ROBUST SPEAKER IDENTIFICATION USING AUDITORY FEATURES AND COMPUTATIONAL AUDITORY SCENE ANALYSIS

Yang Shao[1] *and DeLiang Wang*[1,2]

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

ABSTRACT

The performance of speaker recognition systems drop significantly under noisy conditions. To improve robustness, we have recently proposed novel auditory features and a robust speaker recognition system using a front-end based on computational auditory scene analysis. In this paper, we further study the auditory features by exploring different feature dimensions and incorporating dynamic features. In addition, we evaluate the features and robust recognition in a speaker identification task in a number of noisy conditions. We find that one of the auditory features performs substantially better than a conventional speaker feature. Furthermore, our recognition system achieves significant performance improvements compared with an advanced front-end in a wide range of signal-to-noise conditions.

Index Terms**—** Robust speaker recognition, auditory feature, Gammatone feature, Gammatone frequency cepstral coefficient, computational auditory scene analysis

1. INTRODUCTION

A speaker recognition system typically consists of three stages: feature extraction, speaker modeling, and decision making using pattern classification methods [3, 9]. Usually, short-time cepstral coefficients are extracted as speaker features [9] such as Mel-frequency cepstral coefficients (MFCC) [10], or long-term features such as prosody [20]. For speaker modeling, Gaussian mixture models (GMM) are widely used to model feature distributions of individual speakers [19]. Recognition decisions are usually made based on the likelihood of observing a feature frame given a speaker model. However, when facing distorted speaker features extracted from noisy utterances, such systems usually do not perform well because of mismatch in likelihood calculation [7, 24].

To tackle this robustness problem, speech enhancement methods such as spectral subtraction [7] have been explored for robust speaker recognition. These methods tend to perform well when noise is stationary. RASTA filtering [11] and cepstral mean normalization (CMN) [8] have also been used in speaker recognition but they are mainly intended for convolutive noise. On the other hand, recent studies of robust speech recognition on Aurora [16] have yielded an advanced feature extraction algorithm (AFE) [26], which is standardized by ETSI. ETSI-AFE derives robust MFCC features using a set of sophisticated front-end processes, including speech activity detection and Wiener filtering. An alternative approach to feature enhancement seeks to improve robustness by modeling noise and combining it with clean speaker models [13, 21]. However, these systems cannot deal with novel interference types because of their dependence on the prior information of noise sources.

On the other hand, humans are found to perform better than machines in speaker recognition tasks when input signals are corrupted by background noise such as crosstalk [22]. Furthermore, human subjects are able to select and follow the voice of a particular talker in the presence of multiple speakers as long as the signal-to-noise ratio (SNR) is not exceedingly low [1, 2]. This human ability is due to a perceptual process termed auditory scene analysis (ASA) [1]. Inspired by ASA research, computational auditory scene analysis (CASA) seeks to segregate target speech from a complex auditory scene based on ASA principles [28]. The superior performance of the auditory system in robust speaker recognition motivates us to explore CASA for robust speaker recognition.

Recently, we have proposed a novel auditory feature and a CASA-based robust speaker identification (SID) system [24]. Evaluations show that the auditory feature achieves a recognition performance level that is significantly better than MFCC. The SID system performs substantially better than the baseline system and significantly better than the ETSI-AFE features in a wide range of SNR conditions. In this paper, we continue the study on the auditory features by varying their feature dimensions. In addition, we incorporate dynamic coefficients with the static feature. Finally, we evaluate the novel auditory feature and the SID system under five different noisy conditions. Each of these conditions comprises mixtures with a wide range of SNRs.

The rest of the paper is organized as follows. Section 2 describes auditory features and a robust SID system. Evaluations are presented in Section 3. Section 4 concludes the paper.

2. CASA-BASED ROBUST SPEAKER RECOGNITION

Conceptually, our approach improves noise robustness in two aspects of a SID system: novel robust auditory features in the feature extraction stage; feature enhancement and better likelihood estimation in the speaker scoring stage. Specifically, we employ a CASA system [24] to segregate speech from noise and obtain a binary mask that indicates reliable or corrupted components of an auditory feature. The auditory feature is then enhanced by reconstructing the corrupted components [18, 24, 25].

1-4244-1484-9/08/$25.00 ©2008 IEEE

Additionally, we estimate reconstruction uncertainties [24, 25] and apply them in an uncertainty decoder [6] to calculate speaker likelihoods. This decoder accounts for varied accuracies of the feature enhancement process.

2.1 Auditory Features

Our system first performs auditory filtering by decomposing an input signal into the time-frequency (T-F) domain using a bank of Gammatone filters [28]. Gammatone filters are derived from psychophysical and physiological observations of the auditory periphery and this filterbank is a standard model of cochlear filtering [17]. The impulse response of a Gammatone filter centered at frequency *f* is:

$$g(f,t)=\begin{cases} t^{a-1}e^{-2\pi bt}\cos(2\pi ft), & t\geq 0 \\ 0, & else \end{cases}. \qquad (1)$$

t refers to time; *a*=4 is the order of the filter; *b* is the rectangular bandwidth which increases with the center frequency *f* [17]. We use a bank of 128 filters whose center frequencies range from 50 Hz to 8000 Hz. These center frequencies are equally distributed on the ERB scale [14] and the filters with higher center frequencies have wider bandwidths.

Since the filter output retains original sampling frequency, we down-sample the 128-channel responses to 100 Hz along the time dimension. This yields a corresponding frame rate of 10 ms, which is used in many short-time speech feature extraction algorithms [12]. The magnitudes of the down-sampled outputs are then loudness-compressed by a cubic root operation.

$$G_m[i]=\left|g_{downsample}[i,m]\right|^{1/3}, i=0...N-1, m=0...M-1. \qquad (2)$$

Here, *N*=128, referring to the 128 filter channels. *m* is the frame index; *M* is the number of time frames obtained after down-sampling. The resulting responses $G_m[i]$ form a matrix, representing a T-F decomposition of the input. This T-F representation is a variant of cochleagram [28]. Note that unlike the linear frequency resolution of a spectrogram, a cochleagram retains higher frequency resolution at low frequency range for the same number of frequency components. We base our subsequent processing on this T-F representation.

We call a time frame, *G*[*i*], of the above cochleagram a Gammatone feature (GF). Since it comprises 128 components, the dimension of the GF vector is much larger than that of feature vectors used in a typical speaker recognition system. Additionally, because of the overlap among neighboring filter channels, GF components are largely correlated with each other. In order to reduce dimensionality and de-correlate the components, we apply a discrete cosine transform (DCT) [15] to GF. We call the resulting coefficients Gammatone frequency cepstral coefficients (GFCC) [24]. Specifically, cepstral coefficients, *C*[*j*] *j*=0…*N*-1, are obtained from a GF, *G*[*i*], as follows,

$$C[j]=\sqrt{\frac{2}{N}}\sum_{i=0}^{N-1}G[i]\cos\left(\frac{j\pi}{2N}(2i+1)\right), j=0...N-1. \qquad (3)$$

Rigorously speaking, the newly derived features are not cepstral coefficients because a cepstral analysis requires a log operation between the first and the second frequency analysis for the deconvolution purpose [15]. Here we regard these features as cepstral coefficients because of the functional similarities between the above transformation and that of a typical cepstral analysis in the derivation of MFCC.

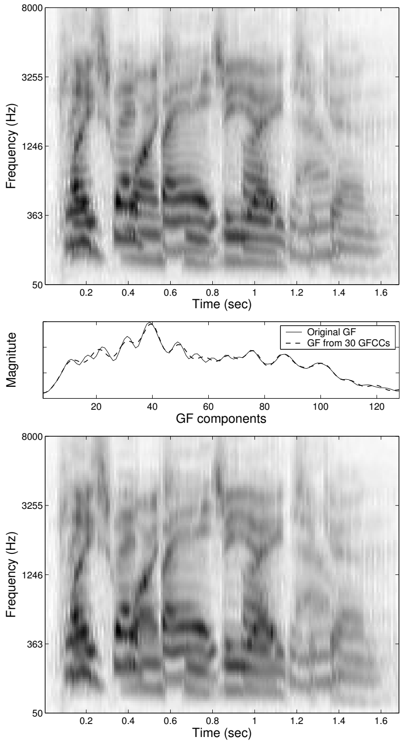

Figure 1. Illustrations of energy compaction by GFCCs. Darker color indicates stronger energy within the corresponding T-F unit.

2.2 Feature dimensions and dynamic features

In our previous study, the lower 23-order GFCC coefficients are used as a feature vector. We chose 23 GFCCs because they are compact and appear to retain most of the information of a GF frame. After performing inverse DCT of GFCCs, we find that the lower 30-order coefficients capture almost all the GF feature information while the GFCCs above the 30th are close to 0 numerically, which means that they provide negligible information. Fig. 1 illustrates a GFCC transformed GF and a cochleagram using 30 GFCCs. The top plot shows a cochleagram of an utterance. The middle plot shows a comparison of a GF frame of the top plot and the resynthesized GF from 30 GFCCs; the original GF is plotted as the solid line and the resynthesized GF by 30 GFCCs is plotted as the dashed line. The bottom plot presents the resynthesized cochleagram using 30 GFCCs. As observed from the figure, the lowest 30-order GFCCs largely retain the information in a 128-dimensional GF. This is due to the "energy compaction" property of DCT [15]. Hence, we use 30-dimensional GFCCs as a feature vector, $Z=(C[j])$, *j*=1…30, in this paper.

Since a typical speaker recognition system uses MFCCs and their first-order dynamic (delta) coefficients. Thus, it is desirable to study how GFCC dynamic features fare for recognition. The delta feature Z_D at time *t* is calculated from a set of neighboring GFCC vectors *Z* around the time frame at *t*.

$$Z_D(t)=\sum_{w=1}^{W}w\cdot\left(Z(t+w)-Z(t-w)\right)\Big/2\sum_{w=1}^{W}w^2, \qquad (4)$$

w is a neighboring window index; *W* denotes the half-window length and it is set to 2 here. In other words, the delta-window is of length 5. The delta coefficients are appended to the 30-dimensional GFCCs, resulting in a 60-dimensional feature vector.

2.3 Speech segregation and robust recognition

To enhance corrupted speaker features under noisy conditions, we apply a pitch-based speech segregation system [24] that performs CASA. 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 rates are used for grouping. As a result, it labels speech-dominated T-F units as reliable in the binary mask and noise-dominated units as unreliable.

In speaker recognition, the probability distribution of an extracted feature vector, produced by a speaker, is modeled as a GMM [19], typically parameterized by diagonal covariance matrices. A binary T-F mask produced by the CASA system indicates whether a GF component is reliable or unreliable. The latter is deemed as missing data since the system does not possess its distribution information. Accordingly, a feature vector is partitioned into reliable components and missing ones. To enhance a corrupted GF, we reconstruct its missing components from a speech prior [18]. Specifically, the missing components are estimated as the expected value conditioned on the reliable data [4, 24, 25]. Reconstruction errors are estimated as GF uncertainties [24, 25]. Enhanced GFs are transformed into GFCC using (3), likewise for uncertainties.

It is shown in [6] that an uncertainty decoder computes the likelihood of observing an enhanced GFCC frame $\hat{Z}$ given mixture component k of a speaker GMM as,

$$\int_{-\infty}^{\infty} p(Z \mid k)p(\hat{Z} \mid Z)dZ = N(\hat{Z};\mu_{Z,k},\sigma_{Z,k}^{2}+\hat{\sigma}_{Z}^{2}) \,. \qquad (5)$$

$\hat{\sigma}_Z^2$ is the diagonal covariances of the DCT transformed GF uncertainties. The non-diagonal covariances are numerically small and thus dropped from computation. Note that clean GFCC Z is integrated out. This uncertainty decoder increases the variances of individual components to account for mask estimation errors [6, 25]. Delta uncertainties are derived from GFCC uncertainties as

$$\hat{\sigma}_D^2(t) = \sum_{w=1}^{W} w^2 \cdot \left(\hat{\sigma}_Z^2(t+w) + \hat{\sigma}_Z^2(t-w)\right) \Bigg/ \left(2\sum_{w=1}^{W} w^2\right)^2 \,. \qquad (6)$$

Second-order dynamic coefficients, known as acceleration features, can be calculated by replacing the GFCCs and their uncertainties in (4) and (6) with the delta coefficients and delta uncertainties respectively.

3. EVALUATIONS

We use the speech materials from the recent speech separation corpus [5]. 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 test utterances with SSN at: −12, −6, 0 and 6 dB. The test set contains 600 utterances in each SNR condition. For a systematic evaluation, we create additional test sets by mixing the clean test utterances with four types of non-stationary noise: speech babble, destroyer operation room noise, F-16 cockpit and factory noise from the Noisex 92 corpus [27]. The mixtures are created at -6 dB, 0 dB, 6 dB and 12 dB SNRs. Speakers are modeled as 64-mixture GMMs. The speech prior comprises 2048 mixture components, and is trained from the pooled training utterances of all speakers.

Feature	-12 dB	-6 dB	0 dB	6 dB
MFCC_D_CMN Baseline	2.83	2.83	4	23.17
GFCC(23) Baseline	3.5	13.83	50	94.83
ETSI-AFE_D	3.5	20.33	58.17	89.5
GFCC(23)	13.33	51.17	87	97.33
GFCC(30)	14.83	54.67	89	97.67
GFCC(30)_D	**9.83**	**58.83**	**92.17**	**98.67**
GFCC(30)_D_A	7.67	37.83	81	97.33

Table 1 Accuracy (%) of robust SID using GFCCs, dynamic features and uncertainty decoding. _D refers to delta feature; _A denotes acceleration feature.

We first evaluate feature dimensions and delta features using the SSN test set and show results in Table 1. The baseline and ETSI-AFE results are taken from our previous study [24]. For the baseline, GFCCs substantially outperform MFCCs. Under other conditions, GFCCs are enhanced and uncertainty decoding is applied. Our robust recognition system substantially outperforms the baseline and the ETSI-AFE. Increasing the number of GFCCs from 23 to 30 further improves SID performance. Delta-augmented GFCCs yield significantly better performance than the static feature alone except at -12 dB condition, where reconstruction does not perform well with few reliable GF components. In addition, we find that including the acceleration feature rather hurts system performance. This is probably because the acceleration window requires 9 frames while estimated binary masks with SSN do not typically contain consecutively reconstructed frames that can provide reliable acceleration feature estimates. Compared with our previous study, we have, on average, significantly improved SID accuracy by increasing the number of GFCCs and incorporating the delta feature.

We then evaluate our system under the additional four noisy conditions. Evaluation results are presented in Table 2. Specifically, we use the 30-dimensional GFCCs and their delta coefficients because they achieved the best overall performance in the preceding experiment. The results in the table corroborate the conclusions in the SSN experiment. Our GFCC feature is substantially better than the MFCC feature, and our recognition system significantly outperforms the ETSI-AFE feature except at 12 dB where the SID performance saturates.

4. CONCLUDING REMARKS

In this paper, we have studied auditory features and a general solution to robust speaker recognition under additive noise conditions. The novel speaker features are derived from auditory filtering and cepstral analysis. Additionally, by using binary T-F masks generated by a CASA system, we enhance the auditory features and estimate their reconstruction uncertainties for better

Babble	**-6 dB**	**0 dB**	**6 dB**	**12 dB**
MFCC_D_CMN Baseline	3.0	10.67	60.33	95.83
GFCC_D Baseline	5.67	39.5	90.67	99.67
GFCC_D	**25.0**	**83.83**	**97.5**	**99.17**
ETSI-AFE_D	19.0	69.83	96.5	99.67
Destroyer	**-6 dB**	**0 dB**	**6 dB**	**12 dB**
MFCC_D_CMN Baseline	2.83	3.33	24.5	71.33
GFCC_D Baseline	3.17	12.5	72.0	96.83
GFCC_D	**16.5**	**76.83**	**97.0**	**98.67**
ETSI-AFE_D	12.83	44.5	76.17	95.0
F16	**-6 dB**	**0 dB**	**6 dB**	**12 dB**
MFCC_D_CMN Baseline	2.83	8.83	9.67	54.17
GFCC_D Baseline	6.17	15.33	57.83	93.83
GFCC_D	**41.67**	**83.5**	**96.5**	**99.17**
ETSI-AFE_D	3.83	37.83	77.5	96.5
Factory	**-6 dB**	**0 dB**	**6 dB**	**12 dB**
MFCC_D_CMN Baseline	2.83	3.33	17.67	65.5
GFCC_D Baseline	8.5	28.83	77.83	98.0
GFCC_D	**46.17**	**87.83**	**97.83**	**99.33**
ETSI-AFE_D	9.5	43.5	79.17	95.67

Table 2: Accuracy (%) of robust SID using GFCCs, dynamic features and uncertainty decoding.

speaker likelihood calculation. Our systematic evaluations show that the proposed feature performs significantly better than a conventional speaker feature. Furthermore, we find that employing CASA as a front-end processor to work in conjunction with uncertainty decoding achieves significant performance improvements over not only conventional speaker features but also advanced robust front-end processing.

It is important to note that our proposed system does not require noise conditions be known *a priori* or assumes a noise model. Hence, the proposed robust speaker recognition system is expected to generalize well to noise types not tested. In addition, our preliminary studies in speaker verification tasks indicate that similar improvements are observed as in the SID evaluations [23].

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).

REFERENCES

[1] A.S. Bregman, *Auditory scene analysis.* Cambridge MA: MIT Press, 1990.

[2] D.S. Brungart, "Information and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 109, pp. 1101-1109, 2001.

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

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

[5] M.P. Cooke and T.W. Lee, "Speech separation and recognition competition," *Available at http://www.dcs.shef.ac.uk/~martin/ SpeechSeparationChallenge.htm*, 2006.

[6] L. Deng*, et al.*, "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.

[7] 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.

[8] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech Signal Proc.*, vol. 29, pp. 254-272, 1981.

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

[10] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Sig. Proc. Mag.*, vol. 11(4), pp. 18-3218-32, 1994.

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

[12] X. Huang, *et al.*, *Spoken language processing.* Upper Saddle River: Prentice Hall, 2001.

[13] T. Matsui and S. Furui, "Speaker recognition using HMM composition in noisy environments," *Computer Speech and Language*, vol. 10, pp. 107-116, 1996.

[14] B.C.J. Moore, *An introduction to the psychology of hearing.* 5th ed., San Diego: Academic, 2003.

[15] A.V. Oppenheim*, et al.*, *Discrete-time signal processing.* 2nd ed., Upper Saddle River, NJ: Prentice-Hall, 1999.

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

[17] R.D. Patterson*, et al.*, "Auditory models as preprocessors for speech recognition," in *The auditory processing of speech: From sounds to words.,* M.E.H. Schouten, Ed., Berlin, Germany: Mouton de Gruyter, pp. 67-83, 1992.

[18] B. Raj*, et al.*, "Reconstruction of missing features for robust speech recognition," *Speech Comm.*, vol. 43, pp. 275-296, 2004.

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

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

[21] R.C. Rose*, et al.*, "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.

[22] A. Schmidt-Nielsen and T.H. Crystal, "Human v.s. machine speaker identification with telephone speech," in *Proc. ICSLP*, 1998.

[23] Y. Shao, *Sequential organization in computational auditory scene analysis*. Ph.D. dissertation, The Ohio State University, 2007.

[24] Y. Shao*, et al.*, "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. ICASSP*, vol. IV, pp. 277-280, 2007.

[25] S. Srinivasan and D.L. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15(7), pp. 2130-2140, 2007.

[26] 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.

[27] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Comm.*, vol. 12(3), pp. 247-251, 1993.

[28] D.L. Wang and G.J. Brown, Ed., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.