INCORPORATING AUDITORY FEATURE UNCERTAINTIES IN ROBUST SPEAKER IDENTIFICATION

Yang Shao[1], *Soundararajan Srinivasan*[2,*] *and DeLiang Wang*[1,3]

[1]Department of Computer Science and Engineering
[2]Biomedical Engineering Department
[3]Center for Cognitive Science
The Ohio State University
Columbus, OH 43210-1277, USA
{shaoy, srinivso, dwang}@cse.ohio-state.edu

ABSTRACT

Conventional speaker recognition systems perform poorly under noisy conditions. Recent research suggests that binary time-frequency (T-F) masks be a promising front-end for robust speaker recognition. In this paper, we propose novel auditory features based on an auditory periphery model, and show that these features capture significant speaker characteristics. Additionally, we estimate uncertainties of the auditory features based on binary T-F masks, and calculate speaker likelihood scores using uncertainty decoding. Our approach achieves substantial performance improvement in a speaker identification task compared with a state-of-the-art robust front-end in a wide range of signal-to-noise conditions.

Index Terms— robust speaker identification, auditory features, uncertainty decoding

1. INTRODUCTION

A speaker recognition system, performing either speaker identification (SID) or speaker verification (SV), typically comprises three processes: feature extraction, pattern classification using speaker modeling, and decision making [1, 7]. Typically, the extracted speaker features are short-term cepstral coefficients such as Mel-frequency cepstral coefficients (MFCC) and perceptual linear predictive (PLP) coefficients, or long-term features such as prosody [16]. For speaker modeling, Gaussian mixture models (GMM) are widely used [15] to model the feature distributions. Such systems usually do not perform well under noisy conditions [6, 18, 21] because the extracted features are distorted by noise, causing mismatched likelihood calculation.

To tackle this noise robustness problem, spectral subtraction has been widely used because of its simplicity [6, 11], but its effectiveness degrades sharply when noise is nonstationary [18]. RASTA filtering [8] and cepstral mean normalization (CMN) have also been widely used but they are mainly designed for convolutive noise. Rose *et al*. [17] use parallel model combination when noise statistics is known *a priori*, which poses restrictions on its applications. On the other hand, recent studies of robust speech recognition on Aurora [12] have yielded an advanced front-end feature extraction algorithm (AFE) [20], standardized by the European Telecommunication Standards Institute (ETSI).

Recently, we have employed a missing data method for robust SID and SV tasks [18]. The basic idea is to decompose the input signal in time-frequency (T-F) and treat the noise-dominant T-F units as missing during recognition. This process requires a binary mask to indicate whether a particular T-F unit is reliable or missing. The binary mask is generated by a computational auditory scene analysis (CASA) system [9]. Our evaluations demonstrate that using binary masks with a T-F representation offers a superior alternative method under nonstationary noise conditions.

In this paper, we first propose two novel speaker features based on an auditory periphery model [13]. Specifically, a Gammatone feature (GF) is obtained from a bank of Gammatone filters, which was originally proposed to model human cochlear filtering. Then, Gammatone frequency cepstral coefficients (GFCC) are derived from GF. We find that such features achieve comparable SID performance to ETSI-AFE features under both clean and noisy conditions. To account for the deviations of noisy features from clean ones, we reconstruct the auditory features from a speech prior based on an estimated binary mask. This missing data method has been employed for robust speech recognition [14, 19]. Additionally, feature uncertainties estimated from reconstruction are utilized by an uncertainty decoder [5] to enhance likelihood calculation in a speaker identification task. Our system achieves substantial improvement over ETSI-AFE features in a wide range of signal-to-noise (SNR) conditions.

The rest of the paper is organized as follows. Section 2 describes the overall system including the proposed auditory feature extraction and uncertainty estimation. SID evaluations are presented in Section 3. Section 4 concludes the paper.

2. SYSTEM DESCRIPTION

Conceptually, our proposed system improves noise robustness in two components of a speaker identification system; novel robust auditory features in the feature extraction component, and feature uncertainty estimation and decoding in the scoring component.

2.1. System overview

Figure 1 presents a diagram of the overall system. Input speech is decomposed into a T-F representation using an auditory filterbank. Specifically we use a Gammatone filterbank [3] to generate a time

*At Research and Technology Center, Robert Bosch LLC, USA.

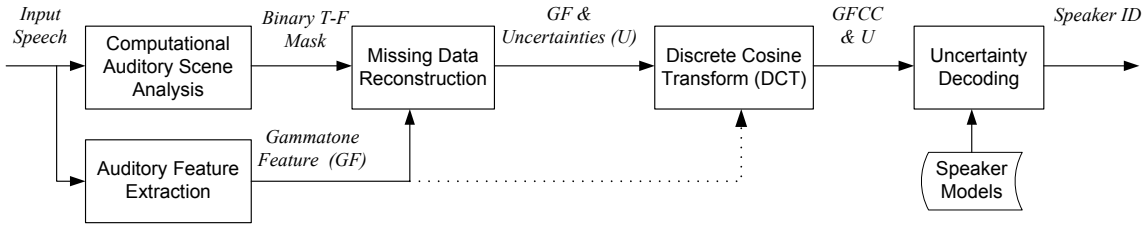

Figure 1. Schematic diagram of the proposed system. Input speech is passed through a computational auditory scene analysis system to produce a binary time-frequency (T-F) mask. Then, extracted Gammatone features (GF) are used in conjunction with the binary mask to reconstruct missing T-F units from a speech prior. GF uncertainties are also estimated in the reconstruction process. GFs and their uncertainties are then transformed into "cepstrum" by the discrete cosine transform (DCT). Finally, uncertainty decoding searches for the best-matched speaker model given the resulting Gammatone frequency cepstral coefficients (GFCC) and uncertainties. The dotted path denotes how GFCCs are extracted from clean speech for the purpose of speaker model training.

sequence of GFs, which are analogous to the discrete Fourier transform (DFT) based spectral coefficients. We also pass the input signal through a CASA system [9], creating a binary T-F mask. An element of this mask indicates whether the corresponding GF component is reliable or corrupted within a time frame.

The corrupted GF feature components are then reconstructed using a speech prior [14], which is derived from a pooled training set. We also estimate uncertainties associated with the reconstructed GF features.

Similar to the cepstral analysis in MFCC extraction, the GF feature is transformed into "cepstrum" by a discrete cosine transform (DCT) [10]. DCT de-correlates the feature components and compacts feature dimensions [10]. The estimated uncertainties can also be transformed into the cepstral domain because of DCT's linearity property. Finally, an uncertainty decoder [5] performs speaker identification using the derived GFCC and the transformed uncertainty estimates.

2.2. Auditory feature extraction

The mixture signal is first analyzed using a 128-channel Gammatone filterbank [3]. Its center frequencies are quasi-logarithmically spaced from 50 Hz to 8 KHz, which models human cochlear filtering [13]. The filterbank outputs are then down-sampled to 100 Hz in the time dimension, corresponding to a frame rate of 10 ms, which is used in many short-term speech feature extraction algorithms. The magnitudes of the down-sampled filterbank outputs are then loudness-compressed using cubic root operation. The resulting GF feature vectors, $G_f(t)$ at time t with component index of frequency f, comprise the T-F representation of the auditory response. This response matrix is called the cochleagram, which is analogous to the spectrogram. Figure 2 presents illustrations of the cochleagram and the spectrogram of a clean speech utterance. Evidently, similar to Mel-scale processing in MFCC extraction, cochleagram provides a much higher frequency resolution at low frequencies than at high frequencies.

A GF feature vector contains 128 components, which is much more than the number of dimensions a typical speaker or speech recognition system uses. Also, because of frequency overlap among neighboring Gammatone filters, individual components are correlated with each other. Hence, we apply DCT on a GF feature vector G to reduce the dimensionality and de-correlate feature components.

$$d(i,j)=\sqrt{\frac{2}{N}}\cos\left(\frac{\pi}{2N}\cdot i\cdot(2j-1)\right)\quad i,j=1...N \tag{1}$$

The elements $d(i,j)$ of a DCT matrix D, are defined in (1). N is the number of dimensions; N =128 in this paper.

$$C=D\times G \qquad D=\{d(i,j)\,|\,i,j=1...N\} \tag{2}$$

Thus, a vector of "cepstral" coefficients C, is obtained by multiplying the DCT matrix D with a GF vector G.

Rigorously speaking, the newly derived features are not cepstral coefficients, since cepstral analysis requires a log operation between the first and the second frequency analysis for the deconvolution purpose. Here we call this feature as Gammatone frequency cepstral coefficients (GFCC) because of the functional similarities between the transformation above and that of cepstral analysis. We use the lowest 23-order GFCCs since they retain the majority information of a GF feature frame, a result of the "energy compaction" property of DCT [10].

2.3. Computational auditory scene analysis

We adapt and apply [18] a pitch-based speech segregation system [9] that performs CASA analysis. This system makes minimal assumptions about the underlying noise and has been shown to significantly improve the SNR of segregated speech under various noisy conditions. This system produces a binary T-F mask as well as estimated pitch tracks. Specifically, it performs voiced speech segregation on a T-F representation derived from Gammatone filterbank filtering and hair-cell transduction. In the low-frequency range, the system generates homogeneous T-F regions based on temporal continuity and cross-channel correlation, and groups them based on periodicity similarity. In the high-frequency range, the envelope of a filter response fluctuates at the pitch rate and amplitude modulation (AM) rates are used for grouping. As a result, it labels speech-dominated T-F units as reliable (1) in the binary mask and noise-dominated units as unreliable (0).

2.4. GF reconstruction and uncertainty estimation

In a typical speaker identification or verification system, the probability distribution of an extracted feature vector, X, produced by a speaker λ, is modeled as a GMM, typically parameterized by diagonal covariance matrices [15]. Under noisy conditions, the aforementioned CASA system produces a binary T-F mask that indicates whether a GF feature component is reliable or corrupted (missing). Thus, the feature vector can be partitioned into reliable components X_r, or unreliable ones X_u. Our previous study [18] employs a missing data method that marginalizes the unreliable

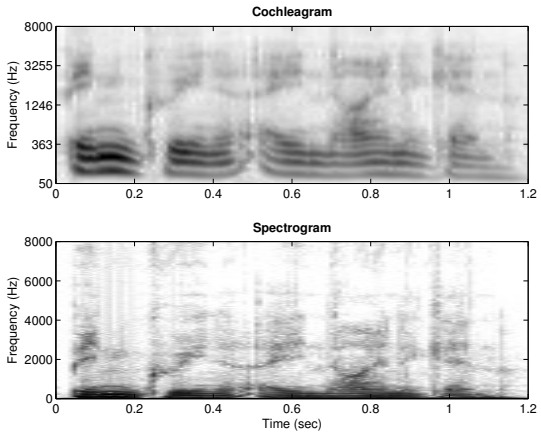

Figure 2. Illustrations of the cochleagram and the spectrogram of a clean speech utterance. Note the asymmetric frequency resolution at low and high frequencies in the cochleagram.

dimensions of the feature distribution to improve robust speaker recognition performance.

Here, we propose using the auditory cepstral feature, GFCC, in conjunction with the binary mask. In order to apply the DCT transform on the corrupted GF, we first reconstruct the missing GF components from a prior speech model, which is similar to the universal background model (UBM) in a typical speaker verification system. Specifically, the speech prior $p(X)$ is modeled as a GMM, and constructed from pooled training data:

$$p(X) = \sum_{k=1}^{M} p(k)p(X \mid k), \tag{3}$$

where M is the number of mixtures, k is the mixture index, and $p(k)$ gives the prior of a mixture, or in other words the mixture weight. $p(X|k)$ is the kth Gaussian distribution with a mean vector μ^k and a diagonal covariance σ^k. Given a binary mask, the components of the mean and variance of each Gaussian can be split into reliable and unreliable ones. We then calculate the *a posteriori* probability of the kth mixture given reliable GF components as in

$$p(k \mid X_r) = \frac{p(k)p(X_r \mid k)}{\sum_{k=1}^{M} p(k)p(X_r \mid k)}. \tag{4}$$

As shown in [4, 19], the unreliable components are estimated as the expected value or the mean conditioned on $X_\mathbf{r}$.

$$\hat{X}_u = \sum_{k=1}^{M} p(k \mid X_r)\mu_u^k \tag{5}$$

μ_u^k refers to the unreliable components of the mean vector of the kth mixture of the speech prior. The reliable components are retained in the reconstruction.

Although (5) gives a good estimate of the unreliable GF components, errors in reconstruction will cause degradation of recognition performance. Thus, estimates of the reconstruction uncertainties would mitigate such degradations by accounting for the reconstruction errors in the speaker likelihood calculation. Specifically, the uncertainties are estimated as,

$$\hat{\sigma} = \sum_{k=1}^{M} p(k \mid X_r)\left\{\left(\begin{bmatrix} X_r \\ \hat{X}_u \end{bmatrix} - \mu^k\right)^2 + \begin{bmatrix} 0 \\ \sigma_u^k \end{bmatrix}\right\} \tag{6}$$

σ_u^k refers to the unreliable components of the diagonal covariance matrix of the kth mixture.

A reconstructed GF feature and its associated uncertainty are then transformed into the GFCC domain using DCT. During the identification process, an uncertainty decoder [5] calculates the likelihood of the reconstructed GFCC given a clean speaker model and the estimated uncertainty. Specifically, the uncertainty is added to the covariance of each mixture of the speaker model.

3. EVALUATION

We evaluate the noise robustness of our proposed auditory features and the uncertainty estimation method in a SID task. The standard MFCC features are used to obtain the baseline performance. We also compare the performance of our proposals with the state-of-the-art robust front-end ETSI-AFE [20].

3.1. Evaluation setup

We use the speech materials from the recent speech separation challenge (SSC) [2]. The training data is drawn from a closed set of 34 talkers, 18 males and 16 female, and consists of 17,000 utterances. We use the speech-shaped noise (SSN) portion of the test set for our SID evaluation. The SSN data was generated by mixing clean utterances with speech-shaped noise at 4 SNRs: −12, −6, 0 and 6 dB. The test set contains 600 utterances in each SNR condition.

The speakers are modeled as 64-mixture GMMs and trained on the training portion of SSC directly. The speech prior model comprises 2048 Gaussian mixtures, and is constructed from the pooled training utterances of all speakers. SID scores are only calculated on the voiced speech frames.

3.2. Evaluation results

Figure 3 presents the SID evaluation results. 'MFCC_D_Z' denotes the baseline SID performance obtained using 24 MFCC features including deltas and after cepstral mean normalization. They are extracted using the HTK toolkit [22]. 'ETSI-AFE' represents the enhanced 24 MFCC features, deltas included, derived from the advanced front-end feature extraction algorithm, which is standardized by the European Telecommunication Standards Institute [20]. 'ETSI-AFE_Z' denotes the cepstral mean normalized 'ETSI-AFE' feature.

'GF' and 'GFCC_C0' are the auditory features described in Section 2.2 with 128 and 23 dimensions respectively. 'GFCC' is the GFCC feature but with the first cepstral coefficient C_0, removed. 'GF_MD' stands for the missing data recognition method using the GF features and estimated binary T-F masks [18].

'GFCC_C0_U' denotes SID performance by the uncertainty decoder using GF reconstruction and estimated uncertainties in the GFCC feature domain as described in Section 2.4. 'GFCC_U' shows the same feature configuration but without C_0. 'GF_U' shows the SID performance when the uncertainty decoder is directly applied in the GF domain, before the DCT transform.

It is observed from Figure 3 that the proposed GF feature performs significantly better than the baseline MFCC feature at low SNR conditions. More importantly, the GFCC features, especially the GFCC without C_0, not only achieve substantial improvement over the baseline feature, but also obtain comparable identification results with the robust features extracted by ETSI-

AFE. Since C_0 relates to the overall energy of a feature frame, it is very susceptible to noise degradation. Thus, removing C_0 is beneficial at low SNR conditions. Note that C_0 has been removed from MFCC and ETSI-AFE features.

The missing data method using marginalization performs significantly better than ETSI-AFE. GF reconstruction and uncertainty decoding in the GF domain further improve SID accuracies. Substantial improvement over ETSI-AFE is obtained after the GF feature and the uncertainty are transformed into the GFCC domain. In summary, GFCC features provide a substantial contribution to the noise robustness of the system.

4. CONCLUSION

In this paper, we have proposed a general solution to robust speaker recognition under additive noise conditions. Novel speaker features are derived from auditory filtering and cepstral analysis. Additionally, by using binary T-F masks generated from a CASA system for speech separation, we estimate the auditory feature uncertainties for better speaker likelihood calculation. Our systematic evaluation shows that the proposed auditory features and uncertainty estimates achieve substantial performance improvement over not only typical speaker features but also the state-of-the-art robust front-end processing.

It is important to note that our proposed system does not assume a noise model. Hence, it should generalize well to additive noise types other than the one tested. Also, the proposed feature extraction and likelihood calculation methods in the system are not restricted to SID tasks. We expect our system to provide a similar performance improvement on SV tasks. Since automatic speech recognition and speaker recognition typically share the same front-end, it is interesting to study the proposed auditory features also in speech recognition tasks.

Acknowledgements. This research was supported in part by an AFOSR grant (FA9550-04-1-0117), an AFRL grant (FA8750-04-1-0093), and an NSF grant (IIS-0534707). We thank G. Hu for his assistance in speech segregation.

5. REFERENCES

[1] J.P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, pp. 1437-1462, 1997.

[2] M. Cooke and T.W. Lee, "Speech separation and recognition competition," *Available at* http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm,

[3] M.P. Cooke, *Modelling auditory processing and organization.* Cambridge U.K.: Cambridge University Press, 1993.

[4] M.P. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, pp. 267-385, 2001.

[5] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variants using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 412-421, 2005.

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

[7] S. Furui, *Digital speech processing, synthesis, and recognition.* New York: Marcel Dekker, 2001.

[8] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2(4), pp. 578-589, 1994.

[9] G. Hu, *Monaural speech organization and segregation*. Ph.D. dissertation, The Ohio State University, 2006.

[10] A.V. Oppenheim, R.W. Schafer, and J.R. Buck, *Discrete-time signal processing.* 2nd ed., Upper Saddle River, NJ: Prentice-Hall, 1999.

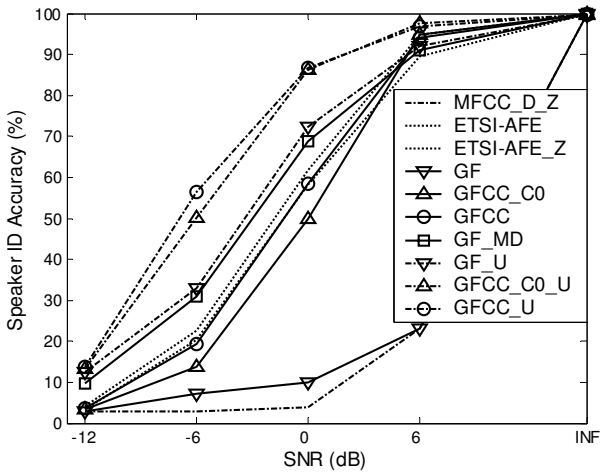

Figure 3. Accuracies of speaker identification in the presence of speech-shaped noise.

[11] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Providing single and multi-channel acoustical robustness to speaker identification systems," in *Proc. ICASSP*, pp. 1107-1110, 1997.

[12] N. Parihar and J. Picone, "Analysis of the aurora large vocabulary evalutions," in *Proc. Eurospeech*, pp. 337-340, 2003.

[13] R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "APU Report 2341: An Efficient Auditory Filterbank Based on the Gammatone Function," *Applied Psychology Unit*, 1988.

[14] B. Raj, M.L. Seltzer, and R.M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Comm.*, vol. 43, pp. 275-296, 2004.

[15] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Comm.*, vol. 17, pp. 91-108, 1995.

[16] D.A. Reynolds, *et al.*, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition," in *Proc. ICASSP*, pp. 784-787, 2003.

[17] R.C. Rose, E.M. Hofstetter, and D.A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. Speech Audio Process.*, vol. 2(2), pp. 245-257, 1994.

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

[19] S. Srinivasan and D.L. Wang, "A supervised learning approach to uncertainty decoding for robust speech recognition," in *Proc. ICASSP*, vol. I, pp. 297-300, 2006.

[20] STQ-AURORA, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," in *ETSI ES 202 050 V1.1.4* European Telecommunications Standards Institute, 2005-11.

[21] N.B. Yoma and M. Villar, "Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm," *IEEE Trans. Speech Audio Process.*, vol. 10(3), pp. 158-166, 2002.

[22] S. Young, D. Kershaw, J. Odell, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.0).* Microsoft Corporation, 2000.