A NOVEL NOISE ROBUST SPEAKER IDENTIFICATION SYSTEM

Fincy Francis[1] and Vishnu Rajan[2]
[1]Embedded System, Sahrdaya College of Engineering and Technology, India
[2]Department of ECE, Jodhpur National University, Rajastan, India
E-Mail: fincythattil@gmail.com

ABSTRACT

Thereare many Speaker Identification algorithms are available today where different auditory feature and extraction techniques are used, but we don't have a method can perform in all acoustic conditions. Here our aim is to develop a high performance and noise robust speaker identification system. The MFCC and GFCC feature components combined are suggested to improve the reliability of a speaker recognition system. The MFCC based speaker recognition provides high accuracy and it is a low complex systems; However they are not very robust at the presence of additive noise and in various different acoustic condition. The GFCC features in recent studies have shown very good robustness against noise and acoustic change. Here we proposing an idea that to integrate both MFCC & GFCC features to improve the overall ASR system performance in low signal to noise ratio (SNR) conditions.

The another aim of this thesis is a detailed evaluation of the parameters used in Automatic speech recognition system such as frame size, number of Gaussian mixtures and GMM technique. In this paper we propose more advanced technique for speaker model creation by using GMM-UBM in order to reduce the error during processing time.

The experiment are conducted on the English Language Speech Database for Speaker Recognition (ELSDR) databases. In order to find out the performance of the system, the test utterances are mixed with noises at various SNR levels to simulate the channel change. The results provide an analytical comparison between MFCC, GFCC and MFCC-GFCC combined features.

Keywords: MFCC features, GFCC feature, combined system.

INTRODUCTION

Speaker recognition is the process of automatically recognizing who is speaking by using the speaker specific information included in speech waves to verify identities being claimed by people accessing systems. By using this technology we can able to make access control for various services by voice. Applicable services include telephone shopping, voice dialing, information and reservation services, voice mail, security control for highly confidential information, database access services, banking over a telephone network and remote access to computers. Speaker recognition technology can be used as a forensics tool.

For real world application noise robust automatic speech recognition systems are essential .We have to remove additive noise, room reverberation and channel/handset variations from the received noisy speech signal. Improving the noise robustness has been a research task for many years.

To reduce the mismatch between training and test conditions [1] speakers can be modeled in multiple noisy environments. CurrentlySpeech enhancement methods are spectral subtraction [2], noise-robust speaker recognition .Computational auditory scene analysis (CASA) can be used to remove noise .Speaker features such as modulation spectral features and those incorporating phase information have shown robustness against reverberation. The Blind DE reverberation algorithms have been used to restore the anechoic signal or the early reflections of reverberant speech. Borgstrom and McCree modeled the effect of reverberation as a channel-wise convolution of short-time spectral envelopes. The National Institute of Standards and Technology (NIST) has conducted a series of speaker recognition evaluations (SRE) since 1996.State-of-the-art systems include joint factor analysis and i-vector based techniques. DEEP neural networks (DNNs) [3] have been adopted in many Automatic Speech Recognition (ASR) systems [5], [7]. Large performance improvements have been reported compared to systems that use Gaussian Mixture Models (GMMs). For noisy speech recognition, DNNs have also obtained comparable performance to the best GMM system with various noise reduction, feature enhancement and model-based compensation methods. However, DNNs are still far from reaching humans' expectations and few methods have been developed to further improve DNNs' noise robustness .To a certain extent, DNNs may be capable of learning some noise-dependent feature normalization effects implicitly through multiple layers of non-linear transformations.

The performance of the automatic speech recognition systems are drops significantly as speech is distorted by interference [9]. The task of improving the robustness of such systems is known as robust speaker recognition. Speech enhancement methods have been explored to achieve noise robustness [13].An alternative approach seeks to improve robustness by modeling noise and combining it with clean speaker models [14].

One of the problems still faced in Speaker Recognition is dealing with the intra-speaker variations. Such variations can arise for multiple reasons such as: recording conditions, environment or mood, etc. One cannot assume a speaker can repeat an utterance in the same manner from trial to trial, that's where score normalization comes into place. Newer techniques have been proposed to normalize using the Z-Score, subtracting the mean and dividing by the standard deviation of the imposter score distribution.

FRONT END PROCESSING AND BLOCK DIAGRAM

VAD

Voice Activity Detection is the process of extracting out silence part from the speech signal otherwise the training might be seriously biased. We can use simple energy based approach to remove the silence part. In this method the frames having average energy is below 0.01 times the average energy of the whole utterance is identified and removed.

Feature Extraction

The auditory features witch shows the highest performance to the ASR systems are GF and GFCC than the other auditory features like MFCC.

The Gammatone Frequency Cepstral Coefficients (GFCC) are auditory feature based on a set of Gammatone filter banks. The GFCC is calculated from the Cochleagram of Gammatone Filter bank. Gammatone Frequency vector can be generated by rectification of each frame of the cochleagram using the cubic root operation. GFCC can be derived from GF by applying discrete cosine transform on it [15].

MelFrequency Cepstral Coefficient is a representation of the short term power spectrum of a sound. Which is based on a linear cosine transform of a log power spectrum on a nonlinear melscale of frequency.

Linear predictive coding is a tool mainly used in audio signal processing and speech processing. To representing the spectral envelope of a digital signal of speech in compressed form, we using the information of a linear predictive model. The basic assumption in LPC is that, in a short period, the *n* th signal is a linear combination of previous *p* signals.

$$\mathrm{x(n)} = \sum_{i=1}^{p} \mathrm{ai\ x((n-1)}$$

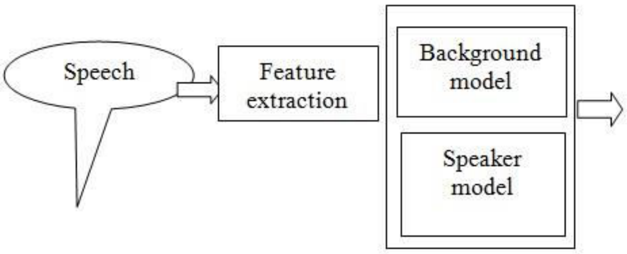


Figure-1. Block diagram of speaker identification system.

Universal Background Model (ubm)

The task to detect a speaker could be defined as two hypothesis tests. The first test is the one in which the speech signal *Z* does come from the hypothesized speaker and the second one where it does not come from the hypothesized speaker.

The *likelihood* of the hypothesis *Hi* given the speech signal can be defined as the probability density function *p* (*Z* | *H i)*. Then we can use a likelihood ratio test given by the two hypotheses to determine the decision. For text independent speaker recognition the most successful model for the creation of likelihood ratio is the Gaussian mixture models (GMM). A GMM could be thought of as a Gaussian distribution describing a one dimensional random variable *X*. The variable *X* is defined as a vector described by the mean and variance. The mixture density for a feature vector, *X* can be defined as:

$$\mathrm{p\ (X \mid \lambda)} = \sum_{i=1}^{m} \mathrm{wi\ pi\ (X)z}$$

This mixture density is a weighted linear combination of unimodal Gaussian densities, (*X*)

$$pi\ (X) = \frac{1}{2\pi^{D/2} \mid \Sigma_{i\mid}\, 1/2}\, e^{!/2\,(x-\mu i)^{T}} \sum_{i}^{-1} (x - \mu i)$$

The UBM is trained using the Expected-Maximization (EM) algorithm. The EM algorithm refines the parameters of the GMM iteratively to increase the likelihood of the estimated model for the feature vectors being observed.

Speaker Model Adaption

The speaker-specific model is adapted from the UBM using the maximum a posteriori (MAP) estimation. The adaptation increases the performance and provides a tighter coupling between the two models.

According to the alignment of the training vectors to the UBM can be computed as follows

$$\mathrm{pr\ (i \mid xt)} = \frac{\mathrm{wi\ pi\ (xt)}}{\sum_{j=1}^{m} wj\ pj\ (xt)}$$

RECOGNITION METHODOLOGY AND SYSTEM DESIGN

Previous studies have shown the accuracy of MFCC under low noise conditions and the robustness of GFCC in noisy environments. It would be beneficial to incorporate the benefits of these two approaches, to reduce or eliminate their individual drawbacks.

Speaker Combined Feature Representation

The strategy we are proposing allows us to combine the feature vector of MFCC and GFCC and use PCA to reduce the feature dimension and remove correlations.

The front-end block diagram of the system is depicted on Figure 1. The system is subdivided into two different subsystems: MFCC and GFCC. Both systems will be running in parallel during the training and test phases. The output of these systems is aggregated and processed using statistical PCA.

These *principal components* are a linear combination of the optimally-weighted observed variables. These optimum basis vectors are the eigenvectors of the covariance matrix of the distribution.

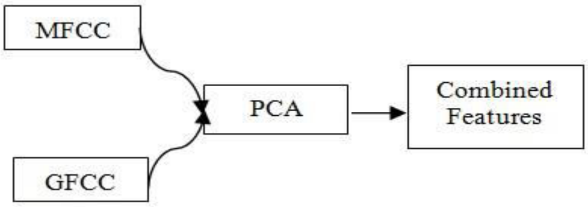


Figure-2. The combined feature representation front-end block diagram.

Experimental Setup

During the evaluation phase, each test segment is scored against the background model and a given speaker model to accept/reject the claim. The same set of tests is performed on both corpora.

The experiment extracts 12-dimensional MFCCs from a pre-emphasized speech signal, mean and variance normalization and writes them to disk in HTK format. The second stage extracts 12-dimensional GFCC's from the same speech signal and stores it to the disk in HTK format as well. The last stage uses the output of the MFCC and GFCC as the input to the PCA function. To complete the experiment, the following steps are executed: UBM training, MAP adaptation, scoring of the verification trials, and computing the performance measures.

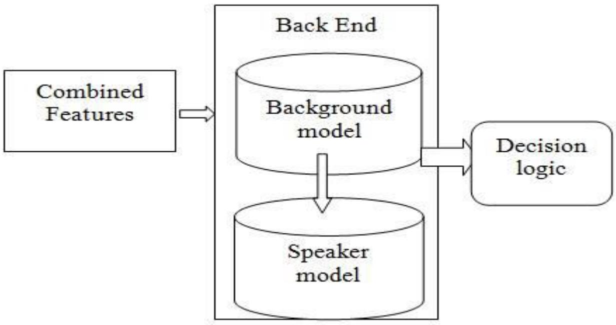


Figure-3. Block diagram of combined features experiment.

For evaluating the performance of the new features in noise, the white Gaussian noise is added to the speech signal in different SNRs from -30dB to 0 dB, respectively.

UBM Training

For the ELSDSR background model training, 18 (8 female and 10 male) speakers were selected. The remaining 4 speakers are used for the test trials. The verification trials consist of 16 trials (4 target vs 12 impostor trials). The GMM was trained using 256 GMM components.

MAP Adaptation

This stage adapts the speaker specific GMM from the UBM using maximum *a posteriori* (MAP) estimation. A MAP adaption relevance factor of 8.0 was used. The ELSDSR Corpus consisted of nine sentences and only eight were used.

Scoring Verification Trials

The verification scores for trials are computed as the log-likelihood ratio between the speaker models and the UBM given the test observations. The National Institute of Standards and Technology (NIST) has created a set of standard performance metrics to score ASR systems.

In statistical hypothesis testing there typically two types of errors, *false positives* and *false negatives*, often considered false alarms. A false positive is when a system incorrectly verifies an impostor as the target during the verification impostor trials. On the other hand a false negative is when the system determines the target as an impostor during the verification target trials. These types of errors are often referred to as false alarms and misses respectively.

DCF is defined as

$$DCF = C_m \, x \, P_{m|t} \, x \, P_t \, x \, C_{fa} \, x \, P_{fa|i} \, x \, (1 - P_t)$$

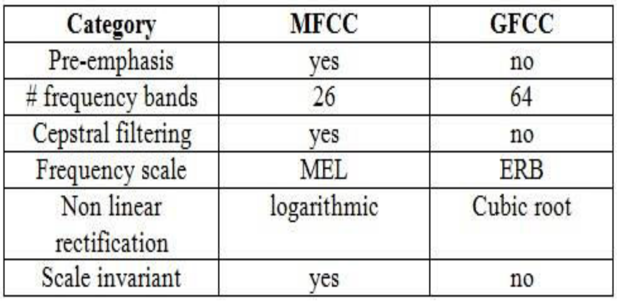
where C_m represent the cost of a miss, $P_{m|t}$ the prior probability of a miss given a target trial, P_t the prior probability of a target trial, C_{fa} the cost of a false alarm and $P_{fa|i}$ the prior probability of a false alarm given an impostor trial. Typical parameter values for the NIST evaluations are $C_m = 10$, $C_{fa} = 1$ and $P_t = 0.01$.

The NIST evaluations have also required the systems to produce a score along with the decision, where higher scores indicate greater likelihood that the correct decision is "true". A very informative way of presenting the system performance is a liner plot of both error rates on a normal scale, denoted by the NIST as the Detection Error Tradeoff (DET) curve.

The resulting curve is linear when the underlying error rates are normal. The Equal Error Rate (EER) is the critical operating area of the curve where the error rates (False Alarms and Misses) are equal.

EVALUATION AND COMPARISON

Table-1. Differences between MFCC and GFCC.



Category	MFCC	GFCC
Pre-emphasis	yes	no
# frequency bands	26	64
Cepstral filtering	yes	no
Frequency scale	MEL	ERB
Non linear rectification	logarithmic	Cubic root
Scale invariant	yes	no

Table-1 shows the difference between different auditory features.

The experiment are conducted on the English Language Speech Database for Speaker Recognition (ELSDR) databases, were the test utterances are mixed with noises at various SNR levels to simulate the channel change.

The English Language Speech Database for Speaker Recognition (ELSDSR) corpus is a dataset designed to provide speech data for the development and evaluation of ASR. Each utterance was recorded to a 16-bit PCM waveform with a sampling frequency of 16 KHz. The suggested training data for each speaker was created with seven paragraphs of text, which contained 11 sentences for a total of 154 utterances collected. The suggested test data was created with two sentences, 44 utterances. Table shows the time duration for both training and test individually.

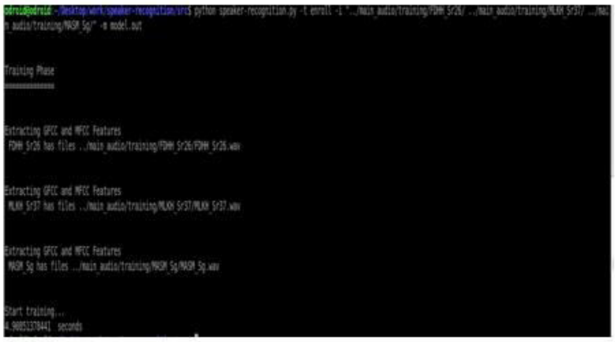
Table-2. ELSDR duration of reading training and test material.

No.	ID	Training Time(sec)	Testing Time(sec)
	Male		
1	MASM	81.2	20.9
2	MCBR	68.4	13.1
3	MFKC	91.6	15.8
4	MKBP	69.9	15.8
5	MLKH	76.8	14.7
6	MMLP	79.6	13.3
7	MMNA	73.1	10.9
8	MNHP	82.9	20.3
9	MOEW	88.0	23.4
10	MPRA	86.8	9.3
11	MREM	79.1	21.8
12	MTLS	66.2	14.05
Average		78.6	16.1
	Female		
13	FAML	99.1	18.7
14	FDHH	77.3	12.7
15	FEAB	92.8	24.0
16	FHRO	86.6	21.2
17	FJAZ	79.2	18.0
18	FMEL	76.3	18.2
19	FMEV	99.1	24.1
20	FSLJ	80.2	18.4
21	FTEJ	102.9	15.8
22	FUAN	89.5	25.1
	Average	88.3	19.6
	Total	1826.6	389.55

The hardware used for testing the performance of this proposed speaker Identification was the Odroid C1 board. The entire algorithm implemented in Odroid C1 by using Python language.

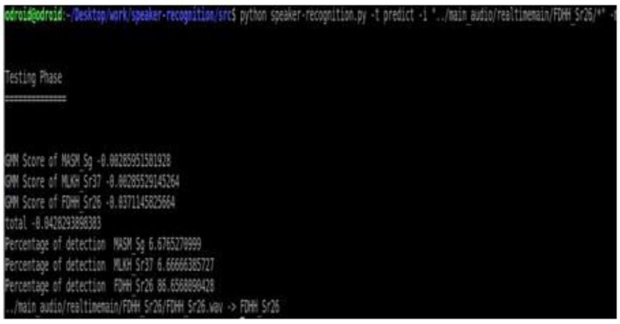
The two main stages in speaker Identification algorithm are Training phase and Testing Phase. In training phase the database of pure speech signal having studio quality is created. In enrollment phase, the speaker's voice is analyzed, then a number of features are extracted to create a voice model of the speaker. Here we extracts both auditory features, GFCC and MFCC from each speech signal. Speaker models can be created by using GMM and MAP adaptation algorithm for each person.

During testing phase, a speech signal which is collected from a real time environment is allowed to perform all the tasks in enrollment phase. Additionally for verification it uses the voice model previously created to compare against a speech utterance.



```
odroid@odroid:~/Desktop/work/speaker-recognition/src$ python speaker-recognition.py -t enroll -i "../main_audio/training/FDHH_Sr26/ ../main_audio/training/MLKH_Sr37/ ../mai
n_audio/training/MASM_Sg/" -m model.out


Training Phase
==============


Extracting GFCC and MFCC Features
 FDHH_Sr26 has files ../main_audio/training/FDHH_Sr26/FDHH_Sr26.wav


Extracting GFCC and MFCC Features
 MLKH_Sr37 has files ../main_audio/training/MLKH_Sr37/MLKH_Sr37.wav


Extracting GFCC and MFCC Features
 MASM_Sg has files ../main_audio/training/MASM_Sg/MASM_Sg.wav


Start training...
4.90853378441  seconds
```

Figure-4. Hardware terminal, performs training phase.



```
odroid@odroid:~/Desktop/work/speaker-recognition/src$ python speaker-recognition.py -t predict -i "../main_audio/realtimemain/FDHH_Sr26/*" -


Testing Phase
=============


GMM Score of MASM_Sg -0.00285951581928
GMM Score of MLKH_Sr37 -0.00285529145264
GMM Score of FDHH_Sr26 -0.0371145825664
total -0.0428293898383
Percentage of detection  MASM_Sg 6.6765270999
Percentage of detection  MLKH_Sr37 6.66666385727
Percentage of detection  FDHH_Sr26 86.6568890428
../main_audio/realtimemain/FDHH_Sr26/FDHH_Sr26.wav -> FDHH_Sr26
```

Figure-5. Hardware terminal, performs testing phase.

Table-3. Summary of the EER achieved for each feature extraction technique.

SNR(dB)	EER % MFCC	EER % GFCC	EER % Combined	DCF
-30	49.929	25	25.303	10
-15	38.859	24	22.848	9.97
-10	27	22.879	17.121	9.49
-5	16	17.252	13.818	7.19
0	12	13.33	10.475	6.27

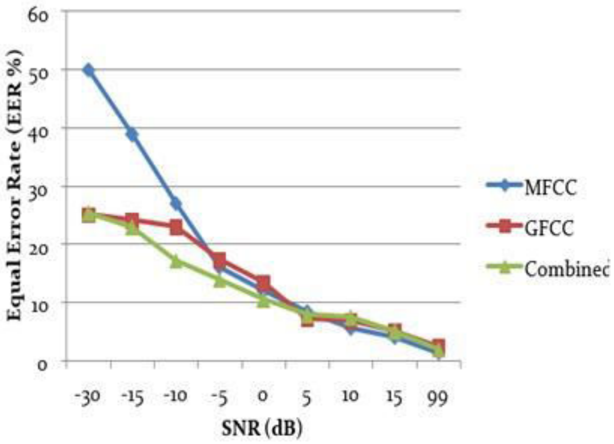


Figure-6. Final total equal error rates for the test trials.

An important finding in our study is that GFCC features outperform conventional MFCC features under noisy conditions. We have conducted an in-depth study on the noise robustness of GFCC and MFCC features. Our experiments first confirm the superior robustness of GFCC relative to MFCC exists on a new corpus. By carefully examining all the differences between them, we conclude that the nonlinear rectification mainly accounts for the noise robustness differences. In particular, the cubic root rectification provides more robustness to the features than the log. In a noisy mixture, there are target dominant T-F units or segments indicative of this energy information. The cubic root operation makes features scale variant (i.e. energy level dependent) and helps to preserve this information. The log operation, on the other hand, does not encode this information. Although the combined system in this chapter significantly outperforms the individual modules .The simple combination strategy in seems to lose its advantage when the performance profiles of the individual modules are similar. In such situations, more sophisticated methods of classifier combination may be needed.

CONCLUSIONS

The performance of Speaker recognition systems has improved due to recent advances in speech processing techniques but there is still need of improvement. In this paper we present the comparison of different parameters used in automatic speech recognition system to increase the accuracy of the system**.** Here w**e** proposing a combined approach for feature extraction and compared with MFCC and GFCC feature extractions algorithms.

The proposed combination feature methodology has shown satisfactory versatility and robustness under ELSDSR dataset. The final results in Table III shows that for the SNR levels tested overall there were significant improvement against the single feature counterparts. The highest improvement against MFCC was found at the -30dB range in which the EER improved 49%.

The results also show that the combined MFCC-GFCC is indeed a viable method to improve recognition rates at low SNR levels.

REFERENCES

[1] Ji Ming, Timothy J. Hazen,James R. Glass and Douglas A. Reynolds. 2007. “Robust Speaker Recognition in Noisy Conditions,” IEEE Transactions On Audio, Speech, And Language Processing, Vol. 15, No. 5.

[2] Ning Wang, P. C. Ching, Nengheng Zheng and Tan Lee. 2011. “Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features,” IEEE Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 1.

[3] Bo Li, Khe Chai Sim. 2014. “A Spectral Masking Approach to Noise-Robust Speech Recognition Using DeepNeural Networks” IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 22, No.8.

[4] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel and Pierre Ouellet. 2011. “Front-End Factor Analysis for Speaker Verification,” IEEE Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 4.

[5] Michael I. Mandel, Scott Bressler, Barbara Shinn-Cunningham and Daniel P. W. Ellis. 2010. “Evaluating Source Separation Algorithms With Reverberant Speech,” IEEE Transactions On Audio, Speech, And Language Processing, Vol. 18, No. 7.

[6] Tiago H. Falk and Wai-Yip Chan. 2010 “Modulation Spectral Features for Robust Far-Field Speaker Identification,” IEEE Transactions On Audio, Speech, And Language Processing, Vol. 18, No. 1.

[7] Tobias May, Steven van de Par, and Armin Kohlrausch. 2012. “Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling,” IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, No. 1.

[8] William Hartmann, Arun Narayanan, Eric Fosler-Lussier and DeLiang Wang. 2013. “A Direct Masking Approach to Robust ASR,” IEEE Transactions On Audio, Speech, And Language Processing, Vol. 21, No. 10.

[9] Gong Y. 2002. Noise-robust open-set speaker recognition using noise-dependent Gaussian mixture classifier. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I-133-I-136.

[10] Yuxuan Wang, Kun Han and DeLiang Wang. 2013. “Exploring Monaural Features for Classification-Based Speech Segregation,” IEEE Transactions On

Audio, Speech, And Language Processing, Vol. 21, No. 2.

[11] Yuxuan Wang and DeLiang Wang. 2013. “Towards Scaling Up Classification-Based Speech Separation,” EEE Transactions On Audio, Speech, And Language Processing, Vol. 21, No. 7.

[12] Zhaozhang Jin and DeLiang Wang. 2011. “Reverberant Speech Segregation Based on Multipitch Tracking and Classification,” IEEE Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 8, November 2011.

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

[14] Matsui T., Kanno T. and Furui S. 1996. Speaker recognition using HMM composition in noisy environments. Computer Speech & Language, Vol. 10, No. 2, pp. 107-116.

[15] Xiaojia Zhao, Yuxuan Wang and DeLiang Wang. 2014. “Robust Speaker Identification in Noisy And Reverberant Conditions’ IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 22, No. 4.