Spectral Subband Centroids for Robust Speaker Identification using Marginalization-based Missing Feature Theory

Aaron Nicolson, Jack Hanson, James Lyons, Kuldip Paliwal
Signal Processing Laboratory
Griffith University, Brisbane, Australia
Email: {aaron.nicolson, jack.hanson, james.lyons}@griffithuni.edu.au, k.paliwal@griffith.edu.au

***Abstract*—Until now, marginalization-based Missing Feature Theory (MFT) for speech classification has been limited to the use of Log Spectral Subband Energies (LSSEs) as features. These features are highly correlated, thus suboptimal for classification with diagonal-covariance Gaussian Mixture Models (GMMs), a common classifier in marginalization-based MFT. In this paper, we propose that Spectral Subband Centroids (SSCs) are more apt for marginalization-based MFT, as they are both decorrelated and spectrally local. Our results show that SSCs as features produce a more robust marginalization-based MFT, diagonal-covariance GMM-based, Automatic Speaker Identification (ASI) system than LSSEs as features, for at all tested SNR values (with Additive White Gaussian Noise (AWGN)). It is also shown that a fully-connected Deep Neural Network (DNN) can accurately estimate the Ideal Binary Mask (IBM) used for MFT.**

***Index Terms*—spectral subband centroids; missing feature theory; speaker identification; deep neural network; ideal binary mask**

I. Introduction

The aim of an Automatic Speaker Identification (ASI) system is to determine a person's identity from a database of known speakers, given a recording of their speech. There are many methods currently used in the literature for implementing a high-performance ASI system, however a common theme amongst many of these approaches is the degradation of their accuracy with the introduction of noise. Cooke *et al* [1], proposed the method of Missing Feature Theory (MFT) to reduce the impact of additive noise on ASI performance.

MFT is based on the knowledge that when information is missing from a speech signal, it is still comprehendible to the listener. The fact that humans can amply comprehend speech that has had fractions of its information removed indicates a significant level of inherent redundancy in the signal's spectro-temporal information. MFT approaches aim to first find the degraded (or unreliable) components of the speech signal's spectrogram, and then to either ignore the unreliable components, or estimate their optimal value.

The two most popular methods of MFT, marginalization and cluster-based reconstruction, deal with the unreliable components in different manners. Marginalization-based approaches integrate over the unreliable components of the spectrogram, thereby effectively using only the reliable components for classification. An immediate drawback of this method is that the features used for classification must intrinsically be spectrally local, meaning that they cannot incorporate information from a range of frequencies. This makes cepstral and other spectrally non-local features incompatible with this approach. Another drawback that stems from this approach is that the classifier itself must be modified to exclude the unreliable components.

The aim of cluster-based reconstruction is to estimate the values of the unreliable components before classification. These values are calculated based on the surrounding reliable components of the spectrogram, using *a priori* knowledge of speech. Since there are no wholly-excluded frequency components, spectrally non-local features such as Mel Frequency Cepstral Coefficients (MFCCs) can be used as features for classification. However, cluster-based reconstruction has been shown to perform worse than marginalization-based methods, due to noise still being incorporated into the classification process [2].

The dominant feature type used in diagonal-covariance Gaussian Mixture Model (GMM)-based ASI systems has been MFCCs [3] [4]. However, as MFCCs are spectrally non-local features, they cannot be used with marginalization-based MFT. This means that MFT-based ASI systems have been restricted to Log Spectral Subband Energies (LSSEs), a feature known to provide inferior system accuracy to MFCCs. The inherent disadvantage of LSSEs is that the features themselves are highly correlated, whereas MFCCs are decorrelated. Diagonal-covariance GMMs are not good at modelling correlated features, leading to inferior LSSE performance.

In this paper, we present for use with marginalization-based MFT a substitute feature for LSSEs, namely Spectral Subband Centroids (SSCs) [5]. SSCs possess several characteristics which make them promising features for MFT. Specifically, they are decorrelated, robust, and spectrally local features. SSCs have also been proven successful as features in ASI systems [6] [7].

Manuscript received June 27, 2012; accepted July 25, 2017.

The reliable components of a spectrogram are identified using a binary spectrographic mask. The Ideal Binary Mask (IBM) can be calculated when both the clean and noisy spectrograms of a signal are given. To demonstrate the practical validity of a MFT-based ASI system using SSCs, estimators are used to approximate the IBM from a given noisy spectrogram.

Section 2 describes the spectrogram used to compute the IBM. Section 3 describes the Spectral Subband Features (SSF) that were used for the comparison, namely LSSEs, MFCCs, and SSCs. Section 4 details the marginalization-based MFT method for diagonal-covariance GMMs. Section 5 presents the two IBM estimators used, namely MMSE STSA w. SPU, and a fully-connected Deep Neural Network (DNN). Sections 6, 7, and 8 are dedicated to the experiment setup, the results and discussion, and the conclusion respectively.

II. Spectrogram and Ideal Binary Mask

A. Spectrogram

The first stage of computing the spectrogram of a speech waveform is short-time analysis. Overlapping analysis frames are typically between 20-40 ms to achieve signal stationarity, with a tapered window applied to mitigate spectral leakage. The Power Spectral Density (PSD) of the i^{th} frame $x_i(n)$ is then estimated from the Discrete Fourier Transform (DFT) using the periodogram method:

$$\hat{P}_i(k) = \frac{1}{N}\left|\sum_{n=0}^{K-1} x_i(n)w(n)e^{-j2\pi kn/K}\right|^2, 0 \le k \le K-1, \quad (1)$$

where K is the DFT length, N the frame length, and $w(n)$ is a windowing function.

Next, the Spectral Subband Energy (SSE) coefficients for the frame are computed from the PSD by using a bank of B triangular-shaped critical band filters spaced uniformly on the mel scale (shown in Fig. 1) as follows:

$$X_i(b) = \sum_k h_b(k)\hat{P}_i(k), \quad 0 \le b \le B-1, \quad (2)$$

where h_b refers to the b^{th} filterbank. The SSEs form the final spectro-temporal representation of the signal.

B. Ideal Binary Mask

When a signal is corrupted by uncorrelated additive noise, the noisy spectrogram $\boldsymbol{X}$ can be modeled as the sum of the clean spectrogram $\boldsymbol{S}$ and the noise spectrogram $\mathbf{N}$:

$$X_i(b) = S_i(b) + N_i(b). \quad (3)$$

In MFT methods, it is assumed that noisy spectrogram components with a SNR at or above a set threshold θ are reliable estimates of the corresponding clean spectrogram components. The SNR of the b^{th} filterbank of the i^{th} frame is calculated as follows:

$$\mathrm{SNR}_i(b) = 10 * \log_{10}\left(\frac{S_i(b)}{X_i(b) - S_i(b)}\right). \quad (4)$$

Conversely, corrupted spectrogram components with a SNR below θ (set to zero in this work) are assumed to be unreliable estimates. An Ideal Binary Mask (IBM) I identifies the reliable and unreliable components of a noisy spectrogram by comparing the SNR from Eq. 4, to the threshold θ:

$$I_i(b) = \begin{cases} 1 & \text{if } \mathrm{SNR}_i(b) \ge \theta \\ 0 & \text{if } \mathrm{SNR}_i(b) < \theta. \end{cases} \quad (5)$$

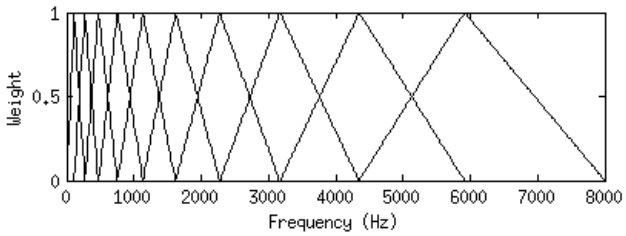

Figure 1. Mel filterbank with 10 triangular-shaped filters.

III. Spectral Subband Features

A. Log Spectral Subband Energies

Previous work on marginalization-based MFT has relied on LSSEs as features for classification. LSSEs are local in both time and frequency, enabling an IBM to identify the reliable components. SSEs from Eq. 1 are scaled by the natural logarithm to form LSSEs:

$$\mathrm{LSSE}_i(b) = \log \sum_k h_b(k)\hat{P}_i(k), \quad 0 \le b \le B-1. \quad (6)$$

However, LSSEs provide inferior diagonal-covariance GMM-based ASI accuracy when compared to MFCCs. The suboptimal classification accuracy is due to the features' high correlation, as illustrated in Fig. 2.

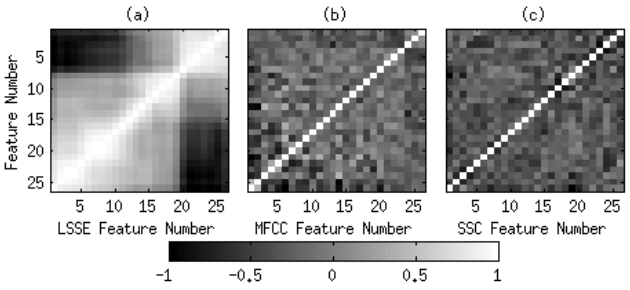

Figure 2. The cross-correlation of features for a) LSSE features, b) MFCC features, and c) SSC features. 26 filterbanks were used.

B. Mel Frequency Cepstral Coefficients

MFCCs decorrelate the LSSEs (shown in Fig. 2) by taking the Discrete Cosine Transform (DCT) over the B filterbanks:

$$\mathrm{MFCC}_i(c) = \sum_{b=0}^{B-1} \mathrm{LSSE}_i(b)\cos\left[\frac{\pi c(b-0.5)}{B}\right], \quad (7)$$

where $0 \le c \le B-1$. However, the DCT *'smears'* the LSSEs throughout all of the MFCCs. This means that

MFCCs cannot be used in marginalization-based MFT as they are not spectrally local.

C. Spectral Subband Centroids

SSCs indicate at what frequency in the filterbank the *'center of mass'* is located. SSCs for a frame are calculated by taking the weighted average of the frequencies present in the subband. The weights are determined by the product of the k^{th} filterbank coefficient, and the k^{th} PSD coefficient:

$$\mathrm{SSC}_i(b) = \frac{\sum_k k h_b(k) \hat{P}_i(k)^\gamma}{\sum_k h_b(k) \hat{P}_i(k)^\gamma}. \quad (8)$$

In this work, a value of $\gamma = 1$ was used. Like LSSEs, SSCs are spectrally local, enabling SSCs to be used as features for marginalization-based MFT classification. SSCs are not only spectrally local, they are also uncorrelated (shown in Fig. 2), like MFCCs. The uncorrelated property of SSCs enables them to be an ideal feature for a diagonal covariance GMM employing marginalization-based MFT.

IV. Marginalization-Based MFT

A common classifier in ASI systems is to model each speaker using a diagonal-covariance Gaussian Mixture Model (GMM). Each speaker model s has a set of M mixtures, with mixture m having mean vector $\boldsymbol{\mu}$, diagonal-covariance matrix $\boldsymbol{\Sigma}$, and $a\ priori$ probability $P(m|s)$. The distribution of random variable $\boldsymbol{x}$ for the m^{th} mixture of speaker model s is $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}|m, s)$. The likelihood that $\boldsymbol{x}$ was from speaker $\boldsymbol{s}$ is given by:

$$f(\mathbf{x}|s) = \sum_{m=1}^{M} P(m|s) f(\mathbf{x}|m, s), \quad (9)$$

where $\boldsymbol{x}$ in this work was represented by either LSSEs, MFCCs, or SSCs.

Using the IBM from Eq. 5, the reliable components of $\boldsymbol{x}$ can be identified. The marginal distribution of $\boldsymbol{x}$ is thus taken over the reliable components $\boldsymbol{x}_r$. $\boldsymbol{x}_r \sim N(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r|m, s)$ gives the marginal distribution for the m^{th} mixture of speaker model s. With only the reliable components $\boldsymbol{x}_r$ being used, the likelihood from Eq. 9 becomes:

$$f(\mathbf{x}_r|s) = \sum_{m=1}^{M} P(m|s) f(\mathbf{x}_r|m, s). \quad (10)$$

V. Ideal Binary Mask Estimates

The Ideal Binary Mask (IBM) is computed from the clean $\boldsymbol{S}$ and noisy $\boldsymbol{X}$ spectrogram, by using Eq. 4 and Eq. 5. However, the clean spectrogram $\boldsymbol{S}$ is not available in practical applications. To estimate the IBM, either the clean spectrogram is estimated, or the IBM is estimated directly.

A. Speech Enhancement

The speech enhancement algorithm used to estimate the clean spectrogram $\widehat{\boldsymbol{S}}$ was the MMSE STSA estimator by Ephraim *et al.* [8]. Speech Presence Uncertainty (SPU) was used in combination with the MMSE STSA estimator. The MATLAB implementation of the algorithm was by Loizou *et al.* [9].

B. Ideal Binary Mask Estimator

To estimate the IBM directly, a Deep Neural Network (DNN) with fully-connected layers was used. Given the corrupted SSE spectrum of a frame, the DNN was tasked with estimating its IBM. The network had 3 hidden layers, with 256 nodes per layer. The hidden layers employed a rectifier activation function. A sigmoid activation function was applied to the output layer. As the state of the output layer neurons was within the interval $[0,1]$, the IBM estimate was found using:

$$\hat{I}_i(b) = \begin{cases} 1 & \text{if } \mathrm{y}_i(b) \geq 0.5 \\ 0 & \text{if } \mathrm{y}_i(b) < 0.5. \end{cases} \quad (11)$$

where $y_i(b)$ was the state of the b^{th} output layer neuron for the i^{th} frame.

The hidden layers of the DNN were initialized using a stacked autoencoder [10], with the $Adam$ algorithm [11] used for gradient descent optimization. Cross entropy was the loss function employed. Each autoencoder layer was trained for 10 epochs, with the complete network fine-tuned for 50 epochs. Early stopping was employed during fine-tuning, with a validation set used to determine error.

VI. Experiment

A. Database

The speech corpus used was the TIMIT [12] database. The TIMIT database has 630 speakers, each with 8 unique and 2 common short sentences recorded at 16kHz. The 8 unique sentences of each speaker were used to train the GMM speaker models.

The 2 common short sentences of each speaker were used for testing. Each test sentence was degraded to a range of SNR levels, by Additive White Gaussian Noise (AWGN). The SNR values were: 30dB, 20dB, 10dB, 0dB, and -10dB. The GMM speaker models, the speech enhancement method, and the DNN, were all tested on the noisy test utterances.

For the DNN training examples, the 8 unique short sentences for each speaker were used, which were corrupted by AWGN at the same SNR values as above. The IBM for each training example was used as the target. 10% of the training data was used as the validation set.

B. Test Details

All speech was framed at 30 ms per frame, with a 10 ms shift. 26 filter banks were used to compute the spectral subband features. The first test involved a comparison of the spectral subband features (LSSEs, MFCCs, and SSCs) for speaker identification. Diagonal-covariance GMMs with 32 mixtures were used as speaker models. LSSEs and

SSCs were also tested using marginalization-based MFT with an IBM.

The second test compared the IBM to the IBM estimates presented in Section V. Speaker models using SSCs, marginalization-based MFT, and diagonal-covariance GMMs with 32 mixtures were used for the comparison.

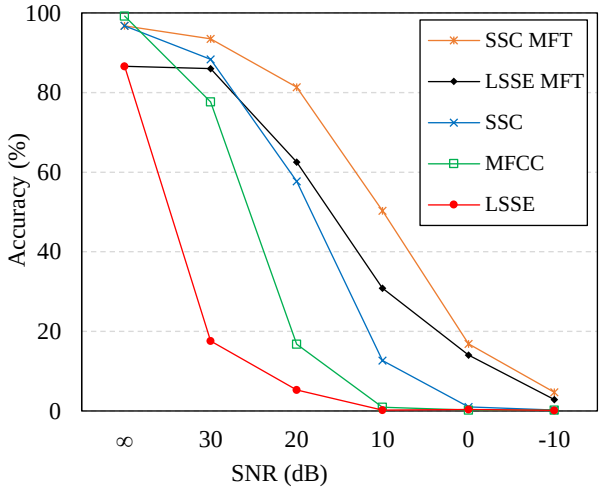

Figure 3. Comparison of a diagonal-covariance GMM-based ASI system (using MFCCs, LSSEs, and SSCs) to a marginalization-based MFT version, using an IBM (for SSCs and LSSEs).

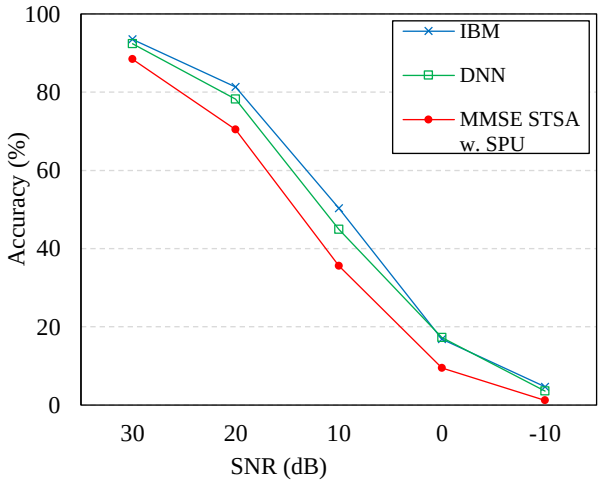

Figure 4. IBM estimate results for a diagonal-covariance GMM-based ASI system using marginalization-based MFT and SSCs as features.

VII. Results And Discussion

Comparing the ASI results from Fig. 3 without marginalization-based MFT shows that the uncorrelated features (MFCCs, and SSCs) achieved a greater accuracy than the correlated features (LSSEs), due to the diagonal-covariance GMM speaker models. The robustness of SSCs to noise was also demonstrated, having a higher accuracy then MFCCs at high SNR values (30 dB, 20 dB, and 10 dB). Applying marginalization-based MFT to a robust feature like SSCs should therefore significantly outperform MFCCs at high SNR values.

Fig. 3 illustrates that in the presence of noise, LSSEs and SSCs as features benefited from the introduction of marginalization-based MFT. Marginalization-based ASI results for LSSEs were better at lower SNR values than ASI results for MFCCs (10 dB, 0 dB, and -10 dB), showing the robustness of the marginalization-based MFT method. However, the best ASI results at all tested SNR values came from SSCs and marginalization-based MFT. Three properties of SSCs allowed them to form a highly robust diagonal-covariance GMM-based MFT ASI system: SSCs are robust, uncorrelated, and local in both time and frequency.

An accurate IBM estimator is needed for marginalization-based MFT to work in a practical sense. Shown in Fig. 4 are the IBM estimate results for marginalization-based MFT and SSCs. MMSE STSA w. SPU, which estimated the clean spectrum for Eq. 4, provided an inaccurate IBM estimate at all tested SNR values. The DNN, which estimated the IBM directly, provided a more accurate estimate of the IBM when compared to MMSE STSA w. SPU. This shows the practical validity of marginalization-based MFT and SSCs for robust ASI when a DNN is used to estimate the IBM.

VIII. Conclusion

In this paper, Spectral Subband Centroids (SSCs) are presented for a marginalization-based MFT, diagonal-covariance GMM-based Automatic Speaker Identification (ASI) system. The current features used in marginalization-based MFT are Log Spectral Subband Energies (LSSEs), however these features perform sub optimally in diagonal-covariance GMMs. SSCs are spectrally local, decorrelated features, and when combined with marginalization-based MFT, are more robust to noise (AWGN) than MFCCs (as well as LSSEs with marginalization-based MFT) at all tested SNR values. It is also shown that a fully-connected Deep Neural Network (DNN) can accurately estimate the Ideal Binary Mask (IBM) used for MFT.

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] B. Raj, S. L. Michael and S. M. Richard, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275-296, 2004.

[3] D. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995.

[4] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980.

[5] K. K. Paliwal, "Spectral subband centroid features for speech recognition," *Acoustics, Speech and Signal Processing*, vol. 2, pp. 617-620, 1998.

[6] T. Kinnunen, B. Zhang, J. Zhu and Y. Wang, "Speaker verification with adaptive spectral subband centroids," *Advances in Biometrics*, pp. 58-66, 2007.

[7] N. P. H. Thian, C. Sanderson and S. Bengio, "Spectral subband centroids as complementary features for speaker authentication," *Biometric Authentication*, pp. 631-639, 2004.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, 1984.

[9] P. C. Loizou, "Implementation and Evaluation of the MMSE Estimator," ch. 7, pp. 237-265, in *Speech enhancement: theory and practice*, CRC Press, 2007.

[10] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, pp. 153-160, 2007.

[11] D. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[12] J. S. Garofolo, "TIMIT acoustic-phonetic continous speech corpus LDC93S1," *Philadelphia: Linguistic Data Consortium*, 1993.

Aaron Nicolson received his BEng degree with Class 1A Honors from Griffith University in 2016. He is studying for a PhD degree at the Signal Processing Laboratory at Griffith University. His research includes machine learning, speech recognition, speaker recognition, and speech enhancement.

Jack Hanson received his BEng (Advanced Studies) degree with Class 1A Honors from Griffith University in 2014. He is studying for a PhD degree at the Signal Processing Laboratory at Griffith University, where he began by learning speech and speaker recognition techniques. Currently, he is working on machine learning applications in structural bioinformatics, and has worked on several publications in this field.

James Lyons received a BEng degree with Honors and a BIT from Griffith University Brisbane, Australia in 2007. He completed a PhD degree in Bioinformatics at Griffith University Brisbane, Australia in 2016. His research interests include Automatic Speech and Speaker recognition, Bioinformatics, protein fold and structural class prediction problems and pattern recognition.

Kuldip K. Paliwal was born in Aligarh, India, in 1952. He received the BS degree from Agra University, Agra, India, in 1969, the MS degree from Aligarh Muslim University, Aligarh, India, in 1971, and the PhD degree from Bombay University, Bombay, India, in 1978. He is a professor in the School of Engineering at Griffith University, Brisbane, Australia. His current research interests include speech coding, speech and speaker recognition, speech enhancement, face recognition, image coding, pattern recognition and machine learning.