Improved Acoustics Modeling for Speech Recognition
Using Transformation Techniques

Carrson C. Fung, Oscar C. Au, Wanggen Wan, Chi H. Yim, Cyan L. Keung

Human Language Technology Center
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong

c.fung@ieee.org, eau@ust.hk, wanwg@hotmail.com, eeyim@ust.hk, keunglui@ust.hk

ABSTRACT
In statistical speech recognition, misclassification often occurs when there is a mismatch between the incoming signal and the acoustics model inside the recognizer. In order to combat this problem, techniques such as Cepstral Mean Subtraction, Vocal Tract Normalization, adaptation and pronunciation model can be used.

In this paper, we proposed a new approach based on transformation technique where the output distribution function in the HMM model, a Gaussian probability density function, could be transformed to match the estimated distribution of the incoming signal by using a memoryless invertible nonlinearity function. Since the new density still has a Gaussian form, the function could be completely characterized by using the Expectation Maximization (EM) algorithm.

1. INTRODUCTION
In many speech recognition systems in use today, the acoustics model is based on some form of a Hidden Markov Model (HMM), e.g. a sub-word model, with states composing of a mixtures of Gaussian density functions to model the different phonetic utterance represented by incoming speech features. This model has to be trained for a long time using speech samples that are closely matched to the testing conditions, otherwise, performance will degrade dramatically. This mismatch can be compensated by techniques such as Cepstral Mean Subtraction (CMS), Vocal Tract Length Normalization (VTL), adaptation, or pronunciation model.

CMS tends to remove the distortion caused by the channel by subtracting a long-term mean of the cepstral features. In VTL, the frequency warping due to vocal tract length difference between different speakers is removed. Adaptation removes the speaker’s style including accent, articulation and physical difference. Lastly, the pronunciation model tries to capture the different pronunciation pattern across various speakers. All these methods tend to focus on only a few mismatches between the trained model and the testing data.

In this paper, we propose a new transformation technique that is more general than the others described above.

This method transforms the existing Gaussian mixture model to some other mixture model that matches better statistically to the testing data than the Gaussian itself. In section 2, the method will be described in details. The results will be presented in the following section. The paper is then concluded in section 4.

2. TRANSFORMATION TECHNIQUE
The transformation technique transforms the Gaussian probability density function to match the estimated distribution of the incoming signal using a memoryless invertible nonlinearity function $g()$. To illustrate, let $V$ denotes a vector of Gaussian random variables with marginal cumulative distribution function $P_V(.)$, and $O$ be a vector of random variable with arbitrary univariate density and a cumulative distribution function $P_O(.)$. According to [2], the nonlinearity can then be determined as:

$$g(o_i) = \sigma_V P_V^{-1}\left( R_V \left( \frac{o_i - \mu_O}{\sigma_O} \right) \right) + \mu_V$$

Using this nonlinearity, the new density function can be expressed as:

$$f_O(o) = (2\pi)^{-i/2} |R_V|^{1/2} \exp \left[ \frac{1}{2} (g(o) - \mu_V)^T R_V^{-1} (g(o) - \mu_V) \right] \prod_{i=1}^{K} |g'(o_i)|$$

where $\mu_V$ and $R_V$ are the mean vector and covariance matrix for the random vector $V$, $g'(o_i)$ is the first order derivative of the function $g(o_i)$ with respect to $o_i$.

In the context of speech recognition, the output probability $b_{jm}(o_i)$ of the new distribution function inside each HMM states becomes:
The mean and variance for the Laplacian were fixed as 0 and 1 throughout the simulation. These values were used initially by the Gaussian as well. Once the Gaussian mean and variance can be used since they will subsequently be updated during the re-estimation process. The mean and variance for the Laplacian were fixed as 0 and 1 throughout the simulation. These values were used initially by the Gaussian as well. Once $g(.)$ is determined, Eq. (3) - (8) can be used for re-estimation.

### 3. RESULTS

In this paper, only a preliminary experiment using Laplacian univariate density is performed. As discussed earlier, Laplacian is considered to be a good function to start with because of its heavy-tailed property.

The simulation consisted of two parts: training and recognition. The recognizer was trained using the clean TIMIT database with 55 phones and was tested using the clean and noisy TIMIT data. The noisy data were created by adding appropriate amount of white noise, which created data with SNR ranging from 20 to 0 dB. These speech data were transformed into the popular Mel-Frequency Cepstral Coefficients with delta-, delta-delta-, and energy values; making a total of 39 components per vector. The number of states per model is 5 with 2 states being non-emitting nodes. The number of mixtures was 5 per state.

Once the training for the new model is done, recognition can be carried out utilizing Eq. (3) during Viterbi decoding to obtain the output probability. All the re-
estimated parameters from training were used to form a new HMM.

Table 1 shows the recognition rate of applying the features into the two different acoustics models. Looking at the results, the system with the transformed acoustics model performs better than the Gaussian one on the average with a gain of 2% in recognition accuracy.

<table>
<thead>
<tr>
<th>SNR</th>
<th>Gaussian Model</th>
<th>Transformed Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>95.34%</td>
<td>95.45%</td>
</tr>
<tr>
<td>20 dB</td>
<td>80.01%</td>
<td>82.86%</td>
</tr>
<tr>
<td>15 dB</td>
<td>69.50%</td>
<td>71.62%</td>
</tr>
<tr>
<td>10 dB</td>
<td>55.21%</td>
<td>57.05%</td>
</tr>
<tr>
<td>5 dB</td>
<td>42.90%</td>
<td>44.67%</td>
</tr>
<tr>
<td>0 dB</td>
<td>29.89%</td>
<td>30.20%</td>
</tr>
</tbody>
</table>

ACKNOWLEDGEMENT

This work is funded in part by a CAG grant and an HKTIIT grant.

5. REFERENCES


4. CONCLUSION

A speech training and recognition system has been built based on a new acoustics model using the transformation technique. According to the results, the transformed model gives a performance improvement of about 2% on the average versus the regular Gaussian mixture model.

Future work will involve distribution functions other than Laplacian in order to obtain other transformation function that have more potentials in giving better results.

Throughout the experiment, the mean and variance for the Laplacian were fixed. These values would allow to vary in the future according to the corresponding re-estimated Gaussian parameters to see if there is any performance gain.