ANALYZING NOISE ROBUSTNESS OF MFCC AND GFCC FEATURES IN SPEAKER IDENTIFICATION

Xiaojia Zhao[1] *and DeLiang Wang*[1,2]

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

ABSTRACT

Automatic speaker recognition can achieve a high level of performance in matched training and testing conditions. However, such performance drops significantly in mismatched noisy conditions. Recent research indicates that a new speaker feature, gammatone frequency cepstral coefficients (GFCC), exhibits superior noise robustness to commonly used mel-frequency cepstral coefficients (MFCC). To gain a deep understanding of the intrinsic robustness of GFCC relative to MFCC, we design speaker identification experiments to systematically analyze their differences and similarities. This study reveals that the nonlinear rectification accounts for the noise robustness differences primarily. Moreover, this study suggests how to enhance MFCC robustness, and further improve GFCC robustness by adopting a different time-frequency representation.

Index Terms— speaker identification, MFCC, GFCC, noise robustness, speaker features

1. INTRODUCTION

Automatic speaker recognition systems perform very well in certain conditions, e.g. without noise, room reverberation, or channel variations. However, such conditions can hardly be met in practice. Real acoustic environments present various challenges to speaker recognition systems. Robustness of speaker recognition systems must be addressed for practical applications.

One challenge is channel/session variation. Recent NIST speaker recognition evaluations (SRE) have mainly focused on addressing the problem of channel variations in speaker verification. State-of-art systems use techniques such as joint factor analysis and i-vector based probabilistic linear discriminant analysis [1], [2]. Another challenge is additive noise. In our daily listening environments, speech often occurs simultaneously with noise sound. To improve noise robustness, Ming *et al.* propose to train speaker models in multiple noisy conditions to alleviate the mismatch with noisy test conditions [3]. Alternatively, one can employ speech enhancement algorithms to clean up noisy speech prior to speaker recognition.

The human ability to perform speaker recognition in noisy conditions has motivated studies of robust speaker recognition from the perspective of computational auditory scene analysis. In one such study [4], we showed that a new speaker feature, *gammatone frequency cepstral coefficients* (GFCC), shows superior noise robustness to commonly used *mel-frequency cepstral coefficients* (MFCC) in speaker identification (SID) tasks (see also [5]). As for the reason, we speculated that the front-end of GFCC, the gammatone filterbank, might be more noise-robust than that of MFCC, the mel-filterbank. In particular, the frequency scales employed in the two filterbanks were believed to be the key difference although no convincing evidence was presented to support this hypothesis.

Why does GFCC appear to be more robust than MFCC, at least for speaker identification? We believe this is an important question for the study of noise-robust speaker recognition. This study was designed to not only answer this question but also gain a deep understanding of the intrinsic noise robustness of GFCC and MFCC features. In this study, we first analyze all of their differences, which helps us to generate a number of assumptions. For each assumption, we design a corresponding set of experiments to test it. In this way, we are able to narrow down possible explanations, which eventually reveal the desired answer.

The rest of the paper is organized as follows. Section 2 describes the detailed differences between GFCC and MFCC. Section 3 presents possible reasons and experimental validations, followed by further exploration in Section 4. We conclude this paper in Section 5. The relationship of this study with prior work is discussed in Section 6.

2. DIFFERENCES IN MFCC AND GFCC DERIVATIONS

MFCC is widely used in both speech and speaker recognition. The HTK toolkit is frequently used to derive MFCC [6], as is in [4]. Therefore, we take a closer look at the HTK version MFCC generation process.

MFCC Extraction (HTK version):

1. Pre-emphasize input signal
2. Perform short-time Fourier analysis to get magnitude spectrum
3. Wrap the magnitude spectrum into mel-spectrum using 26 tri-

The research described in this paper was supported in part by an AFOSR grant (FA9550-12-1-0130).

angular overlapping windows where center frequencies of the windows are equally distributed on the mel scale
4. Take the log operation on the power spectrum (i.e. square of mel-spectrum)
5. Apply the discrete cosine transform (DCT) on the log-mel-power-spectrum to derive cepstral features
6. Perform cepstral liftering

The detailed process of GFCC extraction is listed as follows [4].

GFCC Extraction:

1. Pass input signal through a 64-channel gammatone filterbank
2. At each channel, fully rectify the filter response (i.e. take absolute value) and decimate it to 100 Hz as a way of time windowing. Then take absolute value afterwards. This creates a time-frequency (T-F) representation that is a variant of cochleagram [7]
3. Take cubic root on the T-F representation
4. Apply DCT to derive cepstral features

Broadly speaking, there are two major differences. The obvious one is the frequency scale. GFCC, based on equivalent rectangular bandwidth (ERB) scale, has finer resolution at low frequencies than MFCC (mel scale). The other one is the nonlinear rectification step prior to the DCT. MFCC uses a log while GFCC uses a cubic root. Both have been used in the literature. In addition, the log operation transforms convolution between excitation source and vocal tract (filter) into addition in the spectral domain. Besides these two major differences, there are some other notable differences that are summarized in the following table. Next we analyze each of the differences in Table 1 in detail.

Category	MFCC	GFCC
Pre-emphasis	Yes	No
# of Frequency Bands	26	64
Cepstral Liftering	Yes	No
Frequency Scale	Mel	ERB
Nonlinear Rectification	Logarithmic	Cubic Root
Scale-invariant (w/o 0^{th} coefficient)	Yes	No
Intermediate T-F Representation	Mel Spectrum	Variant of Cochleagram

Table 1: *Differences between GFCC and MFCC*

3. DIAGNOSIS AND EXPERIMENTAL RESULTS

3.1. Experimental setup and benchmark results

The study in [4] was reported on the 2002 NIST SRE dataset with 330 speakers. It is unclear if the noise robustness advantage of GFCC is specific to this dataset. To address this concern, we switch to the TIMIT corpus. Out of the entire 630 speakers, 330 are randomly chosen to match the number of speakers in [4]. Each speaker has 10 utterances and we choose 8 for training and remaining 2 for testing.

We scale clean training and test data to an average sound intensity (see eq. 1 in Sec. 4.3) of 60 dB. Clean test data is then mixed with factory noise from the NOISEX-92 database at a signal-to-noise ratio (SNR) of 0 dB. We use the ideal binary mask (IBM) for voice activity detection [7]. In other words, the frames with at least one reliable T-F unit, labeled as 1 the IBM, are selected for recognition. Twenty-two-dimensional MFCC and GFCC, with 0^{th} coefficient removed, are used for this study. No cepstral mean normalization is performed as the long-term mean of cepstral features is not reliable to represent non-stationary noises such as factory noise. Speakers are modeled using Gaussian mixture models (GMM) adapted from a 1024-component universal background model [8]. No speech separation is performed as the main goal is to evaluate the intrinsic noise robustness of MFCC and GFCC features.

First we establish the benchmark SID performance. The SID performance is shown in Table 2.

Category	MFCC	GFCC
Benchmark performance	3.94	16.36

Table 2: *Benchmark performance of GFCC and MFCC in terms of speaker recognition rates (%)*

As shown in Table 2, the SID performance is rather poor in both cases due to a relatively low SNR and the absence of speech separation. However, it is obvious that GFCC is substantially more noise-robust than MFCC, which is consistent with the findings in [4].

3.2. Difference 1: *Pre-emphasis*

One pre-processing difference between GFCC and MFCC is pre-emphasis. As a common practice, pre-emphasis is adopted in the HTK toolkit by applying a high-pass filter with a setting of [1, -0.97]. High-pass filtering inevitably alters energy distribution across frequencies, as well as the overall energy level. This could have significant impact on the energy-related GFCC features. To verify if pre-emphasis degrades MFCC's noise robustness, we remove it from MFCC and add it to GFCC. The resulting SID performance is shown in the following table.

Category	MFCC	GFCC
w/ Pre-emphasis	3.94 (default)	10.76
w/o Pre-emphais	3.03	16.36 (default)

Table 3: *Impact of pre-emphasis on GFCC and MFCC in terms of speaker recognition rates (%)*

Table 3 suggests that removing pre-emphasis has little impact on MFCC's noise robustness. Adding pre-emphasis to GFCC drops performance as expected, even though the altered GFCC is still substantially more robust. Therefore, pre-emphasis is not the answer.

3.3. Difference 2: *Number of frequency bands/channels*

In the aforementioned MFCC generation, magnitude spectrum is wrapped into 26 mel-bands with triangular overlapping windows. GFCC uses the 64-channel gammatone filterbank as the front-end. It is reasonable to assume that more frequency channels may lead to better noise robustness. We test this assumption by increasing the number of

triangular windows and decreasing the number of channels of the gammatone filterbank. Table 4 presents the results.

Category	MFCC	GFCC
26 Bands/Channels	3.94 (default)	13.33
64 Bands/Channels	2.12	16.36 (default)

Table 4: *Impact of number of frequency bands/channels on GFCC and MFCC in terms of speaker recognition rates (%)*

As shown in Table 4, increasing the number of bands does not improve the robustness of MFCC. Using the 26-channel gammatone filterbank degrades the performance of GFCC by a small amount, but it is still substantially more robust than MFCC. Clearly, the number of frequency bands/channels is not the answer.

3.4. Difference 3: *Cepstral liftering*

A post-processing difference between these two features is cepstral liftering. What it does is to filter cepstral coefficients. A property of DCT operation is "energy compaction" [9], meaning that higher dimensions of cepstral coefficients are numerically very small. This is not a problem, but HTK toolkit purposely amplifies higher dimensional coefficients to balance the magnitudes and variances across dimensions for the sake of displaying parameters, variance flooring, etc. GFCC resolves the same problem by taking the lower 22-dimensional coefficients (without 0^{th} coefficient). Therefore there is no need of cepstral liftering for GFCC. If no cepstral liftering is the reason of GFCC's noise robustness, we expect the robustness of MFCC will be substantially improved by dropping cepstral liftering. The experiments to test this assumption produce results shown in the following table.

Category	MFCC	GFCC
w/ Cepstral Liftering	3.94 (default)	16.36
w/o Cepstral Liftering	3.03	16.36 (default)

Table 5: *Impact of cepstral liftering on GFCC and MFCC in terms of speaker recognition rates (%)*

As illustrated in Table 5, there is no substantial performance change by adding or dropping cepstral liftering for either feature. Hence, cepstral liftering does not appear to be the reason.

3.5. Difference 4: *Frequency scale*

We now evaluate the main hypothesis put forward in [4], frequency scale. It argues that the ERB scale has finer resolution than the mel scale at the low frequency range where speech energy primarily resides. However, recent studies on speaker recognition suggest that high frequency range also contains meaningful speaker information and should not be overlooked [10]. They also show that the linear scale is as robust as the mel scale in certain noisy conditions. This somewhat contradicts the hypothesis. To test this hypothesis, we make the triangular overlapping windows equally distributed on the ERB scale. In the meantime, we make sure the center frequencies of the gammatone filters are equally distributed on the mel scale. If this hypothesis holds true, we expect a performance boost from MFCC and a performance drop for GFCC. Results are presented in Table 6.

Category	MFCC	GFCC
Mel Scale (26 bands)	3.94 (default)	17.88
ERB Scale (64 bands)	3.33	16.36 (default)

Table 6: *Impact of frequency scale on GFCC and MFCC in terms of speaker recognition rates (%)*

The results in Table 6 show that changing the frequency scale does not degrade GFCC's robustness. At the same time, it does not improve MFCC's robustness, either. Note that replacing ERB scale with mel scale even slightly improves the performance of GFCC. It appears that mel scale is a better scale. Hence, scale is not the answer.

3.6. Difference 5: *Log vs. cubic root and scale invariance*

Another main difference between the two features is the nonlinear rectification. It is directly correlated with the scale invariance property. Assuming an input signal *s(t)* is scaled by a constant factor of *k*. Due to the linearity of Fourier analysis and triangular window wrapping, this constant factor is carried to the mel-spectrum. By taking the log operation on the mel-power-spectrum, the constant factor becomes an additive term. It is easy to prove that the DCT of a constant term is all 0 except for the 0^{th} coefficient. Therefore, by excluding 0^{th} coefficient (energy related), MFCC is the same no matter how we scale the input signal. GFCC also carries the constant factor *k* into the intermediate T-F representation because of linearity of gammatone filterbank. Nonetheless, the cubic root operation cannot convert the factor into an additive term. As DCT is linear, this cubic root of *k* will be manifested in the cepstral coefficients. This is why GFCC is not scale-invariant. In other words, we do not get the same GFCC if we scale the input signal by a constant factor. This could make a difference in the noise robustness. We create new MFCC by replacing the log with the cubic root and similarly for GFCC by using the log. To be consistent, for the new MFCC, we drop pre-emphasis and cepstral liftering. We add them on the new GFCC. Results are shown in the following table.

Category	MFCC	GFCC
Log	3.94 (default)	3.03
Cubic Root	28.48	16.36 (default)

Table 7: *Impact of nonlinear rectification on GFCC and MFCC in terms of speaker recognition rates (%)*

As demonstrated in Table 7, the noise robustness of MFCC is dramatically improved by taking the cubic root rectification. Meanwhile, GFCC loses its advantage by switching to the log. The nonlinear rectification is therefore a likely reason for GFCC's superior robustness. It is worth pointing out that the new MFCC is even more robust than the regular GFCC, yielding 12% absolute improvement. With both features undergoing the cubic root operation, the last difference would be the T-F representation prior to the DCT. Our results in Table 7 suggest that mel-power-spectrum is a better choice. We examine this in the next section.

4. FURTHER EXPLORATION

4.1. Study of intermediate T-F representation

MFCC comes from mel-power-spectrum where each element represents power/energy of the corresponding T-F unit. The T-F representation of GFCC is a variant of the cochleagram, which is a T-F representation with each element representing energy of the corresponding T-F unit. GFCC is derived by decimating filter responses instead of calculating energy at each T-F unit. The decimation process potentially throws away too much information. We now derive GFCC directly from the cochleagram. In other words, we apply the cubic root rectification on the cochleagram. We also explore different combinations of frequency scales and the number of frequency bands as shown in Table 8.

Configuration	GFCC-decimation	GFCC-cochleagram	MFCC-cubic root
ERB Scale (64 bands)	16.36 (default)	26.36	20.91
ERB Scale (26 bands)	13.33	23.18	26.36
Mel Scale (26 bands)	17.88	**28.48**	**28.48**
Mel Scale (64 bands)	15.91	20.91	24.55

Table 8: *Impact of T-F representation on GFCC and MFCC in terms of speaker recognition rates (%)*

As seen in Table 8, GFCC derived from the cochleagram improves the SID performance to a comparable level with MFCC. Both obtain optimal performance when used together with the mel-scale and 26 bands/channels. This strongly indicates that the advantage of MFCC in Table 7 is mainly due to the intermediate T-F representation.

4.2. Results on other noises and SNRs

All the previous analysis was made only in one noisy condition (an SNR of 0 dB and factory noise). We have performed the same experiments at an SNR of 6 dB with factory noise and get a similar performance profile. Experiments on two new noises, white noise and speech shape noise, further confirm the findings in this paper. All of these experiments have indicated that GFCC is more noise-robust than MFCC due to the nonlinear rectification. The new MFCC using the cubic root operation substantially improves the SID performance and even outperforms GFCC. Deriving GFCC from the cochleagram substantially improves its robustness and produces comparable or better results than the improved MFCC. The optimal number of bands/channels depends on specific noisy conditions. Similar trends have also been observed using the 2002 NIST SRE dataset.

4.3. The scale variance problem

As we pointed out in Section 3.6, GFCC is not a scale-invariant feature. To perform speaker recognition, we need to scale both training speech and clean test speech (not mixture) to a comparable energy level. We use an utterance-level average sound intensity (in dB) as the measurement of overall energy level given in the following equation.

$$E = 10\log 10\left(\frac{\sum_t s^2(t)}{length(s(t))}\right), \quad (1)$$

where E is the average sound intensity and $s(t)$ is the input signal. It is straightforward to scale training data to a desired intensity (e.g. 60 dB). However, it is not trivial to infer the intensity of the underlying target signal in a mixture and scale it to the same level of the training data. There are two ways to address this scale variance issue. One is to first estimate the input SNR. There are reasonably reliable SNR estimation algorithms in the literature. Given a roughly estimated SNR, we can readily infer the energy ratio between the target and interference and calculate the intensity of the target based on the intensity of the mixture. Another way is to estimate the IBM first. We then average the energy-related T-F representation only in reliable T-F units which are dominated by the target. This average can be used to normalize the entire T-F representation prior to the DCT operation. Training data can be processed similarly where all the T-F units are deemed reliable. Both techniques have been shown to be effective in our study.

5. CONCLUDING REMARKS

To conclude, 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.

Why is the cubic root operation better? It might be the case that some speaker information is embodied through different energy levels. 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.

We have shown that by modifying MFCC extraction, substantial noise robustness improvement is obtained. Since MFCC is widely used in automatic speaker and speech recognition, the findings of this paper should shed new light on effective feature representations for noise robustness.

6. RELATION TO PRIOR WORK

This work presented here focuses on the puzzling question raised in a recent study of robust speaker identification [4]: why is GFCC intrinsically more noise-robust than MFCC? The study in [4] gave two hypotheses without experimental validation. The present study confirms the existence of this phenomenon on a new dataset. Then we have evaluated all the differences systematically, and our analysis reveals the rather surprising answer.

7. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis For Speaker Verification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788 - 798, May 2010.

[2] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka and N. Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *Proc. ICASSP*, 2011, pp. 4832-4835.

[3] J. Ming, T.J. Hazen, J.R. Glass and D.A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15(5), pp. 1711-1723, 2007.

.

[4] X. Zhao, Y. Shao and D.L. Wang, "CASA-Based Robust Speaker Identification," *IEEE Trans. Audio, Speech and Language Processing*, vol.20, no.5, pp.1608-1616, 2012.

[5] Y. Shao, S. Srinivasan, and D.L. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. ICASSP*, 2007, pp. 277-280.

[6] S. Young, *et al.*, *The HTK book (for HTK version 3.0).* Microsoft Corporation, 2000.

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

[8] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19-41, 2000.

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

[10] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, S. Shamma, "Linear versus Mel- Frequency cepstral coefficients for speaker recognition", in *IEEE Workshop on ASRU*, 2011, pp. 559-564.