ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

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, eeau@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.

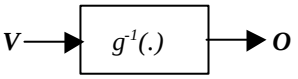
CMS tends to remove the distortion caused by the channel by subtracting a long-term mean of the cepstral features. In VTL, the frequency warping due to vocal tract length difference between different speakers is removed. Adaptation removes the speaker's style including accent, articulation and physical 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 $g^{-1}(.)$. To illustrate, let $\mathbf{V}$ denotes a vector of Gaussian random variables with marginal cumulative distribution function $P_V(.)$, and $\mathbf{O}$ 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:

$$g(o_i) = s_{V_i} P_{\mathbf{V}}^{-1}\left[P_{\mathbf{O}}\left(\frac{o_i - m_{\mathbf{O}}}{s_{\mathbf{O}}} \right) \right] + m_{V_i} \qquad (1)$$



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

$$f_O(\mathbf{o}) = (2\pi)^{-(k/2)} |\mathbf{R}_V|^{1/2} \exp\left\{ -\frac{1}{2} (g(\mathbf{o}) - \hat{\mathbf{i}}_V)^T \mathbf{R}_V^{-1} (g(\mathbf{o}) - \hat{\mathbf{i}}_V) \right\} \prod_{i=1}^{K} |g'(\mathbf{o}_i)| \qquad (2)$$

where m_V and $\mathbf{R}_V$ are the mean vector and covariance matrix for the random vector $\mathbf{V}$, $g'(\mathbf{o}_i)$ is the first order derivative of the function $g(\mathbf{o}_i)$ with respect to $\mathbf{o}_i$.

In the context of speech recognition, the output probability $b_{jm}(\mathbf{o}_t)$ of the new distribution function inside each HMM states becomes:

$$b_{jm}(\mathbf{o}_t) = \sum_{m=1}^{M} c_{jm} f(\mathbf{o}_t) \tag{3}$$

where j denotes the state, m denotes the mixture index with a maximum number of M mixtures, and c_{jm} is the mixture weight. The vector $\boldsymbol{o}_t$ is the observation vector.

Since the new density in Eq. (2) still has Gaussian form, it can be completely characterized by the memoryless nonlinearity, its mean vector and covariance matrix. These parameters, as well as the mixture weight, initial state probability $\boldsymbol{p}_i$, state-transition probability a_{ij}, can all be estimated by using the EM algorithm. The re-estimation formulas for the initial state distribution, state-transition probability distribution, mixture weight, mean, and covariance are given below.

Initial State Distribution:

$$\bar{p}_i = \frac{P(\mathbf{O}, q_0 = i \mid l)}{P(\mathbf{O} \mid l)} \tag{4}$$

State-Transition Probability Distribution:

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T} P(\mathbf{O}, q_{t-1} = i, q_t = j \mid l)}{\sum_{t=1}^{T} P(\mathbf{O}, q_{t-1} = i \mid l)} \tag{5}$$

Mixture Weight:

$$\bar{c}_{jm} = \frac{\sum_{t=1}^{T} \frac{P(\mathbf{O}, q_t = j, m_t = m \mid l)}{P(\mathbf{O} \mid l)}}{\sum_{t=1}^{T} \frac{P(\mathbf{O}, q_t = j \mid l)}{P(\mathbf{O} \mid l)}} \tag{6}$$

Mean:

$$\bar{\boldsymbol{\mu}}_{jm} = \frac{\sum_{t=1}^{T} \frac{P(\mathbf{O}, q_t = j, m_t = m \mid l)}{P(\mathbf{O} \mid l)} g_{jm}(\mathbf{o}_t)}{\sum_{t=1}^{T} \frac{P(\mathbf{O}, q_t = j, m_t = m \mid l)}{P(\mathbf{O} \mid l)}} \tag{7}$$

Covariance:

$$\bar{\mathbf{R}}_{jm} = \frac{\sum_{t=1}^{T} \frac{P(\mathbf{O}, q_t = j, m_t = m \mid l)}{P(\mathbf{O} \mid l)} [g_{jm}(\mathbf{o}_t) - \boldsymbol{\mu}_{jm}][g_{jm}(\mathbf{o}_t) - \boldsymbol{\mu}_{jm}]^T}{\sum_{t=1}^{T} \frac{P(\mathbf{O}, q_t = j, m_t = m \mid l)}{P(\mathbf{O} \mid l)}} \tag{8}$$

These equations are all derived by the EM algorithm and are very similar to the ones without the transformation. The only differences are the output distribution, mean, and covariance as stated in Eq. (3), (7), and (8). The only remaining problem is to find an appropriate *g(.)* function to successfully transform the trained Gaussian models in the HMM to match the distribution of the observation vectors.

Determining the "right" *g(.)* is a difficult task since there is no analytical way of figuring out what it is. Several different distribution functions can be used to see which function would produce a better model under a given condition. One of the functions that were used was the Laplacian function. This function was used because of its heavy-tailed distribution, which can more accurately model non-homogenous noise [3]. Homogeneous can be modeled by Gaussian's.

According to Eq. (1), the first step is to normalize the features by subtracting off the mean and dividing it by the variance. Then the cumulative distribution of the Laplacian is computed with the resulting value as the argument for the inverse Gaussian CDF. After the inverse Gaussian CDF is taken, the mean and variance are then applied according to Eq. (1) to compute the transformation function. Any initial value for 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-

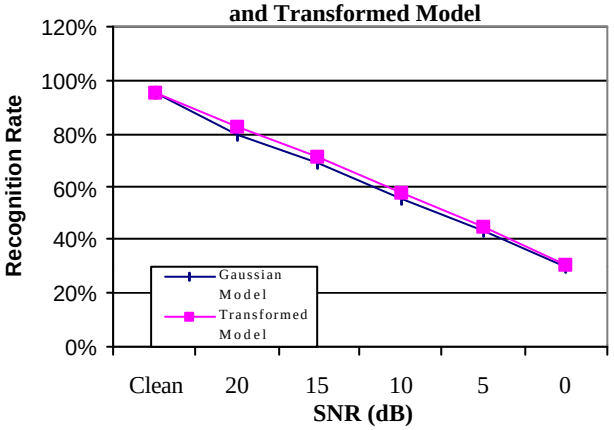
estimated parameters from training were used to form a new HMM.

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 1: Comparison of Recognition Rate between Gaussian Model and Transformed Model

SNR	Gaussian Model	Transformed Model
Clean	95.34%	95.45%
20 dB	80.01%	82.86%
15 dB	69.50%	71.62%
10 dB	55.21%	57.05%
5 dB	42.90%	44.67%
0 dB	29.89	30.20%

Figure 1: Recognition Rate for a HMM based speech recognizer using the Gaussian and Transformed Model



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.

ACKNOWLEGEMENT

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

5. REFERENCES

1. Rabiner, Lawrence and Juang, B.-H. "Fundamentals of Speech Recognition", 1993.
2. Au, O.C. and Thomas, J.B. "On Transformation Noise: Properties and Modeling", *Journal of Franklin Institute*, Vol-330, No. 4, p. 707-720, July 1993.
3. Burley, S. and Darnell, M. "Robust Impulse Noise Suppression Using Adaptive Wavelet De-noising", *ICASSP Proceedings*, 1997.
4. Juang, B.-H. "Maximum-Likelihood Estimation of Mixture Multivariate Stochastic Observations of Markov Chains", *AT&T Technical Journal*, 1984.
5. Huang, X.D., Ariki, Y., Jack, M.A. "Hidden Markov Models For Speech Recognition", 1990.