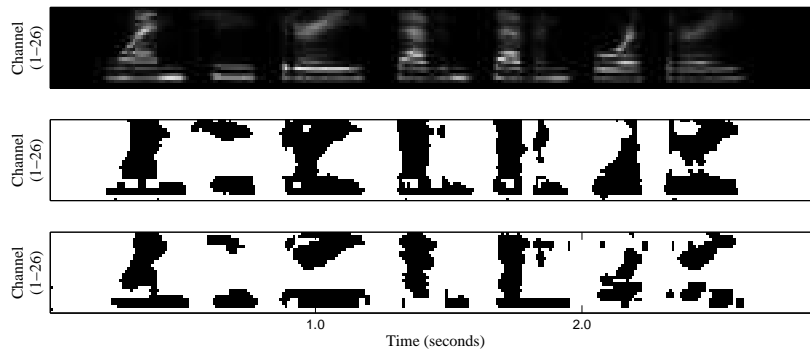


Fig. 6. Examples of Spectral Masks: Panel **a** illustrates the spectrogram of a clean signal (“one two three seven seven four three”). Panel **b** represents the oracle mask of the corresponding signal at a noise level of 0 dB SNR. Panel **c** represents the SSP-based mask for the corresponding noisy signal. The given utterances were from the Aurora-2 database [21], and were corrupted by vehicular noise.

TABLE II

Word-Accuracy Results for HMM-Based Spectral Reconstruction Using Oracle Masks. “none” refers to unprocessed speech signals.  $\mathbf{FB}_T$  refers to estimation along the temporal axis,  $\mathbf{FB}_F$  refers to estimation along the frequency axis, and  $\mathbf{FB}_{2D}$  refers to a combination of both. Results were obtained on the Aurora-2 database [21], and at a resolution of  $R=3$ . Bold entries refer to the maximum results for each condition in the Average case.

SNR (dB)	20	15	10	5	0	-5
Subway Noise						
none	95.30	90.05	70.74	40.47	8.60	0.28
$\mathbf{FB}_T$	98.50	98.13	96.50	91.40	80.38	63.16
$\mathbf{FB}_F$	98.65	98.28	97.33	93.71	84.37	68.47
$\mathbf{FB}_{2D}$	98.59	98.40	97.02	92.32	81.03	63.22
Babble						
none	96.43	91.17	74.55	38.81	6.17	0.15
$\mathbf{FB}_T$	98.28	98.04	97.07	94.86	87.42	72.13
$\mathbf{FB}_F$	98.34	98.07	97.61	96.01	89.18	75.54
$\mathbf{FB}_{2D}$	98.28	98.10	97.37	95.92	91.05	83.95
Vehicular Noise						
none	96.12	89.02	67.31	25.23	0.95	0.00
$\mathbf{FB}_T$	98.39	98.09	97.11	93.68	85.77	71.67
$\mathbf{FB}_F$	98.57	98.48	98.18	94.75	88.10	75.78
$\mathbf{FB}_{2D}$	98.69	98.21	97.55	95.05	90.64	82.70
Exhibition Hall						
none	95.34	88.31	68.90	33.91	4.97	0.46
$\mathbf{FB}_T$	98.58	97.87	96.05	90.68	77.20	57.04
$\mathbf{FB}_F$	98.55	98.09	97.04	93.12	82.38	65.35
$\mathbf{FB}_{2D}$	98.64	98.06	96.85	92.50	82.41	67.60
Average						
none	95.80	89.64	70.38	34.61	5.17	0.22
$\mathbf{FB}_T$	98.44	98.03	96.68	92.66	82.69	66.00
$\mathbf{FB}_F$	98.53	<b>98.23</b>	<b>97.54</b>	<b>94.40</b>	86.01	71.29
$\mathbf{FB}_{2D}$	<b>98.55</b>	98.19	97.20	93.95	<b>86.28</b>	<b>74.37</b>

Here,  $\tau$  is a user-defined parameter. Note that  $\tau=2$  is equivalent to power spectral subtraction (PSS), and  $\tau=1$  is equivalent to magnitude spectral subtraction (MSS). Figure 7 provides word-accuracies for the proposed  $\mathbf{FB}_F$  spectral reconstruction algorithm in the oracle mask case. Included for comparison are spectral subtraction-based imputation for  $\tau=1$  and  $\tau=2$ . It can be observed from Figure 7 that the proposed method greatly outperforms the baseline techniques. Also,

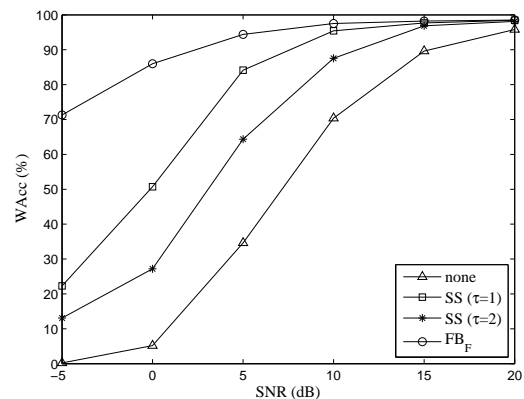


Fig. 7. Word-Accuracy Results for  $\mathbf{FB}_F$  Spectral Reconstruction using Oracle Masks: SS refers to the spectral subtraction-based imputation technique defined in Eq. 49, and “none” refers to unprocessed signals. Results were averaged across all noise types in Set A of the Aurora-2 database.

note that the proposed algorithm was initialized with the MSS estimate (Section II-A). Thus, the increase in performance relative to SS with  $\tau=1$  can be attributed directly to HMM-based reconstruction techniques.

### B. Sensitivity Analysis for $\kappa$

In Section II-A, we introduce a simple linear approximation of  $\sigma_\varepsilon$ , the standard deviation of noise estimation error, as a function of estimated noise magnitude,  $\hat{d}_m$ . The parameter  $\sigma_\varepsilon$  is important to the spectral reconstruction process since it is required during the calculation of observation statistics (Eq. 33). In Section II-A we used a small amount of noise data to set the constant  $\kappa$  to 0.6. In this section we provide sensitivity analysis of  $\kappa$ , and show that the exact value of this parameter is not important with respect to word-accuracy rate, and therefore extensive training is not necessary.

The proposed  $\mathbf{FB}_F$  missing feature-based ASR system was tested on Set A of the Aurora-2 database. The system was used in series with oracle masks. The original value of  $\kappa$  was altered and the resulting effects on word-accuracy rate was observed. Table III provides sensitivity analysis results for  $\kappa$ .  $\Delta\kappa$  refers

to the percent change in  $\kappa$ , the  $\overline{WAcc}$  refers to the word-accuracy averaged across all noise conditions, and  $\Delta\overline{WAcc}$  refers to the change in average word-accuracy relative to that obtained for  $\Delta\kappa=0$ . It can be observed in Table III that changes in  $\kappa$  do not significantly effect the resulting recognition rate of the overall system, making extensive training of  $\kappa$  unnecessary.

TABLE III

Sensitivity Analysis for  $\kappa$ :  $\Delta\kappa$  refers to the percent change in  $\kappa$ ,  $\overline{WAcc}$  refers to the word-accuracy averaged across all noise conditions, and  $\Delta\overline{WAcc}$  refers to the change in average word-accuracy relative to that obtained for  $\Delta\kappa=0$ .

$\Delta\kappa$	$\overline{WAcc}$	$\Delta\overline{WAcc}$
-30.0%	92.13	+1.24
-20.0%	91.90	+0.99
-10.0%	91.57	+0.63
$\pm 0.0\%$	91.00	$\pm 0.00$
+10.0%	90.37	-0.69
+20.0%	89.76	-1.36
+30.0%	89.07	-2.12

### C. Recognition with SPP Masks

Table IV provides word-accuracy results for the proposed missing feature methods, when combined with SPP-based masks. Again, "none" refers to performing recognition on unprocessed speech signals. The resolution of the underlying statistical model was again set to  $R=3$ . As seen in Table IV, the proposed spectral reconstruction algorithms provide greatly improved recognition results for noisy conditions, relative to the baseline system.

The proposed spectral reconstruction framework is compatible with proven noise robust post-processing techniques. To illustrate this, the  $\mathbf{FB}_{2D}$  algorithm was applied to Mel-Filtered spectrographic data and Peak Isolation was applied as a post-processing step [39]. Results were obtained on various noise conditions from the Aurora-2 database [21], and are provided in Table V, using SPP-based masks. The combination of the proposed missing feature algorithm with proven noise robust methods further improves system performance. For comparison, Table V includes comparisons with the ETSI advanced front end (AFE) [41], which is one of the most noise robust front-ends. Even without an oracle mask, the algorithm provides performance similar to the ETSI AFE.

## VI. SUMMARY AND CONCLUSIONS

In this paper, we have presented a novel HMM-based framework for estimation of unreliable spectrographic data. We utilize hidden Markov models to reconstruct corrupted spectral components based on reliable information, unreliable observations, and an underlying signal model, to improve noise robust speech recognition. We derive separate spectral reconstruction methods which exploit intra-channel (across-time) correlation, inter-channel (across-frequency) correlation, or a combination of both.

We reduce the required complexity of HMM processing by implementing downsampled statistical models. The configurations of such downsampled models are designed through

TABLE IV

Word-Accuracy Results for HMM-Based Spectral Reconstruction Using Masks Based on Speech Presence Probability (SPP). "none" refers to unprocessed speech signals.  $\mathbf{FB}_T$  refers to estimation along the temporal axis,  $\mathbf{FB}_F$  refers to estimation along the frequency axis, and  $\mathbf{FB}_{2D}$  refers to a combination of both. Results were obtained on the Aurora-2 database [21], and at a resolution of  $R=3$ . Bold entries refer to the maximum results for each condition in the Average case.

SNR (dB)	20	15	10	5	0	-5
Subway Noise						
none	95.30	90.05	70.74	40.47	8.60	0.28
$\mathbf{FB}_T$	97.51	94.81	87.01	73.26	49.40	27.66
$\mathbf{FB}_F$	97.70	94.96	87.47	73.93	50.17	26.40
$\mathbf{FB}_{2D}$	97.76	94.72	87.63	73.20	47.93	22.97
Babble						
none	96.43	91.17	74.55	38.81	6.17	0.15
$\mathbf{FB}_T$	97.55	95.28	87.73	71.34	43.92	19.74
$\mathbf{FB}_F$	97.55	95.25	88.39	72.88	44.35	19.41
$\mathbf{FB}_{2D}$	97.52	95.10	88.03	73.16	43.71	19.20
Vehicular Noise						
none	96.12	89.02	67.31	25.23	0.95	0.00
$\mathbf{FB}_T$	98.12	96.63	92.81	81.81	60.96	33.55
$\mathbf{FB}_F$	98.03	96.84	93.89	83.15	58.10	28.78
$\mathbf{FB}_{2D}$	98.12	96.75	93.50	83.69	63.02	32.87
Exhibition Hall						
none	95.34	88.31	68.90	33.91	4.97	0.46
$\mathbf{FB}_T$	97.47	94.26	86.36	69.67	44.62	23.11
$\mathbf{FB}_F$	97.47	94.45	87.38	72.29	47.74	23.79
$\mathbf{FB}_{2D}$	97.53	94.08	86.61	71.61	46.74	25.39
Average						
none	95.80	89.64	70.38	34.61	5.17	0.22
$\mathbf{FB}_T$	97.66	95.25	88.48	74.02	49.73	<b>26.02</b>
$\mathbf{FB}_F$	97.69	<b>95.28</b>	<b>89.28</b>	<b>75.56</b>	50.09	24.60
$\mathbf{FB}_{2D}$	<b>97.73</b>	95.16	88.94	75.42	<b>50.35</b>	25.11

TABLE V

Summary of Word-Accuracy Results for HMM-Based Spectral Reconstruction. Results are shown for SPP-based masks (SPP), and with Peak Isolation Post-processing (PKI) [39]. The  $\mathbf{FB}_{2D}$  algorithm was used in all cases. Results for PKI and the ETSI AFE [41] are included for comparison. Results were averaged over all noise conditions included in Set A of the Aurora-2 database [21], and were obtained using model resolution  $R=3$ .

SNR (dB)	20	15	10	5	0	-5
none	95.80	89.64	70.38	34.61	5.17	0.22
SPP	97.69	95.28	89.28	75.56	50.09	24.60
SPP + PKI	<b>97.96</b>	<b>96.17</b>	<b>91.05</b>	<b>78.88</b>	<b>53.61</b>	<b>25.25</b>
PKI	97.81	95.73	89.32	70.99	33.32	10.91
ETSI AFE	98.46	96.96	92.22	79.13	51.11	19.26

the use of tree-structured quantization, and corresponding statistical parameters are adapted accordingly. This avoids expensive retraining of codebooks for various resolutions, and reduces memory requirements. One could train a high resolution model, and test using a downsampled model without the need for retraining.

We show empirically that for our specific task, a model resolution of 3 bits serves as the optimal tradeoff between

performance and complexity. The analysis provided hints at the robustness of the recognizer to small changes in feature values. That is, the recognizer seems not to react to minor variability in features obtained from reconstructed spectrographic data at various resolutions. Instead, it appears that the general spectral shape of estimated data determines recognizer output.

We combine the proposed spectral reconstruction algorithms with oracle masks to provide an upper performance bound for our missing feature analysis. Additionally, we combine the proposed methods with masks based on speech presence probability (SPP), to illustrate more realistic results. Both scenarios provide impressive ASR performance relative to the MFCC baseline system.

The novelty of our work lies in the application of HMM-based MMSE reconstruction to spectral reconstruction, which includes intra-channel and inter-channel approaches. Furthermore, we present novel model downsampling methods which allow for efficient processing without the need for retraining of codebooks.

## VII. ACKNOWLEDGMENTS

This work was supported in part by the NSF.

## REFERENCES

- [1] A. Acero, *Acoustic and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1993.
- [2] H. Hermansky and N. Morgan, *RASTA Processing of Speech*, IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4, pp. 578-589, 1994.
- [3] O. Viikki and K. Laurila, *Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition*, Speech Communication, Vol. 25, pp. 133-147, 1998.
- [4] Q. Zhu and A. Alwan, *Non-linear Feature Extraction for Robust Recognition in Stationary and Non-stationary Noise*, Computer, Speech, and Language, 17(4), pp. 381-402, 2003.
- [5] B. Raj and R. Stern, *Missing Feature Approaches in Speech Recognition*, IEEE Signal Processing Magazine, Vol. 22, Issue 5, pp. 101-116, 2005.
- [6] M. P. Cooke, A. Morris, P. D. Green, *Missing Data Techniques for Robust Speech Recognition*, Proc. ICASSP, vol. 2, pp. 863-866, 1997.
- [7] J. Barker, L. Josifovski, M. Cooke, and P. Green, *Soft Decisions in Missing Feature Data Techniques for Robust Automatic Speech Recognition*, Proc. ICSLP, pp. 373-376, 2000.
- [8] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, *Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data*, Speech Communication, Vol. 34, pp. 267-285, 2001.
- [9] V. Ion and R. Haeb-Umbach, *Uncertainty Decoding for Distributed Speech Recognition Over Error-Prone Networks*, Speech Communication, Vol. 48, pp. 1435-1446, 2006.
- [10] B. R. Ramakrishnan, *Reconstruction of Incomplete Spectrograms for Robust Speech Recognition*, Ph.D. Thesis, CMU, 2000.
- [11] B. Raj, M. L. Seltzer, and R. M. Stern, *Reconstruction of Missing Features for Robust Speech Recognition*, Speech Communication, vol. 43, pp. 275-296, 2004.
- [12] L. R. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE, vol. 77, No. 2, pp. 257-286, 1989.
- [13] A. Viterbi, *Convolutional Codes and their Performance in Communication Systems*, IEEE Transactions on Communications, vol. 19, Issue 5, pt. 1, pp. 751-772, Oct. 1971.
- [14] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
- [15] A. M. Peinado, V. Sanchez, J. C. Segura, and J. L. Perez-Cordoba, *MMSE-Based Channel Mitigation for Distributed Speech Recognition*, Eurospeech, 2001.
- [16] A. M. Peinado, V. Sanchez, J. L. Perez-Cordoba, J. C. Segura, and J. Rubio, *HMM-Based Methods for Channel Error Mitigation in Distributed Speech Recognition*, ICSLP, 2002.
- [17] A. M. Peinado, V. Sanchez, J. L. Perez-Cordoba, and A. de la Torre, *HMM-Based Channel Error Mitigation and its Application to Distributed Speech Recognition*, Speech Communication, vol. 41/4, pp. 549-561, Nov. 2003.
- [18] C. A. Rodbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen, *Hidden Markov Model Based Loss Concealment for Voice Over IP*, Transactions for Audio, Speech, and Language Processing, Vol. 14, No. 5, pp. 1609-1623, 2006.
- [19] B. J. Borgstrom and A. Alwan, *An Efficient Approximation of the Forward-Backward Algorithm to Deal With Packet Loss, With Applications to Remote Speech Recognition*, Proc. of ICASSP, pp. 4425-4428, 2008.
- [20] B. J. Borgstrom and A. Alwan, *HMM-Based Estimation of Unreliable Spectral Components for Noise Robust Speech Recognition*, Proc. Of Interspeech, pp. 1769-1772, 2008.
- [21] D. Pearce, *Enabling New Speech Driven Services For Mobile Devices: An Overview of the ETSI Standards Activities for Distributed Speech Recognition Front-Ends*, AVIOS 2000: Speech Appl. Conf., Vol. 5, pp. 1-6, May 2000.
- [22] R. Martin, *Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics*, IEEE Trans. Speech and Audio Processing, Vol. 9, No. 5, pp. 504-512, 2001.
- [23] I. Cohen, *Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging*, IEEE Trans. Speech and Audio Processing, Vol. 11, No. 5, pp. 466-475, 2003.
- [24] M. L. Seltzer, B. Raj, and R. M. Stern, *A Bayesian Classifier for Spectrographic Mask Estimation for Missing Feature Speech Recognition*, Speech Communication, Vol. 43, pp. 379-393, 2004.
- [25] V. Ion and R. Haeb-Umbach, *Multi-Resolution Soft-Features for Channel-Robust Distributed Speech Recognition*, Proc. of Interspeech, pp. 594-597, 2007.
- [26] W. Kim and R. Stern, *Band-Independent Mask Estimation for Missing Feature Reconstruction in the Presence of Unknown Background Noise*, Proc. ICASSP, pp. 305-308, 2006.
- [27] T. Gerkmann, C. Breithaupt, and R. Martin, *Improved A Posteriori Speech Presence Probability Estimation Based on a Likelihood Ratio with Fixed Priors*, IEEE Trans. on Audio, Speech, and Language Processing, Vol. 16, No. 5, pp. 910-919, 2008.
- [28] D. Rudoy, D. Spendley, and P. Wolfe, *Conditionally Linear Gaussian Models for Estimating Vocal Tract Resonances*, Proc. Interspeech, pp. 526-529, 2007.
- [29] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, *Missing Data Theory, Spectral Subtraction, and Signal-to-Noise Estimation for Robust ASR: an Integrated Study*, Eurospeech, 1999.
- [30] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, New York: Wiley, 2006.
- [31] R. Martin and T. Lotter, *Optimal Recursive Smoothing of Non-Stationary Periodograms*, Proc. Int. Workshop Acoust. Echo and Noise Control (IWAENC), pp. 167-170, 2001.
- [32] J. A. Hartigan, *Clustering Algorithms*, Wiley, 1975.
- [33] S. F. Boll, *Suppression of Acoustic Noise in Speech Using Spectral Subtraction*, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 27, pp. 113-120, 1979.
- [34] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, 1988.
- [35] L. Deng, A. Acero, M. Plumpe, and X. Huang, *Large-Vocabulary Speech Recognition Under Adverse Acoustic Environments*, Proc. of ICSLP, 2000.
- [36] X. Cui and A. Alwan, *Adaptation of Children's Speech with Limited Data Based on Formant-like Peak Alignment*, Computer Speech and Language, Vol. 20, Issue 4, pp. 400-419, 2006.
- [37] Q. Zhu and A. Alwan, *Non-linear feature extraction for robust recognition in stationary and non-stationary noise*, Computer, Speech, and Language, 17(4), pp. 381-402, 2003.
- [38] Y. Ephraim and D. Malah, *Speech Enhancement Using a Minimum Mean-Square Log-Spectral Amplitude Estimator*, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 33, No. 2, pp. 443-445, 1985.
- [39] B. Stroppe and A. Alwan, *A Model of Dynamic Auditory Perception and its Application to Robust Word Recognition*, IEEE Trans. on Speech and Audio Processing, Vol. 5, No. 5, pp. 451-464, 1997.
- [40] H. Hermansky, *Perceptual Linear Predictive (PLP) Analysis of Speech*, JASA, Vol. 87, No. 4, pp. 1738-1752, 1990.
- [41] ETSI Standard Doc., *Speech Processing, Transmission, and Quality Aspects (STQ); Distributed Speech Recognition; Front-end Feature Extraction Algorithms; Compression Algorithms*, ETSI ES 202 050 v1.1.1 (2007-10).