ROBUST SPEAKER IDENTIFICATION USING A CASA FRONT-END

Xiaojia Zhao[1]*, 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
{zhaox, shaoy, dwang}@cse.ohio-state.edu

ABSTRACT

Speaker recognition remains a challenging task under noisy conditions. Inspired by auditory perception, computational auditory scene analysis (CASA) typically segregates speech by producing a binary time-frequency mask. We first show that a recently introduced speaker feature, Gammatone Frequency Cepstral Coefficient, performs substantially better than conventional speaker features under noisy conditions. To deal with noisy speech, we apply CASA separation and then either reconstruct or marginalize corrupted components indicated by the CASA mask. Both methods are effective. We further combine them into a single system depending on the detected signal to noise ratio (SNR). This system achieves significant performance improvements over related systems under a wide range of SNR conditions.

Index Terms— Robust speaker identification, gammatone frequency cepstral coefficient, GFCC, CASA, ideal binary mask

1. INTRODUCTION

The goal of speaker recognition system is to reveal underlying speaker identity based on distinctive speaker features. Cepstral features such as mel-frequency cepstral coefficients (MFCC) and perceptual linear predictive (PLP) coefficients are commonly used speaker features. Typically a speaker's feature space is modeled by a Gaussian mixture model (GMM). Recognition decision is made based on likelihoods of observed features given a speaker model. This approach has shown remarkable performance for speaker recognition in clean conditions. However, when the noise corruption is taken into account, the observed features mismatch clean speaker models, and yield poor recognition performance.

To address this robustness problem, RASTA filtering, cepstral mean normalization, and speech enhancement methods such as spectral subtraction have been explored. Studies of robust speech recognition on Aurora have yielded an advanced front-end feature extraction algorithm (AFE) [13], which is standardized by the European Telecommunication Standards Institute (ETSI). ETSI-AFE adds robustness to conventional MFCC features by sophisticated front-end processes, including speech activity detection and Wiener filtering. Alternative approaches seek to improve robustness by modeling noise and combining it with clean speaker models, or directly modeling speaker in noisy environments. However, these alternatives are heavily dependent on the *a priori* information of noise sources.

In a noisy acoustic environment, human listeners perform robustly in speaker recognition tasks, due to the perceptual process of auditory scene analysis (ASA) [1]. Inspired by ASA research, computational auditory scene analysis (CASA) aims to organize sound based on ASA principles [15]. In this paper, we propose a robust speaker identification (SID) system using CASA as a front-end to perform speech segregation. The output of CASA segregation is a binary time-frequency (T-F) mask that indicates whether a particular T-F unit is dominated by speech or noise. In [12], we proposed Gammatone Feature (GF) and Gammatone Frequency Cepstral Coefficient (GFCC), based on an auditory periphery model, and showed that GFCC performs significantly better than MFCC in terms of speaker identification performance. The proposed system has two modules. To account for the deviations of noisy features from clean ones, the first module enhances the GF by reconstructing corrupted components. The second module performs bounded marginalization on the noisy GF. Each module yields substantial improvement over baseline SID systems. As the two modules perform well in different SNR ranges, we propose a combined system based on a simple SNR detection.

The rest of the paper is organized as follows. Section 2 describes the overall system architecture. Auditory feature extraction and binary mask estimation are discussed in Section 3. Section 4 introduces the two modules and the combined system. SID evaluations are presented in Section 5. Further discussions are given in Section 6.

2. SYSTEM OVERVIEW

The proposed system uses CASA as a front-end processor for robust SID. Figure 1 presents the block diagram of the overall system. Input speech is decomposed using a gammatone filterbank and subsequent time windowing to generate a time sequence of GFs. This T-F analysis results in a cochleagram [15]. Simultaneously, we feed the input signal to a CASA system that computes a binary mask corresponding to the target speech [4]. Each element of this mask corresponds to a T-F unit in the cochleagram, with 1 indicating the T-F unit is dominated by target speech and 0 by noise. The binary mask and input speech are fed to an SNR detector to select either the reconstruction module or marginalization module for SID.

In the reconstruction module, the noise-corrupted components indicated by the CASA mask are reconstructed using a speech prior and the enhanced GF is converted to the cepstral domain by discrete cosine transform (DCT). Subsequently, the obtained cepstral feature, GFCC, is used in conjunction with trained speaker models to derive the underlying speaker identity. In the marginalization module, bounded marginalization is performed on

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

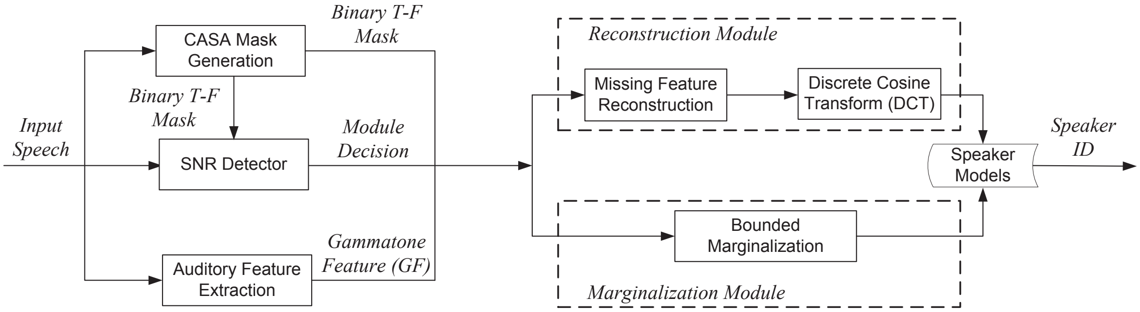

Figure 1. Schematic diagram of a CASA-based robust speaker identification system.

the noisy GF directly to marginalize noise-corrupted components.

Each module provides an SID system by itself. Our experiments suggest that the reconstruction module works better in high SNR conditions and the marginalization module in low SNR conditions. To leverage their respective advantages, our combined system assigns the input signal to one of the modules based on its estimated SNR.

3. FEATURE EXTRACTION AND MASK ESTIMATION

In our previous paper [12], we described how to generate GF and GFCC. First we perform auditory filtering by decomposing an input signal into the T-F domain using a bank of gammatone filters. After decimating fully rectified filter responses to 100 Hz followed by a cubic root operation, a variant of cochleagram is obtained. We call each time slice in this cochleagram GF. By taking DCT, we convert GF to GFCC. Specifically, cepstral coefficients, $C[j], j=0, \ldots, N\text{-}1$ are obtained from a GF as follows

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

Here N=64 refers to the number of frequency channels. $G[i]$ is the ith element of the corresponding GF vector.

As described earlier, a cochleagram is a T-F representation of a signal. With such a representation, a binary mask furnishes the crucial information about whether a T-F unit is dominated by target speech or noise. As a main computational goal of CASA, an ideal binary mask (IBM) is a binary matrix where a mask element equals 1 if the target energy exceeds the noise energy in corresponding T-F unit, and 0 otherwise [14]. The IBM concept is motivated by the auditory masking phenomenon, and is the optimal binary mask in terms of SNR gain.

To estimate the IBM from an input mixture, we employ a recent CASA system that performs feature-based classification [4]. This system is employed because it makes few assumptions about the underlying noises and performs well under various SNR conditions, even with room reverberation. First, we estimate the pitch of the speech signal at each frame using a multipitch tracking algorithm [5]. This algorithm formulates multipitch tracking as a hidden Markov model (HMM), which can produce up to 2 pitch points at each frame. As we deal with noises that are mostly aperiodic, the multipitch tracker tends to output at most one pitch per frame. Given an estimated pitch, a 6-dimensional pitch-based feature vector is extracted for each T-F unit [4]. These features are fed to a multilayer perceptron (MLP) classifier, whose output can be interpreted as the posterior probability of a T-F unit being target dominant. This statistical interpretation naturally leads to the decision threshold of 0.5 in binary mask generation. This process will produce IBM estimation errors. We observe that the bounded marginalization method is more adversely affected by false-alarm errors (0 mistakenly labeled as 1) than by miss errors (1 mistakenly labeled as 0). This observation leads us to choose a higher decision threshold for converting the posterior probabilities as the MLP output to a binary mask than 0.5. We find that raising the threshold from 0.5 to 0.6 yields good results, and we will use this raised threshold in the marginalization module (no change is made in the reconstruction module).

4. COMBINED SYSTEM

4.1. Reconstruction Module

The aforementioned speech segregation system produces a binary T-F mask that indicates whether a GF component is speech dominant or noise dominant. The former one is regarded as reliable while the latter is deemed missing. In order to enhance a noise corrupted GF, we first reconstruct its missing components from a speech prior model, which is a large GMM obtained from pooled training data [10]. With the reconstructed GF, we convert it to GFCC by applying DCT. GFCC can be directly used for recognition in conjunction with trained speaker models [12].

4.2. Marginalization Module

An alternative approach to reconstruction is marginalization, which has shown good performance in robust speech recognition and has been applied to robust speaker recognition [3, 11]. The main idea behind marginalization is to base recognition decision on reliable T-F units. In other words, we want to marginalize unreliable T-F units. To achieve this goal, we can integrate the unreliable T-F units either from minus infinity to positive infinity or in a tighter range, from 0 to the observed feature value. The latter is called bounded marginalization. A systematic evaluation we have done shows that bounded marginalization produces substantially better SID performances than full marginalization. Therefore in the marginalization module, we employ bounded marginalization on GF features.

4.3. Combined System

Between the reconstruction module and the marginalization module, we expect the former to perform better at high SNRs as it is well known that cepstral features outperform spectral features in recognition. On the other hand, marginalization is expected to perform better in low SNR conditions, as reconstruction based on few reliable T-F units likely has poor quality. Also, bounded marginalization makes use of some information from unreliable T-F units. These differing performance trends are indeed confirmed by the evaluation results presented in the next section. To utilize the relative advantages, we propose to combine them into one system using a simple SNR detector.

From a CASA mask, we can resynthsize the target signal out of the mixture [15]. Similarly, we can resynthesize the noise by inverting the binary mask (the complement mask). Because the CASA mask is generated only in voiced frames, the complement mask should be used only in the frames where the original mask has at least one T-F unit labeled as 1. Given the estimated target speech and the estimated interference in voiced frames, SNR is readily calculated. This SNR detector tends to overestimate the mixture SNR because the estimation is taken during intervals when the target speech is voiced, hence having strong energy. This factor will be considered when choosing a decision threshold for distinguishing low and high SNRs. On the other hand, this simple estimator suffices for our purposes as we do not need to precisely detect the mixture SNR, rather only deciding whether the mixture SNR is high or low. If the detected SNR is high, the combined system chooses the reconstruction module to perform SID. Otherwise, it chooses the marginalization module.

5. EVALUATIONS

5.1. Experiment Setup

We employ speech material from the 2002 NIST Speaker Recognition Evaluation corpus [8]. The speaker data is drawn from a randomly selected set of 50 speakers (20 males and 30 females). Each speaker has a roughly 2-minute long telephone recording sampled at 8 kHz. It is divided into 5s long pieces, and 4 of them are used as test utterances and others for training. Totally there are 200 test utterances. To study how the proposed system performs under different types of noisy conditions, the test utterances are mixed with multitalker babble noise (nonstationary), speech shape noise (stationary), and factory noise (nonstationary). Each noise is mixed with test utterances at various SNR levels from -12 dB to 18 dB at 6 dB intervals.

A gammatone filterbank with 64 channels is applied to decompose input signal into cochleagram, and 64-dimensional GF is extracted from it to model speaker dependent characteristics. To reconstruct the noisy GF, a speech prior with 2048 Gaussian components is trained using all the pooled training data. The reconstructed GF is converted to GFCC using DCT. Each speaker is modeled by a 64-component GMM using HTK.

We find the lower 23-order GFCCs largely retain the information in 64-dimensional GFs. This is due to the "energy compaction" property of DCT [7]. Additionally, the 0^{th} cepstral coefficient corresponds to the energy of the whole frame, and is susceptible to noise corruption. Our experiments using the IBM for separation show that removing the 0^{th} coefficient improves the SID performance significantly. Hence, in our experiments we will use 22-dimensional GFCCs.

Considering that the SNR estimator tends to overestimate the input SNR, we set the threshold to 7.5 dB, which can reliably separate individual mixtures into a high-SNR set that selects the reconstruction module for SID and a low-SNR set that selects the marginalization module.

5.2. Evaluation Results

Table 1 shows the SID results of the two modules individually and the combined system for three noises at different SNRs. By assigning the input signal to different modules, the combined system is able to take advantage of both modules. In every noise condition, the combined system outperforms each individual module on average.

We should point out that, in terms of computational complexity, the reconstruction module is faster as it uses 22-dimensional GFCC features, as opposed to 64-dimensional GF features used in the marginalization module. Also, the integration operation in bounded marginalization takes time. These factors lead to the reconstruction module taking about 1/3 of the computing time of the marginalization module.

Babble	**-12 dB**	**-6 dB**	**0 dB**	**6 dB**	**12 dB**	**18 dB**	**Avg.**
REC	3.5	14.5	47	**76.5**	**83**	84.5	51.5
MAR	**4.5**	**16**	**52**	61	76.5	**85.5**	49.25
CMB	4.5	16	52	75.5	82.5	84.5	52.5

Factory	**-12 dB**	**-6 dB**	**0 dB**	**6 dB**	**12 dB**	**18 dB**	**Avg.**
REC	7.5	37	72	**86.5**	**90.5**	**93.5**	64.5
MAR	**21.5**	**49**	**75**	84	88.5	90.5	68.08
CMB	21.5	49	77	85.5	90.5	93.5	69.5

SSN	**-12 dB**	**-6 dB**	**0 dB**	**6 dB**	**12 dB**	**18 dB**	**Avg.**
REC	6	26.5	61	**86.5**	**89.5**	**90.5**	60
MAR	**15.5**	**46**	**63**	77	81.5	83.5	61.08
CMB	15.5	46	63.5	84.5	89	90	64.75

Table 1: SID accuracy (%) of the proposed methods. REC denotes the reconstruction module, MAR the marginalization module and CMB the combined system.

Next we compare our system with several baseline systems. Table 2 gives the SID results of the combined system along with the three baselines. GFCC_22 is a baseline method that uses raw 22-dimensional GFCC features without CASA-based reconstruction [12]. MFCC_22 is the baseline using 22-dimensional MFCC features. It achieves the best performance among a variety of MFCC baselines we have tested. ETSI-AFE_D denotes 12-dimensional ETSI-AFE features plus their delta features. The combined system's SID results are nearly more than 20 percentage points higher than those of MFCC and ETSI-AFE_D baselines. The gain over the GFCC baseline is smaller, reflecting the robustness of GFCC features themselves. Note that, as shown in Table 1, the reconstruction and the marginalization module each already outperforms the baseline systems.

Method	**Babble**	**Factory**	**SSN**	**Average**
Combined System	52.5	69.5	64.75	62.25
GFCC_22	43	53.58	56.33	50.97
MFCC_22	40	38.67	38.75	39.14
ETSI-AFE_D	39.5	45.58	44	43.03

Table 2: SID accuracy (%) of the combined system and baselines. Performance is averaged across different SNR conditions.

5.3. Comparison with a Related System

Pullella et al. [9] recently proposed a system for robust speaker recognition, which also utilizes bounded marginalization to achieve noise robustness. The difference from our marginalization module lies in two aspects. First, we use the gammatone filterbank as the front-end followed by decimation to derive GF features. They use a mel-scale filterbank as the front-end. The second difference is in mask estimation. They compute a binary mask using spectral subtraction, and then feature selection to refine the initial mask. As described earlier, our system uses CASA-based speech segregation to directly estimate the IBM.

Our comparison uses the same experimental setup as in [9]. The speech signals are from the TIDigits corpus [6]. Test utterances are corrupted by white noise and factory noise at -5, 0, 5, 10, 15, and 20 dB.

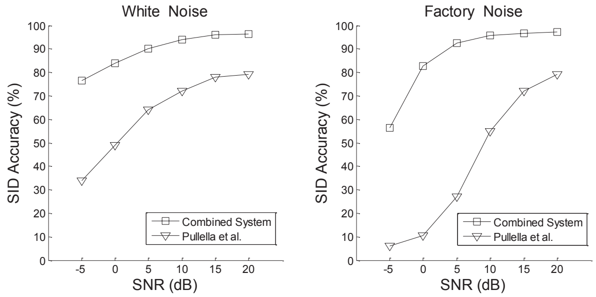

Figure 2. SID accuracy (%) comparisons of the proposed combined system and Pullella et al.'s system.

Figure 2 shows the SID performances of the proposed combined system and Pullella et al.'s system with their respective methods of mask estimation. The comparison shows that our combined system performs much better than their system in both noise conditions, particularly at lower SNR levels. While our system's performance does not vary a lot for the two noises, their system performs considerably worse in the factory noise, presumably because of the ineffectiveness of spectral subtraction for attenuating this nonstationary noise.

6. CONCLUDING REMARKS

Our earlier work used the speech separation and recognition corpus (SSC) [2] as our test data [12], and achieved large performance gains. However, we have found that such gains are somewhat inflated by the large lexicon overlap between training and test material. The SSC corpus has a small vocabulary. Each sentence has a fixed grammar and every word appears in both training and testing data. This situation is similar to the TIDigits corpus discussed in Sec. 5.3. On the other hand, the NIST corpus is a standard dataset for speaker recognition, which is much closer to practical situations.

IBM estimation plays a key role in the SID results, as underscored by the comparison in Figure 2. CASA based speech segregation provides a promising direction to improve SID robustness. As pitch tracking and mask estimation continue to improve, further elevation in SID performance in adverse acoustic conditions can be expected.

To conclude, we have proposed a novel system for robust speaker identification in noisy conditions. By using CASA masks for speech segregation, we can either reconstruct or marginalize unreliable T-F units. Our systematic evaluations show that the reconstruction and marginalization methods and a combined system achieve significant performance improvements over alternative SID systems. It is important to note that our proposed system does not require pretrained noise models, and as a result it is expected to generalize well to noise types not tested in this paper.

Acknowledgements. This research was supported in part by an AFRL grant (FA8750-04-1-0093), AFRL as a subcontractor to RADC Inc. (FA8750-09-C-0067), and an AFOSR grant (FA9550-08-1-0155).

7. REFERENCES

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

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

[3] 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.

[4] Z. Jin and D.L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, pp. 625-638, 2009.

[5] Z. Jin and D.L. Wang, "A multipitch tracking algorithm for noisy and reverberant speech," in *Proc. ICASSP*, pp. 4218-4221, 2010.

[6] R.G. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP*, pp. 328-331, 1984.

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

[8] M. Przybocki and A. Martin, "The NIST Year 2002 Speaker Recognition Evaluation Plan,". Available: http://www.itl.nist.gov/iad/mig/tests/sre/2002/2002-spkrec-evalplan-v60.pdf, 2002.

[9] D. Pullella, M. Kühne, and R. Togneri, "Robust speaker identification using combined feature selection and missing data recognition," in *Proc. ICASSP*, pp. 4833-4836, 2008.

[10] 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.

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

[12] Y. Shao, and D.L. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proc. ICASSP*, pp. 1589-1592, 2008.

[13] 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.

[14] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines,* P. Divenyi, Ed., Norwell MA: Kluwer Academic, pp. 181-197, 2005.

[15] D.L. Wang and G.J. Brown, Ed., *Computational auditory scene analysis: Principles, algorithms, and applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.