ROBUST SPEAKER IDENTIFICATION USING COMBINED FEATURE SELECTION AND MISSING DATA RECOGNITION

Daniel Pullella, Marco Kühne and Roberto Togneri

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

ABSTRACT

Missing data techniques have been recently applied to speaker recognition to increase performance in noisy environments. The drawback of these techniques is the vulnerability of the recognizer to errors in the classification of time-frequency points as corrupt or reliable. In this paper we propose the combination of missing data processing and feature selection to reduce these errors. The formation of a set of speaker discriminative features allows time-frequency reliability masks to be refined via the removal of the non-discriminative frequency sub-bands. The reduced set is selected dynamically using multi-condition training and an estimate of the global SNR allowing for efficient top-down processing. Experimental results show that the combined technique achieves significant improvement over traditional bottom-up processing thus demonstrating the validity of the approach.

Index Terms— speaker identification, robustness, feature selection, missing data

1. INTRODUCTION

Speaker recognition is an important problem in modern communications with applications to security and access control as well as personalization. In practice speaker recognition is adversely affected by the presence of noise and acoustic variabilities in the speech to be processed. Missing data processing is an effective technique for compensating against arbitrary disturbances within a speech signal. These approaches rely on the construction of a time-frequency (TF) mask to label each TF point as either speech or noise dominant. Past research [1, 2] has demonstrated, in the context of speech recognition, that missing data methods can provide extremely high robustness if a perfectly constructed TF mask is available. Based on this observation research has largely concentrated on the accurate estimation of the TF reliability mask. However, producing the ideal TF mask under practical non-stationary noises remains an extremely difficult task. Techniques used to estimate the ideal mask produce two types of errors: the inclusion of unreliable points and the exclusion of reliable points. The weakness of traditional approaches to missing data is that the recognizer has no protection from these errors, particularly where true unreliable components are assigned a high reliability.

To solve this problem recent research has introduced the idea of combining *bottom-up* (BU) methods such as auditory scene analysis with *top-down* (TD) methods which utilize trained acoustic models. Examples specific to speech recognition include the multisource decoder [3], and the two-stage speech separation hypothesis testing approach [4]. Top-down processing is utilized for speaker recognition in the universal compensation technique [5], where a search is performed over the feature space of each model within a set of noise corrupted models to find components that best match the input spectrum. Although this gives the best possible subset for recognition, the required exhaustive search over the feature space can be computational for high feature dimensions.

In this paper we propose a novel combination of bottom-up missing data processing and top-down feature selection to achieve efficient robust speaker identification. This technique is based on the uneven distribution of speaker specific information in the frequency domain, allowing the removal of non-discriminative frequency subbands and the formation of a reduced feature set. Multi-condition training is used as in [5] to dynamically select the most discriminative features for a set of speakers given an estimate of the global signal-to-noise ratio (SNR) of the evaluation environment. This subset is applied to the missing data mask resulting in the removal of unreliable inclusion errors in non-discriminative sub-bands. This technique calculates the discriminative ability of each feature based on the trained speaker models, and thus can perform efficient top-down processing independent of the feature dimension. Experimental evaluation was conducted by comparing the performance of the combined missing data feature selection approach to standard bottom-up methods for speech corrupted by stationary and non-stationary additive noises. The results show that our combined approach performs significantly better than these bottom-up only methods.

The remainder of this paper is organized as follows. Section 2 describes the proposed system including the feature selection theory and its integration with missing data processing. The evaluation of the system is presented in Section 3. Conclusions and future work are discussed in Section 4.

2. SYSTEM OVERVIEW

This work proposes a new technique to perform speaker recognition in arbitrary noise based on the novel combination of missing data speech processing with model based feature selection. Missing data processing is firstly applied to the input speech producing a binary TF mask which captures the effect of noise corruption. Using the trained speaker models a subset of speaker discriminative features is defined dynamically based on the approximate noise conditions of the test environment. By using this subset to refine the reliability mask, unreliable inclusion errors within the non-discriminative frequency bands are removed, and the performance of the missing data recognizer is enhanced (see Fig. 1).

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

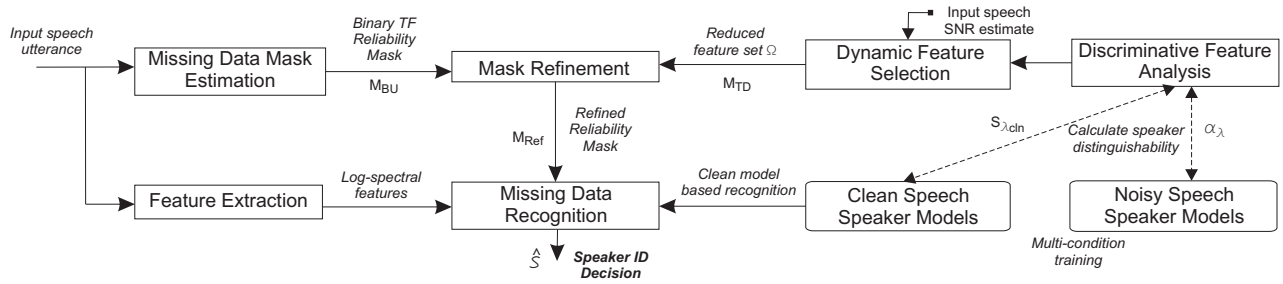

Fig. 1. The combined feature selection missing data speaker identification system. Bottom-up noise estimation produces a binary TF reliability mask. Top-down selection of the most speaker discriminative features is used to refine this mask and improve recognition.

2.1. GMM-based Speaker Identification

Identification is performed using Gaussian Mixture Models (GMMs) which represent each speaker as a weighted sum of M diagonal D-variate Gaussian densities [6]. For an observation vector denoted $\mathbf{x} = (x_1, x_2, ..., x_D)'$ and a given speaker represented by model λ, the probability density of the observation is

$$p(\mathbf{x}|\lambda) = \sum_{k=1}^{M} c_k \prod_{f=1}^{D} \mathcal{N}(x_f; \mu_{kf}, \sigma_{kf}^2), \tag{1}$$

where c_k is the weight of the k^{th} mixture and $\mathcal{N}(x_f; \mu_{kf}, \sigma_{kf}^2)$ is a uni-variate Gaussian distribution with mean μ_{kf} and variance σ_{kf}^2. Each speaker model is defined by the set of mean vectors, variance vectors and weights from all the component distributions:

$$\lambda = \{c_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2 | k = 1, 2, ..., M\}. \tag{2}$$

Model parameters are trained using maximum likelihood (ML) estimation via the expectation-maximization (EM) algorithm. For a set of $\mathcal{S}$ speakers $\mathcal{S} = \{1, 2, ..., N\}$ clean speech training produces the corresponding GMMs $\lambda_{j_{\text{cln}}}$, $j = 1, 2, ..., N$. The identification decision $\hat{\mathcal{S}}$ for an utterance consisting of a sequence of observation vectors $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)'$ is achieved by maximizing the total log-likelihood according to

$$\hat{\mathcal{S}} = \underset{1 \le j \le N}{\operatorname{argmax}} \sum_{t=1}^{T} \log p(\mathbf{x}_t | \lambda_{j_{\text{cln}}}). \tag{3}$$

2.2. Missing Data Mask Estimation

Missing data processing is based on a TF representation of the speech signal, where the effect of noise is modeled by the corruption of individual TF points. This requires the construction of a TF mask indicating the reliability of each TF point. The ideal TF mask is produced according to the 0 dB SNR criterion, where TF points are assigned as reliable if the speech energy exceeds the noise energy:

$$M(t, f) = \begin{cases} 1 & \text{if} \quad X(t, f) > N(t, f), \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Here $M(t, f)$ is the TF mask value at time t and frequency f, while $X(t, f)$ and $N(t, f)$ are the speech and noise energies respectively. However, the construction of this mask requires a priori knowledge of the noise and is not producible in practice. In this system spectral subtraction is used to estimate the binary TF reliability mask:

$$M_{\text{BU}}(t, f) = \begin{cases} 1 & \text{if} \quad Y(t, f) - \bar{N}(t, f) > \bar{N}(t, f), \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $Y(t, f)$ is the power spectrum of the noisy speech signal and $\bar{N}(t, f)$ is the estimated noise power spectrum. The 0 dB SNR criterion is used due to its superior performance compared to the negative energy criterion [1]. The estimated noise power is calculated by averaging the power spectrum of the first 10 frames of the utterance.

2.3. Dynamic Feature Selection

The feature selection subsystem utilizes information in the trained speaker models to identify the features with the most speaker distinguishability for the given test utterance. Discriminative analysis is used to calculate for each filter-bank feature $f \in \mathcal{F} = \{1, 2, ..., D\}$ the speaker distinguishability (quantified by the *feature selectivity* $S(f)$) in clean speech, and multi-condition training is employed to estimate the robustness of each feature in the given environment.

The use of diagonal covariance models translates to an assumption of statistical independence between feature components. The F-Ratio can thus be used to calculate the selectivity of a feature based on its distribution for each speaker [7, 8]. F-Ratio analysis is applied to each feature of speaker model λ by approximating its distribution with the single constituent mixture of maximal weighting:

$$\tilde{\lambda} = \{c_\tau, \boldsymbol{\mu}_\tau, \boldsymbol{\sigma}_\tau^2 | \tau = \underset{k}{\operatorname{argmax}}\, c_k\}. \tag{6}$$

The selectivity for feature f of speaker model $\tilde{\lambda}_{\text{cln}}$ is

$$S_{\tilde{\lambda}_{\text{cln}}}(f) = \frac{(\mu_{\tau f_{\text{cln}}} - \bar{\mu}_{\tau f_{\text{cln}}})^2}{\sigma^2_{\tau f_{\text{cln}}}}, \tag{7}$$

where $\mu_{\tau f_{\text{cln}}}$ and $\sigma^2_{\tau f_{\text{cln}}}$ are the mean and variance respectively of the distribution for feature f of model $\tilde{\lambda}_{\text{cln}}$, and $\bar{\mu}_{\tau f_{\text{cln}}}$ is the average distribution mean for feature f over all models $\tilde{\lambda}_{j_{\text{cln}}}, j = 1, 2, ..., N$.

Noise in the testing speech causes a mismatch between the model based clean speech selectivities and the true speaker distinguishability of the features. To solve this problem multi-condition training is used to estimate the robustness of each feature in the given noise condition and hence modify its selectivity. White noises of varying SNR are added to the training speech data allowing the construction of noise corrupted speaker models $\lambda_{j_{\text{snr}}}, j = 1, 2, ..., N, \text{snr} \in [-\infty, \infty]$. Given an estimate of the global SNR of the noisy speech, the closest set of white noise models is used to alter the selectivity values prioritizing features that are robust.

To achieve this an attenuation factor $\alpha_{\tilde{\lambda}}(f)$ is defined based on the distance between the distribution mean for the clean speech models $\mu_{\tau f_{\text{cln}}}$ and the noise corrupted speech models $\mu_{\tau f_{\text{snr}}}$ at the estimated global SNR:

$$\alpha_{\tilde{\lambda}}(f) = \gamma^{-|\mu_{\tau f_{\text{cln}}} - \mu_{\tau f_{\text{snr}}}|}. \quad (8)$$

Here γ is a strictly positive constant resulting in $\alpha_{\tilde{\lambda}}(f)$ values in the interval $[0, 1]$. The magnitude of γ determines how much attenuation a noise affected feature receives. This technique dynamically adapts the clean speech model selectivity values using a global estimate of the input utterance SNR. The features whose distributions show large invariance to the noise have $\alpha_{\tilde{\lambda}}(f) \approx 1$, and the features whose distributions are extremely distorted by the noise have $\alpha_{\tilde{\lambda}}(f) \approx 0$.

The dynamic selectivity for each feature distribution is

$$S_{\tilde{\lambda}}(f) = \alpha_{\tilde{\lambda}}(f) \times S_{\tilde{\lambda}_{\text{cln}}}(f), \quad (9)$$

and the overall selectivity $S(f)$ for each feature f is obtained by summing the distribution selectivities over all speaker models:

$$S(f) = \sum_{j=1}^{N} S_{\tilde{\lambda}_j}(f) = \sum_{j=1}^{N} \alpha_{\tilde{\lambda}_j}(f) \times S_{\tilde{\lambda}_{j_{\text{cln}}}}(f). \quad (10)$$

2.4. Bottom-up Refinement via Top-down Selection

Refinement of the TF reliability mask proceeds by first forming the feature subset based on the selectivity values. For binary masking the subset Ω is created by including the κ features with the highest selectivity values. Formally this is the subset of features $\Omega \subseteq \mathcal{F}$ with cardinality $|\Omega| = \kappa$ such that

$$J(\Omega) = \max_{Z \subseteq \mathcal{F}, |Z| = \kappa} J(Z), \quad (11)$$

where $J(.)$ is the selection criterion function defined as the sum of the feature selectivities over all included features:

$$J(Z) = \sum_{f \in Z} S(f). \quad (12)$$

This subset is then expressed as a TF mask

$$M_{\text{TD}}(t, f) = \begin{cases} 1 & f \in \Omega, \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

which allows each point within the bottom-up mask to be refined based on the corresponding point in the (top-down) selection mask. The reliability and selection masks are combined according to

$$M_{\text{Ref}}(t, f) = M_{\text{BU}}(t, f) * M_{\text{TD}}(t, f), \quad (14)$$

where the bottom-up reliability mask is given by $M_{\text{BU}}(t, f)$, the top-down selection mask by $M_{\text{TD}}(t, f)$ and $*$ is the binary AND operator (see Fig. 2).

2.5. Missing Data Recognition

Marginalization based recognition is implemented as it facilitates the combination between bottom-up and top-down processing. Let the refined binary TF mask vector corresponding to observation vector $\mathbf{x}_t$ be $\mathbf{m}_t = (m_{t1}, m_{t2}, ..., m_{tD})' = (M_{\text{Ref}}(t, 1), M_{\text{Ref}}(t, 2), ..., M_{\text{Ref}}(t, D))'$. The probability density for vector $\mathbf{x}_t$ produced by model λ using bounded marginalization becomes

$$p(\mathbf{x}_t|\lambda) = \sum_{k=1}^{M} c_k \prod_{f=1}^{D} \left(m_{tf} \mathcal{N}(x_f; \mu_{kf}, \sigma_{kf}^2) + (1 - m_{tf}) \int_{x_{\text{low}}}^{x_{\text{high}}} \mathcal{N}(\tilde{x}_f; \mu_{kf}, \sigma_{kf}^2) d\tilde{x}_f \right). \quad (15)$$

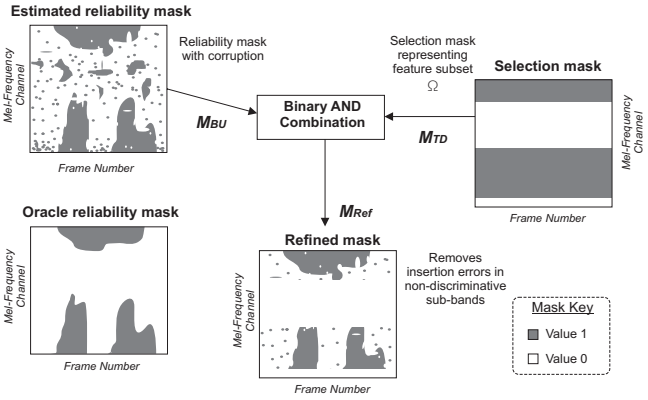

Fig. 2. Refinement of an estimated binary reliability mask using feature selection. For a perfect (oracle) reliability mask the produced refined mask closely approximates the original estimated mask. For an imperfect (estimated) reliability mask unreliable inclusion errors occurring within the non-discriminative features are removed.

To complement the top-down feature selection process the standard missing data recognition strategy is modified: bounded marginalization is only performed for an unreliable TF point in the refined mask if this point is included in the selection mask. The integration bounds are therefore defined by

$$[x_{\text{low}}, x_{\text{high}}] = \begin{cases} [0, x_f] & f \in \Omega, \\ [-\infty, \infty] & \text{otherwise.} \end{cases} \quad (16)$$

Thus the missing data recognizer uses the knowledge provided by the feature selection subsystem about the discriminative ability of a particular feature in the estimated noise conditions. A TF point labeled as discriminative by the selection subsystem is assumed to have some useful information, even if it is unreliable according to the bottom-up missing data strategy. However, if the feature is labeled as non-discriminative then the point is fully marginalized and has no effect on recognition. Using the bounded marginalization density defined by (15) speaker recognition is performed as in (3).

3. EVALUATION

The system was evaluated via closed-set text-independent speaker identification experiments with a 31 speaker subset (21 males and 10 females) of the TiDigits database. For each speaker 50 of the available 77 connected digits speech utterances were randomly selected for model training, and the remaining 27 utterances were used for testing. The Hidden Markov Model Toolkit (HTK) [9] was used to construct the GMMs, where diagonal covariance matrices were assumed for each of the 16 mixtures. The speech utterances were framed using a 25 ms Hamming window with a 10 ms frame step. A 48-channel HTK mel-filterbank was used to produce log-spectral feature vectors for each frame. A baseline system using cepstral features was also evaluated. For this system the speaker models were created using a vector of the first 24 mean normalized MFCCs derived from the 48-channel HTK mel-filterbank.

Additive noise conditions were simulated by corrupting each speaker's testing utterances with stationary white noise and non-stationary factory noise at SNRs ranging from 20 dB to -5 dB. White noise at these SNR levels was also added to each speaker's training speech data to produce the multi-condition noise models re-

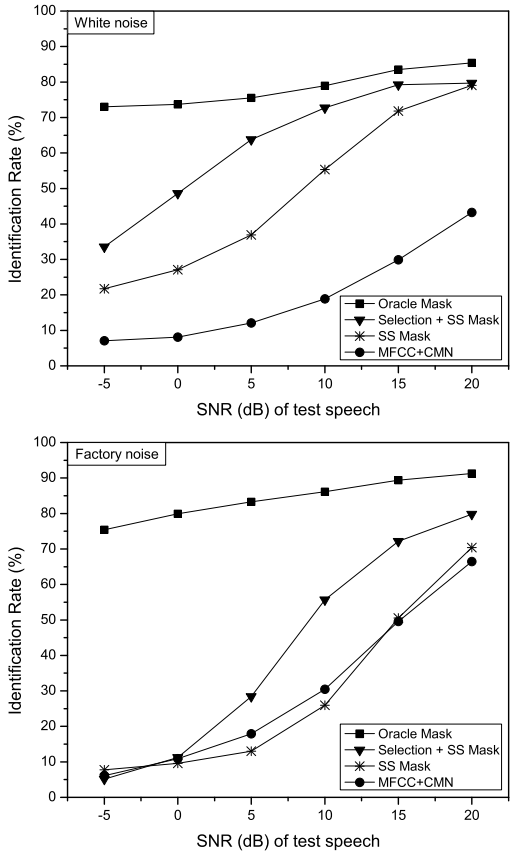

Fig. 3. Speaker identification performance in white noise (top) and factory noise (bottom). Results are shown for bottom-up only systems using a priori and spectral subtraction masking (Oracle Mask, SS Mask), the combined system (Selection + SS Mask) and the cepstral baseline (MFCC+CMN).

quired by the feature selection process. Selection was performed with global SNR estimates corresponding to the true SNR of the noisy speech, and an attenuation factor of $\gamma = 6$ which was determined empirically using a smaller validation speaker set.

3.1. Results and Discussion

To determine a suitable reduced feature set size experiments were performed using oracle missing data reliability masking. Full results are omitted for brevity, but with a feature subset size of $\kappa \geq 18$ the performance of the combined system was within 1% of the bottom-up system for white noise, and within 5% for factory noise. The decrease in performance for factory noise is possibly due to the white noise based multi-condition training used to derive the feature selectivities.

The combined system with 18-best feature selection was then evaluated for spectral subtraction masking (as in (5)). The results show significant performance improvement when the combined missing data feature selection system is used over a single bottom-up stage (see Fig. 3). Under stationary white noise the spectral subtraction technique is able to provide a good estimation of the binary TF reliability mask, and so its use alone produces recognition rates far exceeding those of the MFCC-CMN baseline. As the SNR decreases the estimated TF mask becomes more sparse which causes unreliable inclusion errors to have a larger negative effect on recognition. By applying the top-down selection mask a large number of these unreliable inclusion errors are removed, and this results in the observed increase in performance (up to 27% absolute) compared to the use of only the bottom-up mask.

For the non-stationary factory noise spectral subtraction is unable to provide an accurate mask estimation, and the standard bottom-up system performs similarly to the MFCC-CMN baseline. However, the use of the feature selection subsystem still results in large performance improvements (up to 25% absolute) for SNRs above 0 dB. At lower SNRs the severity of the distortion results in few TF points which are dominated by speech. These points are extremely unlikely to be identified by the spectral subtraction algorithm, and so refinement of the reliability mask by the selection mask can do little to help in this case.

The results demonstrate the validity of using top-down feature selection to refine imperfect missing data masks for speaker recognition. However, this is a preliminary study and as such has several limitations. These include the assumption of perfect estimation of the global SNR and the use of relatively simple discriminant analysis techniques to determine the feature subsets.

4. CONCLUSIONS

We have proposed the combination of missing data and feature selection for robust speaker identification. The formation of a subset of discriminative filter-bank features allows the refinement of binary TF reliability masks by the removal of unreliable inclusion errors for non-discriminative bands. Experimental evaluation illustrated that the combined approach can approximate the performance of traditional bottom-up missing feature methods when a priori noise knowledge is available. However when the reliability mask is imperfectly estimated the combined system significantly outperforms the traditional bottom-up only approach. In future work we aim to improve the feature selection method such that the subset is chosen based on recognizer feedback, to test the system under a wider variety of noise conditions with more realistic speech data, and to extend the system to use soft masking decisions.

5. REFERENCES

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

[2] M. Cooke, A. Morris, and P. Green, "Missing data techniques for robust speech recognition," in *Proc. ICASSP*, 1997, vol. 2, pp. 863–866.

[3] J. Barker, M. Cooke, and D. Ellis, "Integrating bottom-up and top-down constraints to achieve robust asr: The multisource decoder," in *CRAC Workshop*, 2001, pp. 63–66.

[4] S. Srinivasan and D. Wang, "Robust speech recognition by integrating speech separation and hypothesis testing," in *Proc. ICASSP*, 2005, vol. 1, pp. 89–92.

[5] 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.

[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. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Signal Process.*, vol. 23, no. 2, pp. 176–182, 1975.

[8] J. Wolf, "Efficient acoustic parameters for speaker recognition," *JASA*, vol. 51, no. 6B, pp. 2044–2056, 1972.

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