EEM214B Final Project Report: Speaker Verification in Noisy Environment

Shubham Agarwal[1]*, Deepak Muralidharan*[1]

[1]UCLA, Electrical Engineering

shubhamagarwal@ucla.edu, deepakmuralidharan2308@gmail.com

Abstract

Speech allows humans to decipher important information about the speaker. One of them is personal identity. From voice itself, we are able to identify the person we are talking to over phone. Though it seems easy for humans, it has been very difficult for a machine to achieve the accuracy specially in a noisy environment. With wide applications in the field of security, automatic speaker recognition is currently an active area of research in speech processing. In this project we present a comparison of various state-of-the-art algorithms available for speaker verification. Also, we propose a new method using GFCCs coupled with pre and post processing of the speech signal for noise removal which gives better results in a noisy environment.

Index Terms: Speaker recognition, human-computer interaction, Speaker Verification

1. Introduction

Speaker verification (SV) can be defined as the process of validating the claimed identity of an individual using his/her speech. State-of-the-art Speaker Verification systems perform reasonably well when the spoken utterances are clean, i.e., free from any sort of distortions caused by external factors. However, the accuracy of the SV systems start degrading heavily when the speech signals are distorted due to background or environmental noise. Robustness towards noise is crucial for several speaker verification applications, especially in mobile and hand-held devices where the background environments are uncontrolled, time-varying and unpredictable. In this paper, we work on the task of speaker verification in the presence of varying amounts of babble noise (5dB, 10dB and clean speech). We assume that all our experiments are performed in the mismatched conditions i.e. the training data consists of clean, noiseless speech but the test data is noisy. Our aim is to explore noise-robust speaker specific features, pre-processing and post-processing techniques for speech signal to provide a high speaker verification accuracy. All our results are compared against baseline MFCC. Section 2 discusses the pre-processing techniques that were followed to remove noise from the noisy speech in the test set, section 3 discusses about some of the normalization/ post-processing methods that we implemented on our processed features. And in section 4 we mention the different feature extraction approaches that we used. Finally, sections 5 and 6 talk about conclusion, discussion and future work.

Mel Frequency Cepstral Coefficients (MFCC) have been widely used as primary features for speaker recognition in many research papers. The MFCCs generally provide a good representation of the linguistic content of the speech [1]. MFCCs use a mel-scale filter bank, such that the filtering becomes close to human auditory system. Human are better at identifying small changes in pitch at low frequencies than at higher frequencies. MFCCs along with their first and second derivatives have been the state-of-the-art features for speaker recognition for a long time since their inception in 1980s. We treat the prediction accuracy of MFCCs as a baseline. The prediction error obtained using MFCCs is given in Table 1.

SNR	Pred. Error Males (%)	Pred. Error Females (%)
Clean Speech	0	0
10dB	24	28
5dB	56	58

Table 1: Prediction error with MFCC (Baseline)

2. Pre-processing

Since speaker identification systems work well with clean speech, pre- processing the speech signal to remove noise seems to be one way to go forward. To the best of the authors' knowledge, there are two ways to do this, one involves estimating the noise from the speech signal and then removing it. The other way is to apply signal processing techniques like adaptive filtering or wiener filtering which have been widely used in other areas. We tried these algorithms with clean as well as noisy speech data.

2.1. Noise Estimation from signal

First we tried estimating the noise directly from the signal. Since in our speech samples, the speaker starts talking after a short interval of time, we tried to get the noise from the front portion of the speech file. We then subtracted the noise in Fourier domain. With this approach, the results were not very good and we got around 40% error which was much higher than the baseline.

Next we tried taking only the voiced portions of speech for speaker identification. Here also, the results were not very good. This could be because if we consider only the voiced portions, the unvoiced speech segments are removed leading to loss of information.

2.2. Wiener Filtering

Wiener filtering is a widely used algorithm for noise removal in signal processing. Here, we design a filter to have an estimate of the signal without noise. We use a Wiener filter based on Decision-Directed approach proposed by Scalart[2]. The results using this Wiener Filter pre-processing on the MFCCs are given in Table 2.

SNR	Pred. Error Males (%)	Pred. Error Females (%)
Clean Speech	0	0
10dB	24	12
5dB	42	36

Table 2: Weiner pre-processing applied with MFCC

We see that wiener filtering alone reduces the error in prediction significantly for the female speakers. For male speakers, the effect is not that consistent, at 10dB, the error percentage is same but it reduces for 5dB SNR.

3. Normalization/ Post-processing

In speaker verification applications, the information that is extracted from the speech needs to be speaker specific and robust to noise and various effects. Prior literature has examined ways of removing linear effects by applying methods such as Cepstral Mean subtraction[8]. However, under additive noise, the performance of CMS degrades substantially as the average vocal tract configuration information pertaining to the speaker might also also lost.

As an improvement to this, Cepstral Mean and Variance Normalization (CMVN) was proposed which involved normalizing the distribution of single cepstral features by subtracting the mean and scaling by their standard deviation [7]. In our project, we combined CMVN with a relatively new method called feature warping (explained below) which warps the distribution of a cepstral stream to a standardized distribution over a specified time interval, and this provided us with a very high performance boost by around nearly 20%. The algorithms for CMVN and feature warping have been explained below.

3.1. Cepstral Mean and Variance Normalization

- The cepstral coefficients are linearly transformed to have the same segmental statistics (zero mean, unit variance), regardless of the noise condition.
- The method does not require prior knowledge of the noise statistics, it adapts quickly to changing noise conditions, and it is independent of voice activity detection.
- However, as we discussed before, there are some disadvantages. The method does not work well for short utterances and the performance might degrade as noise increases.

3.2. Feature Warping

- As explained in the introduction to this section, this method warps the distribution of a cepstral feature stream to a standardized distribution over a specified time interval.
- The process begins by deriving the complete set of cepstral coefficients from the speech segment.
- Each cepstral coefficient is then analyzed independently as a separate feature stream over time for use in the warping process.
- A (typically three second) window of features is extracted from the feature stream and processed in the warping algorithm to determine a mapped feature for the initial cepstral feature in the middle of the window.

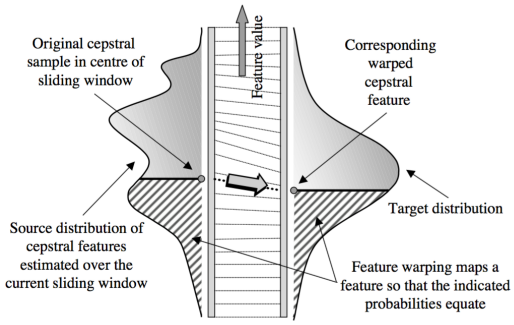

Figure 1: *Warping of features according to target distribution shape*

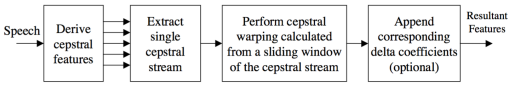

Figure 2: *Block Diagram of the parameterization process*

- The sliding window is shifted by a single frame each time and the analysis is repeated.

3.2.1. Advantages of feature warping

The robustness of the feature processing to slowly changing additive noise characteristics and linear channel effects are attractive traits. Improved warping may be achieved by selection of a more appropriate target distribution that may also be speaker specific. The additional advantage of the approach is that it may be cascaded with other feature enhancement techniques such as some forms of modulation spectrum processing and non-linear neural network mapping approaches.

We found that on an average, all our feature extraction methods (MFCC, GFCC, PLP etc.) gave a performance boost when the CMVN-feature warping post processing/normalization was applied. Hence, we believe that post-processing plays a crucial role in speaker verification in order to notch up the accuracy by a few percentage.

4. Feature Extraction Methods

The feature extraction stage corresponds to the enumeration of knowledge sources in a speech signal. The raw speech signal is reduced to a set of parameters/coefficients where the speaker-specific properties are emphasized and all the redundant information is suppressed. For our task of text independent speaker verification, we have mainly examined low-level features which convey physiological information about the speaker. The ideal feature for speaker verification would be the one which has a high inter-speaker variability and low intra-speaker variability. Though short term spectral features (eg. MFCC) are often preferred for speaker verification due to their high accuracy and real-time extraction, they often degrade very badly in the presence of noise. This motivates us to shift towards noise robust features for speaker verification. Also, we have to remind the reader that all our experiments are performed under a mis-

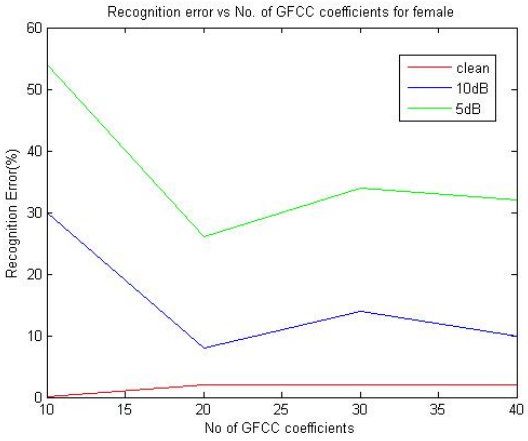

Figure 3: *Accuracy vs No. of GFCC coefficients for female*

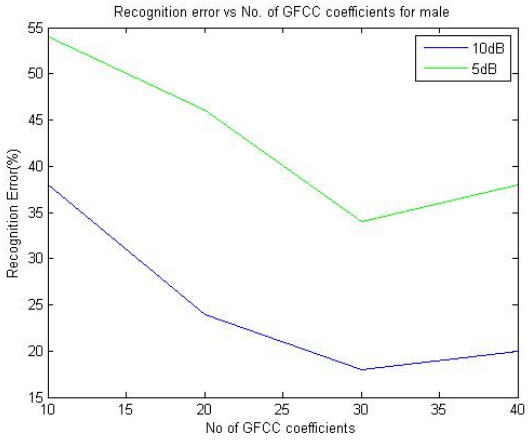

Figure 4: *Accuracy vs No. of GFCC coefficients for male*

matched setting i.e. the training data is clean but the test data contains added noise to it. For classification, we use the Gaussian Mixture Model(GMM) classifier. This is a popular supervised learning classifier for pattern recognition tasks. Having established the background and the motivation, we now give a brief overview of all the feature extraction methods that we tried and the corresponding SV accuracies we got for each of them.

4.1. Pitch

Pitch of a person is directly related to his/her voicing source and have been traditionally considered good for speaker identification. There are a lot of pitch tracking algorithms available in speech. Here, we use Robust Adaptive Pitch Tracking (RAPT) algorithm which claims to be working well for noisy data. RAPT uses Normalized cross-correlation function for pitch calculation which is supposed to perform better on noisy data. For us the pitch feature did not help in the accuracy of speaker recognition. It might be due to the presence of babble noise which adversely affects the low frequency components. Using pitch obtained from RAPT algorithm in Voicebox[9][10], we got around 90% prediction error for both male and female irrespective of the noise level. We also tried extracting the pitch from only the voiced segments but this also did not help much. Owing to this, we decided not to include pitch for further analysis.

4.2. Gammatone Frequency Coefficient Cepstra (GFCC)

Though MFCCs were very good features in recognizing the speech, they were unable to match the accuracy of the human auditory system when presented with noisy data. One of the differences was the mel-frequency bank which was different from what humans use to filter the speech. To mitigate this, GFCCs were introduced which use a gammatone filter bank instead of using the triangular mel-scale filters. We used a 32 channel gammatone filter bank in our experiments. The calculation of GFCC comprises following steps: 1. Pre-emphasis 2. Windowing (we use a hamming window) 3. FFT 4. GTFB 5. Equal Loudness 6. Logarithmic Compression 7. DCT 8. Conversion to GFCC After these steps we get GFCC coefficients. Since the number of coefficients is also a hyper-parameter, we tested the SV accuracy with varying the number of coefficients. The results are shown in the Figures 3 and 4.

We see that for male the lowest error is obtained with 30 coefficients while for female the lowest error occurs at 20 coefficients. If we want to use a common system for male and female speakers, we found that having 23 coefficients was giving us the best results. The results are given in Table 3. We see an improvement at all noise levels from the baseline using GFCCs.

SNR	Pred. Error Males (%)	Pred. Error Females (%)
Clean Speech	0	0
10dB	8	16
5dB	28	38

Table 3: Prediction Error using GFCCs

4.3. Power Normalized Cepstral Coefficients (PNCC)

Prior research has examined the efficiency of PNCC features for the task of noise robust automatic speech recognition (ASR)[10]. Hence, we looked at that approach in our project. The algorithm for extracting PNCC features from the speech signal is explained below. (Refer Figure 5 for the PNCC block diagram.)

- Pre-emphasis filter of form $\mathrm{H(z)} = 1 - 0.97\mathrm{z}^{-1}$ is applied to boost the high frequency components.
- Short time Fourier Transform is taken with Hamming Window of duration 25.6 ms, 10 ms overlap between frames, and DFT of size 1024.
- Spectral power in 40 analysis bands is calculated by weighting of magnitude squared STFT outputs with frequency response associated with 40 channel gammatone-shaped filter bank. This results in a short time spectral power P$[m, l]$ where m = frame index and l = channel index.
- The running average of P$[m, l]$, the power observed in a single analysis frame is computed, to estimate medium-time power Q $[m, l]$. Here Q [m,l] is used only for noise estimation and compensation (used to modify information based on short time power estimates P[m,l]).
- Two Nonlinear time varying operations: Asymmetric Noise Suppression and Temporal Masking for noise subtraction are performed using longer duration temporal analysis.

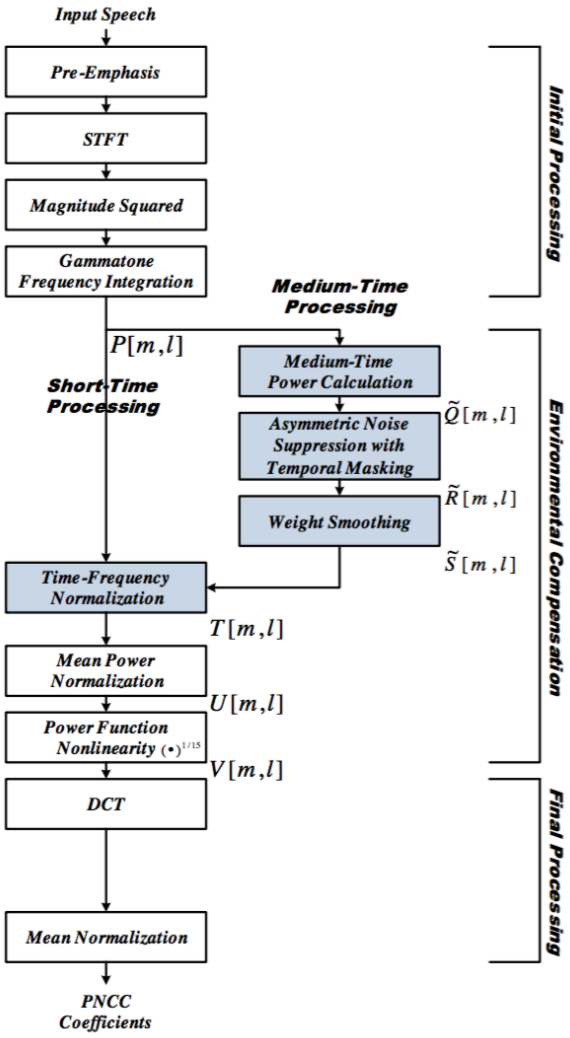

Figure 5: *PNCC Algorithm*

- This is followed by mean power normalization to reduce the impact of amplitude scaling.
- Power law non-linearity with exponent 1/15 is used instead of a log non-linearity as in MFCC.
- Discrete Cosine Transform (DCT) followed by mean normalization is applied which results in the PNCC coefficients.

We tested the SV accuracy for PNCC features with and without feature warping and cepstral mean variance normalization. We found that feature warping and CMVN boosted the accuracy of the vanilla PNCC algorithm by nearly 20%. The results for speaker verification with PNCC features are shown in Tables 4 and 5.

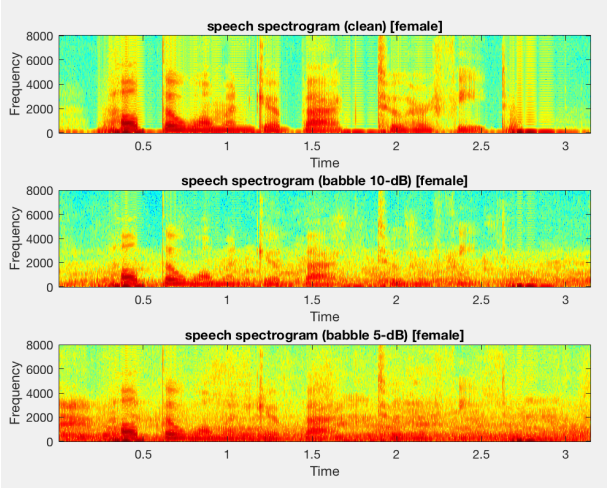

Figure 6: *Speech spectrogram for female speaker*

SNR	Pred. Error without feature processing (%)	Pred. Error with feature processing (%)
Clean Speech	0	2
10dB	14	14
5dB	26	22

Table 4: PNCC prediction error results For Male Speaker

SNR	Pred. Error without feature processing (%)	Pred. Error with feature processing (%)
Clean Speech	0	2
10dB	10	4
5dB	38	12

Table 5: PNCC prediction error results For Female Speaker

We found that PNCC features seem to work better for speaker verification in female speakers compared to male speakers. This might be due to the fact that the fundamental frequency of male speakers is smaller compared to female speakers, and since babble noise is higher at lower frequencies, the PNCC features might not work that well for male speakers. This is evident from the PNCC spectrograms for male and female plotted in Figures 7 and 9 where we can see that the PNCC features are more robust (do not change significantly) for female as the noise SNR varies. However, for male, there is a significant difference in PNCC spectrogram for babble noise at 5dB and 10dB which indicates that some features are lost/getting distorted as the noise increases.

4.4. Perceptual Linear Prediction (PLP)

PLP uses psycho-physical concepts for auditory signal spectral analysis. Since the spectrum is represented as an all pole model, it is highly biased towards the voiced segments. Although PLP was developed as a feature to reduce the variability of speech across speakers, it has become a very popular technique for speaker recognition along with MFCCs. A more detailed explanation of the algorithm can be found in [11]. In our experiments

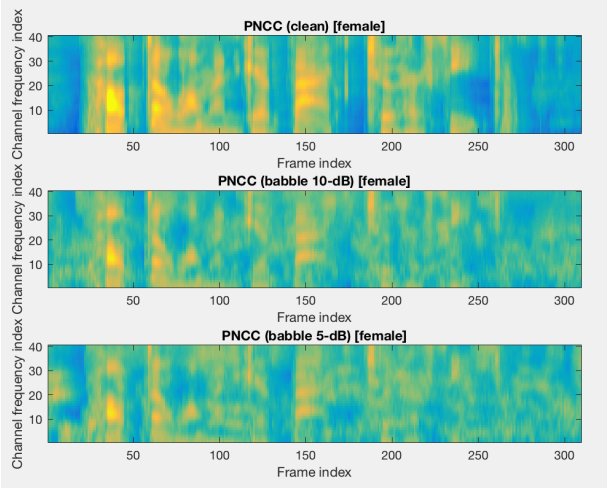

Figure 7: *PNCC spectrogram for female speaker*

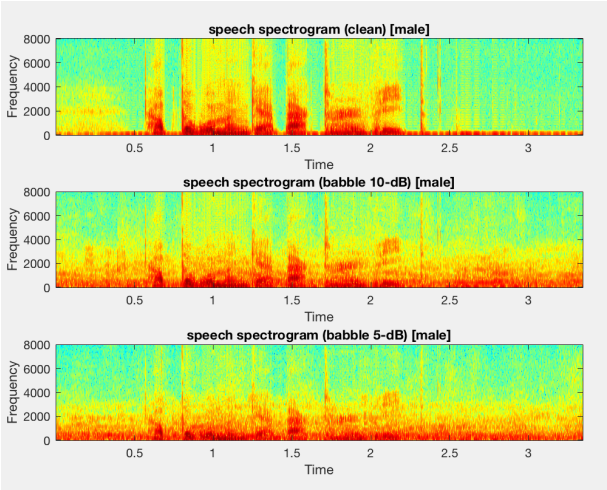

Figure 8: *Speech spectrogram for male speaker*

PLP performed almost equal to the baseline. It was not as effective as the GFCC or PNCC coefficients. Table 6 shows our results:

SNR	Pred. Error Males (%)	Pred. Error Females (%)
Clean Speech	4	4
10dB	42	46
5dB	62	60

Table 6: Percentage prediction error using PLP coefficients alone.

4.5. MFCC/ GFCC DNN Bottleneck Features

In recent times, neural networks have been used extensively for the task of speech recognition and speaker verification [12]. However, it has been long been known that since neural networks involve a lot of parameters, it is difficult to train them for mismatched conditions and they break down in the presence of noise. Still, we explored the use of bottleneck features or the activation outputs from the last feature layer of the DNNs (trained for speaker classification). Our hypothesis was that

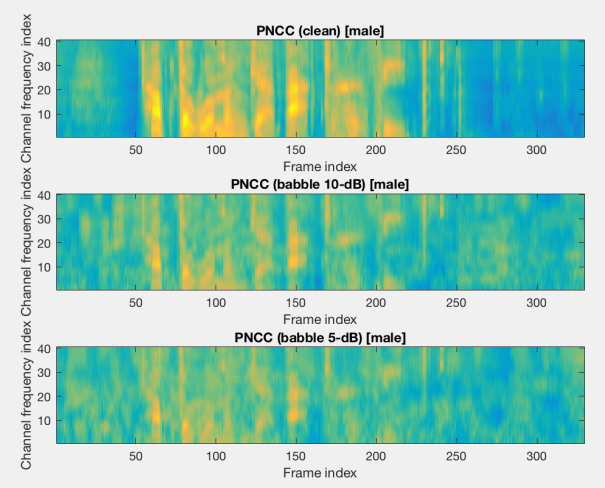

Figure 9: *PNCC spectrogram for male speaker*

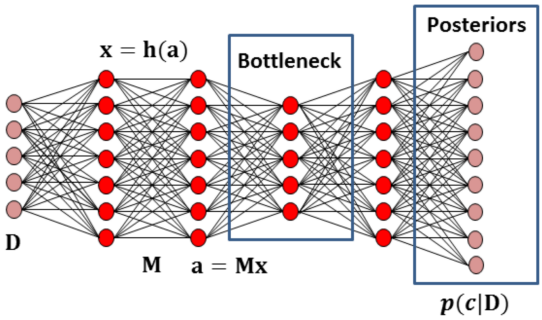

Figure 10: *MFCC/GFCC Bottleneck features*

since these features have been trained specifically for the task of speaker classification, these bottleneck features would contain more speaker specific information which have more inter-speaker variability and less intra-speaker variability. The algorithm for our neural network bottleneck feature extraction is presented below. (Refer Figure 10 for the MFCC bottleneck feature extraction)

- Extracted MFCC/GFCC features for each utterance with window duration of 25.6 ms and 10 ms overlap between frames.
- Concatenated features from 10 consecutive frames together and assigned label for each data point as the speaker class.
- Then we created a fully connected feed-forward neural network with 3 hidden layers and 20 neurons in each hidden layer.
- After training the neural network, we extract the features from the last feature layer (bottleneck features).
- As noted above, our hypothesis was that this feature would have more speaker-specific information compared to the standard MFCC/GFCC features.
- Sent these features instead of MFCC/GFCC features as inputs into our training task using GMMs.

The speaker verification results for the bottleneck features derived from MFCC/GFCC coefficients is reported in Tables 7,8,9 and 10. As we can clearly see, the bottleneck features perform reasonably well when there is no noise but break down severely in the presence of noise. Hence, we realized that neural networks cannot be used in mismatched conditions. For neural networks to work, we would need to have a large amount of training data both with and without noise.

SNR	Pred. Error without feature processing (%)	Pred. Error with feature processing (%)
Clean Speech	4	4
10dB	72	66
5dB	82	74

Table 7: MFCC Bottleneck Prediction error results For Male Speaker

SNR	Pred. Error without feature processing (%)	Pred. Error with feature processing (%)
Clean Speech	6	6
10dB	70	60
5dB	70	64

Table 8: MFCC Bottleneck Prediction error results For Female Speaker

SNR	Pred. Error without feature processing (%)	Pred. Error with feature processing (%)
Clean Speech	4	4
10dB	54	50
5dB	76	70

Table 9: GFCC Bottleneck Prediction error results For Male Speaker

SNR	Pred. Error without feature processing (%)	Pred. Error with feature processing (%)
Clean Speech	4	4
10dB	66	62
5dB	74	72

Table 10: GFCC Bottleneck Prediction error results For Female Speaker

5. Conclusion

We found that GFCC along with post-processing with CMVN and feature warping gave the best results. Also, pre-processing using Wiener filtering was found to be good for only female speakers. For male speakers, the results were not as good as using GFCCs with CMVN and feature warping. Overall we suggest using GFCC features along with post-processing with CMVN and Feature Warping for the speaker verification task.

SNR	% Pred. Error Males (change)	% Pred. Error Females (change)
Clean Speech	4(+4)	2(+2)
10dB	10(-14)	4(-24)
5dB	26(-30)	20(-38)

Table 11: Results with GFCC coefficients and post processing with cmvn and feature warping.

6. Future Work

Although the results we achieved are good, still we are very far from the performance of the human auditory system. We propose the following for future analysis:

- Here we used only the babble noise for testing. It would be great to see the results with speech corrupted with other noises using the proposed scheme.
- Passing a combination of features as an input for neural network while computing bottleneck features.
- Understanding the importance of unvoiced portions of speech could be very helpful in understanding of human auditory system and recognition.

7. References

[1] Geiger, Jrgen T., et al. "Using linguistic information to detect overlapping speech." *INTERSPEECH. 2013*.

[2] Scalart, Pascal. "Speech enhancement based on a priori signal to noise estimation." Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 *IEEE International Conference on. Vol. 2. IEEE, 1996.*

[3] C. Kim , R. M. Stern, "Power Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," *ICASSP 2012*.

[4] Y. Shao, Z. Jin, D. Wang, S. Srinivasan, "An auditory based feature for Robust Speech Recognition," *ICASSP 2009*.

[5] F. Richardson, D. Reynolds, N. Dehak, *A unified deep neural network for speech and language recognition*, arXiv 2015.

[6] J. Pelecanos, S. Sridharan, "Feature warping for Robust Speaker Verification," in *Speaker Recognition Workshop 2001*.

[7] F. Liu, R. M. Stern,X. Huang,A. Acero, "Efficient Cepstral Normalization for Robust Speech Recognition,*1993*.

[8] Westphal, Martin. "The use of cepstral means in conversational speech recognition." *EUROSPEECH. 1997.*

[9] Talkin, David. "A robust algorithm for pitch tracking (RAPT)." *Speech coding and synthesis 495 (1995): 518.*

[10] Ferrer, Luciana, et al. A noise-robust system for NIST 2012 speaker recognition evaluation. *SRI INTERNATIONAL MENLO PARK CA SPEECH TECHNOLOGY AND RESEARCH LAB, 2013.*

[11] Hermansky, Hynek. "Perceptual linear predictive (PLP) analysis of speech." *the Journal of the Acoustical Society of America 87.4 (1990): 1738-1752.*

[12] Richardson, Fred, Douglas Reynolds, and Najim Dehak. "Deep neural network approaches to speaker and language recognition." *Signal Processing Letters, IEEE 22.10 (2015): 1671-1675.*