AN AUDITORY-BASED FEATURE FOR ROBUST SPEECH RECOGNITION

Yang Shao[1], *Zhaozhang Jin*[1], *DeLiang Wang*[1,2]

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

Soundararajan Srinivasan

Research and Technology Center
Robert Bosch LLC.
Pittsburgh, PA 15212, USA
Soundar.Srinivasan@us.bosch.com

ABSTRACT

A conventional automatic speech recognizer does not perform well in the presence of noise, while human listeners are able to segregate and recognize speech in noisy conditions. We study a novel feature based on an auditory periphery model for robust speech recognition. Specifically, gammatone frequency cepstral coefficients are derived by applying a cepstral analysis on gammatone filterbank responses. Our evaluations show that the proposed feature performs considerably better than conventional acoustic features. We further demonstrate that integrating the proposed feature with a computational auditory scene analysis system yields promising recognition performance.

***Index Terms*—** Robust speech recognition, auditory feature, gammatone frequency cepstral coefficients, computational auditory scene analysis.

1. INTRODUCTION

In everyday listening conditions, the acoustic input reaching our ears is often a mixture of multiple concurrent sound sources. While human listeners are able to segregate and recognize a target signal under such conditions, robust automatic speech recognition remains a challenging problem [1, 14]. Automatic speech recognizers (ASRs) are typically trained on clean speech and face the mismatch problem when tested in the presence of interference.

To tackle this robustness problem, speech enhancement methods, such as spectral subtraction, have been utilized for robust speech recognition. These methods tend to perform well when noise is stationary. RASTA filtering [11] and cepstral mean normalization [14] have also been widely applied but they are mainly intended for convolutive noise. An alternate approach involves the joint decoding of the speech mixture based on knowledge of all the sources present in the mixture [8, 9]. These model-based systems rely heavily on the use of *a priori* information of noise sources. Hence, they have limited ability to handle novel interferences.

On the contrary, human listeners are capable of recognizing speech when input signals are corrupted by noise [6]. Furthermore, for a cochannel signal that has comparable energies from both talkers, human listeners can readily select and follow one speaker's voice [3]. Even in more adverse scenarios such as a cocktail party, listeners can select and follow the voice of a particular talker as long as the signal-to-noise ratio (SNR) is not exceedingly low [2]. The human ability in these complex acoustic environments is accounted for by a perceptual process called auditory scene analysis (ASA) [2]. Inspired by ASA studies, computational auditory scene analysis (CASA) seeks to segregate target speech from a complex auditory scene [24]. Such systems have been integrated with ASRs to perform robust recognition [4, 19, 21].

In CASA research, the gammatone filterbank has been widely used to transform signals into the time-frequency (T-F) domain [24]. This filterbank was originally designed to model human cochlear filtering [17]. Recently, we have proposed an auditory feature based on gammatone filtering for robust speaker recognition [20]. We have found that this auditory feature, namely GFCC (gammatone frequency cepstral coefficients), performs substantially better than conventional Mel-frequency cepstral coefficients (MFCCs). In addition, the auditory feature coupled with CASA-based speech segregation [12] and uncertainty decoding [7] yields significant improvements in robust speaker recognition over features derived by an advanced front-end feature extraction algorithm, ETSI-AFE [22].

In this paper, we study GFCC for robust speech recognition. GFCCs are derived by a cepstral analysis from the gammatone feature (GF) obtained from a bank of gammatone filters. We compare the proposed feature with MFCCs on a test set with speech-shaped noise (SSN) [5]. Besides, we also compare GFCC with the perceptual linear predictive (PLP) cepstral coefficients [10]. Note that PLP analysis is also motivated by perceptual models. In addition to SSN, we evaluate on another test set that includes four noise types from the Noisex 92 corpus [23]. Finally, based on the GFCC feature, we explore the idea of incorporating a CASA system [12] that segregates voiced speech from background noise as a front-end processor for further improvement.

The rest of the paper is organized as follows. Section 2 describes the auditory feature. Section 3 describes CASA-based speech segregation and robust speech recognition. Evaluations are presented in Section 4. Section 5 concludes the paper.

2. AUDITORY FEATURES

A standard model for T-F analysis in CASA system involves a bank of gammatone filters [24]. Gammatone filters are derived from psychophysical 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:

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

t refers to time; $a = 4$ is the order of the filter; b is the rectangular bandwidth which increases with the center frequency f. We use a bank of 128 filters whose center frequency ranges from 50 Hz to

8000 Hz. These center frequencies are equally distributed on the ERB scale [15] and the filters with higher center frequencies respond to wider frequency ranges.

Note that the filter output retains original sampling frequency. To obtain a frame rate used in typical speech processing applications, we down-sample the 128 channel responses to 100 Hz along the time dimension, resulting in the corresponding frame rate of 100 Hz. The magnitudes of the down-sampled outputs are then loudness-compressed by a cubic root operation. The resulting responses $G_c[m]$ form a matrix, representing a T-F decomposition of the input. m is the frame index and c is the channel index. We call this T-F representation a cochleagram [24], analogous to the widely used spectrogram. Note that unlike the linear frequency resolution of a spectrogram, a cochleagram provides a much higher frequency resolution at low frequencies than at high frequencies. We base our subsequent processing on this T-F representation.

We call a time frame of the above cochleagram a GF feature. Here, a GF vector comprises 128 frequency components. Note that the dimension of a GF vector is much larger than that of feature vectors used in a typical speaker recognition system. Additionally, because of overlap among neighboring filter channels, the gammatone features are largely correlated with each other. Here, we apply a discrete cosine transform (DCT) [16] to a GF in order to reduce its dimensionality and de-correlate its components. The resulting coefficients are called GFCCs [20]. Specifically, for frame m, GFCCs $C_i[m]$ are obtained from GFs $G_c[m]$ as follows:

$$C_i[m] = \sqrt{\frac{2}{N}} \sum_{c=0}^{N-1} G_c[m] \cos\left(\frac{i\pi}{2N}(2c+1)\right), \; i = 0, ..., N-1. \tag{2}$$

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 purpose of deconvolution [16]. Here we regard these features as cepstral coefficients because of the functional similarities between the above transformation and that of a typical cepstral analysis.

After performing inverse DCT of GFCCs, we find that by including up to 30 coefficients almost all the GF feature information is captured while the GFCCs above the 30th are close to 0 numerically. 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 its 30 GFCCs. The bottom plot presents the resynthesized cochleagram from the top plot using 30 GFCCs. As observed from the figure, the 30 lowest order GFCCs retain the majority information in a 128-dimensional GF. This is due to the "energy compaction" property of the DCT [16]. Hence, we use this 30-dimensional GFCCs as a feature vector in this paper. The 1st order coefficient is the summation of all the GF components—it relates to the overall energy of a GF frame and is susceptible to noise degradation. Thus, we remove C_1 from the feature vector. The static GFCC feature is:

$$Z[m] = \Big\{ C_i[m] \,\Big|\, i = 2, ..., 30 \Big\}. \tag{3}$$

Besides, a dynamic feature that is composed of delta coefficients is calculated to incorporate temporal information. Specifically, a vector of delta coefficients Z_D at time frame m is calculated as

$$Z_D[m] = \sum_{w=1}^{W} w\Big(Z[m+w] - Z[m-w]\Big) \Big/ 2\sum_{w=1}^{W} w^2, \tag{4}$$

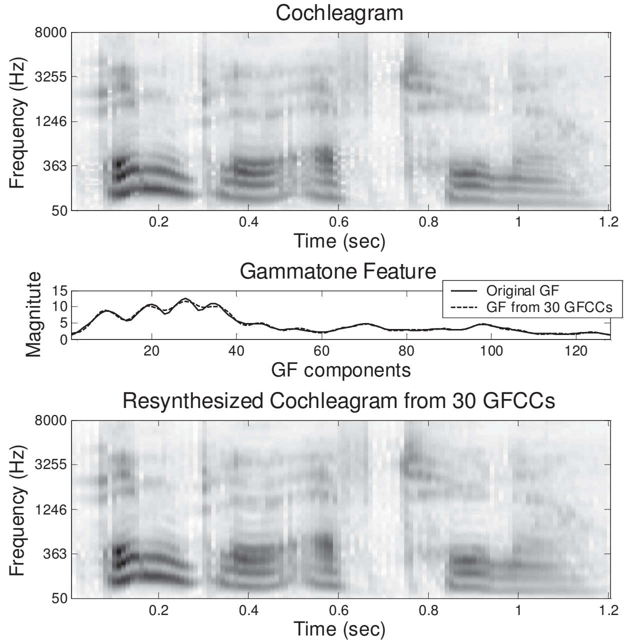

Fig. 1. Illustrations of GF and GFCC.

where w is a neighboring window index; W denotes the half-window length and it is typically set to 2.

3. SPEECH SEGREGATION AND ROBUST ASR

Under adverse environments, our GFCC feature will be corrupted by background noise. Previous studies have shown that CASA-based speech segregation provides robust recognition results [4, 19, 21]. To enhance the GFCC feature in noise, we employ a CASA system [12] that performs voiced speech segregation and estimates a binary T-F mask.

CASA systems make minimal assumptions about the underlying noise and have shown significant SNR improvement on segregated speech under various noisy conditions. 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. In the binary mask, it labels speech-dominated T-F units as reliable (1) and noise-dominated units as unreliable (0).

In speech recognition, a pronunciation unit is usually modeled as a hidden Markov model (HMM) [14]. The feature distribution within a HMM state is typically modeled as a Gaussian mixture model (GMM) [14], usually parameterized by diagonal covariance matrices. A binary T-F mask produced by the CASA system indicates whether a GF component is reliable or not. 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 [18, 20, 21]. Reconstruction errors are estimated as GF uncertainties. Enhanced GFs are then transformed into GFCCs using (2), likewise for uncertainties.

We use an uncertainty decoder [7, 21] to determine the content of an noisy utterance using enhanced GFCC feature frames. Here, only the diagonal covariance $\hat{\sigma}_Z^2$ of the DCT transformed GF uncertainties are used. The non-diagonal covariances are numerically small and thus dropped from computation. This uncertainty decoder increases the variances of individual components to account for the mask estimation errors. Delta uncertainties are derived from GFCC uncertainties as

$$\hat{\sigma}_D^2[m] = \sum_{w=1}^{W} w^2 \left(\hat{\sigma}_Z^2[m+w]+\hat{\sigma}_Z^2[m-w]\right) \Big/ \left(2\sum_{w=1}^{W} w^2\right)^2 \tag{5}$$

4. EVALUATIONS

4.1. Evaluations Using Speech-Shaped Noise

We first evaluate the GFCC feature (see Section 2) and the robust recognition method (see Section 3) on the speech-shaped noise (SSN) test set of the speech separation and recognition task [5]. The utterances in the corpus follow a sentence grammar:

$command $color $preposition $letter $number $adverb.

There are 4 word choices each for $command, $color, $preposition and $adverb, 25 choices for $letter (A-Z except W), and 10 choices for $number (1-9 and zero). For example, a valid utterance could be "Place blue at F 2 now". The possible choices in each position are roughly uniformly distributed in the corpus. The training data consists of a total of 17,000 clean utterances from 34 speakers. The SSN test data is created by mixing clean utterances with SSN at 4 SNRs: -12 dB, -6 dB, 0 dB and 6 dB. Each SNR condition contains 600 utterances. For recognition, whole-word HMM-based speaker-independent models are trained on clean speech. Each word model comprises 8 states and 32 Gaussian mixtures with diagonal covariance in each state. For missing data reconstruction, we use the speech prior model with a mixture of 2,048 Gaussian densities to reconstruct missing features of GF. The uncertainty decoder also uses diagonal covariance for uncertainties. During the recognition process, given estimated uncertainties and clean ASR models, the uncertainty decoder calculates the likelihood of reconstructed 58-dimensional GFCC_D features and transcribes the speech.

Fig. 2 compares recognition performances of different features. MFCC_D_A is the 36-dimensional MFCC feature including delta and acceleration coefficients (with DC component removed). Similarly, PLP_D_A denotes the 36-dimensional cepstral coefficients derived by PLP analysis. ETSI_D_A represents the enhanced 36-dimensional MFCC feature derived by ETSI-AFE. GFCC_D is the 58-dimensional GFCC feature, with deltas included. Enhanced GFCC_D shows the results using feature reconstruction and uncertainty decoding (described in Section 3).

The GFCC feature outperforms the MFCC and the PLP features across all SNR conditions. There is a considerable advantage of using GFCC when the noise level is moderate (e.g., at 0 and 6 dB). Under the clean condition, GFCC is still the best among the three. Strictly speaking, the above comparison is not entirely fair due to different feature dimensionalities. However, it has been noted that additional MFCC coefficients do not improve performance [14]—in other words, the performance is not expected to improve much if one increases the feature dimensions of MFCC and PLP in the experiment. We should note that these three features are on noisy data without enhancement.

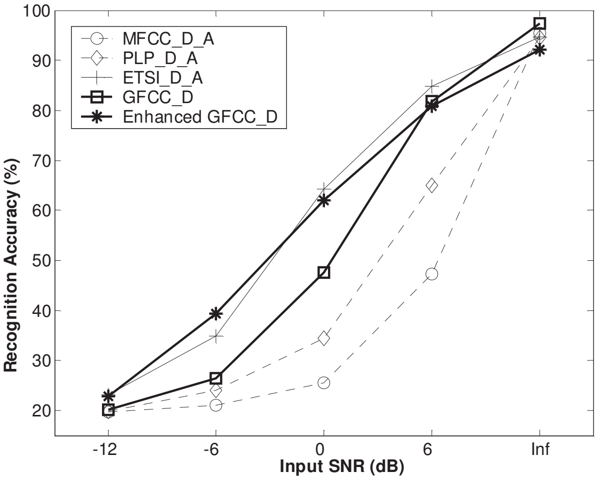

Fig. 2. Evaluation of various features on the SSN test set.

We then compare the enhanced GFCC feature with the ESTI feature. These two methods enhance features in different ways—the former utilizes missing data reconstruction coupled with an uncertainty decoder and the latter applies a Wiener filter based denoising approach. They yield comparable performance as can be observed in Fig. 2. Nevertheless, we still consider the enhanced GFCC_D as a promising method. Firstly, only voiced speech is currently segregated from a noisy utterance and unvoiced portions totally rely on reconstruction. This harms the performance and explains why the enhanced GFCC has the worst result under the clean condition. Secondly, this method can deal with a more general acoustic background (even in the presence of another speech) due to the nature of CASA-based segregation. In fact, it has been successfully applied to a two-talker speech recognition task [19], in which ESTI fails to work.

4.2. Evaluations Using Other Noise types

We also experiment under four other non-stationary noisy conditions. They are factory noise, speech babble, destroyer operation room and F-16 cockpit from the Noisex 92 corpus [23]. Test set is created by mixing the clean test utterances from the previous SSN task with the four noise recordings. Mixtures are created at -6 dB, 0 dB, 6 dB and 12 dB SNRs. Thus, we have 600 test utterances from 34 speakers for each of the 16 noisy conditions.

Evaluation results are reported in Table 1. Similar conclusions are drawn: The GFCC feature performs consistently better than the MFCC and the PLP features. This feature also outperforms the ETSI feature at 12 dB. The enhanced GFCC considerably improves the recognition accuracy under moderate SNR conditions. It also yields comparable performance to the ETSI feature.

5. CONCLUSIONS

We have investigated a robust feature, GFCC, for speech recognition, which is derived from an auditory filterbank. Our evaluations show that the GFCC feature outperforms the MFCC and the PLP features. We have further explored the idea of incorporating a CASA system that performs voiced speech segregation as a front-end for robust recognition. It is encouraging to observe that our results are

Table 1. Evaluation of various features on four noisy conditions. Numbers in the table show recognition accuracy in percentage (%).

Factory	**-6 dB**	**0 dB**	**6 dB**	**12 dB**
MFCC_D_A	20.39	22.19	35.36	61.31
PLP_D_A	22.78	30.92	54.61	78.44
ETSI_D_A	29.75	52.81	76.50	88.00
GFCC_D	27.89	47.28	78.28	93.42
Enhanced GFCC_D	34.72	59.53	78.08	85.69
Babble	**-6 dB**	**0 dB**	**6 dB**	**12 dB**
MFCC_D_A	22.81	29.83	47.28	67.97
PLP_D_A	27.78	40.83	63.61	81.75
ETSI_D_A	33.61	57.39	80.11	90.39
GFCC_D	24.19	41.83	74.03	91.81
Enhanced GFCC_D	33.42	55.28	76.94	86.47
Destroyer	**-6 dB**	**0 dB**	**6 dB**	**12 dB**
MFCC_D_A	20.53	26.19	43.14	67.67
PLP_D_A	23.14	35.89	62.97	84.17
ETSI_D_A	34.97	60.47	81.78	90.44
GFCC_D	26.11	48.44	78.25	92.78
Enhanced GFCC_D	30.53	55.61	77.22	86.47
F16	**-6 dB**	**0 dB**	**6 dB**	**12 dB**
MFCC_D_A	19.89	22.11	32.75	59.69
PLP_D_A	22.50	29.44	49.44	76.19
ETSI_D_A	28.61	55.92	79.89	90.03
GFCC_D	23.14	41.00	70.92	91.86
Enhanced GFCC_D	34.36	57.83	76.56	85.36

comparable to those using the ETSI-AFE enhanced feature despite the fact that only voiced speech is segregated in our method. Future work that incorporates unvoiced speech segregation [13] will likely lead to further improvements in ASR performance.

Acknowledgements. This research was supported in part by an AFOSR grant (FA9550-08-1-0155) and an NSF grant (IIS-0534707).

6. REFERENCES

[1] J. B. Allen, *Articulation and Intelligibility*. San Rafael, CA: Morgan & Claypool, 2005.

[2] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[3] D. S. Brungart, “Information and energetic masking effects in the perception of two simultaneous talkers,” *J. Acoust. Soc. Amer.*, vol. 109, pp. 1101–1109, 2001.

[4] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, “Robust automatic speech recognition with missing and unreliable acoustic data,” *Speech Comm.*, pp. 267–285, 2001.

[5] M. Cooke and T. W. Lee, “Speech separation and recognition competition.” [Online]. Available: http://www.dcs.shef.ac.uk/martin/SpeechSeparationChallenge.htm

[6] C. J. Darwin, “Listening to speech in the presence of other sounds,” *Phil. Trans. R. Soc. B*, vol. 363, pp. 1011–1021, 2008.

[7] L. Deng, J. Droppo, and A. Acero, “Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion,” *IEEE Trans. Speech Audio Processing*, pp. 412–421, 2005.

[8] A. N. Deoras and M. Hasegawa-Johnson, “A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel,” in *Proc. IEEE ICASSP*, 2004, pp. 861–864.

[9] M. J. F. Gales and S. J. Young, “Robust continuous speech recognition using parallel model combination,” *IEEE Trans. Speech Audio Processing*, pp. 352–359, 1996.

[10] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *J. Acoust. Soc. Amer.*, pp. 1738–1752, 1990.

[11] ——, “RASTA processing of speech,” *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 578–589, 1994.

[12] G. Hu and D. L. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” *IEEE Trans. Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[13] ——, “Segregation of unvoiced speech from nonspeech interference,” *J. Acoust. Soc. Amer.*, vol. 124, pp. 1306–1319, 2008.

[14] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Upper Saddle River, NJ: Prentice Hall PTR, 2001.

[15] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. San Diego, CA: Academic Press, 2003.

[16] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[17] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, “An efficient auditory filterbank based on the gammatone function,” Appl. Psychol. Unit, Cambridge, UK, APU Rep. 2341, 1988.

[18] B. Raj, M. L. Seltzer, and R. M. Stern, “Reconstruction of missing features for robust speech recognition,” *Speech Comm.*, pp. 275–296, 2004.

[19] Y. Shao, S. Srinivasan, Z. Jin, and D. L. Wang, “A computational auditory scene analysis system for speech segregation and robust speech recognition,” *Computer Speech and Language*, in press.

[20] Y. Shao and D. L. Wang, “Robust speaker identification using auditory features and computational auditory scene analysis,” in *Proc. IEEE ICASSP*, 2008, pp. 1589–1592.

[21] S. Srinivasan and D. L. Wang, “Transforming binary uncertainties for robust speech recognition,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, pp. 2130–2140, 2007.

[22] 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*, 2005.

[23] 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, pp. 247–251, 1993.

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