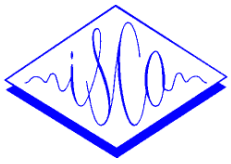

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

MULTILINGUAL SPEECH RECOGNITION WITH LANGUAGE IDENTIFICATION

*Bin Ma, Cuntai Guan, Haizhou Li and Chin-Hui Lee**

InfoTalk Technology, Republic of Singapore
{bin.ma, cuntai.guan, haizhou.li}@infotalkcorp.com
*School of Computing, National University of Singapore
chl@comp.nus.edu.sg

ABSTRACT

This paper presents a new approach to multilingual speech recognition. The proposed algorithm combines both language identification (LID) and speech recognition into a single process. It is shown to be effective for multilingual grammar-based speech recognition where the language information is not available prior to recognition. The idea is to make use of acoustic-phonetic and lexical information in each language to reduce possible mismatch caused by potential difference in acoustic and recording conditions when the training utterances for each language were collected. By doing so, it is shown that, with the help of LID information, the word error rate of a mixed Mandarin and English speech recognition system is greatly reduced. The same formulation can also be used to enhance language identification accuracy.

1. INTRODUCTION

Spoken dialogue systems have received great interests in recent years by helping business reduce operational cost and upgrade customer services, especially in telecommunications, enterprise service automation and customer relation management (CRM). Mixed-lingual and multilingual 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 here is to carry out speech recognition in the mixed-lingual vocabulary without prior knowledge of language information. In other words, it becomes necessary to make both language identification and word recognition decisions at the same time on short utterances or speech segments.

It is known that language characteristics reside in a spoken utterance at different levels, such as acoustic-phonetic-prosodic [1,4], phonotactic [2,3], lexical and grammatical structures [5]. Intensive research efforts on language identification (LID) have been made to explore the effectiveness of different language characteristics and their combinations.

In some earlier studies, it is shown that acoustic-prosodic features reflect the phonetic pronunciation pattern that is considered as the underlying representation of phonological unit in a language [2,7]; the lexical rules governing a language are encoded in the phonotactics among phonemes, which can be modeled by phoneme *n*-gram and are viewed as additional language discriminative evidence [3]. It is also shown that the combination of phonotactic and acoustic-posodic information provides promising language discriminative ability with low computational cost [3].

Some other recent studies, using large vocabulary speech recognizer, demonstrate that lexical and grammatical structures are the most effective source of language discriminative knowledge at the cost of more expensive computation [5]. Inspired by this finding, in view of the fact that grammatical and lexical based speech recognition is needed anyway in the case of multilingual spoken dialogue applications, we believe that grammar-based LID provides a complementary source of information to improve mixed-lingual recognition accuracy.

One of the major difficulties in using LID for multilingual speech recognition is that it is often required to use long utterances to warrant high LID accuracy in conventional approaches. However short utterances are typical for speech recognition. We propose a likelihood score normalization scheme to compensate for possible mismatch in the acoustic models for each individual language. We also propose a classification-based LID approach that is quite effective even for short utterances. The two techniques were combined and tested on a large vocabulary word recognition task with Mandarin and English as the intended languages. Experiments indicated that the proposed techniques improve the LID accuracy and greatly reduce the word error rate. The same formulation is equally applicable to speech recognition with more than two languages.

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.

$$\hat{W} = \arg\max_{w} P(W \mid X) \qquad (1)$$

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

$$\hat{W} = \arg\max_{w} \sum_{i} P(W \mid X, L_i)$$
$$= \arg\max_{w} \sum_{i} \frac{P(X \mid W, L_i) \cdot P(W, L_i)}{P(X)} \quad (2)$$

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

$$\hat{W} = \arg\max_{w} \sum_{i} P(X \mid W, L_i) \cdot P(W, L_i) \quad (3)$$

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:

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

where the language dependency is now expressed in terms of the set of the acoustic phone models for that language. 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.

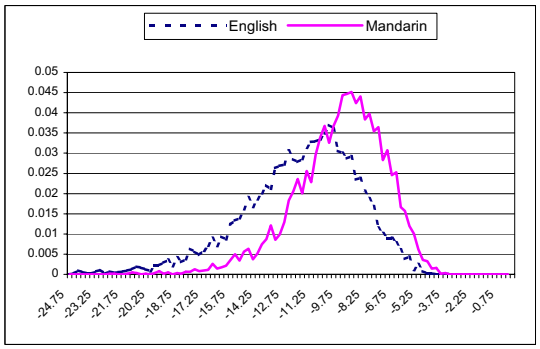

Figure 1: *Distribution of Likelihood Probability (English and Mandarin) for the same utterances*

2.2. Proposed Approach

Rewriting (2), we need to evaluate the following:

$$\hat{W} = \arg\max_{w} P(W \mid X) = \arg\max_{w} \sum_{i} P(W \mid X, L_i)$$
$$= \arg\max_{w} \sum_{i} \frac{P(X \mid \Lambda_{L_i}, W) \cdot P(W \mid L_i) \cdot P(L_i)}{\sum_{j,k} P(X \mid \Lambda_{L_j}, W_k) \cdot P(W_k \mid L_j) \cdot P(L_j)} \quad (5)$$

To simplify the maximization of Equation (5), one simple way is to do language ID first followed by word recognition in the identified language, i.e.

$$\hat{W} = \arg\max_{j} P(W_j \mid X) \approx \arg\max_{j} [\max_{i} P(W_{ij} \mid X, L_i)] \quad (6)$$

Considering large vocabulary speech recognition, the above maximization can be done in the following steps: (1) one or more recognition hypotheses are generated by evaluating likelihood scores using all the acoustic models of all the languages involved; (2) language identification is performed as described in the next section; and (3) the recognition result is the best word hypothesis of the recognized language.

2.3. Language Identification

For a speech signal X, a language identifier tries to identify the language from a set of languages represented by the individual acoustic model set of phones. Assuming q is a sequence of phones, we can evaluate the following:

$$\hat{L} = \arg\max_{L} P(L \mid X) = \arg\max_{L} \sum_{q} P(L \mid X, q)$$
$$= \arg\max_{L} \sum_{q} \frac{P(X \mid \Lambda_L, q) \cdot P(q \mid L) \cdot P(L)}{\sum_{i} P(X \mid \Lambda_{L_i}, q) \cdot P(q \mid L_i) \cdot P(L_i)} \quad (7)$$

Now assume X could be partitioned into M independent phone-segments as specified by any q, then we have

$$\hat{L} = \arg\max_{L} P(L \mid X) = \arg\max_{L} \sum_{q} P(L \mid X, q)$$
$$= \arg\max_{L} \sum_{q} \prod_{j} P(L \mid X_j, q_j) \quad (8)$$
$$= \arg\max_{L} \sum_{q} \prod_{j} \frac{P(X_j \mid \Lambda_L, q_j) \cdot P(q_j \mid L) \cdot P(L)}{\sum_{i} P(X_j \mid \Lambda_{L_i}, q_j) \cdot P(q_j \mid L_i) \cdot P(L_i)}$$

It is not easy to fully calculate the maximization equation in (8), so certain approximation is necessary. There are two sums in this equation. The first sum can be fulfilled by picking the top few phone sequences for each language in consideration and score all of them with each set of language-dependent acoustic models. The second sum in the denominator is more difficult but could again be approximated by N-best phone scores from a corresponding cohort set of the phone [8].

2.4. Equivalent Phone Class for Language Identification

To approximate the second sum in the denominator in Equation (8), we proposed to define a common set of phones across all the languages of consideration, called "equivalent class". For each utterance that is used for evaluation Eq. (8), 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

$$p_k, p_l \in \Theta,$$
$$\Theta = \{p_k, p_l \mid \left| P(X \mid \Lambda_{L_i}, p_k) - P(X \mid \Lambda_{L_j}, p_l) \right| < \varepsilon\} \quad (9)$$

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.

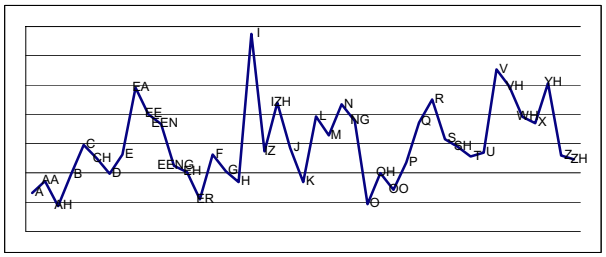

Figure 2*: Mean Distribution of Acoustic Probability for Mandarin sound /I/ over all Mandarin models*

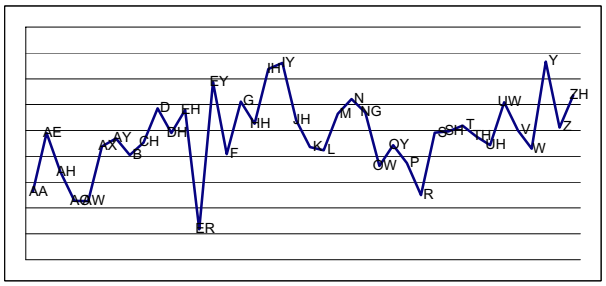

Figure 3*: Mean Distribution of Acoustic Probability for Speech segments of Mandarin sound /I/ over all English models*

3. EXPERIMENT AND RESULTS

3.1. Database

We implement the experiments on a bilingual large vocabulary isolated word recognition task. There are 42,000 Mandarin and 35,000 English words, respectively, in the recognition grammar. The testing set consists of 500 utterances of isolated words for each of English and Mandarin. The average frame numbers (12.5 ms for each frame and silence portion is excluded) for the testing utterances are 56 for Mandarin and 64 for English. Such short duration utterances are usually difficult to achieve high language identification accuracy.

3.2. Conventional Multilingual Speech Recognition

In this bilingual recognition task, we perform LID first and then do word recognition according to the result of the LID. To each utterance in the testing set, Viterbi search is used to calculate the likelihood probabilities with acoustic models and word list grammars of Mandarin and English, respectively. LID result is decided by comparing the two likelihood probability scores.

As mentioned in Section 2.1, there is an inherent "bias" among the different languages. As a result, this method shows a biased performance. Table 1 lists the LID and ASR accuracies in the tests. We can see that most of English utterances were wrongly recognized in the LID test.

	Mandarin	English	Average
LID (%)	100.0	7.6	53.8
ASR (%)	84.8	6.0	45.4

Table 1: *Conventional Multilingual ASR*

3.3. Multilingual Recognition with the Proposed Approach

By incorporating the normalization effect in Eq. (8), we obtained in Figure 4 the distributions of the acoustic likelihood scores with English and Mandarin acoustic models on the same set of utterances used to obtain the plots in Figure 1. It clearly shows that the acoustic scores are now aligned very well between English and Mandarin. Using the LID algorithm as given in Eq. (8), the results got much better as shown in Table 2.

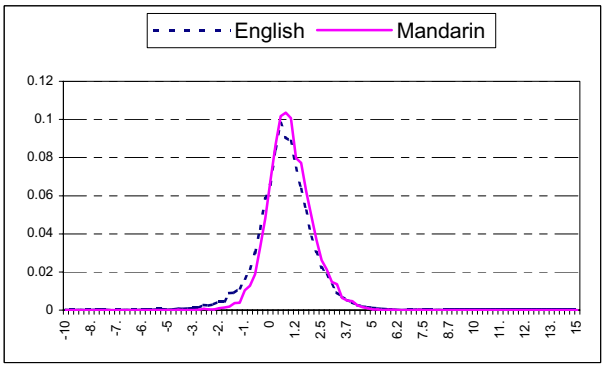

Figure 4*: Distribution of Normalized Acoustic Scores*

	Mandarin	English	Average
LID (%)	98.0	76.6	87.3
ASR (%)	83.8	64.8	74.3

Table 2: *LID and ASR Results with Equation (8)*

3.4. Classifier for Language ID

We examined the acoustic score distributions of all the testing utterances with Mandarin & English acoustic models. It was found that an even better LID result can be achieved if we design a classifier to classify the testing utterances further based on the acoustic scores coming from Equation (8). In our experiments, a simple linear classifier is trained with another set of 500 Mandarin and 500 English utterances. Scatter plots of Mandarin versus English raw and normalized scores, with and without applying Eq. (8), are shown on left hand side of Figure 5. It is clear that the language separation is better for the normalized scores. The separation is even better for longer utterances. The cases with average scores computed from 3 utterances are plotted on the right hand side of Figure 5. Another interesting fact is that the logarithm of normalized scores tend to be centered around 0, reflecting the fact of the nature of using such *log generalized likelihood ratio score* for designing a language classifier is beneficial.

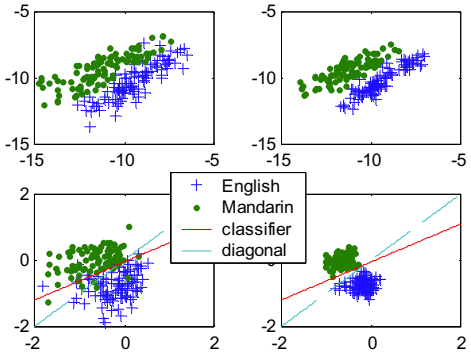

Figure 5: *Scatter plots for raw scores (upper), for normalized scores (lower)*

The LID & ASR recognition open test results are shown in the first two rows of Table 3. On average for each test utterance, we only have about 0.75 second speech data. To compare the LID performance over various speech utterance sizes, we evaluated the LID performance using 2 and 3 utterances. These results are shown in the rows 3 and 5, respectively in the Table 3. In our bilingual experiments, 3 utterances or about 2.2 second of speech can make a perfect separation of the two languages.

	Mandarin	English	Average
LID (%) 1 Utt	89.8	94.0	91.9
ASR (%) 1 Utt	77.6	78.8	78.2
LID (%) 2 Utt	99.6	99.8	99.7
ASR (%) 2 Utt	84.8	84.4	84.6
LID (%) 3 Utt	100.0	100.0	100.0
ASR (%) 3 Utt	85.2	84.4	84.8

Table 3: *LID and ASR results with a linear LID Classifier*

The corresponding ASR results after performing LID is also listed in Table 3. Some interesting results were observed. First, the ASR word accuracies for English and Mandarin are now more or less balanced in all cases. Second, an improved LID performance clearly enhanced the ASR accuracies as shown in the cases with LID using two utterances. Once the LID performance saturated, there is no more clear improvement for ASR. Further word error reduction would come from better acoustic modeling in each language and maybe also from jointly modeling the phones of the two languages.

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 an inherent bias in acoustic scores from different languages, it is also theoretically 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 such acoustic score bias causes a serious performance degradation. We then propose two approaches to improve our system. First, we propose a score normalization scheme by incorporating *N*-best scores from competing phone models into evaluating the normalized likelihood scores. We also propose to use a linear classifier to perform language identification before speech recognition is carried out. Although this could potentially be sub-optimal, we found good LID accuracy could be achieved even with less than 1 second of speech. Perfect LID performance is obtained when using about 3 input isolated word utterances (only slightly over 2 seconds of speech data). We also found that by performing LID first, the performance of our Mandarin-English word recognition task was greatly enhanced. We intend to explore other LID techniques, such as language identification without word recognition, in the future.

REFERENCES

[1] M.A. Zissman, K. M. Berkling, "Automatic language identification", Multi-lingual Interoperability in Speech Technology, Leusden, Sept. 1999.

[2] Y. K. Muthusamy, E. Barnard, and R.A. Cole, "Reviewing automatic language recognition", IEEE Signal Processing Magazine, Oct 1994.

[3] Jiri Navratil, ``Spoken Language Recognition – A Step Towards Multilinguality in Speech Processing", *IEEE Trans. Speech and Audio Proc*., 9(6), pp. 678-685, 2001.

[4] M.A. Zissman, "Comparison of four approaches to automatic language identification", IEEE Trans. Speech and Audio Proc., 4(1), pp31-44, 1996.

[5] T. Schultz, I. Rogina and A. Waibel, "LVCSR-based Language Identification", Proc. Int. Conf. Acoust. Speech, Signal Proc, Atlanta, May 1996.

[6] Y.Yan, E. Barnard, and R.A.Cole, "Development of an approach to automatic language identification based on phone recognition", Computer Speech Language 10(1), pp 37-45, 1996.

[7] T. J. Hazen and V. W. Zue, "Segment-based automatic language identification", J. Acoust. Soc. Amer., 101(4), pp. 2323-2331, 1997.

[8] R. A. Sukkar and C.-H. Lee, "Vocabulary Independent Discriminative Utterance Verification for Non-Keyword Rejection in Subword Based Speech Recognition," IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 6, pp. 420-429, Nov. 1996.