A CRITERION FOR THE ENHANCEMENT OF TIME-FREQUENCY MASKS IN MISSING DATA RECOGNITION

Daniel Pullella and Roberto Togneri

School of Electrical, Electronic and Computer Engineering
The University of Western Australia
{daniel,roberto}@ee.uwa.edu.au

ABSTRACT

Despite their effectiveness for robust speech processing, missing data techniques are vulnerable to errors in the classification of the input speech signal's time-frequency points. A direct method for the removal of these mask errors is through the top-down optimization of the estimated mask, however this requires a measure to evaluate the mask quality without a priori noise knowledge. In this paper we propose the normalized likelihood confidence as such a criterion for robust speaker recognition. In this approach the accuracy with which an estimated mask classifies time-frequency points as corrupt or reliable is related to its likelihood score confidence. This is based on the conceptual effect of binary mask errors on the model likelihood distributions produced by accumulated marginalization densities. Experimental results confirm a relationship between the normalized likelihood distance and the accuracy of the time-frequency mask produced by various estimation strategies.

Index Terms— robust speaker recognition, time-frequency masking, missing data

1. INTRODUCTION

Speech processing systems typically suffer a decrease in performance when noise and acoustic variabilities are present in the speech. Missing data methods have been shown to be effective for providing robustness to environmental distortions for both speech and speaker recognition tasks [1, 2]. In these techniques a time-frequency (TF) mask is constructed which labels each individual TF point as speech or noise dominated. A priori noise knowledge allows the construction of oracle masks which perfectly separate speech from noise, and as a result these masks can provide extremely high robustness to noise corruption. Past research has focused on producing an accurate estimation of the oracle reliability mask [3], however this is difficult in practice due to the presence of non-stationary noises. The weakness of traditional approaches to missing data is that the recognizer has no protection from errors in the estimated mask, particularly where truly unreliable components are assigned a high reliability.

Recent research in robust speaker recognition has attempted to improve the estimation of the reliability mask by utilizing information in the trained models. An example is the universal compensation technique [4] where a search is performed within multicondition spectra to produce the model specific feature subset which maximizes recognition performance. An alternate approach, which avoids the potentially expensive calculation of model specific subsets, is to use *top-down* optimization to remove errors within an initial estimated *bottom-up* TF mask with the goal of reproducing the ideal oracle mask. This approach requires the existence of a criterion to evaluate the quality of a given TF mask at each stage of the refinement process without the presence of a priori noise knowledge. Past research has related the difference in sound class dependent statistics between oracle and estimated TF masks to the difference in their recognition accuracies [5]. However, the calculation of this distance is dependent on the availability of the oracle mask, and thus cannot be used when attempting to reproduce the oracle mask from an initial inaccurate estimate.

This paper proposes normalized likelihood distances as a novel measure for determining the quality of estimated TF reliability masks in arbitrary unknown noise conditions. Conceptually the criterion is based on the differences between the model likelihood distributions of the perfect oracle mask and inaccurately estimated masks over a sufficiently wide frame context. This is enabled by the nature of the speaker recognition task in that a given speech sample corresponds to only one model over its entire duration. In conjunction with the discriminative properties of missing data likelihoods, this allows the accuracy with which an estimated mask classifies the TF points as speech or noise dominated to be related to its likelihood score confidence. Formulating the mask confidence as the normalized likelihood distance thus provides a criterion for measuring the quality of reliability masks without the requirement of oracle noise knowledge. The measure is evaluated experimentally for estimated TF masks of varying accuracy, and the results confirm the ability of the normalized likelihood distance to discriminate between oracle masks and corrupted estimates.

The remainder of this paper is organized as follows. Section 2 describes the proposed measure including an overview of missing data marginalization and the theoretical calculation of the normalized segment likelihood distance. Section 3 presents an evaluation of the measure and a discussion of the results. Conclusions and future work are outlined in Section 4.

2. METHOD OVERVIEW

2.1. GMM Identification with Marginalization

In Gaussian Mixture Model (GMM) speaker identification each speaker is represented as a weighted sum of M Gaussian densities [6]. Given a speaker represented by model λ, and a spectral observation vector $\vec{x} = (x_1, x_2, ..., x_D)'$ the observation likelihood is

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} c_i \mathcal{N}(\vec{x}; \vec{\mu}_i, \Sigma_i), \quad (1)$$

This work was supported in part by the Samaha Research Scholarship (F8046) of The University of Western Australia.

where c_i is the weight of the ith mixture and $\mathcal{N}$ is a D-variate Gaussian with mean vector $\vec{\mu}_i \in \mathbb{R}^D$ and covariance matrix $\Sigma_i \in \mathbb{R}^{D\times D}$. Each speaker is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities as in [6].

Let the binary TF mask vector corresponding to observation $\vec{x}$ be $\vec{m} = (m_1, m_2, ..., m_D)'$. Based on this mask vector the components of a given observation can be labeled as reliable (r) or unreliable (u) according to $\vec{x}_r = \{x_f | m_f = 1, f = 1, 2, \ldots, D\}$ and $\vec{x}_u = \{x_f | m_f = 0, f = 1, 2, \ldots, D\}$. This allows a corresponding separation of the model parameters (refer to [7]), and the marginal probability density $p(\vec{x}_r|\lambda)$ is obtained by integrating over the distribution of the unreliable components:

$$p(\vec{x}_r|\lambda) = \sum_{i=1}^{M} c_i \mathcal{N}(\vec{x}_r; \vec{\mu}_{r_i}, \Sigma_{rr_i}) \int_{\vec{x}_l}^{\vec{x}_h} \mathcal{N}(\vec{x}_u; \vec{\mu}_{u|r_i}, \Sigma_{u|r_i}) d\vec{x}_u, \quad (2)$$

where $\vec{\mu}_{u|r_i} \in \mathbb{R}^{D_u}$ and $\Sigma_{u|r_i} \in \mathbb{R}^{D_u \times D_u}$ are the conditional mean and conditional covariance respectively as defined in [8]. For a group of S speakers with corresponding GMMs λ_s, $s \in \mathcal{S} = \{1, 2, ..., S\}$ the identification decision for an utterance $X = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T)$ is obtained by maximization of the marginal log-likelihoods accumulated over all observations:

$$\hat{s} = \underset{1\le k\le S}{\operatorname{argmax}} \sum_{t=1}^{T} \log p(\vec{x}_{t_r}|\lambda_k). \quad (3)$$

2.2. Bounded Marginalization and Mask Errors

In bounded marginalization recognition the lower and upper integration bounds are set to 0 and the observed component value respectively such that $[\vec{x}_l, \vec{x}_h] = [\vec{0}, \vec{x}_u]$ [8]. Here the bounded integral effectively uses the unreliably declared TF points to provide counter-evidence by penalizing models whose energy is much greater than the observed feature value.

Binary estimates of the oracle mask may contain two types of errors: inclusion errors and deletion errors. Conceptually, when compared to the oracle mask, estimated masks with either type of error experience a decrease in the likelihood of the true model and an increase in the likelihood of one or more imposter models. For inclusion errors the true model likelihood is decreased due to the mismatch between the trained distributions and the observed values of the inclusion error components, while imposters may benefit from a decrease in the likelihood penalty (since some truly noisy components no longer contribute to the multi-variate integral). For deletion errors the true model likelihood decreases since observed components which are matched to the trained distribution are now used in the integral contribution increasing the counter-evidence. Conversely likelihoods are increased for imposter models whose feature distributions have lower energy than the observed deleted components due to the removal of the mismatch.

2.3. Likelihood Confidence Hypothesis

The conceptual effect of binary mask errors allows the formulation of a relationship between the distribution of accumulated likelihoods over all models and the accuracy of an estimated mask compared to the true oracle mask. A hypothesis is stated as follows:

1. Over a sufficient number of observations, the oracle reliability mask produces both a large likelihood for the winning model (often the true model), and a large likelihood difference between this winning model and its nearest competitors.
2. The correction of all errors within an estimated mask should increase the likelihood score of the winning model and the likelihood difference between the winning model and its nearest competing models in comparison to these values from the initial estimated mask.

This is based on the critical assumption that an accurately estimated oracle mask will have a higher *likelihood confidence*, and will possess a winning model with a higher likelihood score compared to an inaccurately estimated oracle mask.

2.4. Normalized Segment Likelihood Distances

Let an observation segment $\mathcal{X}$ be defined as the set of K observations $\mathcal{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_K\}$, with corresponding reliability mask vectors $\mathcal{M} = \{\vec{m}_1, \vec{m}_2, \ldots, \vec{m}_K\}$. The likelihood score of the segment $\mathcal{X}$ on a given model λ is computed as the sum of the marginal likelihoods over all observations:

$$L_\lambda(\mathcal{X}|\mathcal{M}) = \sum_{k=1}^{K} \log p(\vec{x}_{k_r}|\lambda), \quad (4)$$

where $p(\vec{x}_{k_r}|\lambda)$ is defined as in (2), and the reliability partitioning is determined by $\vec{m}_k$. The distribution of segment likelihoods for all speaker models is defined as

$$\mathcal{D}(\mathcal{X}|\mathcal{M}) = \cup_{s\in\mathcal{S}}\{L_{\lambda_s}(\mathcal{X}|\mathcal{M})\}. \quad (5)$$

The likelihood confidence of the mask segment is quantified by the distance between the maximum likelihood within the distribution and the mean likelihood of the N-nearest competing models. The maximum likelihood within the distribution set is given by

$$L_{\max}(\mathcal{X}|\mathcal{M}) = \max_{s\in\mathcal{S}} L_{\lambda_s}(\mathcal{X}|\mathcal{M}), \quad (6)$$

and $s_{\max}$ denotes the speaker whose model produces this maximal likelihood. Formally the N-nearest competitor speakers are the subset of speakers $Z \subseteq \mathcal{S}$ with $|Z| = N$ such that

$$Z = \underset{|V|=N, V\subseteq\mathcal{S}, s_{\max}\notin V}{\operatorname{argmax}} Q(V), \quad (7)$$

where $Q(.)$ is the sum of the segment likelihoods from all models whose speakers are included in the set:

$$Q(V) = \sum_{s\in V} L_{\lambda_s}(\mathcal{X}|\mathcal{M}). \quad (8)$$

The distribution of competitor segment likelihoods is thus given by

$$\mathcal{D}_{\text{comp}}(\mathcal{X}|\mathcal{M}) = \{L_{\lambda_s}(\mathcal{X}|\mathcal{M}) | s \in Z\}, \quad (9)$$

and the likelihood distance (LD) of the mask segment is

$$LD = L_{\max}(\mathcal{X}|\mathcal{M}) - E[\mathcal{D}_{\text{comp}}(\mathcal{X}|\mathcal{M})]. \quad (10)$$

To prioritize mask segments which approximate the oracle segment over highly corrupted segments, the likelihood distance LD is normalized by the absolute value of the maximal likelihood to produce the normalized segment likelihood distance (NSLD):

$$NSLD = \frac{L_{\max}(\mathcal{X}|\mathcal{M}) - E[\mathcal{D}_{\text{comp}}(\mathcal{X}|\mathcal{M})]}{|L_{\max}(\mathcal{X}|\mathcal{M})|^\kappa}, \quad (11)$$

where κ is a bias factor controlling the relative weight of the absolute value of the maximal likelihood compared to the likelihood distance for the criterion (see Fig. 1).

3. EVALUATION

3.1. Experimental Setup

The criterion was evaluated on text-independent speaker identification using a 95 speaker set from the TIMIT database. For each

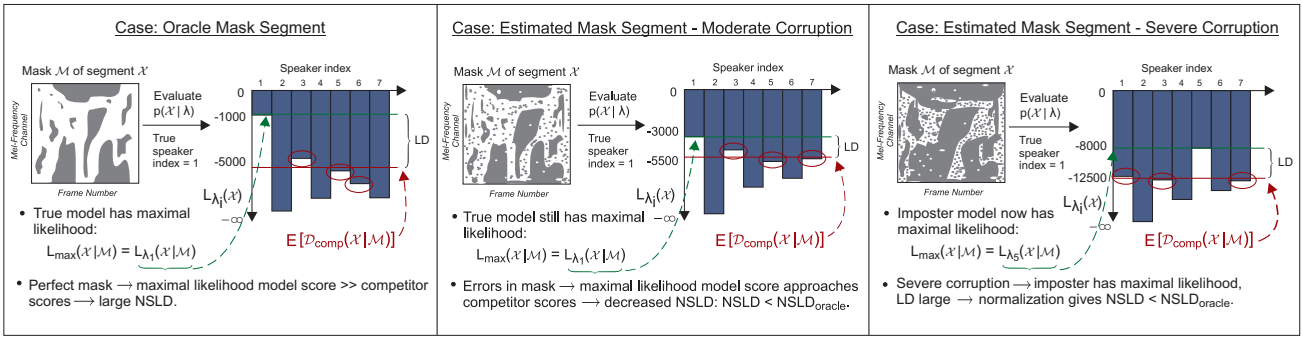


Fig. 1: The normalized likelihood confidence for TF masks of varying accuracy. For oracle mask segments the maximal distribution likelihood exceeds the competitor likelihoods. Moderate corruption produces a decreased maximal-to-competitor likelihood distance compared to the oracle segment. Extreme corruption may produce maximal likelihoods from imposter models, and normalization is thus required.

speaker the SI and SX sentences used for model training and segments from the SA sentences used for testing. Averaged NSLD values were calculated for 21 randomly selected test utterances from the data set. The Hidden Markov Model Toolkit (HTK) [9] was used to construct spectral feature vectors from a 48-channel mel-filterbank as well as 4 mixture full covariance GMMs to model the speakers.

Additive white noise from the NOISEX database was used to corrupt test utterances mixed at SNRs of 20, 10 and 0 dB. In each noise condition the behavior of the normalized likelihood distance measure was evaluated for varying degrees of mask corruption compared to the oracle mask. For an observed noisy TF component $x_f \in \vec{x}_t$ with corresponding clean speech and noise spectral energies of x_f^s and x_f^n respectively, the oracle mask value is given by

$$m_f^{\text{oracle}} = \begin{cases} 1 & \text{if} \quad x_f^s > x_f^n, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

The response of the NSLD measure to random inclusion and deletion errors in the oracle mask was examined, where P_{inc} and P_{del} are the respective probabilities of including each noise dominated component and deleting each speech dominated component. Since practical estimation techniques typically produce masks containing block corruption rather than random errors, the criterion was also evaluated for spectral subtraction mask estimation [10]. For estimated speech and noise spectral energies $\bar{x}_{tf}^s$ and $\bar{x}_{tf}^n$ respectively, the spectral subtraction mask value is defined as

$$m_f^{\text{ss}(\theta)} = \begin{cases} 1 & \text{if} \quad 10\log_{10}\left(\frac{\bar{x}_f^s}{\bar{x}_f^n}\right) > \theta, \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

where the energy estimates are produced as in [7], and θ is the energy threshold in dB. To investigate the effect of inclusion errors in spectral subtraction masks, the behaviour of the NSLD was also measured for combined oracle and spectral subtraction masking where component wise multiplication was used to remove inclusion errors:

$$m_f^{\text{o*ss}(\theta)} = m_f^{\text{oracle}} * m_f^{\text{ss}(\theta)} \tag{14}$$

3.2. Results

NSLD values were calculated for segments consisting of $K = 50$ frames using a competitor set size of $N = 5$ and a normalization bias of $\kappa = 1$. These values were determined empirically based on separate experiments which are omitted for brevity. For random corruption it is observed that the normalized likelihood distance peaks for the oracle mask ($P_{\text{inc}} = P_{\text{del}} = 0$) in conditions of 20 dB and 10 dB (see Fig. 2(a) and 2(b)). As the probability of deletion and inclusion is increased the NSLD decreases from this peak value until sufficient corruption is introduced such that there is a switch in the model producing the maximal likelihood. For the 10 dB case this occurs for inclusion and deletion corruption probabilities of $P_{\text{inc}} \approx 0.5$ and $P_{\text{del}} \approx 0.7$ respectively, and additional corruption beyond these values causes the NSLD to increase as the distance between the new maximal likelihood and its competing likelihoods increases.

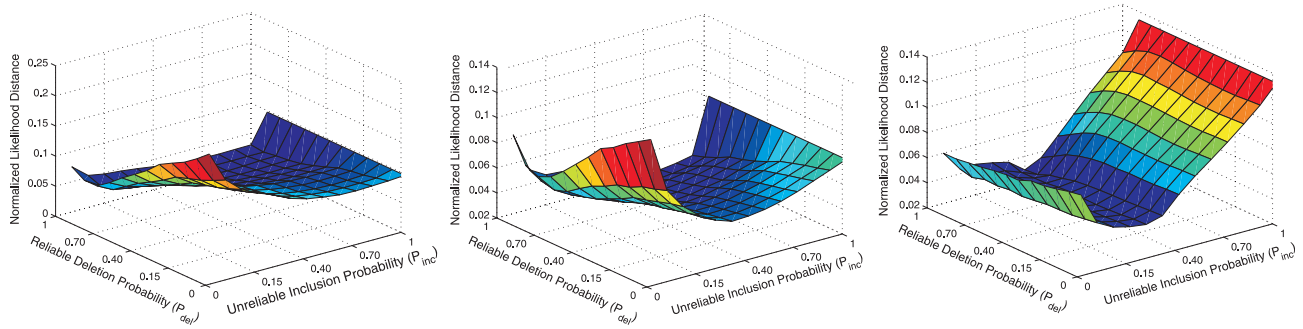
In the 0 dB condition the peak NSLD value is obtained for the unity mask, and only a small amount of inclusion corruption causes inflection (see Fig. 2(c)). Due to the strength of the noise, oracle segments have small maximal-to-competitor likelihood distances and so less inclusions are required to cause a switch in the model producing the maximal likelihood. Although additional corruption may decrease the value of this maximal likelihood the mean competitor likelihood decreases more rapidly allowing the corrupted segment NSLD to exceed the oracle segment NSLD. In this case the NSLD can correctly discriminate between the true oracle mask and inclusion error corrupted masks only for $0 \leq P_{\text{inc}} \leq 0.6$.

A relationship between the quality of the estimated mask and the obtained NSLD values was also confirmed for spectral subtraction masking. In all noise conditions significantly lower NSLD values were observed for spectral subtraction estimated masks compared to the true oracle mask (see Fig. 3). When inclusion errors were removed in the spectral subtraction masks the NSLD values increased to become approximately equal to the oracle NSLD values. The difference between NSLD values for spectral subtraction and combined oracle spectral subtraction masks decreased as θ increased, and this is due to a reduction in the number of inclusion errors which are removed by the oracle combination.

3.3. Discussion

The results presented demonstrate the validity of using the normalized likelihood confidence as a measure for determining the quality of an estimated mask segment. For both randomly corrupted oracle masks and spectral subtraction estimated masks a decrease in NSLD was observed compared to the true oracle mask.

Although not explicitly shown in this investigation, it should be noted that the behavior of the NSLD measure is dependent on both the segment size and the number of competitors considered. The segments should consist of a large enough number of frames such that reliable deletions occurring in low speech energy regions cannot bias the likelihood confidence towards imposter models. The trade-off in choosing the size of the competitor likelihood distribution is between



(a) White Noise 20dB (b) White Noise 10dB (c) White Noise 0dB

Fig. 2: Normalized segment likelihood distances averaged over all segments for random inclusion and deletion error corruption of the oracle mask. Results are presented for additive white noise distortion at 20 dB (a), 10 dB (b) and 0 dB (c).

obtaining maximum discriminability between oracle and corrupted segments and introducing vulnerability by including too many outlier models in the distance calculation. The optimal competitor set size will depend on the type of mask errors encountered and may therefore differ depending on the specific mask estimation technique used.

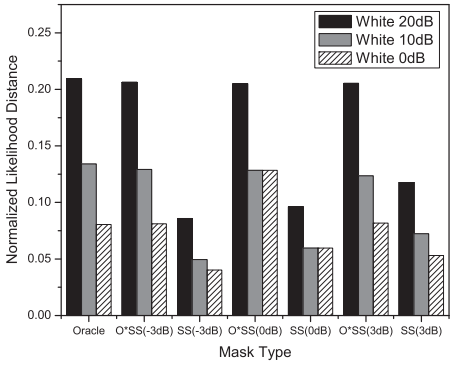


Fig. 3: Normalized segment likelihood distances for the oracle mask (Oracle), standard spectral subtraction masks (SS) and combined oracle spectral subtraction masks (O*SS). Energy thresholds of $\theta = -3, 0$ and 3 dB are used.

The advantage of using the NSLD to measure mask quality is that its calculation does not require knowledge of the oracle mask. This motivates its use as a criterion to evaluate modifications made to segments of an estimated mask by a top-down refinement scheme. However use of the metric in this way is subject to some limitations. Firstly, it cannot be guaranteed that the oracle mask segment has the largest NSLD value compared to any arbitrary mask segment. The evaluation demonstrates that for binary estimated masks with reasonable corruption compared to the oracle mask the NSLD decreases and decreases in proportion to the magnitude of this corruption. The primary requirement of top-down mask optimization is to correct segments where corruption causes the maximal likelihood value to occur for an imposter model, but at low SNRs segments with extreme inclusion corruption can be favored by the criterion (as in Fig. 2(c)). This emphasizes the need to constrain the optimization using an initial bottom-up mask estimate to avoid these superficial cases. Secondly, the lack of a closed form solution to maximizing the NSLD suggests that optimization must proceed through successive modification and evaluation of possible candidate segments. The large space of possible solutions for the optimized segment may impose limits on the techniques which can be applied.

4. CONCLUSIONS

This work has proposed normalized likelihood distances as a criterion for measuring the quality of a TF mask for missing data speaker recognition. Based on the properties of bounded marginal densities, the normalized likelihood confidence is used to quantify the corruption in an estimated reliability mask without the need for a priori noise knowledge. Experimental evaluation confirmed the existence of a relationship between the averaged NSLD values and the accuracy of the reliability mask estimate for both random corruption of the ideal mask and practical SNR-based estimation. Future work will focus on the utilization of the measure within an optimization method for the refinement of estimated reliability masks.

5. REFERENCES

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, no. 3, pp. 267–285, 2001.

[2] Y. Shao and D. Wang, "Robust speaker recognition using binary time-frequency masks," in *Proc. ICASSP*, 2006, vol. 1, pp. 645–648.

[3] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, pp. 181–197. Kluwer Academic, Norwell MA, 2005.

[4] J. Ming, D. Stewart, and S. Vaseghi, "Speaker identification in unknown noisy conditions - a universal compensation approach," in *Proc. ICASSP*, 2005, vol. 1, pp. 617–620.

[5] J. Gemmeke, B. Cranen, and L. Bosch, "On the relation between statistical properties of spectrographic masks and recognition accuracy," in *Proc. IASTED SPPRA*, 2008.

[6] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, 1995.

[7] M. Kühne, D. Pullella, R. Togneri, and S. Nordholm, "Towards the use of full covariance models for missing data speaker recognition," in *Proc. ICASSP*, 2008, pp. 4537–4540.

[8] A. Morris, M. Cooke, and P. Green, "Some solutions to the missing feature problem in data classification, with application to noise robust asr," in *Proc. ICASSP*, 1998, vol. 2, pp. 737–740.

[9] S. Young et al., *Hidden Markov Model Toolkit (HTK) Version 3.2.1 User's Guide*, Cambridge University Engineering Department, Cambridge, MA, 2002.

[10] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Proc. ICASSP*, 1998, vol. 1, pp. 121–124.