A Statistical Approach to Mel-Domain Mask Estimation for Missing-Feature ASR

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

***Abstract*— In this letter, we present a statistical approach to Mel-domain mask estimation for missing feature (MF)-based automatic speech recognition (ASR). Mel-domain time-frequency masks are of interest, since MF systems have been shown successful in that domain. Time- and channel-specific reliability measures are derived as posterior probabilities of active speech using a 2-state speech model. Since closed form distributions for Mel-domain spectra do not exist, they are instead modeled as χ^2 processes with empirically-determined degrees of freedom. Additionally, we present HMM-based decoding to exploit temporal correlation of spectral speech data. The proposed mask estimation algorithm is integrated with an example MF-based ASR front-end from [14], and is shown to outperform the spectral subtraction (SS)-based method from [10] in terms of word-accuracy, when applied to the Aurora-2 database.**

Index Terms: mask estimation, noise robust ASR, missing features, speech presence uncertainty, χ^2 random variables.

I. Introduction

In recent years, missing feature (MF) methods have received attention as an approach to noise robust automatic speech recognition (ASR) [1]. MF techniques can be generally grouped into two main categories: front-end spectral reconstruction [14], [15], and back-end marginalization [2]. Each category requires a mask which provides hard or soft measures of reliability for time- and frequency- specific components. Furthermore, ASR front ends generally include perceptually motivated frequency warping, such as Mel-filtering, to emphasize discriminative information, and MF spectral reconstruction techniques have been shown to be successful in those domains [1],[14],[15].

In this letter, we present a statistical approach to Mel-domain mask estimation for missing feature (MF)-based automatic speech recognition (ASR). As opposed to previous solutions to soft-decision mask estimation, such as [11], which uses a tunable sigmoid function to map SNR-related measures to the range $[0, 1]$, the proposed method provides a statistical approach offering intuitive probabilistic mask values. Additionally, the proposed mask estimation method relies on a simple training process which requires only clean speech. In contrast, similar work, such as [3] and [12], includes extensive training to determine the empirical distribution of classifiers in various noise types and levels. In addition, the mask estimation technique in [9] requires training of phonetic class-dependent vector quantizer codebooks.

This work was supported in part by the NSF.

II. Statistical Framework

A. The Linear Frequency Domain

In this study, an additive noise model is assumed which is expressed after short-time spectral analysis as:

$$Y(n,k) = X(n,k) + D(n,k), \tag{1}$$

where $Y(n,k)$ is the observed signal, $X(n,k)$ is the underlying clean speech, $D(n,k)$ is the corruptive noise, and n and k denote time and frequency channel indices, respectively. Real and imaginary components of $X(n,k)$ and $D(n,k)$ are assumed to be observations of independent zero-mean Gaussian processes. We model high-level speech activity with a two-state Markov model, wherein state H_1 denotes time-frequency components corresponding to active speech, and state H_0 corresponds to time-frequency components comprised solely of noise. Following the Gaussian framework, power spectral coefficients of the observed signal are exponentially conditionally distributed for each state:

$$p\left(|Y(n,k)|^2 \middle| H_0\right) = \frac{1}{\sigma_d^2(k)} \exp\left(-\frac{|Y(n,k)|^2}{\sigma_d^2(k)}\right) \tag{2}$$

$$p\left(|Y(n,k)|^2 \middle| H_1\right) = \frac{1}{\sigma_x^2(k) + \sigma_d^2(k)} \exp\left(-\frac{|Y(n,k)|^2}{\sigma_x^2(k) + \sigma_d^2(k)}\right).$$

Note that the conditional distributions of Eq. 2 are special cases of the χ^2 distribution, which is given as:

$$p(x) = \frac{\lambda^{k/2}}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp\left(-\frac{\lambda x}{2}\right), \text{ for } x \geq 0, \tag{3}$$

where k denotes the degree of freedom, Γ denotes the Gamma function, and λ serves as a size parameter. Specifically, the size parameter is related to the second noncentral moment by λ=$1/\sigma^2$. Note that the distributions of Eq. 2 correspond to k=2.

B. The Mel-Filtered Domain

Automatic speech recognition (ASR) front ends generally include perceptually motivated frequency warping to emphasize discriminative information. For example, the Mel-filterbank is designed to approximate the human auditory system. As discussed in [6], the Mel-filterbank can be approximated in the spectral domain by a set of triangular filters $w_m(k)$, where m denotes Mel-channel index, resulting in Mel-domain power-spectra expressed as:

$$|\hat{Y}(n,m)|^2 = \sum_{k=c_{m-1}}^{c_{m+1}} w_m(k)\,|Y(n,k)|^2, \tag{4}$$

where c_m denotes the center frequency of the m^{th}-channel Mel-filter. Also, let N_m denote the number of Mel-channels used.

It can be observed in Eq. 4 that Mel-filtering involves the weighted sum of exponentially-distributed processes. The distribution of the sum of weighted exponential random variables with distinct, possibly unequal, variances is often referred to as the generalized χ^2 distribution. Although a closed form expression for such distributions does not exist, several studies have proposed approximations which utilize Pearson curves, or moment matching to simpler χ^2 distributions [7].

We propose to model the channel observation $|\hat{Y}(n,m)|^2$ by a simpler χ^2 distribution with k_m degrees of freedom, where k_m is determined using the ratio of noncentral moments. Noncentral moments of the χ^2 random variable from Eq. 3 are known to be:

$$E[x^m] = 2^m \frac{\Gamma(m + k/2)}{\Gamma(k/2)} \tag{5}$$

Using Eq. 5, channel-specific degrees of freedom can be determined as:

$$\begin{aligned} k_m &= 2\left(\frac{E\left[|\hat{Y}(n,m)|^4\right]}{E\left[|\hat{Y}(n,m)|^2\right]^2} - 1\right)^{-1} \\ &\approx 2\left(\frac{\sum_{n=1}^{T}|\hat{Y}(n,m)|^4}{\left(\sum_{n=1}^{T}|\hat{Y}(n,m)|^2\right)^2} - 1\right)^{-1} \end{aligned} \tag{6}$$

where T denotes the number of frames used during training.

Furthermore, the χ^2 distribution size parameter is set equal to appropriate "weighted average" noise and signal variances, defined as in [8]:

$$\begin{aligned} \hat{\sigma}_d^2(m) &\triangleq \frac{\sum_{i=c_{m-1}}^{c_{m+1}} w_m(i)\,\sigma_d^2(i)}{\sum_{i=c_{m-1}}^{c_{m+1}} w_m(i)}, \\ \hat{\sigma}_x^2(m) &\triangleq \frac{\sum_{i=c_{m-1}}^{c_{m+1}} w_m(i)\,\sigma_x^2(i)}{\sum_{i=c_{m-1}}^{c_{m+1}} w_m(i)}. \end{aligned} \tag{7}$$

Here, the variance of the linear-frequency noise processes are approximated as $\sigma_d^2(k) \approx \hat{N}_k^2$ where $\hat{N}_k$ is the local noise estimate of the k^{th} channel. The conditional distributions of the m^{th}-channel Mel-domain power spectral coefficient is then approximated as:

$$\begin{aligned} p\left(|\hat{Y}(n,m)|^2 \middle| H_0\right) &= \left(2^{k_m/2}\Gamma(k_m/2)\right)^{-1} \\ &\times \left(\frac{|\hat{Y}(n,m)|^2}{\hat{\sigma}_d^2(m)}\right)^{(k_m/2-1)} \exp\left(-\frac{|\hat{Y}(n,m)|^2}{2\hat{\sigma}_d^2(m)}\right), \end{aligned} \tag{8}$$

$$\begin{aligned} p\left(|\hat{Y}(n,m)|^2 \middle| H_1\right) &= \left(2^{k_m/2}\Gamma(k_m/2)\right)^{-1} \\ &\left(\frac{|\hat{Y}(n,m)|^2}{\hat{\sigma}_d^2(m) + \hat{\sigma}_x^2(m)}\right)^{(k_m/2-1)} \exp\left(-\frac{|\hat{Y}(n,m)|^2}{2\left(\hat{\sigma}_d^2(m) + \hat{\sigma}_x^2(m)\right)}\right), \end{aligned} \tag{9}$$

III. Mask Estimation

A. Soft-Decision Masks

In this section we utilize the statistical framework developed in Section II to infer the probability of active speech in individual time-frequency bins. A Bayesian approach allows the posterior probability of active speech to be expressed as:

$$P\left(H_1 \middle| |\hat{Y}(n,m)|^2\right) = \frac{\Lambda_m(n)}{1 + \Lambda_m(n)}, \tag{10}$$

where the Generalized Likelihood Ratio (GLR) is defined as:

$$\Lambda_m(n) = \frac{p_1 p\left(|\hat{Y}(n,m)|^2 \middle| H_1\right)}{(1-p_1)\,p\left(|\hat{Y}(n,m)|^2 \middle| H_0\right)}, \tag{11}$$

and p_1 denotes the steady-state probability of state H_1. Substitution of Eq. 8 and 10 into Eq. 11 leads to:

$$\begin{aligned} \Lambda_m(n) = \frac{p_1}{1-p_1}&\left(\frac{\hat{\sigma}_d^2(m)}{\hat{\sigma}_d^2(m) + \hat{\sigma}_x^2(m)}\right)^{(k_m/2-1)} \\ &\times \exp\left(\frac{|\hat{Y}(n,m)|^2 \hat{\sigma}_x^2(m)}{2\hat{\sigma}_d^2(m)\left(\hat{\sigma}_d^2(m) + \hat{\sigma}_x^2(m)\right)}\right) \end{aligned} \tag{12}$$

Analogous to the linear frequency domain parameters presented in [4], we define Mel-domain a priori ($\hat{\xi}_m(n)$) and posteriori ($\hat{\gamma}_m(n)$) SNRs as:

$$\hat{\xi}_m(n) \triangleq \frac{\hat{\sigma}_x^2(m)}{\hat{\sigma}_d^2(m)}, \quad \hat{\gamma}_m(n) \triangleq \frac{|\hat{Y}(n,m)|^2}{\hat{\sigma}_d^2(m)}. \tag{13}$$

Since the term $\sigma_x^2(m)$ is hidden, we approximate the a priori SNR according to the maximum likelihood approach presented in [5]:

$$\hat{\xi}_m(n) \approx \max\{\bar{\gamma}_m(n) - 1, 0\}, \tag{14}$$

where the smoothed a posteriori SNR is determined recursively as:

$$\bar{\gamma}_m(n) = \delta\bar{\gamma}_m(n-1) + (1-\delta)\,\hat{\gamma}_m(n), \tag{15}$$

and where $0 \ll \delta < 1$ is the forgetting factor. Eq. 12 then reduces to:

$$\begin{aligned} \Lambda_m(n) = \frac{p_1}{1-p_1}&\left(\frac{1}{1 + \hat{\xi}_m(n)}\right)^{(k_m/2-1)} \\ &\times \exp\left(\frac{\hat{\gamma}_m(n)\,\hat{\xi}_m(n)}{2\left(1 + \hat{\xi}_m(n)\right)}\right) \end{aligned} \tag{16}$$

Substitution of Eqs. 16 into Eq. 10 reveals time- and channel-specific probabilities of active speech, which comprise the speech presence uncertainty mask.

B. HMM-Based Decoding

In Section III-A, we proposed speech presence uncertainty masks comprised of posterior probabilities conditioned on single Mel-domain power spectrum observations $|\hat{Y}(n,m)|^2$. In this section we apply HMM-based decoding to exploit the well-known temporal correlation of spectral speech data. Following [13], we define forward and backward variables recursively as:

$$\begin{aligned}\alpha_m^i(n) &= P\left(H_i \middle| |\hat{Y}(1,m)|^2, ..., |\hat{Y}(n,m)|^2\right) \\ &= \left[\sum_{j=0}^{1} a_{ji}\alpha_m^j(n-1)\right] p\left(|\hat{Y}(n,m)|^2 \middle| H_i\right), \\ \beta_m^i(n) &= P\left(H_i \middle| |\hat{Y}(n,m)|^2, ..., |\hat{Y}(n+N_{la},m)|^2\right) \\ &= \sum_{j=0}^{1} a_{ij}\beta_m^j(n+1)\, p\left(|\hat{Y}(n+1,m)|^2 \middle| H_j\right),\end{aligned} \tag{17}$$

where a_{ij} represents the transitional probability from state H_i to state H_j, and N_{la} denotes the number of look-ahead frames used during backwards estimation. Note that with the assumption of a 2-state Markov model, the steady-state probability p_1 can be obtained from transitional statistics as p_1=$a_{01}/(a_{01}+a_{10})$.

We define the ratio of forward variables as:

$$\Phi_m(n) = \frac{\alpha_m^1(n)}{\alpha_m^0(n)} = \frac{a_{01}+a_{11}\Phi_m(n-1)}{a_{00}+a_{01}\Phi_m(n-1)}\Lambda_m(n), \tag{18}$$

and the ratio of backward variables as:

$$\begin{aligned}\Psi_m(n) &= \frac{\beta_m^1(n)}{\beta_m^0(n)} \\ &= \frac{a_{01}+a_{11}\Psi_m(n+1)\Lambda_m(n+1)}{a_{00}+a_{01}\Psi_m(n+1)\Lambda_m(n+1)}.\end{aligned} \tag{19}$$

Since Eqs. 18 and 19 are expressed recursively, boundary conditions are provided as:

$$\Phi_m(1) = \frac{p_1}{1-p_1}, \quad \Psi_m(n+N_{la}) = \frac{p_1}{1-p_1}. \tag{20}$$

Note that a similar expression to Eq. 18 was presented in [16], but was applied on a global, channel-independent basis, for the task of voice activity detection (VAD). The authors of [17] also apply similar theory to the VAD task. Using Eq. 11 and Eqs. 17-19, the decoded GLR is defined as:

$$\tilde{\Lambda}_m(n) = \frac{\alpha_m^1(n)\,\beta_m^1(n)}{\alpha_m^0(n)\,\beta_m^0(n)} = \Phi_m(n)\,\Psi_m(n). \tag{21}$$

Substitution of Eq. 21 into Eq. 10 reveals a soft-decision speech presence uncertainty mask which exploits temporal correlation of spectral speech data.

TABLE I
PARAMETERS UTILIZED DURING MASK ESTIMATION

Parameter	Value	Description
δ	0.25	Forgetting factor from Eq. 15
a_{01}	0.15	Transitional probability, $H_0 \rightarrow H_1$ (Sec. III-B)
a_{10}	0.35	Transitional probability, $H_1 \rightarrow H_0$ (Sec. III-B)
η_{th}	0.75	Thresholding parameter (Sec. III-C)

C. Binary Masks

Although soft-decision masks are generally useful in noise robust speech processing, the majority of missing feature data imputation techniques for noise robust ASR, such as [14] and [15], require binary masks which differentiate between reliable and unreliable Mel-domain components. Probabilistic values derived in previous sections can be mapped to binary values via hard-thresholding with parameter η_{th}. Figure 1 provides illustrative examples of Mel-domain mask estimation. Panel **a** shows the clean utterance "nine one nine six nine five one," from the Aurora-2 database. Panel **b** provides the soft-decision mask from Sec. III-A, for the corresponding speech signal degraded by vehicle noise at 5 dB SNR. Panel **c** shows the mask after HMM-based decoding, and panel **d** provides the resulting binary mask.

IV. EXPERIMENTAL RESULTS

To assess the accuracy of proposed Mel-domain mask estimation, we apply it to MF-based ASR. As an illustrative example we utilize the spectral reconstruction technique from [14], which exploits the underlying compressibility of speech data to infer unreliable spectral components. The front end feature extraction includes 13 MFCCs, along with the log-energy, derivatives, and second derivatives. The overall ASR system was applied to the Aurora-2 database, and 16-state, 3-mixture word models were used for recognition.

As a baseline we implemented a Mel-domain version of the mask estimation technique proposed in [10], which determines reliable time-frequency components as the intersection of two criteria:

(i) The Negative Energy Criterion: $|\hat{Y}(n,m)| - |\hat{D}(n,m)| > 0$

(ii) The SNR Criterion: $\left(|\hat{Y}(n,m)| - |\hat{D}(n,m)|\right)^2 > 1/2|\hat{Y}(n,m)|^2$

where $\hat{D}(n,m)$ represents the Mel-filtered spectral noise estimate. As opposed to other prior mask estimation methods, [10] provides a fair baseline, since neither it nor the proposed technique require extensive training using noisy speech, or large codebooks.

Table II provides word-accuracy results for MF-based ASR using proposed mask estimation in combination with the spectral reconstruction technique from [14]. Results for spectral subtraction-based mask estimation from [10] are included as reference. Results for the use of oracle masks, which require knowledge of the clean speech signal, are included as a performance bound. Additionally, results for the baseline system (BL) without spectral reconstruction are included. It

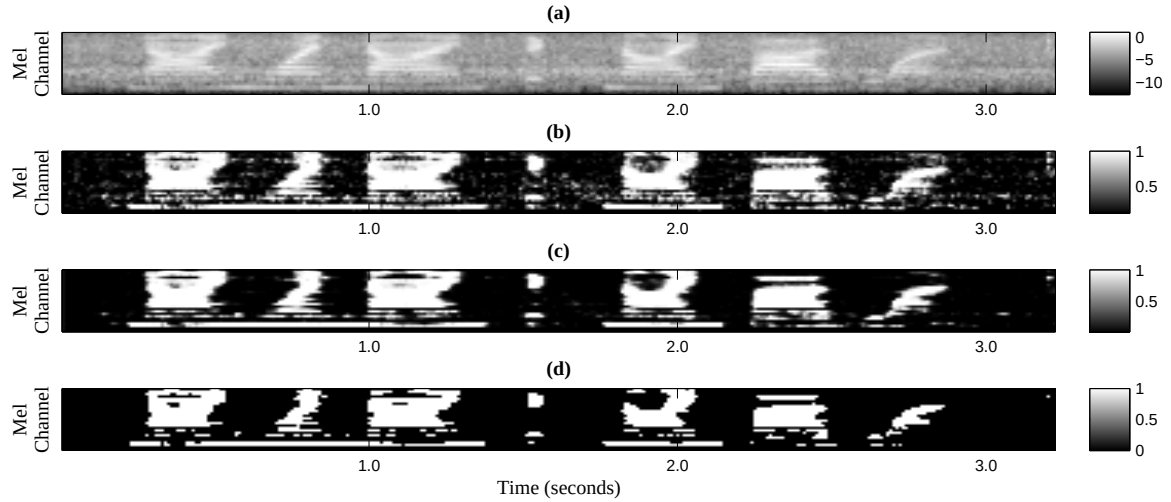

Fig. 1. Examples of Mel-Domain Mask Estimation: Panel **a** shows the clean utterance "nine one nine six nine five one," from the Aurora-2 database. Panel **b** provides the soft-decision mask from Sec. III-A, for the corresponding speech signal degraded by vehicle noise at 5 dB SNR. Panel **c** shows the mask after HMM-based decoding, and panel **d** provides the resulting binary mask.

TABLE II

WORD-ACCURACY RESULTS FOR MISSING FEATURE ASR USING COMPRESSIVE SENSING (CS)-BASED SPECTRAL RECONSTRUCTION [14]: RESULTS ARE AVERAGED ACROSS SETS A AND B OF THE AURORA-2 DATABASE. RESULTS ARE INCLUDED FOR SS-BASED MASKS [10], PROPOSED MASKS, AND ORACLE MASKS.

SNR (dB)	20	15	10	5	0	−5	Ave.
BL	95.5	84.5	62.6	38.1	18.5	10.2	51.6
SS-Based [10]	97.2	95.0	88.6	72.1	39.6	11.3	67.3
Proposed	97.6	95.9	91.0	79.4	53.4	21.8	73.2
Oracle	97.6	96.9	95.1	91.0	80.9	58.7	86.7

should be noted that all reported results include cepstral mean subtraction (CMS). Also, N_{la} was set to include the entire current utterance. As can be observed, the proposed mask estimation method provides significant improvements in word-accuracy relative to that of [10].

V. CONCLUSIONS

In this paper, we present a statistical approach to Mel-domain mask estimation in which reliability measures are derived as conditional probabilities of active speech using a Bayesian approach. Mel-domain power spectra are modeled as χ^2 processes with empirically-determined degrees of freedom. The proposed mask estimation algorithm is applied to the compressive sensing-based MF spectral reconstruction technique from [14], and is shown to outperform the baseline method from [10] in terms of word-accuracy.

REFERENCES

[1] B. Raj and R. Stern, *Missing Feature Approaches in Speech Recognition*, IEEE Signal Processing Magazine, Vol. 22, Issue 5, pp. 101-116, 2005.

[2] M. P. Cooke, A. Morris, P. D. Green, *Missing Data Techniques for Robust Speech Recognition*, Proc. ICASSP, vol. 2, pp. 863-866, 1997.

[3] W. Kim and R. M. Stern, *Band-Independent Mask Estimation for Missing-Feature Reconstruction in the Presence of Unknown Background Noise*, ICASSP, pp. 305-308,2006.

[4] R. J. McAulay and M. L. Malpass, *Speech Enhancement Using a Soft-Decision Noise Suppression Filter*, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 28, No.2, pp. 137-145, 1980.

[5] Y. Ephraim and D. Malah, *Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator*, IEEE Trans. on Acoustics Speech, and Signal Processing, Vol. 32, No. 6, pp. 1109-1121, 1984.

[6] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*

[7] H. Solomon and M. A. Stephens, *Distribution of a Sum of Weighted Chi-Square Variables*, American Statistical Association, Vol. 72, No. 360, pp. 881-885, 1977.

[8] R. Martin and T. Lotter, *Optimal recursive smoothing of non-stationary periodograms*, in Proc. International Workshop on Acoustic Echo Control and Noise Reduction, pp. 4346, 2001.

[9] M. V. Segbroeck and H. V. hamme, *Vector-Quantization based Mask Estimation for Missing Data Automatic Speech Recognition*, ICSLP, 2004.

[10] 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, pp. 24072410, 1999.

[11] P. Green , J. Barker , M. Cooke , and L. Josifovski, *Handling Missing and Unreliable Information in Speech Recognition*, Proc. AISTATS, 2001.

[12] M. L. Seltzer, B. Raj, R. M. Stern, *A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition*, Speech Communication, Vol. 43, pp. 379393, 2004.

[13] 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.

[14] B. J. Borgstrom and A. Alwan, *Utilizing Compressibility in Reconstructing Spectrographic Data, with Applications to Noise Robust ASR*, IEEE Signal Processing Letters, Vol. 16, Issue 5, pp. 398-401, 2009.

[15] B. J. Borgstrom and A. Alwan, *Missing Feature Imputation of Log-Spectral Data For Noise Robust ASR*, Workshop on DSP in Mobile and Vehicular Systems, 2009.

[16] J. Sohn, N. S. Kim, and W. Sung, *A statistical model-based VAD*, IEEE Signal Processing Letters, vol. 16, No. 1, pp. 1-3, 1999.

[17] W. Q. Syed and H.-C. Wu, *Speech waveform compression using robust adaptive voice activity detection for nonstationary noise*, EURASIP Journal on Audio, Speech, and Music Processing, 2008.