HMM-Based Estimation of Unreliable Spectral Components for Noise Robust Speech Recognition

Bengt J. Borgström and Abeer Alwan

Department of Electrical Engineering,
University of California, Los Angeles
jonas@ee.ucla.edu, alwan@ee.ucla.edu

Abstract

This paper presents a novel approach for reconstructing unreliable spectral components, which utilizes HMM-based missing feature algorithms, and applies them to noise robust speech recognition. The proposed technique uses the forward-backward algorithm to estimate corrupt spectrographic data based on nearby reliable features, noisy observations, and on an underlying statistical model. The estimation process can be applied based on intra-channel information, intra-feature information, or a combination of both. The overall system is shown to provide vast improvements for the Consonant Challenge Database [1], for both MFCCs and PLP features, when using an oracle mask. Moreover, through downsampling of statistical models [2], the required complexity of the system is greatly reduced with negligible effects on results.

Index Terms: Automatic Speech Recognition, Noise Robustness, Missing Features, Hidden Markov Models.

1. Introduction

Many applications within communications or signal processing deal with the task of handling unreliable data. During the transmission of digital data over an error-prone channel, packets may become corrupt due to channel effects. When acoustic or image signals are recorded, data may be corrupt due to environmental noise. In each case, system performance can be improved by reconstructing unreliable data prior to further processing.

In [3]-[4], estimation techniques are applied to reconstruction of unreliable spectral coefficients of speech sampled in noisy acoustic conditions for the purpose of noise robust speech recognition. In [5], HMM-based methods are applied to estimate corrupt packets for Distributed Speech Recognition (DSR). Similarly, in [6], HMM-based methods are utilized to reconstruct parameters for packet-based speech coding.

In this paper we present a novel approach for reconstructing unreliable spectral components using HMM-based decoding. We develop estimation techniques utilizing intra-channel data, intra-feature data, or a combination of both. Additionally, using downsampling of statistical models previously proposed in [2], we are able to reduce the required complexity by a factor greater than 800. We apply the proposed techniques to the Consonant Challenge Database [1] to illustrate their effectiveness. Furthermore, we show the robustness of the presented spectral estimation framework to pre- and post-processing by applying it in series with proven noise robust algorithms.

Section 2 discusses the role of HMMs in the estimation process and describes efficient downsampling of statistical models. In Section 3, the proposed estimation techniques are applied to noise robust speech recognition. Experimental results are shown in Section 4. In Section 5, concluding remarks are given.

2. Estimation of Missing Features

2.1. The Role of HMMs in the Estimation Process

In speech communication or recognition systems, features are extracted from input time waveforms for further analysis. Let $\mathbf{f}(n) = [x_1(n), x_2(n), \ldots, x_M(n)]^T$ represent the feature vector processed at time n. Within many applications, such features may become corrupt and thus unreliable, and estimation can be used to reconstruct the comprising components. This study explores the use of HMMs to model signals during the estimation process.

Due to the discrete nature of HMM states, features must be quantized, at least implicitly, prior to the estimation process. Let the quantizer of x_m be represented by Q_m, and let the set of corresponding centroids be referred to as $\{c_m^1, c_m^2, \ldots, c_m^N\}$, where N is the number of centroids in Q_m. In order to apply estimation methods, separate HMMs are constructed for each of the quantizers Q_m, for $1 \leq m \leq M$, to model the feature trajectories.

Let the HMM applied to the output signal from Q_m be referred to as Λ_m=($\mathbf{A}_m, \mathbf{B}_m, \vec{\pi}_m$), where $\mathbf{A}_m$ provides transitional statistics, $\mathbf{B}_m$ provides observation statistics, and $\vec{\pi}_m$ provides steady-state statistics [7]. The steady-state probabilities of Λ_m can be determined empirically from training data:

$$\vec{\pi}_m(i) = \frac{\text{no. of samples quantized to centroid } c_m^i}{\text{total no. of samples}}. \quad (1)$$

The transitional probabilities of Λ_m can similarly be determined from training data as:

$$\mathbf{A}_m(i,j) = \frac{\text{no. of samples transitioning from } c_m^i \text{ to } c_m^j}{\text{no. of samples quantized to } c_m^i}. \quad (2)$$

Formulation of the components of $\mathbf{B}_m$ is dependent on the specific application of the system. In general, observation statistics are defined as:

$$b_{j,m}(o_m(n)) = P\left(Q_m\{x_m(n) + \eta_m(n)\} = c^j | o_m(n)\right), \quad (3)$$

where $\eta_m(n)$ is the hidden noise process. The derivation of observation statistics for estimation of unreliable spectral components will be further discussed in Section 3.

Given an HMM-based framework, various decoding techniques exist. One such technique is the forward-backward (FB)

algorithm, which determines the optimal feature vector estimate $\mathbf{f}(n)$ given the first and last reliable features at temporal indices $n-k_1$ and $n+k_2$, and given the series of observations $o_m(n-k_1+1),\ldots,o_m(n+k_2)$. The estimate of component $x_m(n)$, using the FB algorithm, is determined as [7]:

$$\widehat{x}_m(n)=\sum_{i=1}^{N} c_m^i \gamma_m^i(n), \tag{4}$$

where:

$$\gamma_m^i(n)=\frac{\alpha_m^i(n)\,\beta_m^i(n)}{\sum_{j=1}^{N}\alpha_m^j(n)\,\beta_m^j(n)}. \tag{5}$$

The set of values γ_m^i represents the distribution of $x_m(n)$, conditioned on reliable features, as well as on noisy observations:

$$\gamma_m^i(n)=P(x_m(n)=c^i|x_m(n-k_1),x_m(n+k_2), o_m(n-k_1+1),\ldots,o_m(n+k_2)). \tag{6}$$

The values $\alpha_m^i(n)$ and $\beta_m^i(n)$, known as the forward and backward variables, respectively, can be determined recursively:

$$\alpha_m^i(n)=\left[\sum_{j=1}^{N}\mathbf{A}_m(j,i)\,\alpha_m^j(n-1)\right]b_{j,m}(o_m(n)), \tag{7}$$

$$\beta_m^i(n)=\sum_{j=1}^{N}\mathbf{A}_m(i,j)\,\beta_m^j(n+1)\,b_{j,m}(o_m(n+1)). \tag{8}$$

The HMM-based estimation techniques discussed thus far have previously been applied to channel mitigation for remote speech recognition [5] and speech communication [6]. The computational load induced by such approaches (see Section 3.3), may prove them to be too complex for resource-constrained or delay-sensitive applications.

2.2. Downsampling of Statistical Models

In [2], we propose a framework for efficient HMM-based missing feature estimation based on downsampling of underlying statistical models. In this paper, we use quantizers with less resolution to configure 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.

Let Q_m^R represent a R-bit quantizer for component m, with centroids $\left\{c_m^{R,1},c_m^{R,2},\ldots,c_m^{R,2^R}\right\}$. Using tree-structure quantization, centroids can be mapped according to:

$$c_m^{8,i}\Rightarrow c_m^{R,j},\ \text{for } 1\le R<8,\ \text{where } j=\left\lfloor\frac{i}{2^{8-R}}\right\rfloor. \tag{9}$$

In this manner, centroids can easily be regrouped as clusters without requiring expensive retraining of the quantization codebook.

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

$$\vec{\pi}_m^R(i)=\sum_{k=0}^{2^\tau-1}\vec{\pi}_m^{R+\tau}(2^\tau i-k), \tag{10}$$

and:

$$\mathbf{A}_m^R(i,j)=\frac{1}{2^\tau}\left[\sum_{k=0}^{2^\tau-1}\sum_{l=0}^{2^\tau-1}\mathbf{A}_m^{R+\tau}(2^\tau i-k,2^\tau j-l)\right], \tag{11}$$

where τ is chosen as τ=8-R.

The observation statistics of Λ_m^R, denoted as $b_{i,m}^R(o_m)$, are application-specific, and thus an explicit general formula can not be derived.

3. 1D and 2D HMM-Based Spectral Reconstruction

In this paper, we develop a novel framework for HMM-based estimation for reconstruction of unreliable spectral components of speech degraded by noise, in order to provide improvements for noise robust speech recognition. An interesting aspect of reconstructing unreliable spectral components is the possibility of utilizing statistics across two dimensions during the estimation process. In many similar applications, such as speech communication and remote speech recognition, unreliable data results from corrupt or dropped packets [5], [6], and thus correlation along the frequency axis is not available. In the current study, we apply HMM-based estimation utilizing intra-channel and/or intra-feature data to reconstruct corrupt spectral coefficients.

It is important to note that reconstruction is carried out in the spectral or Mel-filtered spectral domain, since this allows for further transformation into the cepstral domain without propagation of unreliable information. Additionally, it has previously been shown that recognition of cepstral features outperforms that of spectral features [8].

Traditionally, missing feature analysis for robust speech recognition has required two separate tasks, namely mask estimation and feature estimation. Mask estimation determines the reliability of each spectral component of an utterance based on the characteristics of current speech and noise signals, and supplies the recognizer with either hard decisions or soft reliability metrics [3],[4]. Feature estimation then reconstructs those components deemed unreliable, based on neighboring reliable features, and based on an underlying signal model. This paper will focus on the latter component, and will thus assume the availability of an "oracle mask" [4].

3.1. Utilizing Correlation Across Feature Vectors

Following the notation from Sections 2.1 and 2.2 , let $\mathbf{f}(n)=[x_1(n),x_2(n),\ldots,x_M]^T$ represent the feature extracted from an utterance at time n, so that $x_i(n)$ represents the component corresponding to the i^{th} channel. A spectrographic representation can then expressed as:

$$\mathfrak{S}(n,i)=[\mathbf{f}(1),\mathbf{f}(2),\cdots,\mathbf{f}(T)] =\begin{bmatrix} x_1(1) & x_1(2) & \cdots & x_1(T)\\ \vdots & \vdots & \ddots & \vdots\\ x_M(1) & x_M(2) & \cdots & x_M(T)\end{bmatrix}. \tag{12}$$

The effect of noise on spectrographic data of speech varies vastly with respect to the spectral characteristics of the noise. A simple yet often unrealistic method to modeling additive noise

is to assume a flat spectral distribution [7]. In deriving observation statistics for the current application, we use a similar simplified approach since it leads to Gaussian random variables. However, we apply it on a component-by-component basis, thus not requiring the strict constraint of spectral flatness across all channels. Assume the additive noise corrupting channel m can be locally or globally modeled by a Gaussian random process with mean μ_m and variance σ_m^2. If the observed data feature is $o_m(n)$, the corresponding observation probability distribution is given by:

$$b_{i,m}^R(o_m(n)) = P\left(c_m^{R,i}|o_m(n)\right) = \int_{z_1}^{z_2} \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{-\frac{(\eta_m+\mu_m-o_m(n))^2}{2\sigma_m^2}} d\eta_m, \quad (13)$$

where z_1 and z_2 represent the upper and lower boundaries of centroid $c^{R,i}$. Also, η_m is the hidden noise process. Note that previous techniques in [4] do not exploit information within noisy observation data, as does the proposed framework.

Once observation statistics have been determined, the approximated HMM-based techniques derived in Section 2.2 can be applied along the time axis to reconstruct unreliable components of $\mathfrak{S}$ according to Equations 4-7 [5], [6]. In this application, the probabilities $\gamma_m^i(n)$ refer to the distribution of spectral component $x_m(n)$ conditioned on past and future reliable components within channel i, as well as on noisy spectrographic data within channel i. This technique is referred to as $\mathbf{FB_T}$.

3.2. Utilizing Correlation Across Frequency Channels

We wish to exploit the strong inter-channel correlation present within speech during spectral reconstruction. It is interesting to note that [9] also noticed cross-correlation among frequencies, which was exploited for formant tracking. When performing estimation along the frequency axis, the stationarity assumed by the steady-state and transitional statistics of $\vec{\pi}_m^R$ and $\mathbf{A}_m^R$, respectively, do not hold. It is therefore necessary to introduce the probabilities:

$$P_{m:n}^{i:j} = P\left(Q_m\{x_m+\eta_m\} = c_m^i | Q_n\{x_n+\eta_n\} = c_n^j\right), \quad (14)$$

which define the intra-feature transitional statistics of spectrographic data across channels. Note that in Equation 14, $|m-n|$=1 must hold in order for $P_{m:n}^{i:j}$ to be valid. Modified forward and backward variables can be expressed for the estimation of missing data along the frequency axis:

$$\delta_m^{R,i}(n) = \left[\sum_{j=1}^N P_{m:m-1}^{i:j}\delta_m^{R,j}(n-1)\right] b_{i,m}^R(o_m(n)), \quad (15)$$

$$\epsilon_m^{R,i}(n) = \sum_{j=1}^N P_{m:m+1}^{j:i}\epsilon_m^{R,j}(n+1)\, b_{j,m}^R(o_m(n+1)). \quad (16)$$

In this application, referred to as $\mathbf{FB_F}$, $\gamma_m^{R,i}(n)$ represents the distribution of $x_m(n)$ conditioned on reliable components and noisy observations from feature vector $\mathbf{f}(n)$, which is analogous to Equation 6. The corresponding expression is given by:

$$\gamma_m^{R,i}(n) = \frac{\delta_m^{R,i}(n)\,\epsilon_m^{R,i}(n)}{\sum_{j=1}^N \delta_m^{R,j}(n)\,\epsilon_m^{R,j}(n)}. \quad (17)$$

Table 1: *Operations Required by Proposed Spectrogram Reconstruction Techniques, as a Function of Model Resolution:* $\mathbf{FB_T}$ *refers to intra-channel HMM-based estimation,* $\mathbf{FB_F}$ *refers to intra-feature estimation, and* $\mathbf{FB_{2D}}$ *refers to the combination of both.*

Operation		multiplications	additions
R=8	$\mathbf{FB_T}$	3.387×10^7	3.353×10^7
	$\mathbf{FB_F}$	4.432×10^7	4.392×10^7
	$\mathbf{FB_{2D}}$	7.818×10^7	7.745×10^7
R=3	$\mathbf{FB_T}$	40,224	29,960
	$\mathbf{FB_F}$	51,888	38,920
	$\mathbf{FB_{2D}}$	92,112	68,880

Additionally, we 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, $\gamma_m^{R,i}(n)$ is the distribution of $x_m(n)$ conditioned on data contained in both channel i and feature vector $\mathbf{f}(n)$:

$$\gamma_m^{R,i}(n) = \frac{\alpha_m^{R,i}(n)\,\beta_m^{R,i}(n)\,\delta_m^{R,i}(n)\,\epsilon_m^{R,i}(n)}{\sum_{j=1}^N \alpha_m^{R,j}(n)\,\beta_m^{R,j}(n)\,\delta_m^{R,j}(n)\,\epsilon_m^{R,j}(n)}. \quad (18)$$

For the conditional distributions given by Equations 17 and 18, the corresponding estimates $\hat{x}_m(n)$ are determined via Equation 4.

3.3. Complexity Analysis

The methods developed of Section 2.2 can greatly reduce the required complexity of the proposed estimation system. Note that the order of complexity for the forward-backward decoding algorithm is $O\left(2^N\right)$, 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 number of required operations are provided in Table 1. The sample utterance was 1.2 seconds in duration, and was degraded by 8-talker babble noise at -2 dB.

As can be concluded from Table 1, the computational load induced by the proposed techniques is vastly reduced as the resolution R is decreased. When the resolution is set to R=3, the order of complexity is reduced by a factor of $>$800, relative to the standard of R=8.

4. Experimental Results

The algorithms developed in Section 3 were applied to the Consonant Challenge Database [1] to illustrate their effectiveness. The database consists of vowel-consonant-vowel (VCV) utterances degraded by 6 noise types. Word-accuracy results obtained using HMM-based spectral reconstruction are provided in Table 2. MFCC refers to the baseline system with no spectral reconstruction. If the resolution is not specified in parentheses, then R=3. As can be concluded from rows 2 and 3, downsampling results in negligible effects on system performance.

The proposed spectral estimation framework is robust to the feature type. To illustrate this, we apply it in series with known noise robust feature extraction algorithms. In Table 2, PLPCC refers to perceptual linear predictive cepstral coefficients [11], which is a feature extraction technique applied to

Table 2: *Word-Accuracy Results for HMM-Based Estimation Techniques, Obtained on the Consonant Challenge Database [1]. Bold-faced entries refer to the best results obtained for each test and for each noise condition. In each case, training was performed on clean data, representing a mismatched condition.*

Estimation Algorithm	clean (baseline)	competing talker (-6 dB)	8-talker babble (-2 dB)	speech-shaped (-6 dB)	factory noise (0 dB)	mod. spch-shaped (-6 dB)	3-talker babble (-3 dB)	Average (2-7)	Relative Increase
MFCC	87.24	8.59	6.25	7.03	4.95	11.98	7.29	7.68	N/A
$\mathbf{FB_T}$ (R=3)	87.24	**22.14**	17.19	16.41	12.76	21.35	14.58	17.41	126.63
$\mathbf{FB_T}$ (R=8)	87.24	20.05	16.41	17.97	13.54	18.49	14.06	16.75	118.14
$\mathbf{FB_F}$	87.24	20.83	**20.83**	**21.88**	14.89	19.27	**19.01**	**19.44**	**153.13**
$\mathbf{FB_{2D}}$	87.24	20.83	19.27	19.53	**16.41**	**21.88**	16.67	19.09	148.57
PLPCC	83.85	9.11	10.68	**9.38**	5.21	16.67	10.68	10.29	33.98
$\mathbf{FB_T}$	83.85	**27.34**	**40.10**	7.03	17.19	**31.25**	**37.76**	**26.78**	**248.68**
$\mathbf{FB_F}$	83.85	22.40	34.90	7.55	**31.77**	26.30	24.22	24.52	219.21
$\mathbf{FB_{2D}}$	83.85	24.22	38.02	5.47	16.67	27.08	29.69	23.53	206.32
PLPCC + PK-ISO	80.99	11.46	9.11	7.81	7.81	14.84	8.07	9.85	28.26
$\mathbf{FB_T}$	80.99	**33.33**	39.32	**9.11**	17.19	**33.59**	**36.72**	28.21	267.32
$\mathbf{FB_F}$	80.99	24.48	**44.53**	8.59	**47.92**	29.69	31.77	**31.16**	**305.73**
$\mathbf{FB_{2D}}$	80.99	29.17	40.89	6.51	21.09	32.29	33.33	27.21	254.30

the input speech signal prior to reconstruction. The proposed framework is robust to post-processing techniques as well. PK-ISO refers to peak isolation [10], which is performed in the cepstral domain after spectral estimation.

Each proposed spectral reconstruction technique provides significant improvements in word-accuracy, for the variety of baseline systems tested. Reconstruction of MFCC coefficients produces better results when estimation exploits intra-feature data relative to intra-channel data. However, reconstruction of PLPCC coefficients produces better results when estimation is applied along the time axis. This could be due to the well known temporally smooth characteristics of LPC-type features, which are often utilized for speech coding. Analysis of transitional probabilities $\mathbf{A}_m^8(i,j)$ resulted in an average conditional transition entropy rate of 4.32 bits for PLPCC features, and an average conditional transition entropy rate of 5.77 bits for MFCC features. This result supports the above claim.

5. Conclusions

In this paper we present a framework for HMM-based estimation of unreliable spectral components for noise robust speech recognition. The proposed techniques are able to utilize intra-feature and/or intra-channel correlation. Additionally, when model downsampling is applied, the computational load is greatly reduced. The proposed methods are shown to provide significant improvements for MFCCs and PLPCCs, when applied to the Consonant Challenge Database [1], with and without post-processing. Future work will consider other databases to find out the generalization of results.

6. References

[1] $http://www.odettes.dds.nl/challenge_IS08/index.html$

[2] B. J. Borgström 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, to appear, 2008.

[3] 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.

[4] 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.

[5] A. Peinado, V. Sanchez, J. Perez-Cordoba, and A. de la Torre, *HMM-Based Channel Error Mitigation and its Application to Distributed Speech Recognition*, Speech Communication, vol. 41, pp. 549-561, Nov. 2003.

[6] C. Rødbro, M. Murthi, S. Andersen, and S. 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.

[7] 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.

[8] A. Acero, *Acoustic and Environemntal Robustness in Automatic Speech Recognition*, Kluwer, 1993.

[9] D. Rudoy, D. Spendley, and P. Wolfe, *Conditionally Linear Guassian Models for Estimating Vocal Tract Resonances*, Proc. Interspeech, pp. 526-529, 2007.

[10] Q. Zhu amd 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.

[11] H. Hermansky and N. Morgan, *RASTA Processing of Speech*, IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4, pp. 578-589, 1994.