MULTILINGUAL PRONUNCIATION MODELING FOR IMPROVING MULTILINGUAL SPEECH RECOGNITION

Jilei Tian1, Juha Häkkinen2, Olli Viikki1

1Nokia Research Center, Speech and Audio Systems Laboratory, Tampere, Finland
2Nokia Mobile Phones, Tampere, Finland

{Jilei.Tian, Juha.M.Hakkinen, Olli.Vikki}@nokia.com

ABSTRACT

Multilinguality aspects are becoming increasingly important in the Automatic Speech Recognition (ASR) systems. It is apparent that coping with large variability of the speech signal is an even bigger challenge in multilingual ASR systems than it has been in conventional monolingual systems. In this paper, we address the importance of combining multilingual pronunciation modeling and acoustic model adaptation. To compensate the pronunciation variability across various speakers, multilingual pronunciation modeling method is proposed. Due to the limited processing power and memory resources available in many systems, we also propose a pruning scheme that removes pronunciation variants from the vocabulary based on the statistical scores obtained during the deployment of the system. To further compensate the mismatches between the multilingual acoustic models and the speaker’s pronunciation, on-line MAP acoustic model adaptation is applied. Experimental results with 25 languages indicate the usefulness and efficiency of the joint use of these techniques both in clean and noisy conditions.

1. INTRODUCTION

The use of multilingual technology (e.g., different languages use the same set of acoustic models [1]) enables an ASR system to simultaneously support multiple languages even within limited memory resources [8]. It is apparent that coping with large variability of the speech signal becomes even a bigger challenge in multilingual ASR systems than it has been in conventional monolingual systems. Speaker variability occurs both within a language (e.g., dialects) and between languages (e.g., the mismatch between multilingual acoustic models and the language of the current speaker). It is necessary to design methods that are capable of compensating the effects in both cases. Over the recent years, acoustic model adaptation techniques [4][9] have successfully been applied in HMM-based ASR. It is obvious that acoustic model adaptation techniques also play a significant role in multilingual ASR to reduce these mismatches since the language-specific details cannot be characterized very accurately. In addition to the acoustic variability, there is also a great deal of pronunciation variability involved in multilingual ASR systems which needs to be compensated.

Non-native vocabulary items and the mismatches between the speakers and the original speaker-independent acoustic models are two major reasons resulting in performance degradation in multilingual speech recognition systems. Prior knowledge of the language or speech-based language identification are commonly used in multilingual ASR [2]. It is, however, not feasible in command and control type of applications, where only a small amount of speech material is available, and where the vocabulary may be inherently multilingual (consider, e.g., a navigation system). A statistical multilingual text-to-phoneme mapping method was presented in [5]. The presented method allows for pronunciation variation at pheme level. In this paper, a different approach is taken. We combine multilingual pronunciation modeling and on-line acoustic model adaptation to maximize the speaker-specific recognition accuracy within constrained system resources.

When the multilingual pronunciation approach, where one pronunciation variant is generated for each language supported and the acoustic model adaptation are enabled, the system converges to a single, most likely, pronunciation for each entry, and the complexity of the system reduces to the same level with monolingual ASR systems.

The rest of the paper is organized as follows. Section 2 gives a brief overview of the multilingual ASR framework in which the new methods are integrated. Then we outline the multilingual pronunciation modeling principles and pruning algorithm. This is followed by experimental results confirming the usefulness of the proposed techniques in Section 3. Finally, conclusions are given in Section 4.

2. MULTILINGUAL PRONUNCIATION MODELING

2.1. Multilingual ASR system

Figure 1 illustrates an architecture proposed for multilingual ASR systems [8]. The multilingual ASR engine consists of three key units: automatic language identification, on-line pronunciation modeling, and multilingual acoustic modeling modules. The assumption is that the vocabulary items are given in the textual form. First, the Language Identification (LID) module detects the language of the vocabulary item [3]. Once this has been determined, an appropriate on-line text/grapheme-to-phoneme conversion is applied to obtain the phoneme...
Finally, the recognition model for each vocabulary item is constructed by concatenating the multilingual acoustic monophone models. Using these basic modules the recognizer can, in principle, automatically cope with multilingual vocabulary items without any assistance from the user.

2.2. Language identification

One of the key challenges is to apply the correct pronunciation modeling scheme for each of the vocabulary entries. To choose the appropriate text-to-phoneme conversion scheme for a given entry, one first needs to know or determine the language identity of the entry. LID can be done automatically from the written form of each entry. Automatic LID has several problems that necessitate the use of multilingual pronunciation modeling methods.

- The automatic LID is prone to errors resulting in erroneous transcriptions. Typically, the correct language identity is among the $n$-best ($n$ being about 3-5) language candidates provided by the LID module.
- LID is an ambiguous process as it is not always possible to say what is the language identity of a certain entry. For instance, the name "Peter" appears in many European languages and it is pronounced in somewhat different ways in different languages.
- Speaker can utter a non-native entry in a different way than the language that the entry actually belongs to. The recognition accuracy of non-native entries can be improved by generating pronunciation variants using the speaker's native text-to-phoneme scheme instead of that of the entry.

2.3. Multilingual pronunciation generation

Multilingual pronunciation variants are generated according to the $n$-best classification of LID. It is also advantageous to take the speaker's native language, and use it as the default language for all entries in addition to the $n$-best languages (see Figure 2). When the number of pronunciation variants increases, their confusability may increase as well. Therefore, only a certain number (4 or 5) of the best $n$-best pronunciation variants should be used.

$$\text{Pron} = (W_j, \text{Lang}_j)$$

$$= \arg \min_{(W_j, \text{Lang}_j)} P(W_j, \text{Lang}_j)$$

(1)

The two items on the right side of Equation (1) are updated as follows.

Figure 1. Architecture for a multilingual ASR system.

Figure 2. Phoneme sequences generation for multilingual pronunciations.

The recognizer regards the additional pronunciations as normal entries. Hence, they increase the complexity of the system accordingly. This is the major disadvantage associated with the generation of multiple pronunciation variants. There is always a maximum number of vocabulary items that the recognizer can process in real-time. The processing power and memory resources available in the implementation platform typically set this limit. Therefore an on-line pruning scheme that automatically limits the number of active pronunciation variants is needed.

2.4. Multilingual pronunciation pruning

We propose a pronunciation-pruning scheme that removes less likely pronunciation variants from the vocabulary space. A leading assumption is that the pronunciation of a single user does not normally change significantly over time and there is always one (or two) pronunciation variants that tend to get recognized more frequently than the other variants. Each pronunciation variant is associated with a recognition statistics score that characterizes how often this particular variant has been correctly recognized. The scores are continuously updated for the correctly recognized vocabulary items. The final score for pruning is formed by combining the score of the LID module and the recognition statistics score. It is now possible to prune unnecessary pronunciation variants from the active vocabulary space without degrading the overall recognition rate. In the following, a brief description of the pruning algorithm is given.

First, LID is used to select $n$-best list of language for each entry to limit the number of languages. For a given entry ($W$) and its language tag ($\text{Lang}$), a corresponding pronunciation ($\text{Pron}$) is uniquely defined. Equation 1 is used to determine the best candidate for pruning.

$$\text{Pron} = (W_j, \text{Lang}_j)$$

$$= \arg \min_{(W_j, \text{Lang}_j)} P(W_j, \text{Lang}_j)$$

(1)

The two items on the right side of Equation (1) are updated as follows.
where $N_k(W_j)$ is the count of correct recognitions of the vocabulary entry $W_j$ until time instance $k$, and $M$ is the active vocabulary size.

$$P_k(W_j) = \frac{N_k(W_j)}{\sum_i N_k(W_j)}, \quad (2)$$

where $N_k(W_j)$ is the count of correct recognitions of the vocabulary entry $W_j$ until time instance $k$. By substituting equation (2) and (3) into (1), we have

$$P(W_j) = P_k(W_j) \cdot P(Lang_j | W_j),$$

$$= \frac{P_k(W_j) \cdot P(W_j)}{\sum_i P_k(W_i) \cdot P(W_i)}, \quad (4)$$

The pruning starts after a certain number of correct recognition results have been observed in order to have enough statistical information available, or after the maximum vocabulary size is exceeded. The pronunciation variant of each entry with the lowest score is removed from the vocabulary space. Each entry is guaranteed to have at least one pronunciation variant. If there are two pronunciation variants for an entry whose scores are close to each other and higher than the threshold value $T$, then both are retained in order not to remove potentially valid pronunciation variants. If there are two pronunciation variants for an entry whose recognition statistics scores are the same, then the pronunciation with the smallest LID-based score is removed. If all entries have only one pronunciation variant or all the alternative variants are associated with the score value, which is above the threshold, then the pruning algorithm informs that no more pronunciation pruning can be made.

The LID scores are initially given by the LID module to indicate the probability of the language that the given entry belongs to. In the recognition phase, the recognition result states are used to update the scores. When an entry is verified to be correctly recognized, then all the LID scores related to the entry are updated by equation (5):

$$\text{score}_{i+1} = \text{score}_i + \alpha \cdot (1.0 - \text{score}_i), \quad \text{for matched case}$$

$$\text{score}_{i+1} = \text{score}_i - \beta \cdot \text{score}_i, \quad \text{for unmatched case}, \quad (5)$$

where $0 \leq \text{score}_i \leq 1$, and $i$ denotes the count of successful recognition results. Parameters $\alpha$ and $\beta$ are updating factors between 0 and 1. For the correct (matched) pronunciation, the score increases and all other scores of the same entry decrease.

### 3. EXPERIMENTAL RESULTS

#### 3.1. Description of the test system

Experimental evaluation of the proposed methods was carried out using 25 languages (Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Icelandic, Italian, Latvian, Norwegian, Polish, Portuguese, Romanian, Russian, Slovakian, Slovenian, Spanish, Swedish, Turkish, and Ukrainian). As in [8], a set of 12 MFCC coefficients and log-energy, together with their first- and second-order time derivatives, were extracted from a continuous-time speech signal sampled at 8 kHz. A feature vector normalization scheme was then applied on the features. The log-energy and its time derivatives were mean and variance normalized, and for the rest of the coefficients only mean subtraction was applied. The acoustic modeling was based on monolingual continuous density monophone HMMs. Acoustically training data was not available for 18 out of the 25 languages used in the evaluation. The multilingual HMMs were trained on English, German, Finnish, French, Dutch, Danish and Spanish speech data. Only clean speech was used in the training phase. The performance evaluation, however, was carried out both under clean and noisy conditions. Noisy test data was obtained by artificially mixing noise to the clean test utterances. Four kind of noise (car, café, car noise with background speech, and airport hall), and a sample of rock music were used at SNRs randomly chosen between $+20$ and $+5$dB. The SNR distribution was set to be uniform. For the acoustic model adaptation experiments, the order of the input utterances was randomized so that the same vocabulary item was never repeated. For each speaker, the acoustic model adaptation was started from the original multilingual speaker-independent HMMs.

#### 3.2. Results

First, a set of experiments was carried out to determine the effect of the multilingual pronunciation modeling and the pruning algorithm on the recognition accuracy. In these tests, the language identity of each vocabulary item was specified either manually (denoted by MAN), or using the automatic LID algorithm. LID produced both 1-best and 4-best language tags with and without speakers' native (or default) language. They are denoted by LID1, LID1def and LID4, respectively.

In order to obtain a reference for testing the automatic system, a human expert assigned the language identity to each vocabulary item. As shown by Table 1, poor recognition rates were obtained by using a single pronunciation with automatic LID1, mainly because of erroneous language identity decisions. By adding the default language, the recognition performance drop could be almost compensated. When the default language is invalid, the recognition performance corresponding to 4-best LID is also close to the reference result. It is interesting to note that the use of the pruning method consistently results in even better performance than without pruning. The explanation is that confusion is reduced when keeping only the correct pronunciations in use after pruning. Table 1 also gives the recognition results in noisy environments. It is in line with the conclusion drawn from the clean testing.

Another set of experiments was carried out on a database specifically designed for adaptation tests. The database contains a relatively high number of repetitions (between 20 and 30). In addition, highly variable noise conditions, and in many cases severe language mismatches (i.e., strong foreign accent), are present in the recordings. By using a higher number of repetitions, it is possible to more reliably evaluate the performance gains due to the multilingual pronunciation pruning algorithm and on-line acoustic model adaptation.
Table 1. Recognition rates with and without multilingual pronunciation pruning, and with and without acoustic model adaptation in clean. Noisy rates are shown only with pruning enabled.

<table>
<thead>
<tr>
<th></th>
<th>No adaptation</th>
<th>Adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25 Lang</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAN</td>
<td>LID1</td>
</tr>
<tr>
<td>NoPruning, clean</td>
<td>93.77</td>
<td>87.51</td>
</tr>
<tr>
<td>Pruning, clean</td>
<td>93.77</td>
<td>87.51</td>
</tr>
<tr>
<td>Pruning, noise</td>
<td>85.71</td>
<td>78.83</td>
</tr>
</tbody>
</table>

Figure 3. Recognition rates with and without multilingual pronunciation pruning, and acoustic model adaptation.

Figure 3 shows the results obtained on the adaptation database. Note that the absolute recognition scores are not directly comparable to the ones presented in Table 1, because of the differences in the vocabulary and ambient noises. The trends, however, are well in line with the above observations. The results justify the multilingual pronunciation pruning scheme. In order to see the effect of multilingual pronunciation generation and the pruning scheme, the acoustic model adaptation was not used at first. Clearly the use of multilingual pronunciation modeling significantly improves the recognition rate that even slightly exceeds the reference result (manual LID). The application of the pruning scheme further enhances the recognition performance.

On-line acoustic adaptation proved to be very effective for reducing environmental and speaker-specific mismatches. When a single pronunciation was used, the performance decreased due to the incorrect LID decisions. When the multilingual pronunciation modeling was enabled, the recognition performance improved considerably. The best performance was obtained when both multilingual pronunciation pruning and on-line acoustic adaptation are used.

4. CONCLUSIONS

Multilingual pronunciation modeling and on-line acoustic model adaptation were proposed for maximizing the speaker-specific recognition accuracy in multilingual speaker-independent isolated word recognition systems. In addition to acoustic model mismatches, pronunciation variation plays a significant role in multilingual ASR. The generation of multiple pronunciation variants for each vocabulary entry improves the recognition performance. The improvement comes at the cost of improved system complexity, which is an important factor in many real-world systems. A pronunciation variant pruning algorithm was presented for coping with the complexity. Experimental results in clean and noisy recognition conditions confirmed the usefulness of the proposed techniques, and the best results were obtained when both techniques were combined.

5. REFERENCES