EEM214B Final Project: Speaker Recognition in Noisy Environment

Deepak Muralidharan and Shubham Agarwal

UCLA Electrical Engineering Department

6/6/2016

Overview

- Introduction
- Babble noise characteristics
- Pre-processing
- Normalization/ Post-processing
- Feature Extraction Methods
- Results
- Future Work

Introduction

- Speaker verification (SV) is the process of validating the claimed identity of an individual using his/her speech.
- State-of-the-art SV 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 such systems degrade severely when speech signals are distorted due to the presence of environmental or background noise.
- Robustness towards environmental noise is crucial for several SV applications, especially in hand-held devices where the background environments are uncontrolled, time-varying and unpredictable.

Babble noise characteristics

- More dB at lower frequencies, less dB at higher frequencies.
- Babble is known to be the most challenging distortion in speech systems, due to its speaker/speech like characteristics.
- One of the main factors impacting the nature of babble is the number of speakers in babble noise.
- From our results, we could say that babble noise affects speaker verification performance in male speaker more compared to female speaker.
- Possible reasons might be that since the fundamental frequency of male speakers is lesser compared to female speakers, babble noise which has a higher dB at lower frequencies distorts the speaker-specific information for male more compared to female.

Pre-processing

Background

- Recognition systems work well with speech without noise.
- Extensive research has been done on noise removal in digital signals.

Challenges

- Differentiating noise and speech.
- Classifying and removing noise from different sources.
- May loose the speech signal itself.

Noise removal pre-processing techniques

Wiener Filtering

- This is effective in removing additive noise and microphone blurring.
- Gave good results with 10 dB babble noise. Not so good when noise SNR increased to 5dB.

Estimating the noise from signal

- We tried estimating the noise by taking the front part of the signal.
- Performance was not as good as the Weiner filter.

Noise removal pre-processing techniques

Signal Normalization

- Directly normalizing the input signal coefficients.
- Good results compared to the baseline but weiner filtering gave better results.

SNR	Predicion Error Male	Predicion Error Female
Clean Speech	0	0
10dB	24	12
5dB	42	36

Table 1: Wiener pre-processing applied with MFCC

Normalization/ Post-processing

High performance boost!

Gave us a very good improvement in the presence of 5 dB and 10 dB noise (around 30 % decrease in error rate).

Two main steps

- Cepstral Mean and Variance Normalization
- Feature Warping

Cepstral Mean and Variance Normalization

Algorithm

The cepstral coefficients are linearly transformed to have the same segmental statistics (zero mean, unit variance), regardless of the noise condition.

Advantages

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, there are some disadvantages

- Does not work well for short utterances.
- Performance might degrade as noise increases.

Feature Warping

Algorithm

- The feature warping method warps the distribution of a cepstral feature stream to a standardised 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.
- The sliding window is shifted by a single frame each time and the analysis is repeated.

Feature Warping

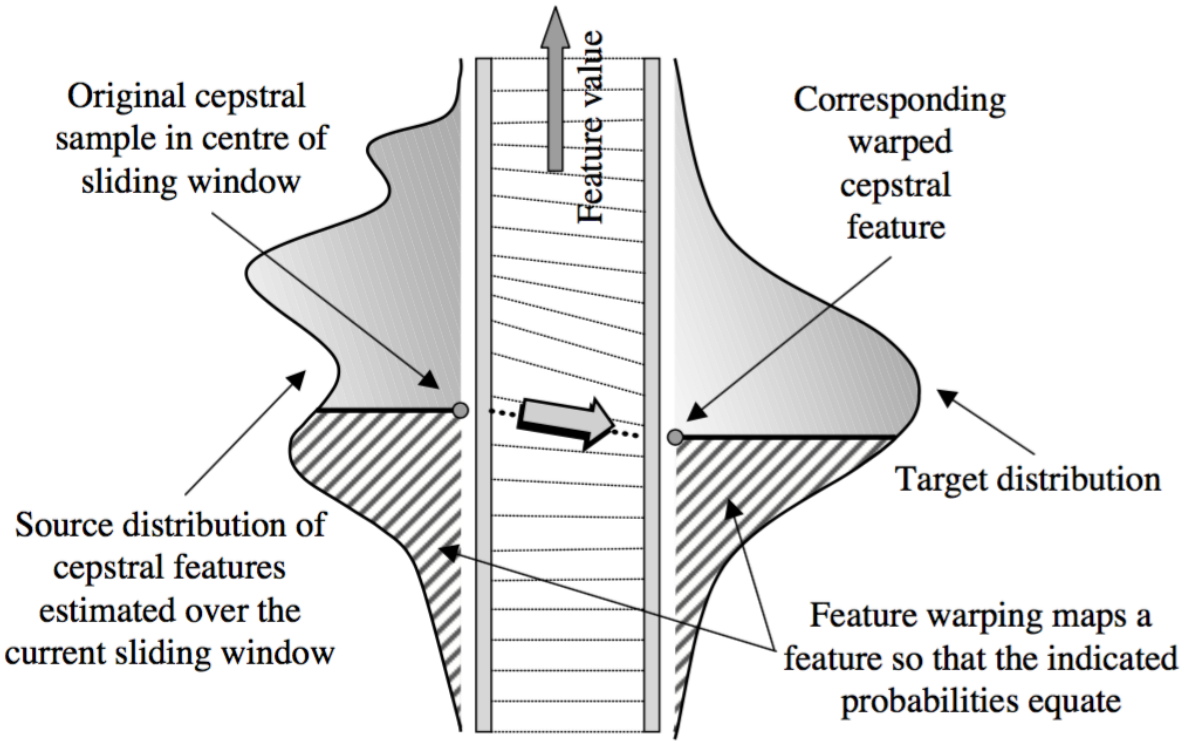

Figure 1: Warping of features according to a target distribution shape

Feature Warping

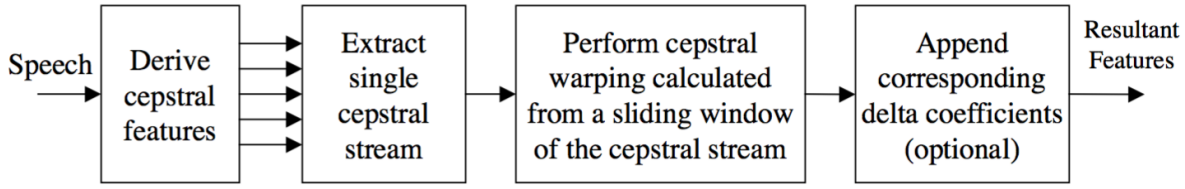

Figure 2: Block diagram of the parameterisation process

Feature Warping

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.

Feature Extraction Methods

- Mel Frequency Cepstral Coefficients (baseline)
- Pitch
- Gammatone Frequency Coefficient Cepstra (GFCC)
- Power Normalized Cepstral Coefficients (PNCC)
- Perceptual Linear Prediction (PLP)
- MFCC Bottleneck Features (last feature layer of NN)
- GFCC Bottleneck Features (last feature layer of NN)

Mel Frequency Cepstral Coefficients (baseline)

Algorithm

- Break the signal into frames
- Do windowing for each frame
- Take the Fourier transform
- Map the power spectrum to mel-scale
- Take logarithm.
- Apply DCT for decorrelation.
- The amplitudes of resulting signal are MFCC

Mel Frequency Cepstral Coefficients (baseline)

Results (for Male speaker)

SNR	Prediction Error without feature processing (%)	Prediction Error with feature processing (%)
Clean Speech	0	0
10dB	24	28
5dB	56	58

Table 2: MFCC prediction error results For Male Speaker

Mel Frequency Cepstral Coefficients (baseline)

Results (for Female Speaker)

SNR	Prediction Error without feature processing (%)	Prediction Error with feature processing (%)
Clean Speech	0	0
10dB	38	10
5dB	60	48

Table 3: MFCC prediction error results For Female Speaker

Pitch

- Widely used feature in literature for identification tasks.
- Directly related to the voice excitation source
- Tried the RAPT(Robust algorithm for pitch tracking) algorithm available in voicebox.
- Not much improvement in speaker recognition in present of babble noise.
  About 90% prediction error for both 10dB and 5dB SNR.
- Taking out only the voiced portions also gave similar results

Gammatone Frequency Coefficient Cepstra (GFCC)

- Uses a bank of gammatone filters to obtain cepstral coefficients.
- Close to human choclea with high frequency resolution at ower frequencies.
- One instance of filtering gives us GFs. Highly correlated due to frame overlap.
- DCT is then used for decorrelation.

$$C_i(m) = \sqrt{2/m} \sum_{c=0}^{N-1} G_c(m) cos(\frac{i\pi(2c+1)}{2N})$$

i=0...N-1
where, m is the frame index and
N is the total number of coefficients.

Gammatone Frequency Coefficient Cepstra (GFCC)

Results (for Male speaker)

SNR	Prediction Error without feature processing (%)	Prediction Error with feature processing (%)
Clean Speech	0	2
10dB	8	10
5dB	28	20

Table 4: GFCC prediction error results For Male Speaker

Gammatone Frequency Coefficient Cepstra (GFCC)

Results (for Female Speaker)

SNR	Prediction Error without feature processing (%)	Prediction Error with feature processing (%)
Clean Speech	0	2
10dB	16	6
5dB	38	28

Table 5: GFCC prediction Error results For Female Speaker

Results

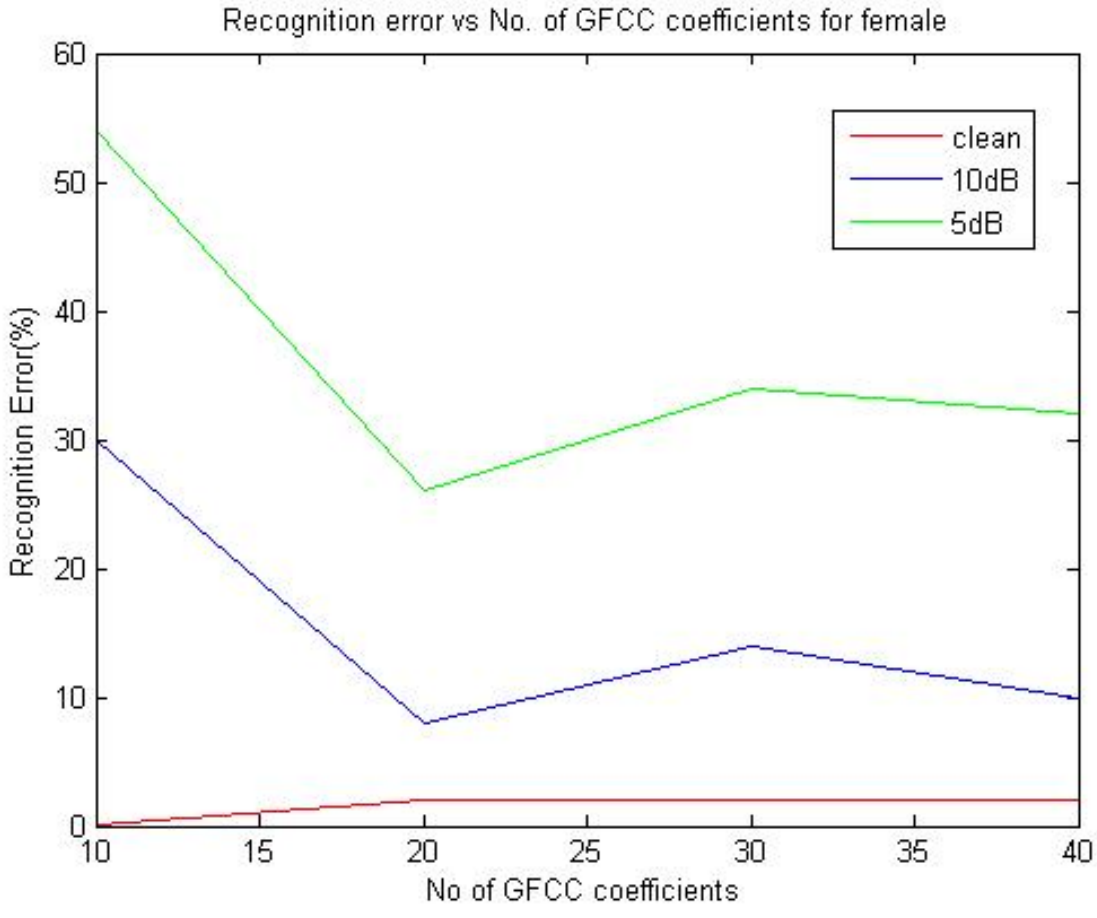

Figure 3: Recognition error vs Number of GFCC co-efficients for female

Results

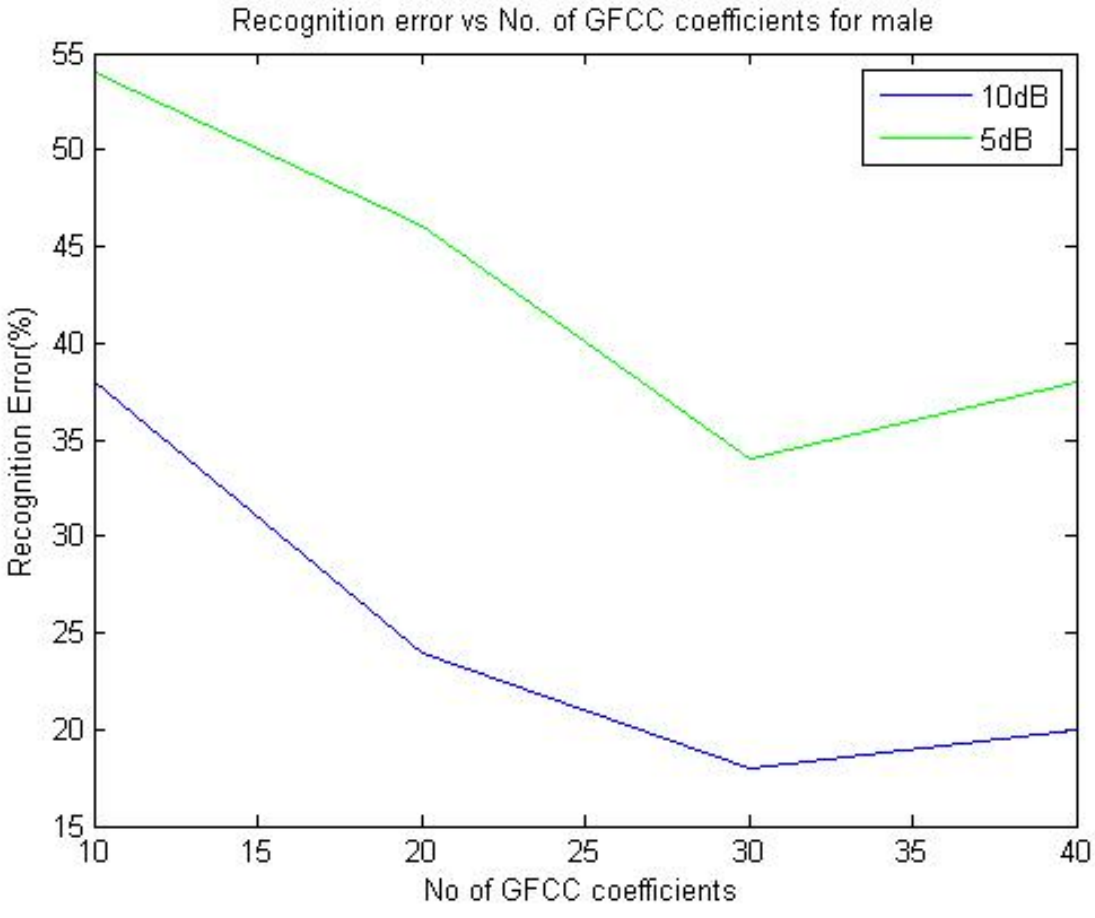

Figure 4: Recognition error vs Number of GFCC co-efficients for male

Power Normalized Cepstral Coefficients (PNCC)

Algorithm

- Pre-emphasis filter of form $H(z) = 1 - 0.97z^{-1}$.
- Short time Fourier Transform with Hamming Window of duration 25.6 ms, 10 ms overlap between frames, and DFT of size 1024.
- Spectral power in 40 analysis bands by weighting of magnitude squared STFT outputs with frequency response associated with 40 channel gammatone-shaped filter bank.
- Output (of the previous step) is short time spectral power $P[m, l]$ where m = frame index and l = channel index.
- Compute running average of $P[m, l]$, the power observed in a single analysis frame, to estimate "medium-time power" $Q[m, l]$.

Power Normalized Cepstral Coefficients (PNCC)

Algorithm (contd.)

- Note: Q $[m, l]$ used only for noise estimation and compensation (used to modify information based on short time power estimates P$[m, l]$).
- Nonlinear time varying operations: Asymmetric Noise Suppression and Temporal Masking for noise subtraction performed using longer duration temporal analysis:
- Mean power normalization to reduce impact of amplitude scaling.
- Power law non-linearity with exponent $1/15$.
- Discrete Cosine Transform.
- Mean Normalization.

Power Normalized Cepstral Coefficients (PNCC)

Results (for Male speaker)

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

Table 6: PNCC prediction error results For Male Speaker

Power Normalized Cepstral Coefficients (PNCC)

Results (for Female Speaker)

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

Table 7: PNCC prediction error results For Female Speaker

Power Normalized Cepstral Coefficients (PNCC)

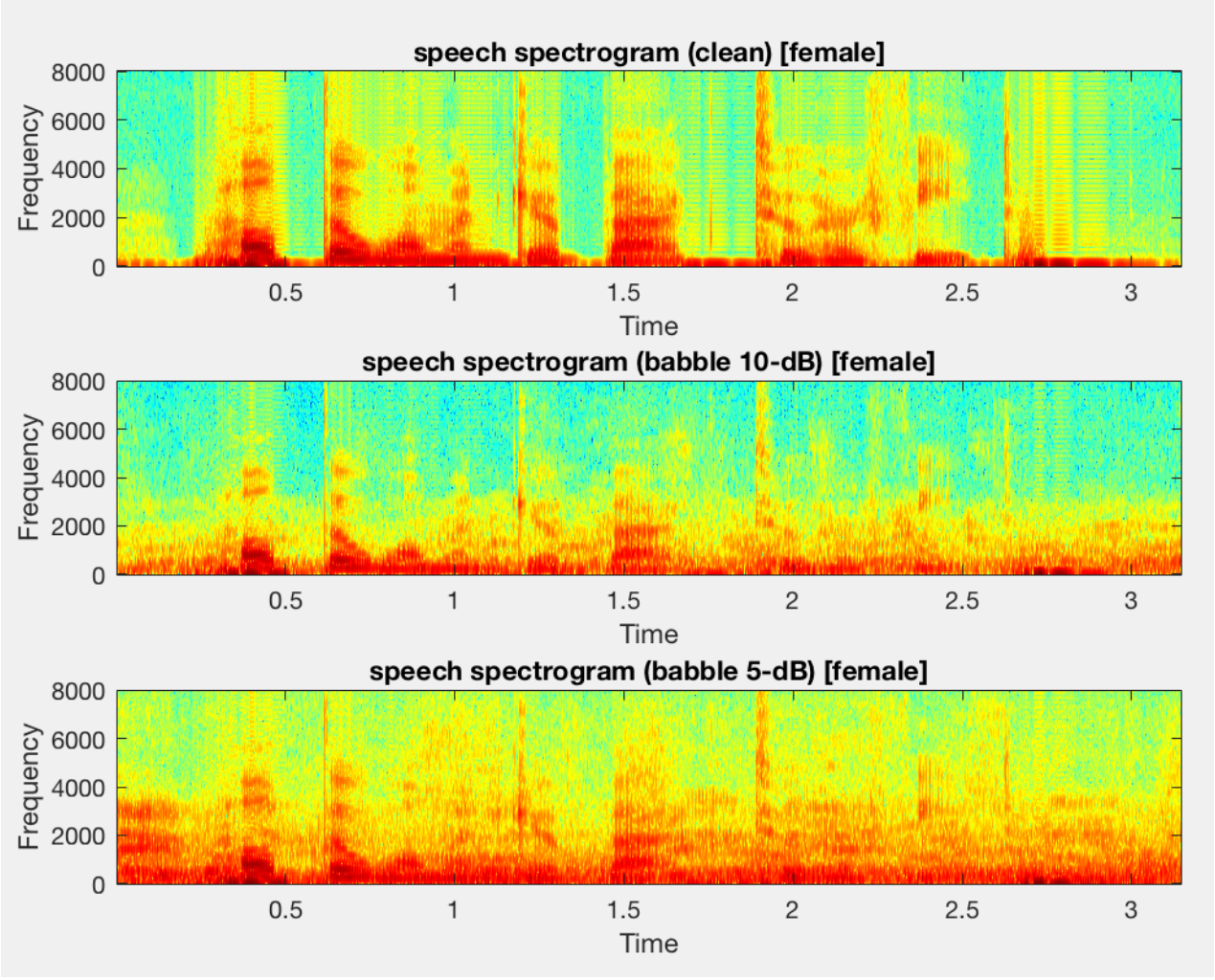

Figure 5: Speech spectrogram for female speaker

Power Normalized Cepstral Coefficients (PNCC)

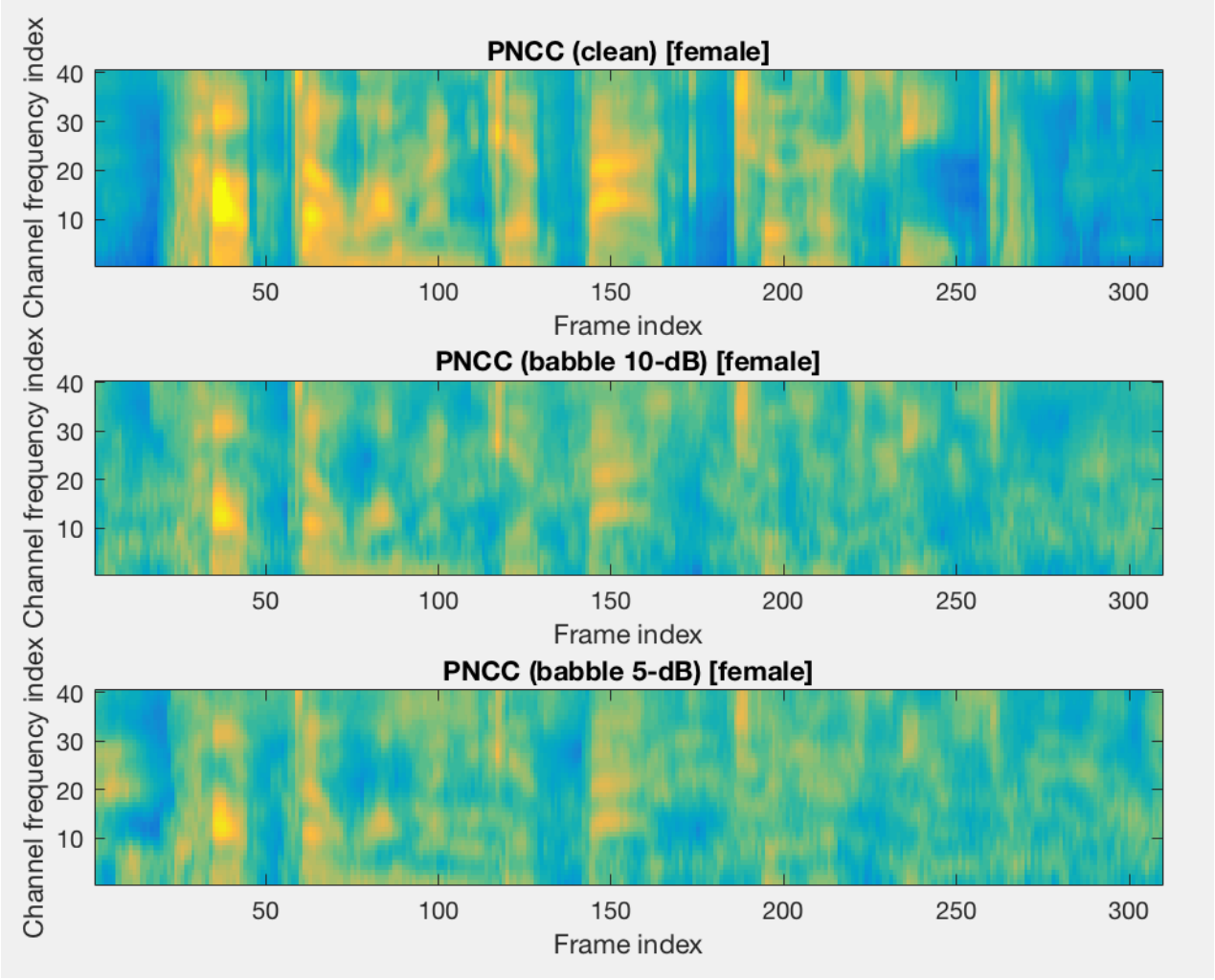

Figure 6: PNCC spectrogram for female speaker

Power Normalized Cepstral Coefficients (PNCC)

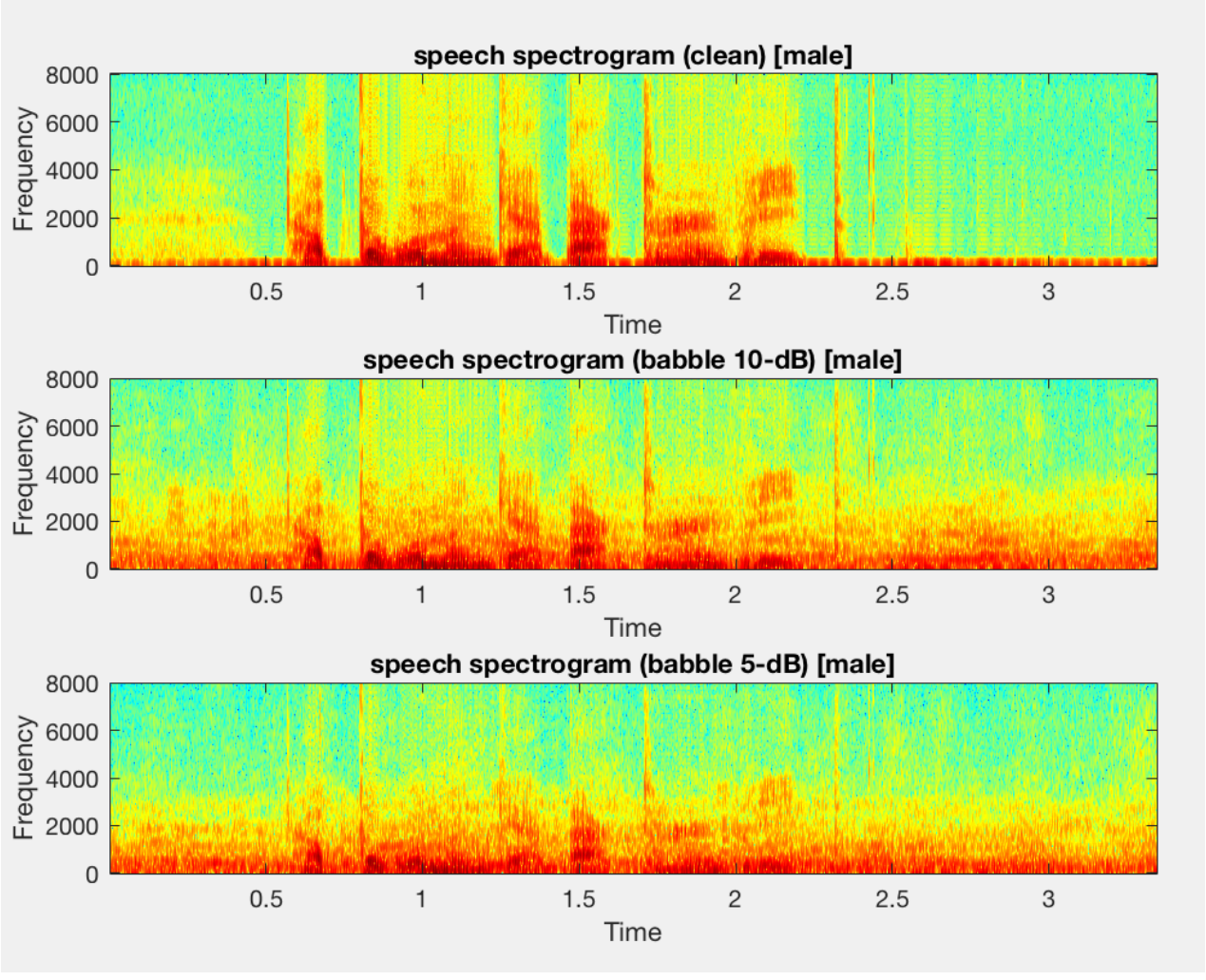

Figure 7: Speech spectrogram for male speaker

Power Normalized Cepstral Coefficients (PNCC)

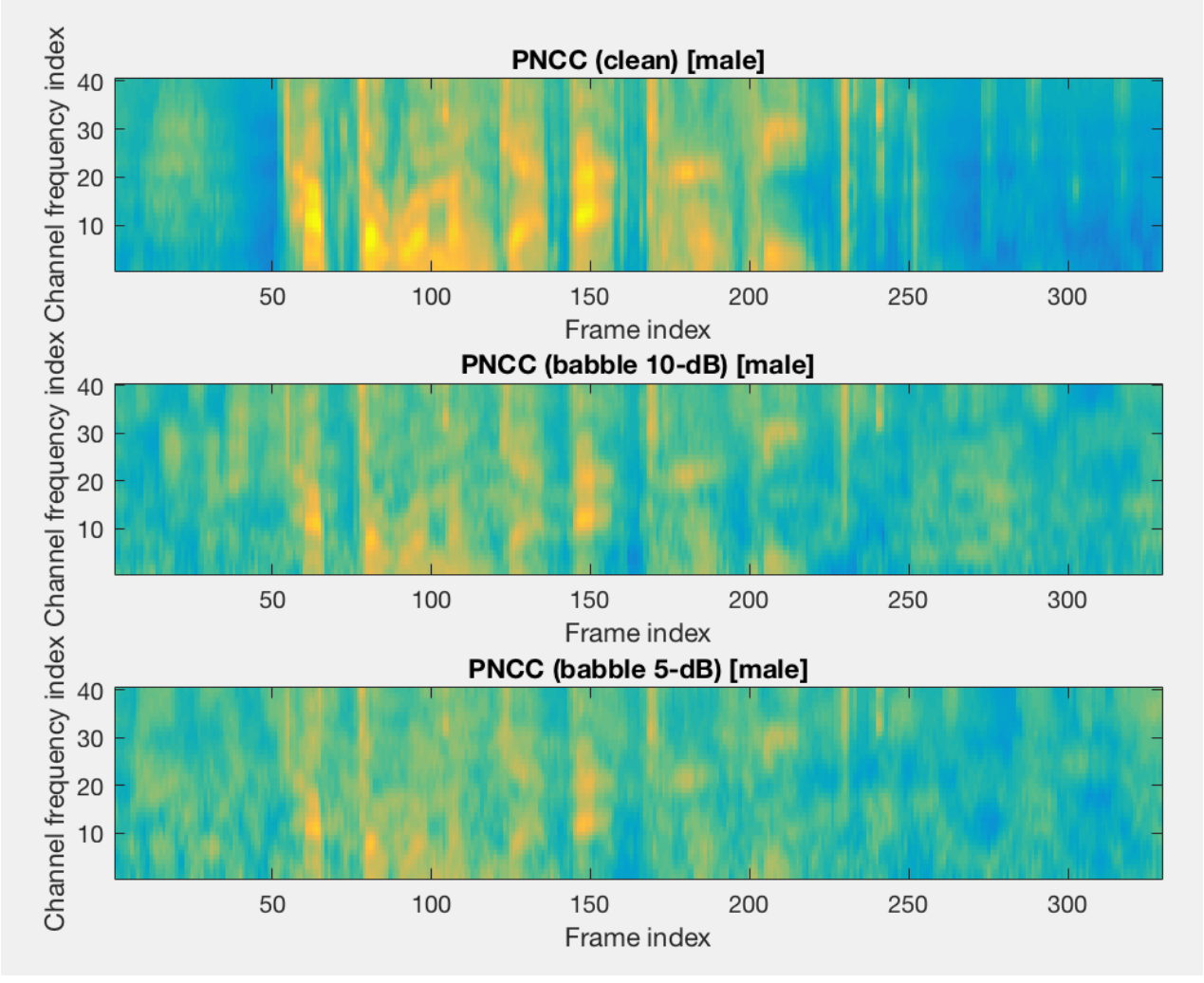

Figure 8: PNCC spectrogram for male speaker

Perceptual Linear Prediction (PLP)

Algorithm

- Frame Blocking
- Pre-emphasis and windowing
- Critical Band Analysis
- Equal loudness pre-emphasis band
- Intensity-Loudness conversion
- Inverse DFT
- LPC analysis
- LPC to cepstrum conversion
- We get PLP coefficients

Perceptual Linear Prediction (PLP)

Results (for Male speaker)

SNR	Prediction Error without feature processing (%)	Prediction Error with feature processing (%)
Clean Speech	4	2
10dB	42	32
5dB	62	52

Table 8: PLP prediction error results For Male Speaker

Perceptual Linear Prediction (PLP)

Results (for Female Speaker)

SNR	Prediction Error without feature processing (%)	Prediction Error with feature processing (%)
Clean Speech	4	2
10dB	46	38
5dB	60	50

Table 9: PLP prediction error results For Female Speaker

MFCC/GFCC Bottleneck Features

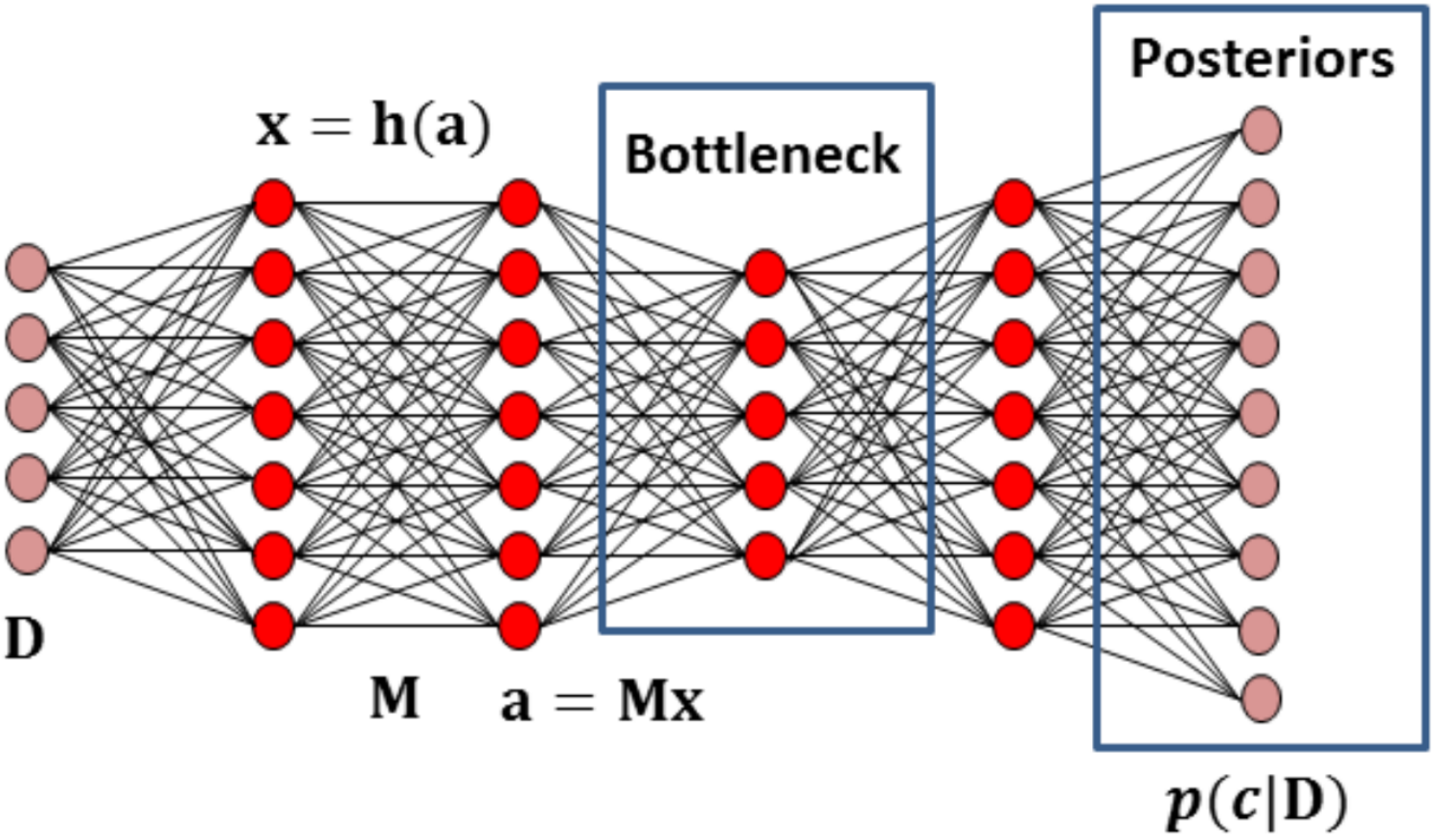

Figure 9: MFCC/GFCC Bottleneck Features from DNN.

MFCC/GFCC Bottleneck Features

Algorithm

- 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).
- 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 into our training task using GMMs.

MFCC Bottleneck Features (last feature layer of NN)

Results (for Male speaker)

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

Table 10: Prediction error results For Male Speaker

MFCC Bottleneck Features (last feature layer of NN)

Results (for Female Speaker)

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

Table 11: Prediction error results For Female Speaker

GFCC Bottleneck Features (last feature layer of NN)

Results (for Male speaker)

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

Table 12: Prediction error results For Male Speaker

GFCC Bottleneck Features (last feature layer of NN)

Results (for Female Speaker)

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

Table 13: Prediction error results For Female Speaker

Future Work

Noise modeling using deep neural networks

- Given enough utterances of the same training data in the noised condition (input) and clean condition (output), a non-linear mapping of noise can be estimated and modeled using a DNN.
- This information can then be sent as features along with the standard features(MFCCs, GFCCs etc.) extracted from the noisy speech, which might help in normalization of the noise information thus giving more noise robust features.
- Might produce better results in mismatched conditions.

Results

- We found that GFCC along with post-processing with CMVN and feature warping gave the best results.
- 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	Prediction Error Male	Prediction Error Female
Clean Speech	4(+4)	2(+2)
10dB	10(-14)	4(-24)
5dB	26(-30)	20(-38)

Table 14: Results for GFCC with post processing.

References

- Kim C, Stern RM. *Power Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition*. ICASSP 2012.
- Shao Y, Jin Z, Wang D, Srinivasan S. *An auditory based feature for Robust Speech Recognition*. ICASSP 2009.
- Richardson F, Reynolds D, Dehak N. *A unified deep neural network for speech and language recognition*. arXiv 2015.
- Hermansky H. *Perceptual linear predictive (PLP) analysis of speech*. 1989.
- Pelecanos J, Sridharan S. *Feature warping for Robust Speaker Verification*. Speaker Recognition Workshop 2001.
- Liu F, Stern RM, Huang X, Acero A. *Efficient Cepstral Normalization for Robust Speech Recognition*. 1993.