# THE EFFECTS OF ROOM ACOUSTICS ON MFCC SPEECH PARAMETER

*Yue Pan and Alex Waibel*

Interactive System Laboratories
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213, USA
{ypan, waibel}@cs.cmu.edu

## ABSTRACT

Automatic speech recognition systems attain high performance for close-talking applications, but they deteriorate significantly in distant-talking environment. The reason is the mismatch between training and testing conditions. We have carried out a research work for a better understanding of the effects of room acoustics on speech feature by comparing simultaneous recordings of close talking and distant talking speech utterances. The characteristics of two degrading sources, background noise and room reverberation are discussed. Their impacts on the spectrum are different. The noise affects on the valley of the spectrum while the reverberation causes the distortion at the peaks at the pitch frequency and its multiples. In the situation of very few training data, we attempt to choose the efficient compensation approaches in the spectrum, spectrum subband or cepstrum domain. Vector Quantization based model is used to study the influence of the variation on feature vector distribution. The results of speaker identification experiments are presented for both close-talking and distant talking data.

## 1. INTRODUCTION

The performance of automatic speech/speaker recognition systems is significantly degraded by acoustic mismatches between training and testing conditions. Existing ASR systems, which designed and trained on close talking conditions, deteriorates rapidly when using a distant microphone in a meeting room. These systems all include a signal processing front-end that converts a speech waveform into feature parameters. One of most used features is Mel-based Cepstrum Coefficients (MFCCs). Many approaches to compensate the effect or to mitigate the mismatch have been studied at each steps of the processing. For example, in time domain, inverse filtering [1] and microphone array beam-forming [2][5] have been used to obtain better quality speech; in spectrum and its subband domain, spectral subband centroids (SSCs) parameter [3] is developed; and in cepstral domain, mean subtraction normalization [4], and neural networks mapping [2] are proposed. Model level adaptation also has been addressed extensively [5]

In this paper, our research goals are: (a) get a better understanding of characteristics room acoustic phenomenon and find the factors independently or jointly affecting MFCC speech features; (b) thus provide guiding information for efficient adaptation in different signal processing steps; (c) and study the cluster of speech vector distribution variations through a VQ based the speaker identification system with the influence of recognition results. Experiment data is collected by recording speech signal through two channels at the same time, one for close-talking signal, and the other one for distant-talking. Our approach focus on analysis of the distortion of the spectrum or MFCC derived spectrum between the two simultaneous signals, while avoiding direct analysis of their cepstrum.

It has been known there are two major sources in room acoustics causing the performance degradation in distant-talking speech recognition: background noises and reverberation. In our meeting room environment, background noise is considered as stationary most of the time. By the knowledge of the properties of the current existing noise spectrum, an adaptation could be made to signal or acoustic model at the low signal noise ratio subband.

Reverberation is a kind of multi-path effects, the situation in which there are several propagation paths from speaker to listener. As convolutional noise, it can be modeled by an FIR system. Because its response function varies as the distance to microphone and temperature [1] etc., thus time domain methods are not good options for this application due to the lack of reference signal. Our study shows traditional channel normalization such as cepstral mean subtraction is also not as good as it is in usual channel normalization. (e.g. telephone channel).

## 2. CHARACTERISTICS OF ROOM ACOUSTICS

For the meeting room speech recognition task, there is a fixed distant microphone and users could speak freely at different positions. When the speech recognition system is trained by close clean speech, the study of characteristics of room acoustics is very useful for robust feature extraction to get a satisfactory performance.

### 2.1. Database

The Database contains a set of two-channel speech utterance of 10 speakers. Channel one is connected to a remote microphone. Channel two is for a directional close microphone. The distant microphone was set a higher gain so that the speech signal it catches has the same mean energy as the close microphone. The data was recorded in Interactive System Laboratory meeting room at 16kHz sampling rate, 16 bits per sample, using a PC sound card. As the usual condition in meeting rooms or offices, noise sources are computer fans and air conditioning systems and moderate reverberation is present. All Speakers were asked to read the same prepared text. At the beginning of each recording session, a few seconds of the silence (background noise) are recorded and can be used for later adaptation. Both microphones record simultaneously. The tiny time delay between signals of two microphones due to the distance can be compensated by sample shifting. This simultaneous dual-channel recordings through two microphones, give us an exact reference of the effects introduced by the room acoustics.

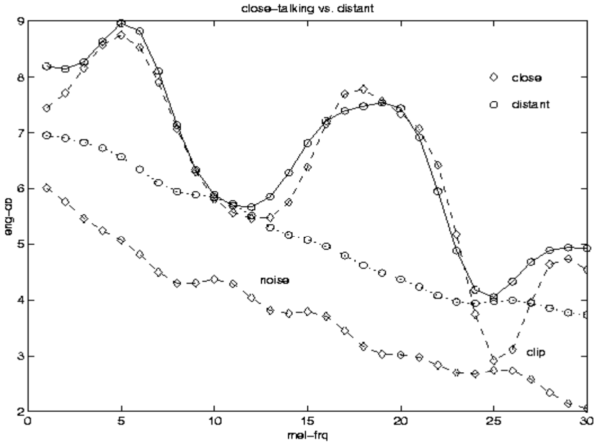

**Figure 1:** 30 Mel-Spectra coefficients of close talking and distant-talking. The lower two dotted line are the noise mean spectrum. The upper two lines are the MFCC spectrum at a sample frame.

## 2.2. Mel-scale Spectrum Derived from MFCC

Cepstrum is the inverse Discrete Cosine Transform (DCT) of the logarithm of the short-term power spectrum of the signal. The advantage of using such coefficients is that they reduce the dimension of a speech spectral vector. Also by applying the transform, the original Mel-scale spectrum gets smoothed. In our study, MFCC vector distortion is illustrated in the MFCC Spectrum, by applying once more DCT transform on the MFCCs. In other words, we restore Mel-scale spectrum from cepstral parameters. The reason is we want to explain the properties of distortion in spectrum domain, with clear physical meanings. Under Euler distance measurement, the distortion between MFCC Spectrums is the same as that of the direct MFCC vectors.

## 2.3. Background Noises

First, we inspect the influence of the background noise in MFCC derived spectrum. The energy of the noise mean spectrum of distant microphone is much higher than that of the close microphone. Figure 1 illustrates the MFCC Mel spectrum of close and distant talking of the same frame. The lower two dotted lines are the mean noise spectrum. At a given time, the signal energy in different frequency bands is different, so is the signal-to-noise ratio (SNR). The low energy regions are more corrupted or masked by the noise spectrum. And it is exactly these corrupted local regions caused the biggest mismatch of cepstrum vectors between close and distant talking speech. Thus schemes of adaptation on subband are suitable and efficient to mitigate the distortion and reduce the vector mismatch.

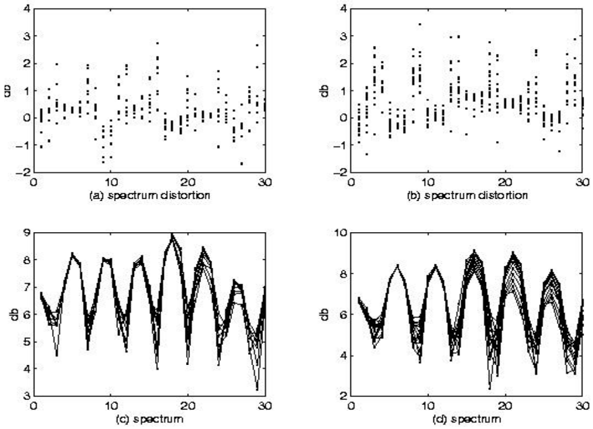

**Figure 2:** The first 30 components of DFT power spectrum of *150 ms* speech at two different pitches. The lower plots are the close-talking spectrum and the upper plots show the spectrum distortion between close and distant-talking data at the corresponding frequency.

## 2.4. Reverberation verse Pitch

In theory, multi-path effect of the reverberation has the property like comb filter. For finer details, it is required to analyze directly in DFT spectrum.

If *f0* is the frequency of the pitch period, the DFT spectrum has peaks at integer multiples of *f0.* In Figure 2, the lower two plots illustrate frames of the DFT power spectrum with two different pitches. The upper two plots present the corresponding distortion between close and distant talking speech for those frames. Only the first 30 of 256 DFT spectrum components are plotted (i.e. frequency band of 0 - 1KHz), since the finer information at higher frequency spectrum will be hided by the Mel scale filters at the successive step in MFCC. We can see the low energy components have much larger distortion for the reason of noise. In order to isolate the pure reverberation system response, we only inspect the peaks in the spectrum where SNR is high. That is the value exactly at *f0* and its multiples. Figure 3 illustrates the statistical distribution of distortions only at these positions. The lower line is the mean distortion at pitch and its multiples, comparing to the upper line – the mean distortion computed on all the frequencies.

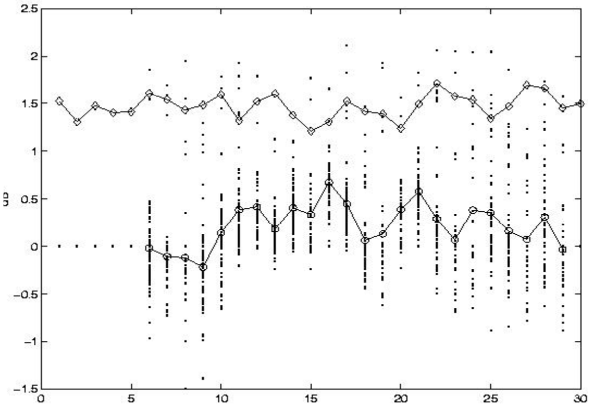

**Figure 3:** distribution of the spectrum distortion for vowels at different pitch and its multiples, for *5 sec* male speaker utterance A small dot means the distortion at the frequency in a frame speech. The solid line of '°' is the mean of these dots. The line of '◊' is the mean distortion computed on the whole spectrum without the consideration of pitch. There is no plot for the first 5 components, because the pitch cannot be lower than a minimum frequency.

In sum, the pitch of speech signal plays an import role in the spectrum variations. The comb properties of reverberation and the periodic component of speech make up together, producing an effect like alias in signal processing.

The principle behind most popular methods to ameliorate the effects of channel variability as Cepstral Mean Subtraction is based upon the behavior of the cepstrum under convolutinal distortion and the assumption that the channel function does not vary significantly over the duration of the utterance. As the pitch shifts during speech utterances, the distant-talking MFCC feature changes too even if the close-talking spectrum envelop keeps the same. Since the pitch changes within utterance, cepstral mean normalization becomes ineffective. In other words, the spectrum distortion is the function of pitch, so there are variations of the distant talking even keeping the room acoustic transfer function and reference spectrum envelope unchanged.

In this circumstance, when there is no reference adapting data, how to quickly adapt the system to get satisfactory performance becomes a serious question. Obviously only in-frame information is not enough for effective adaptation. A time delay neural networks may be tried, hoping to extract reverberation robust feature on temporal spectrum structure.

## 3. EXPERIMENT

Close-set speaker identification experiments are conducted on the database. A vector quantizer (VQ) model is chosen, because it uses cepstrum distortion as the classification criterion, which is consistent to our analysis method. The codebook also gives us a good view of the distribution of the speech vectors in vector space. In practice, this nonparametric method has been proved good in speaker ID system.

### 3.1. VQ Based Speaker Identification System

Our vector codebook based Vector quantization (VQ) model [6] has a codebook of 25 vectors for each speaker. Codebook vectors stand for clusters of the vectors' statistical distribution. Accumulation of distortion of test segment to the codebook is the criterion to assign speaker id. *30s* of each speaker is used as training set, another *120s* as test set. Test utterances are separated into *1 sec* segments. 1-2 seconds of silence (background noise) is manually labeled for each recording session and is used for back ground noise adaptation. 3-5 seconds simultaneous speech is reserved for reverberation adaptation. The preprocessing step for the baseline systems are:

- High pass boost.
- 32ms (512 samples) frame size, 10 ms shifting.
- Hamming window.
- 30 Mel-scale from 100 Hz – 8000Hz
- 13 cepstral coefficients

### 3.2. Spectrum Clipping and Cepstral Mapping

**Codebook vector clipping** Follow the analysis in the previous section, we know noise-masking causes the subband mismatch. It is easier to adapt the codebook to distant data by applying the same noise mask, than restoring missing information for distant data. Thus, in our system, the design is the Spectrum Clipping, as doing the following operations on each item in model's codebook.

1. Estimate noise spectrum and determine a noise-masking threshold.
2. Apply DCT transform on vector, which recover Mel-spectrum from cepstrum.
3. Clip Mel spectrum according to noise masking threshold. If the energy value of a Mel component is less the corresponding noise threshold, mask it by the noise value; else keep it unchanged.
4. Apply inverse DCT on clipped Mel spectrum, which compress Mel spectrum back to cepstrum vector.

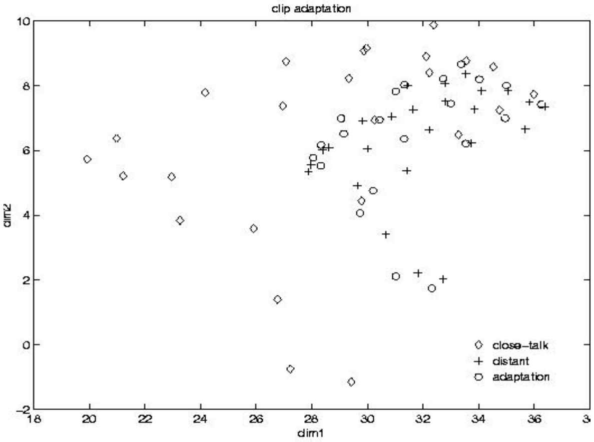

**Figure 4:** Distributions for codebook vectors for one single male speaker. '◊' represents the vectors of close-talking model. '+' represents the vectors of distant-talking model. '°' represents the model after noise level clipping.

In such way, vector codebook is adapted to fit input data. The effectiveness of clipping operation is illustrated in Figure 4. Close clean speech, cepstrum has a wider distribution than close speech before adaptation. After the introducing of noise clipping adaptation, the two distribution matches better.

**NN network mapping** 13 3-layer MLP neural networks, one for each dimension of the cepstrum, are trained on the reserved simultaneous recording data. Each MLP has 9 input nodes, 4 hidden nodes and one output node. The hidden nodes are sigmoid perceptrons and the output node is a linear perceptron. The input is distant talking cepstrum sequence of 9 frames and these networks try to map them to a close-talking vector. MLPs is used in the experiment for compensate for reverberation.

| **Training/Model** | **Test data** | **Accuracy (%)** |
|---|---|---|
| Close-talking | Close-talking | 98.50 |
| Distant-talking | Distant-talking | 95.61 |
| Close-talking | Distant-talking | 38.89 |
| Clip Close-talking | Distant-talking | 86.94 |
| Close-talking | Mapping distant | 87.90 |
| Clip Close-talking | Clip+Mapping distant | 88.57 |

**Table 1:** The speaker identification results on different training and testing conditions.

## 3.3. Results

As we can see from table 1, the effects of reverberation and noise are reflected in great degradation in recognition rates when no adaptation applied. From the table, with training data, we can see simply subband clipping of the codebook vector is an efficient way for quick adaptation.

## 4. CONCLUSIONS

Effects of room acoustics on speech parameters are the result of a combination of background noise and reverberation. Stationary background noise introduces masking at low energy frequency subband and therefore spectrum subband adaptation is quite efficient and effective. In terms to the reverberation, it influences the high-energy spectrum as well. If in the case of having reference signals, it is good to do mapping at time domain preprocessing and spectrum. But when it is hard to get reference signal or the acoustic conditions change frequently, Frame-based mapping could be of problem. Our future research will focus on model based robust temporal spectrum structure of spectrum for practical distant talking speech recognition.

## 5. ACKNOWLEDGEMENTS

The author would like to thank Dr. Jie Yang for his inspiring suggestions and active support. This research is partly supported by the Defense Advanced Research Project Agency under contract number DAAD17-99-C-0061. Yue Pan's research is supported in part by the National Science Fundation under grant SBR-9720374.

## 6. REFERENCES

1. M. Omura, M. Yada, H. Saruwatari, S. Kajita, K, Takeda and F. Itakura, "Compensating of Room Acoustic Transfer Functions Affected by Change of Room Temperature", *Proc. ICASSP, pp 941-944, 1999.*

2. U. Bub, M. Hunke and A. Waibel, "Knowing Who to Listen to in Speech Recognition: Visually Guided Beamforming", *Proc. ICASSP, pp 848-851, 1995.*

3. S. Tsuge, T. Fukada and H. Singer "Speaker Normalized Spectral Subband Parameters for Noise Robust Speech Recognition", *Proc. ICASSP, pp.285-288, 1999.*

4. A. Garcia and R. Mammone, "Channel-Robust Speaker Identification Using Modified-Mean Cepstral Mean Normalization with Frequency Warping", *Proc. ICASSP, pp.325-328, 1999.*

5. P. Raghavan, RJ. Renomeron, C. Che, D-S, Yuk and JL, Flanagan, Speech Recognition in a Reverberant Environment using Matched Filter Array (MFA) Processing and Linguistic-Tree Maximum Likelihood Linear Regression (LT-MLLR) Adaptation", *Proc. ICASSP, pp. 777-780, 1999.*

6. H. Gish and M. Schimidt. "Text-Independent Speaker Identification". *IEEE SP Magazine, October, pp. 18-32, 1994.*