

LIKELIHOOD PROBABILITY MISMATCH ANALYSIS AND NORMALIZATION IN MULTILINGUAL SPEECH APPLICATIONS

Bin Ma, Cuntai Guan, Haizhou Li

InfoTalk Technology, Republic of Singapore
{bin.ma, cuntai.guan, haizhou.li}@infotalkcorp.com

ABSTRACT

In this paper, with a multilingual speech recognition system, we exam the HMM likelihood scores among the different acoustic models and observe that there exist scoring mismatches. The mismatches might come from different recording environments in which the training data for each language were collected, or come from different acoustic modeling structures. This analysis helps us understand the gaps among the likelihood probabilities on these acoustic models. Based on the observation of the differences of likelihood probability scores from different languages, we study a simple frame based likelihood probability normalization method to balance the likelihood scores of multiple acoustic models in the recognition system. Experiments show that this normalization method is effective to compensate the likelihood probability biases that come from different training corpora and different acoustic structures.

1. INTRODUCTION

Mixed-lingual and multilingual speech applications are of strong demand with the emerging need for globalization and growing international business interflow. It is quite common, especially in Asia, that more people now speak in mixed-language interchangeably even in one sentence. The challenge of such spoken dialogue systems is to carry out speech recognition in the mixed-lingual vocabulary without prior knowledge of language information, while several acoustic models of different languages are involved in the same time in the recognition system.

Since the acoustic models of different languages are usually trained separately with separated training corpora, the acoustic mismatches among the acoustic models of different languages possibly cause biases in the acoustic likelihood probabilities – which are used in most of the state-of-art speech recognition decoders. The biases might come from mismatching acoustic environments in which the training data for each language were collected, or come from different acoustic modeling structures. It could also come from the fact that acoustic models were trained with different size of the training set or different acoustic resolution for modeling phones in each individual language. Directly using these biased likelihood scores leads to such a problem that some languages are more “weighted” and thus result in unbalanced recognition results.

To deal with the mismatching acoustic environment problem, compensation algorithms have been widely used in robust speech recognition where compensation algorithms are accomplished in the signal, feature and model spaces [1,2,3,4,5] to reduce the various distortions caused by different acoustic environments. Here, compensation methods are also needed to normalize the biased likelihood probability scores.

In this paper, with the bilingual speech recognition system, including Mandarin and English, we analyze the mismatching likelihood scores among the different acoustic models and acoustic modeling structures. Based on the observation of the gaps among the likelihood probabilities from the acoustic models of different languages, a simple frame-based likelihood probability normalization method is used in speech recognition system to balance the likelihood scores among the multiple acoustic models. Experiments show that this normalization method is effective to compensate the likelihood probability biases among the languages and is helpful to balance recognition accuracies of the languages in the multilingual speech recognition system.

2. MULTILINGUAL AUTOMATIC SPEECH RECOGNITION (ASR)

2.1 Problem Statement in Multilingual Speech Recognition

Given a speech signal X , a speech recognizer tries to identify the word from a set of words in a language by maximizing the posteriori probability, i.e.

$$\hat{W} = \arg \max_w P(W | X) \quad (1)$$

In a multilingual speech recognition system, the vocabulary is extended to more than one language, each of which is represented by the individual acoustic model set.

$$\begin{aligned} \hat{W} &= \arg \max_w \sum_i P(W | X, L_i) \\ &= \arg \max_w \sum_i \frac{P(X | W, L_i) \cdot P(W, L_i)}{P(X)} \end{aligned} \quad (2)$$

In a maximum likelihood framework, the above formula could be simplified as

$$\hat{W} = \arg \max_w \sum_i P(X | W, L_i) \cdot P(W, L_i) \quad (3)$$

There exists a difficulty in performing this maximization because not all words are allowed in each language (some phones existing in one language are missing in others). Conventionally, we will carry out word recognition for each language and the winning word in a language takes all:

$$\hat{W} = \arg \max_j P(W_j | X) \approx \arg \max_{ij} P(X | W_{ij}, \Lambda_i) \quad (4)$$

where the language dependency is now expressed in terms of the set of the acoustic phone models for a language of focus. The likelihood score in Equation (4) clearly depends on the acoustic models used in evaluating the score and any score ‘‘bias’’ is likely to be reflected in the comparison making recognition result incorrect if no score compensation is performed.

This fact is demonstrated in the following. In Figure 1 we plot the empirical distribution of the average frame likelihood scores generated from evaluating a set of common utterances on the available sets of English and Mandarin acoustic models. The score plots in Figure 1 clearly show a biased preference towards recognizing Mandarin over English words. Obviously, this offset will lead to unbalanced recognition performance between the two languages. This bias might come from mismatched recording environments in which the training data for each language were collected. It could also come from the fact that acoustic models were trained with different sizes of the training set or different acoustic resolution for modeling phones in each individual language.

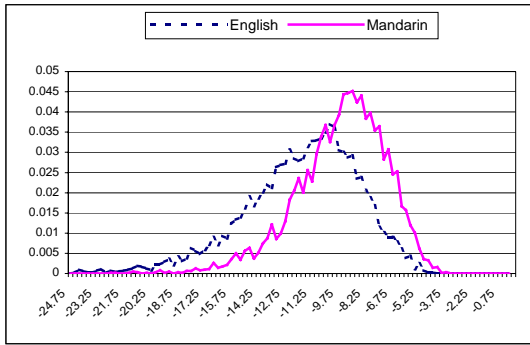


Figure 1: *Distribution of Likelihood Probability (English and Mandarin) for the same utterances*

2.2 Likelihood Probability Normalization

Suppose that $X = \{x_t, t=1, \dots, T\}$ is speech signal for a word W_{ij} of the language i , and $Q_{ij} = \{q_{ijt}, t=1, \dots, T\}$ is the corresponding phone sequence for the W_{ij} obtained via Viterbi search:

$$\begin{aligned} \hat{W} &\approx \arg \max_{ij} P(X | W_{ij}, \Lambda_i) \\ &= \arg \max_{ij} \prod_{t=1}^T P(x_t | q_{ijt}, \Lambda_i) \end{aligned} \quad (5)$$

It is not trivial to find the maximization of (5) explicitly, but the observations in section 2.1 motivate a simple way to find some sub-optimal solutions. Here we propose to add a linear language-specific weight to each frame of the likelihood probability. Let $\bar{\omega}_i$ be the weight of languages i , and rewriting equation (5), we obtain:

$$\begin{aligned} \hat{W} &\approx \arg \max_{ij} \prod_{t=1}^T \bar{\omega}_i \cdot P(x_t | q_{ijt}, \Lambda_i) \\ &= \arg \max_{ij} \sum_{t=1}^T [\omega_i + \log(P(x_t | q_{ijt}, \Lambda_i))] \end{aligned} \quad (6)$$

Where $\omega_i = \log(\bar{\omega}_i)$.

2.3 Compensation Weight Estimation

The simplest way to estimate the compensation weights is to calculate the frame-based average log probability for each language in the training process, and then compensate the biases among the different languages by adding a compensation factor to the language of lower acoustic likelihood score. In the Mandarin-English bilingual system, the frame-based average score for the log probability is calculated from the training process as follows:

$$\begin{aligned} \eta_{\text{mandarin}} &= \frac{1}{T_m} \sum_{t=1}^{T_m} \log(P(x_t | q_{mt}, \Lambda_m)) \\ \eta_{\text{english}} &= \frac{1}{T_e} \sum_{t=1}^{T_e} \log(P(x_t | q_{et}, \Lambda_e)) \end{aligned} \quad (7)$$

Where m denotes the acoustic model of Mandarin, and e denotes the acoustic model of English. The bias score between the Mandarin and English is obtain as:

$$\Delta = \eta_{\text{mandarin}} - \eta_{\text{english}} \quad (8)$$

So in the Mandarin-English bilingual system, ω_{mandarin} is set to 0, and ω_{english} is set to Δ .

2.4 Equivalent Phone Class for Compensation Weight Estimation

An alternative way to estimate the compensation weights is to use part of training data in which not all the phoneme segmentations are involved. Only those phonemes in Mandarin and English that have similar acoustic characteristics are considered.

A common set of phones across the languages of consideration, called ‘‘equivalent class’’ is defined. For each utterance that is

used in the equation (7), only speech segments that fall into the equivalent class are of interests.

The equivalent class should be defined based on acoustic characteristics of all the phones in the multilingual context. Specifically, a pair of phones in two languages belongs to one equivalent phone class Θ as is defined as follows

$$p_k, p_l \in \Theta,$$

$$\Theta = \{p_k, p_l \mid |P(X \mid \Lambda_{L_k}, p_k) - P(X \mid \Lambda_{L_l}, p_l)| < \varepsilon\} \quad (9)$$

Figures 2 and 3 show the average likelihood scores for the segments of Mandarin speech /I/ computed with Mandarin (Figure 2) and English acoustic models (Figure 3). From the two plots, it can be seen that the sound /I/ gives the highest scores by the Mandarin model “I” and English model “IY”. So Mandarin /I/ and English /IY/ can be classified as belonging to the same equivalent phone class. Similar behavior can be observed when the English sound segments of /IY/ were evaluated on the Mandarin and English models, respectively. Thus a set of equivalent classes can be designed.

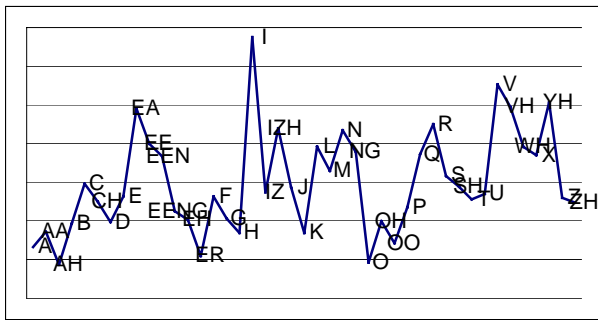


Figure 2: Mean Distribution of Acoustic Probability for Mandarin sound /I/ over all Mandarin models

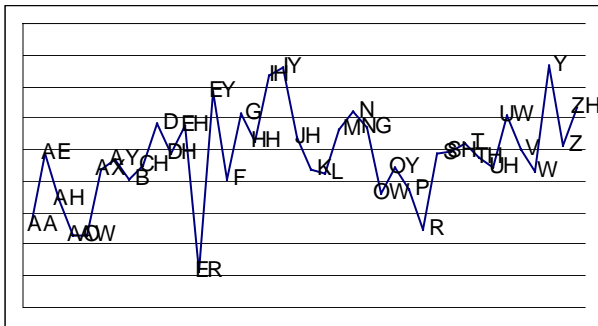


Figure 3: Mean Distribution of Acoustic Probability for Speech segments of Mandarin sound /I/ over all English models

3. EXPERIMENT AND RESULTS

3.1 Training Database

We implement the experiments on a bilingual (Mandarin and English) telephony speech recognition system. The training data

are collected from fix-line and mobile telephony systems. The acoustic models of Mandarin and English are trained separately with Mandarin training corpus and English training corpus individually. In all the experiments, the training corpora for these two languages are kept unchanged.

3.2 Structure of Acoustic Modeling

To show the effects of different acoustic modeling structures on likelihood probability bias in the multilingual system, two sets of acoustic model for each language are built. These two sets of model are mainly different in the acoustic resolution.

- **Model Structure 1**

Context-dependent phone tying model. All the context-dependent sub-word units with the same central phoneme share one set of 128 Gaussian mixture components.

- **Model Structure 2**

Phonetic decision-tree based state-tying model. 1500 tied states, each with 16 Gaussian mixture components, are constructed based on the phonetic decision-tree.

The above-mentioned acoustic modeling structures are in different parameter sharing schemes and different acoustic resolutions. They might cause unbalanced recognition results by bringing different scores of likelihood probability.

3.3 Testing Set

The experiments are implemented in the mix-language isolated word task. The recognition grammar consists of a mixed word list of both Mandarin and English. The testing set consists of 1000 utterances of isolated words for each of Mandarin and English.

To each utterance in the testing set, Viterbi search is used to calculate the likelihood probabilities with acoustic models of two languages and the mix-language word list grammar. The recognition result is a word either in Mandarin or in English. Since there exists likelihood probability bias with the acoustic models of Mandarin and English, the performance before the likelihood probability normalization is unbalanced. The accuracy of English testing set is quite low at this case.

3.4 Likelihood Probability Normalization

In the first set of experiments, **Model Structure 1** is used as the acoustic modeling structures of both Mandarin and English. Since the acoustic modeling structures of Mandarin and English are same, the likelihood probability bias mainly comes from the different training corpora. In Figure 4, we plot the recognition performance in Word Error Rate (WER) of Mandarin, English and the average of the two languages, while different compensation weights are added on the likelihood probability calculation of English acoustic modeling.

In another set of experiments, Mandarin acoustic model is trained under **Model Structure 2**, while the acoustic modeling structure of English remains unchanged. The recognition performances (WER) by adding different compensation weights on the likelihood probability calculation of English acoustic modeling are shown in the Figure 5.

Comparing the plots in Figure 4 and Figure 5, we can find that even with same training data sets, the behaviors of the likelihood probability compensation under different acoustic modeling structures are not the same. The acoustic modeling structure and acoustic resolution have impact on the likelihood probability bias.

4. CONCLUSION

Multilingual speech recognition is becoming an important practical problem when more and more spoken dialogue applications are being deployed in Asia. Due to the fact that there could exist a bias in acoustic scores from different languages, it is interesting to find ways to compensate for this score bias, which could come from different acoustic and recording conditions, different sizes of the training set for each language and different acoustic and phonetic resolution in modeling each of the languages of interest. We have studied a multilingual word recognition task with Mandarin and English as the two intended languages. We found that such acoustic score bias causes a serious performance degradation on English. We then introduce a simple score normalization scheme by adding compensation factors to the acoustic likelihood probability calculation of the acoustic model of each involved language. We find that this simple normalization method can help to resolve the problem of unbalanced recognition results.

Acknowledgement

The authors would like to thank Prof Chin-Hui Lee of National University of Singapore for his valuable inputs and fruitful discussions during the work of this paper.

5. REFERENCE

- [1] A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition", In, Proc. ICASSP, volume 2, 849-852, 1990.
- [2] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", Journal Acoust. Soc. Am., 55(6), 1304-1312, 1974.
- [3] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures", In Proc. ICASSP, 353-356, 1996.
- [4] M. Rahim and B.-H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition", IEEE Trans. on Speech and Audio Processing, 4(1), 19-30, 1996.
- [5] A. Sanker and C.-H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition", IEEE, Trans. on Speech and Audio Processing, 4(3), 190-202, 1996.

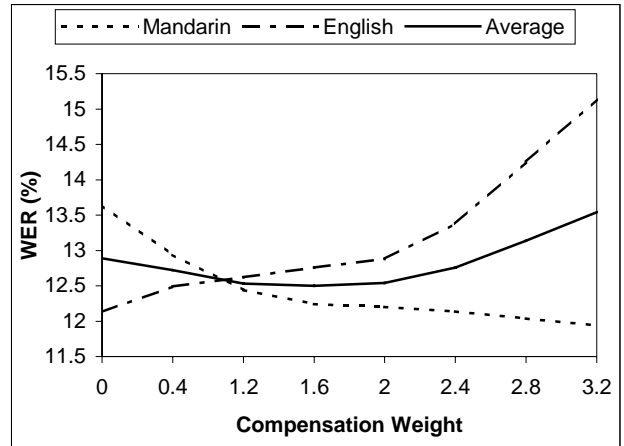


Figure 4: WER with different compensation weights on English acoustic likelihood probability under the same acoustic structures of Mandarin and English

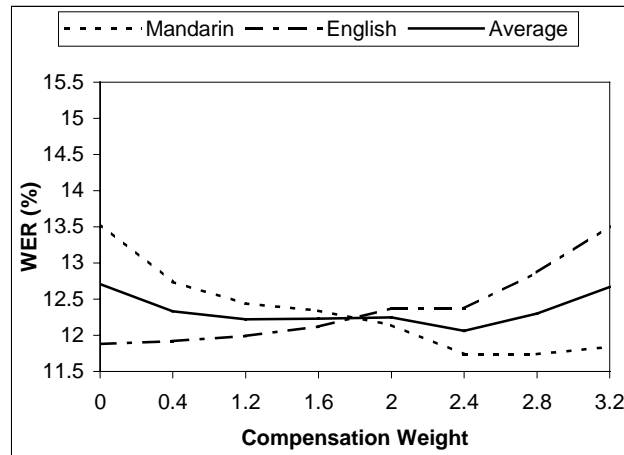


Figure 5: WER with different compensation weights on English acoustic likelihood probability under different acoustic structures of Mandarin and English