ROBUST ANALYSIS AND WEIGHTING ON MFCC COMPONENTS FOR SPEECH RECOGNITION AND SPEAKER IDENTIFICATION

Xi Zhou[1,2], *Yun Fu*[1,2,3], *Ming Liu*[1,2], *Mark Hasegawa-Johnson*[1,2], *Thomas S. Huang*[1,2]

[1]Beckman Institute, University of Illinois at Urbana-Champaign (UIUC), Urbana, IL 61801, USA
[2]Dept. of ECE, University of Illinois at Urbana-Champaign (UIUC), Urbana, IL 61801, USA
[3]Dept. of Statistics, University of Illinois at Urbana-Champaign (UIUC), Urbana, IL 61801, USA

ABSTRACT

Mismatch between training and testing data is a major error source for both Automatic Speech Recognition (ASR) and Automatic Speaker Identification (ASI). In this paper, we first present a statistical weighting concept to exploit the unequal sensitivity of Mel-Frequency Cepstral Coefficients (MFCC) components to against the mismatch, such as ambient noise, recording equipment, transmission channels, and inter-speaker variations. We further design a new Kullback-Leibler (KL) Distance based weighting algorithm according to the proposed weighting concept to real-world problems in which the label information is often not provided. We examine our algorithm in ASR with mismatch by different speakers and also in ASI with mismatch by channel noises. Experimental results demonstrate the effectiveness and robustness of our proposed method.

1. INTRODUCTION

The Mel-Frequency Cepstral Coefficients (MFCCs) are commonly used in speech features for Automatic Speech Recognition (ASR) [1] and Automatic Speaker Identification (ASI) [2] in the past decades. In laboratory environments, MFCC associated with statistical modeling, by means of Hidden Markov Model (HMM) or Gaussian Mixture Model (GMM), has achieved a high performance in both ASR and ASI. However, in the real applications, the performance degrades rapidly when there is a substantial mismatch between the training and testing conditions since the MFCC features are comparatively sensitive to additive mismatch distortion and each component of MFCC is in a different way to react against the unexpected mismatches. Variable mismatch environments, such as ambient noise, recording equipment, and transmission channels, result in a severe degradation of recognition performance [3]. Inter-speaker variations also increase the variability to the speech signal further degrades ASR performance.

To tackle the mismatch problem, a number of techniques have been presented for several years. In principle, the techniques can be classified into three categories according to the discussion in [4]: (1) *normalization* schemes try to reduce the mismatch by normalizing the acoustic features; (2) *adaptation* techniques use one or more transformations of the acoustic models to adapt it to the specific circumstances; (3) *weighting* methods implemented in decoding exploit the unequal sensitivities of acoustic feature components to against mismatch.

The weighting technique is attractive because it only needs a trivial change in the decoding process without changing on the feature and the model [4]. Furthermore, even adding the white noise to the speech signal, the effect is not the same at all frequency bins. (The SNR is different for each frequency bins.) The differences at the frequency bins cause the differences of cepstral coefficients from the filter banks, and therefore each component of MFCC features has different robustness to the mismatch when exposed to a mismatch environment [5]. Hence, the weighting method is reasonably applicable to deal with the mismatch problem in MFCC features.

In this paper, we present a novel weighting method to deal with the unequal sensitivity of MFCC components to against the mismatch. Firstly, we characterize each component of MFCC by quantifying its relative robustness to mismatch according to a provided statistical weighting concept. Then the weight for each component of MFCC is estimated through a new Kullback-Leibler Distance (KLD) [6] based weighting algorithm which does not require the label information to be available. Finally those different weightings are applied to the corresponding likelihood components in decoding. In the experiments, we examine our algorithm in ASR with mismatch by different speakers and also in ASI with mismatch by different channels. The results provide evidences to demonstrate that our methods outperform the equal weight setting steadily in both ASR and ASI under mismatch conditions.

The paper is organized as following: Section 2 quantifies the robustness of MFCC components by calculating a likelihood ratio. In Section 3, the new KLD based weighting algorithm is proposed and Section 4 introduces the method of applying the exponential weightings in decoding. The experimental results are shown in section 5. Finally, the conclusion and the future work are presented.

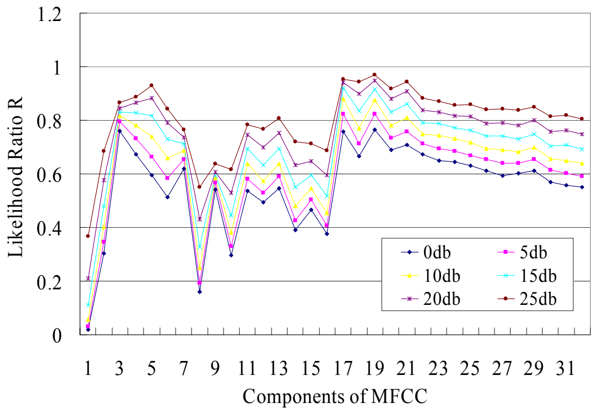

Fig. 1 Likelihood ratio of MFCC components at 6 different mismatch levels.

2. ROBUSTNESS OF MFCC COMPONENTS

MFCC features are comparatively sensitive to additive mismatch distortion. Just adding the white noise to the speech signal will affect the speech power spectrum at all the frequencies. However, the effect is not the same at all frequency bins; in another word, the signal-to-noise ratio is not the same in all the frequency bins. Moreover, the noise robustness of the dynamic MFCC feature, which is always used along with static MFCC feature in ASR and ASI, is better than that of static feature [7] [8]. Actually, when exposed to mismatch environments, each component of MFCC features has different sensitivity to the mismatch.

To quantify the relative sensitivity to the noise of each feature component, we calculate the likelihood ratio of each component between training and corresponding testing data. Let R_j denotes the likelihood ratio of the j-th component:

$$R_j = \frac{L\left(Y_j \mid \theta_j\right)}{L\left(X_j \mid \theta_j\right)} \tag{1}$$

where $X_j = \left(x_1^j, x_2^j, \cdots, x_n^j\right)$ is a sequence of MFCC's j-th component in the training condition, while $Y_j = \left(y_1^j, y_2^j, \cdots, y_n^j\right)$ is the corresponding features in the testing condition. Let $L\left(X_j \mid \theta_j\right)$ and $L\left(Y_j \mid \theta_j\right)$ denote the output likelihood of X and Y respectively, where θ is estimated based on X under maximum likelihood criterion. We use a single Gaussian probability density function to simplify our analysis. For the multi-mixture of Gaussian density function, we can follow the same but somewhat more complicated analysis. It follows

$$\begin{aligned}
L\left(X_j \mid \theta_j\right) &= \prod_{i=1}^{n} N\left(x_i^j; \mu_j, \sigma_j\right) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_j^{\,2}}} \exp\left\{-\frac{1}{2}\left(\frac{x_i^j - \mu_j}{\sigma_j}\right)^2\right\} \\
L\left(Y_j \mid \theta_j\right) &= \prod_{i=1}^{n} N\left(y_i^j; \mu_j, \sigma_j\right) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_j^{\,2}}} \exp\left\{-\frac{1}{2}\left(\frac{y_i^j - \mu_j}{\sigma_j}\right)^2\right\}
\end{aligned} \tag{2}$$

Then R_j can be re-written as

$$\begin{aligned}
R_j &= \exp\left\{\log L\left(Y_j \mid \theta_j\right) - \log L\left(X_j \mid \theta_j\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\sum_{\mathrm{i=1}}^{\mathrm{n}}\left(\left(\frac{y_i^j - \mu_j}{\sigma_j}\right)^2 - \left(\frac{x_i^j - \mu_j}{\sigma_j}\right)^2\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\sum_{\mathrm{i=1}}^{\mathrm{n}}\left(\frac{y_i^j - x_i^j}{\sigma_j}\right)^2\right\}
\end{aligned} \tag{3}$$

It is obvious to see that when R is equal to 1, the mismatch makes no influence on the data. The robustness of the feature is decreased when R decreasing.

We analyze the sensitivity of each component of MFCC on TIMIT dataset [11]. The clean speech data in TIMIT are used as the training data, while by adding Gaussian white noises to the clean speech data in TIMIT and controlling the SNRs from 0dB to 25dB, incrementing at a step of 5dB, we get six kinds of mismatch training data. Then we calculate the likelihood ratio R for each component of MFCC separately.

Figure 1 describes the likelihood ratio of MFCC components (1 to 16) as well as the corresponding dynamic components (17 to 32). We observe that

(1) The likelihood ratio changes a lot for these components which reveals the underlying unequal sensitivity to the mismatch;

(2) The likelihood ratio decreases corresponding to the decreasing SNRs for all the components of MFCC;

(3) The likelihood ratio of static MFCC components is smaller than that of its static counterpart, which supports the conclusion that the dynamic features are more resilient to noise than the static features.

3. KULLBACK-LEIBLER DISTANCE (KLD) BASED WEIGHTING

In the real application, it's hard or even impossible to get the corresponding utterances of training data in the test conditions. So we cannot calculate the weight of each component base on the likelihood ratio discussed at above section. It is also time consuming to label speech data for every new training condition.

From a statistical point of view, reducing the mismatch between training and testing conditions means to match the distributions of the features. The similarity between the distributions of training data and testing data reveals the robustness to the mismatch. Therefore, we introduced a KLD based weight method which needs no labeling information for both training and testing data.

KLD is a measure (a 'distance' in a heuristic sense) between reality distribution, p, and approximating model, q, and is defined as the integral for continuous functions where p and q are dimensional probability distributions. K–L distance, denoted by $D(p \| q)$, is the 'distance' when model q is used to approximate real distribution, p.

$$D(p \| q) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{4}$$

We adopt KL distance to measure the sensitivity of mismatch for each component, $d_i = D(p_i \| q_i)$ where p_i denotes the distribution of i-th component on the testing feature and q_i denotes the distribution defined by the model of i-th component.

Obviously, large distance means large difference between training and test feature in a special dimension, therefore the weight of this dimension should be less. We simply choose the weight $w_i = 1/d_i$.

4. EXPONENTIAL WEIGHTINGS IN DECODING

As discussed in the previous section, the likelihood decreases at each dimension due to the mismatch between the training and testing distributions. Since the sensitivity of MFCC's component is not the same against the mismatch, we propose to weight the log likelihood of each feature component differently in decoding to exploit their uneven noise robustness.

Assuming the feature components are mutually independent (actually it is assumed by most of the ASR and ASI systems that use the diagonal covariance matrices), the output likelihood of an observation in a Gaussian mixture density function can be splitted into the sum of one dimensional Gaussian components at the exponential term as Equation 5.

$$P(o_t \mid \theta) = \sum_{k=1}^{M} c_k \exp\left\{ \sum_{j=1}^{D} \log\left[N\left(o_t^j; \mu_k^j, \sigma_k^j\right)\right]\right\} \tag{5}$$

where k is the k-th Gaussian mixture index; c_k is the mixture weight. The acoustic likelihood components can be computed with different exponential weightings as:

$$P(o_t \mid \theta) = \sum_{k=1}^{M} c_k \exp\left\{ \sum_{j=1}^{D} w_j \log\left[N\left(o_t^j; \mu_k^j, \sigma_k^j\right)\right]\right\} \tag{6}$$

where w_j is the weight of the j-th component of MFCC.

5. EXPERIMENTS

5.1 ASR on TIDIGITS Corpus Database

To demonstrate the effectiveness of the proposed method for compensating the inter-speaker variations, we conduct large scale speech recognition on TIDIGITS corpus database [12] which consists of isolate digit and connected digits. The database contains 326 speakers (111 men, 114 women, 50 boys and 51 girls), each speaker producing 77 digit sequences including 22 isolated digit sequences. Data are equally splitted between training set and testing set for each category. The sequences can include 11 different digits, from "zero" to "nine", plus "oh". The data have been sampled at 20 kHz and digitalized with a resolution of 16 bits. All speech files were preprocessed into MFCC components, using HTK [9]. The feature includes 12 cepstral coefficients, the first derivatives, and the second derivatives, giving 36 components in total. The frame size was 25 ms and the shift was 10 ms.

We use the training data of man only to train HMMs, and then test on all the four categories (man, woman, boy, girl) of TIDIGITS corpus. The HMMs are composed of left-to-right no skip HMM which has 5 states and 8-component Gaussian mixture model for each state. HTK toolkit is used to train and evaluate the baseline system without weighting. The KLD based weighting method is used to estimate the weights for 36 components of MFCC.

Table 1 shows the Word Error Rate (WER) of our weighing method comparing with the system without weighing. Clearly, the performance is significantly degraded on mismatching condition where the models are trained with man's data and tested with another. The worst performance number happens at man-girl mismatch condition. After using our weighting method, the performance in mismatch condition (man-woman, man-boy, and man-girl) increased steadily. Meantime, in the match case (man-man), the performance only decreased a little (from 0.37% to 0.41%) from the originally system.

Table 1 Comparing the performance of weighting and without weighting in ASR

WER (%)	Man	Woman	Boy	Girl
Baseline	0.37	9.96	18.18	26.05
Weighting	0.41	8.15	16.54	24.27

5.2 ASI on TIMIT and NTIMIT Database

TIMIT [11] contains 6300 sentences spoken by 630 speakers. NTIMIT [13] is the TIMIT corpus re-played through the telephone network. In the speaker identification system, we use diagonal covariance Gaussian Mixture Model (GMM) for each speaker. The speaker dependent GMM is adapted from a Universal Background Model (UBM) through MAP criterion. The detail of the baseline system can be seen in [10]. The data of TIMIT are used for training and each 3-second sentence of NTIMIT is used for testing separately. We extract MFCC features from all the speech data. The feature includes 16 cepstral coefficients and the first derivatives, giving 32 components in total. The frame size was 20 ms and the shift was 10 ms. Again the KLD based weighting method is used for estimating the weights for 32 components of MFCC.

Figure 2 shows the performance of the proposed weighting method comparing with the method without weighting. With the increasing number of the speaker set, the Identification Error Rate (IER) of both methods decreased. However, our weighting method always keeps about half of the IER against the original method. When the total 630 speakers are included, the original method ends up with 7.26% IER, while the IER of our weighting method is only 3.48% and achieved 52% relative error reduction.

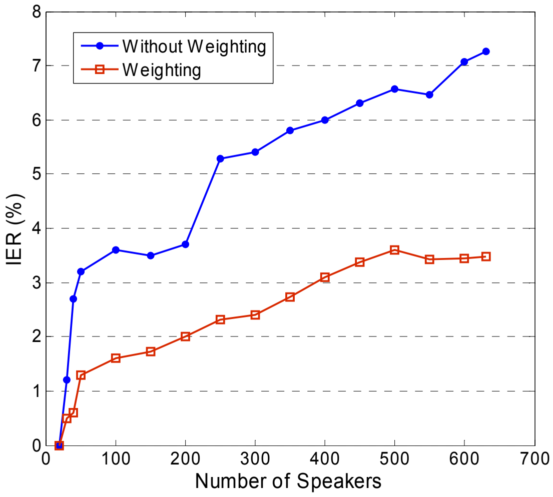

Fig. 2 Performance comparing the performance of with and without weighting in ASI

6. CONCLUSION

In this paper, we present the novel weighting method to deal with the unequal sensitivity of MFCC components reacting to the common problem–mismatch. Furthermore, we introduce a new KL Distance based weighting algorithm according to the proposed weighting concept to handle real-world problems in which the label information is often not provided. Our algorithm is demonstrated to be effective in both ASR with mismatch by different speakers and ASI with mismatch by different channel noises. The proposed method is very general which can be extended to other fields of study relative to pattern classification and feature representation, such as image and image processing.

7. ACKNOWLEDGEMENT

This work was funded in part by the Disruptive Technology Office VACE III Contract NBCHC060160 issued by DOI-NBC, Ft. Huachuca, AZ; and in part by the National Science Foundation Grant CCF 04-26627. The views and conclusions are those of the authors, not of the US Government or its Agencies.

8. REFERENCES

[1] H. Hermansky, “Mel Cepstrum, Deltas, Double Deltas, .. - What Else is New?” in *Proc. Robust Methods for Speech Recognition in Adverse Condition*, 1999.

[2] D. Reynolds and R. Rose, “Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models,” *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995.

[3] J. de Veth, B. Cranen, L. Boves “Acouslic Features and Distance Measure to Reduce Vulnerability of ASR Performance due to the Presence of a Communication Channel and/or Background Noise,” In: JLC. Junqua, *G.* van Noord (Eds.), *Robustness in Language and Speech Terknol*ogv. Kluwer Academic Publishers, Doordrccht, the Netherlands, pp. 9-45. 2001.

[4] C. Yang, F.K. Soong and T. Lee “Static and Dynamic Spectral Features: Their Noise Robustness and Optimal Weights for ASR,” *IEEE conf. on ICASSP’05*, vol.1, pp. 241- 244, 2005.

[5] K.K. Paliwal, “Spectral Subband Centroids as Features for Speech Recognition”, *IEEE conf. on ICASSP’98*, vol. 2, pp. 617-620, 1998.

[6] V. Krishnamurthy and J. Moore “On-line Estimation of Hidden Markov Hodel Parameters Based on the Kullback-Leibler Information Measure,” *IEEE Trans. on Signal Processing*, vol. 41, no. 8, pp. 2557-2573, 1993.

[7] S. Furui, “Speaker-Independent Isolated Word Recognition using Dynamic Features of Speech Spectrum,” *IEEE Trans.on Acoust. Speech Signal Proc.,* vol.34, pp.52-59, 1986.

[8] F. K. Soong and A. E. Rosenberg, “On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition,” *IEEE Trans .on Acoust. Speech Signal Proc.,* vol. 36, pp.871-879, 1988.

[9] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, *The HTK Book*, Cambridge Univ., 1996.

[10] D. A. Reynolds , T. F. Quatieri, and R. B. Dunn, “Speaker Verification using Adapted Gaussian Mixture Models,” *Digital Signal Processing*, vol. 10, pp.19-41, 2000.

[11] TIMIT, http://www.mpi.nl/world/tg/corpora/timit/timit.html

[12] R. G. Leonard, “A Database for Speaker-Independent Digit Recognition”, *IEEE conf. on ICASSP’84*, vol. 3, p. 42.11, 1984.

[13] W. Fisher, V. Zue, J. Bernstein, and D. Pallet. “An Acoustic-Phonetic Data Base". *J. Acoust. Soc. Amer. Suppl.(A)*, 81,S92,1986.