ONLINE UNSUPERVISED MULTILINGUAL
ACOUSTIC MODEL ADAPTATION FOR
NONNATIVE ASR

Sethserey Sam¹, Eric Castelli², Laurent Besacier³

¹Computer Science Department, Institute of Technology of Cambodia, BP 86, Bvd. Pochentong Phnom Penh, Cambodia, Tel. +855 12 509 383, Fax. +855 23 880 369, Email: sam.sethserey@itc.edu.kh
²International Research Institute MICA - HUST - CNRS/UMI2954 - Grenoble INP - Hanoi University of Science and Technology, 1, Dai Co Viet st., Hanoi, Viet Nam, Tel. +84 43 868 30 87, Fax. +84 43 868 35 51, Email: eric.castelli@mica.edu.vn
³GETALP Group, LIG Laboratory, UMR CNRS 5524 BP 53, 38041, Grenoble Cedex 9, France, Tel. +33 476 635 695, Fax. +33 476 514 985, Email: laurent.besacier@imag.fr

Received Date: January 16, 2012

Abstract

Automatic speech recognition (ASR) is currently one of the main research interests in computer science. Hence, many ASR systems are available in the market. Yet, the performance of speech and language recognition systems is poor on nonnative speech. The challenge for nonnative speech recognition is to maximize the accuracy of a speech recognition system when only a small amount of nonnative data is available. Recent studies on nonnative speech recognition were focused on supervised context in which spoken languages (L2) and speakers’ mother tongue languages (L1) are known in advance. In this paper, we want to study the adaptation approach of nonnative speech in which both L1 and L2 are unknown in advance. Such new approach is called online-unsupervised multilingual acoustic model adaptation. Thus, “unsupervised” means we don’t know in advance the nonnative speech utterance (it’s L1 and L2); and “online” means the adaptation is made during the decoding. Thus, the proposed approach decomposes into two stages. The first stage, contained language observer module, aims to recover the linguistic information (spoken languages and the origins of the speakers) of the unknown speech utterances to be decoded. The second stage is to adapt the multilingual acoustic model based on knowledge provided by language observer module. It is clear that the multilingual acoustic model must contain the acoustic units of L2 and L1. In this study, we report on the acoustic model adaptation for improving the recognition of nonnative speech in English, French and Vietnamese, spoken by speakers of different origins. Results degradation around 7% of baseline systems’ phone error rates (PERs) obtained from the experiments demonstrate the feasibility of the method.

Keywords: Hybrid Interpolation, Language Recognition, Multilingual Acoustic Modeling, Nonnative ASR, Unsupervised Adaptation

Introduction

Automatic speech recognition technologies are now integrated into many systems. However, nonnative speakers continue to suffer high error rate in ASR systems. Due to the insufficient training data, it is difficult or unfeasible to bootstrap a nonnative system of a language spoken by speakers of different origins (for example, English spoken by French speakers is different from English spoken by Vietnamese speakers). However, using the native ASR model to recognize nonnative speech is not a good solution either due to the difference between native and non-speech models trained.

Recently, different acoustic model adaptation techniques using little or no adaptation data have been proposed to improve the nonnative speech recognition performance, such as MLLR, supervised interpolation between two acoustic models, and polyphone decision...
tree specialization [1]. Those techniques were applied in the supervised context where nonnative speech (i.e. the speaker origin) was known in advance.

This paper focuses on improving multilingual acoustic models for automatic phonetic transcription of speech such as “multilingual meeting” in the unsupervised context.

There are several challenges in “multilingual meeting” speech: 1) it contains dialogues between native and nonnative speakers; 2) there is not only one spoken language but several languages spoken by speakers from different origins; 3) it is difficult to collect sufficient data to bootstrapping transcription systems. To solve these challenges, we propose a process of multilingual acoustic model adaptation called "online unsupervised adaptation". In the adaptation, we studied several approaches for adapting multilingual acoustic models in unsupervised way (spoken languages and the origins of the speakers are not known in advance) and no additional data is used during the adaptation process. Moreover, the adaptation takes place during the decoding process. Thus, our proposed approach comprises two modules. The first module called "the language observer" is to recover the linguistic information (spoken languages and the origins of the speakers) of the speech utterances to be decoded. The second module is to adapt the multilingual acoustic model based on knowledge provided by the language observer. To evaluate the usefulness of our unsupervised multilingual acoustic model adaptation, we use the test data, which are extracted from multilingual meeting corpus, containing the native and nonnative speech of three languages: English (EN), French (FR) and Vietnamese (VN).

The rest of the paper is organized as follows. Section 2 presents the multilingual meeting corpus setup from which we extract the test data. Section 3, 4 and 5 present the multilingual acoustic-phonetic recognizer (baseline), the language observer and the acoustic model adaptation techniques respectively. The experiment results are presented in Section 6. We finish with a brief conclusion in Section 7.

**Multilingual Meeting Corpus**

We extract the test data from the MICA meeting speech corpus recorded in the professional meeting room of MICA international research institute [2]. This corpus contains around 5 hours of transcribed speech in 4 languages EN (English), FR (French), VN (Vietnamese) and KH (Khmer: Cambodia’s official language). This multilingual meeting corpus involves the speech from 14 speakers (1 English, 5 French, 5 Vietnamese and 3 Cambodian). In the corpus dialog, we discover that speakers use their native or nonnative languages to communicate, according to whom they speak with. Table 1 presents the distribution of the languages spoken by speakers with different native languages (66 % of the corpus is nonnative speech). It is important to mention that the quality of the recording speech is also analyzed based on signal to noise ratio (SNR) before the corpus creation.

Due to the lack of nonnative speakers of Khmer, we extract only the native and nonnative speech data of EN, FR and VN from the above corpus, and we select only the utterances longer than 5 seconds for our experiments. Each speech segment contains one language only (native or nonnative but no code-switching). Table 2 presents the quantity of test data that has been used in the experiments.

Note that, in the content of Table 2, the labels in capital letters denote the spoken language of the speech segments and the labels in lowercase denote the native language of the speakers (for example, ENfr means English spoken by native French speakers).
Table 1. Total Duration of Transcribed Speech (Value in Seconds)

<table>
<thead>
<tr>
<th>Speaker Origin</th>
<th>#Speakers (Sex)</th>
<th>Lang-KH</th>
<th>Lang-VN</th>
<th>Lang-FR</th>
<th>Lang-EN</th>
<th>#Total by speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spk-KH</td>
<td>3(M)</td>
<td>570</td>
<td>452</td>
<td>1822</td>
<td>3452</td>
<td>6296</td>
</tr>
<tr>
<td>Spk-VN</td>
<td>2(F), 3(M)</td>
<td>0</td>
<td>1747</td>
<td>1177</td>
<td>675</td>
<td>3599</td>
</tr>
<tr>
<td>Spk-FR</td>
<td>1(F), 4(M)</td>
<td>0</td>
<td>1550</td>
<td>2797</td>
<td>1370</td>
<td>5717</td>
</tr>
<tr>
<td>Spk-EN</td>
<td>1UK (M)</td>
<td>0</td>
<td>590</td>
<td>584</td>
<td>911</td>
<td>2085</td>
</tr>
<tr>
<td>#Total by language</td>
<td>14</td>
<td>570</td>
<td>4339</td>
<td>6380</td>
<td>6408</td>
<td>17697</td>
</tr>
</tbody>
</table>

Table 2: Quantity of Testing Data (Value in Seconds) Used in Our Experiments

<table>
<thead>
<tr>
<th>Language</th>
<th>Speech (Native/Nonnative)</th>
<th>Qty.</th>
<th>Total by Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>ENen</td>
<td>239</td>
<td>1195</td>
</tr>
<tr>
<td></td>
<td>ENfr</td>
<td>715</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ENvn</td>
<td>241</td>
<td></td>
</tr>
<tr>
<td>FR</td>
<td>FRfr</td>
<td>241</td>
<td>980</td>
</tr>
<tr>
<td></td>
<td>FREN</td>
<td>253</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FVRn</td>
<td>486</td>
<td></td>
</tr>
<tr>
<td>VN</td>
<td>VNvn</td>
<td>235</td>
<td>933</td>
</tr>
<tr>
<td></td>
<td>VNEN</td>
<td>215</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VNfr</td>
<td>483</td>
<td></td>
</tr>
<tr>
<td>Total:</td>
<td></td>
<td></td>
<td>3108</td>
</tr>
</tbody>
</table>

Thus, we have a test data set of around 52 minutes where the nonnative speech represents 69% of the total. Meanwhile, we would like to emphasis that even we study the adaptation techniques for nonnative speech but native speech are also considered for comparing and evaluating the performance of the proposed adaptation approaches. Moreover, we did not reserve any additional data for the adaptation process, because our objective is to adapt the multilingual acoustic models by using only the information from the testing data in its process.

Multilingual Acoustic-phonetic Recognizer (Baseline System)

All recognition experiments described in this paper use the Sphinx3 decoder [3]. Our multilingual acoustic-phonetic recognizer (MultPR) covers three languages: English, French, and Vietnamese. The multilingual acoustic modeling (Mult-AM) is created by combining the existing monolingual acoustic models of EN, FR and VN trained respectively on three corpora: WSJ [4], BREF120 [5], and VNSpeechCorpus [6]. The combination of acoustic models is simply achieved based on the ML-sep combination method of [7]. It means that there is no data to share across languages among the three monolingual acoustic models. Moreover, our Mult-AM is a context independent acoustic
model that contains 124 acoustic units: EN (40 phonemes), FR (43 phonemes) and VN (41 phonemes). Each acoustic unit is represented by a Hidden Markov Model (HMM) with 3 states and 16 Gaussian components per state. The acoustic parameters mentioned above are selected based on the previous studied in acoustic model adaptation [1][16].

For multilingual language modeling and lexical modeling, we simply create respectively a flat LM of phones (phone loop grammar) and a phone list, for the 124 phonemes. In Figure 1, each phoneme, in the baseline system output, is presented in the SAMPA format proposed by John Wells [8] and is appended with the label of language that the phoneme belongs to.

\[
\text{Multilingual acoustic-phonetic recognizer (MultPR)}
\]

Figure 1. Example of a MultPR output

**Language Observer**

**Multilingual Acoustic-Phonetic Recognizer Followed by Vector Space Modeling (MultPR-VSM)**

To generate the posterior score of the three involved languages, we study the language identification system by using a well-known phonotactic model called phone recognizer followed by vector space modeling (PR-VSM). This approach was proposed in [9].

\[
\text{Non-native speech} \quad \rightarrow \quad \text{MultPR} \quad \rightarrow \quad \text{MultPR-Front-end} \quad \rightarrow \quad \text{VSM-Back-end} \quad \rightarrow \quad \text{Lang-EN} \quad \downarrow \quad \text{Lang-FR} \quad \downarrow \quad \text{Lang-VN}
\]

Figure 2. Block diagram of the PR-VSM: MultPR Front-end followed by VSM Back-end proposed by Li, Ma and Lee [9]

In fact, PR-VSM consists of two parts: a phone recognizer (PR) front-end and a vector space modeling (VSM) back-end as illustrated in Figure 2. In this paper, the PR front-end is the multilingual acoustic-phonetic recognizer (MultPR, mentioned in Section 3).

In the VSM training stage, 6 hours of spoken speech utterances (2 hours per language) have been extracted from the following corpora: WSJ (EN), BREF120 (FR) and VNSpeechCorpus (VN). These spoken utterances, with their language label, are tokenized by MultPR (mentioned in Figure 1) and then converted into a collection of spoken document vectors based on the bag-of-sound phonotactic approach [10]. Finally, this
collection of language-labeled spoken document vectors is used to design language classifiers (to group document vectors in their “language categories” (EN, FR and VN)) by using support vector machines [11] developed in the text categorization community.

In the testing stage, an unknown test utterance (mentioned in Table 2) is converted to a query vector (using basically the same procedure as that used for extracting the document vectors during the VSM training stage), so that the language posterior scores are generated as in the case of text document classification [12].

To summarize, we can briefly calculate the language posterior scores of the phonotactic MultPR-VSM in language observer as:

\[
P(L_i) = \frac{P(T \mid VSM(L_i))}{\sum_{j=1}^{N} P(T \mid VSM(L_j))}
\]

where \( P \) is the language posterior score, \( L_i \) is one of the three languages (EN, FR or VN). \( T \) is the phoneme sequence, which is the result of MultPR. \( N \) is the total number of languages.

Analysis of Language Observer Quality

Before we decide to use MultPR-VSM as a tool to detecting the language and speaker origin, we need to evaluate its posterior scores. We evaluate the performance based on two criteria:

- Conventional LID error rate: here the maximum of the MultPR-VSM posterior scores is used to choose the most likely language.
- LID+ORG: we also propose a slightly different measure, where a test is considered as successful if the maximum of the MultPR-VSM posterior scores gives either the spoken language or the speaker’s native language (in the latter case, the second most likely language must be the spoken language, otherwise we consider the test has failed). For instance, if the language posterior scores are \( P(\text{FR}) = 0.5, P(\text{EN}) = 0.4 \) and \( P(\text{VN}) = 0.1 \), and if the utterance reference is English spoken by a French speaker, then the LID metric indicates an incorrect identification, while LID+ORG indicates a correct one.

![Figure 3. Evaluation of MultPR-VSM based on LID / LID+ORG accuracy (%)](image)
As shown in Figure 3, MultPR-VSM gives generally poor accuracy for nonnative speech, compared to native speech in term of LID evaluation. However, the native and nonnative accuracy are very competitive in our LID+ORG evaluation. Thus, it is possible to say that, in the case of nonnative speech, MultPR-VSM gives us not only spoken language information but also the original language information of the speakers.

Online Unsupervised Multilingual Acoustic Model Adaptation

Maximum Likelihood Linear Regression (MLLR)

MLLR is simple and known to be robust for unsupervised adaptation as well as effective for small amount of adaptation data [13]. The first pass hypothesis of the current test utterance is used to guide the estimation of the MLLR transformation matrices which are then used to produce the adapted mean vectors of the acoustic modeling. There are several types of MLLR adaptation; but in this paper, all MLLR adaptations used in our experiments are based on a global mean-only transformation [14], which is simple, fast and gives competitive performance compared to other types. Figure 4 illustrates the MLLR adaptation process. In the figure, it is shown that MLLR readapts the multilingual acoustic model without the help of the language observer module. We consider MLLR as the baseline adaptation and its performance will be used to compare with our proposed adaptations, which uses the language observer as the adaptation controller.

![Figure 4. MLLR, an online-unsupervised acoustic model adaptation process](image)

Acoustic Model Interpolation (INTER)

In the previous study [15], it was shown that speakers borrow acoustic and phonetic features from their native languages in their nonnative speech. This motivates us to create a nonnative acoustic model of a language L2 (target language) spoken by a speaker who has L1 (source language) as the mother tongue language by simply interpolating the native acoustic models of target language (L2) and source language (L1). Such approach demands no additional data in its adaptation process. We investigate adaptation using nonnative speaker cross-lingual acoustic model merging and interpolation (hybrid-interpolation) [16], which is one of the most useful interpolation techniques in nonnative ASR.

In each hybrid interpolation process [16], when the Euclidean distance (Equation 2) between a Gaussian in certain state of the target model (referred to as target Gaussian) and the associated Gaussian in certain state of the source model (referred to as source Gaussian), is below a threshold, their means, variances and mixture weights will be interpolated (Equation 3). Otherwise, merging is performed: for the source Gaussians that
are far from their associated target Gaussians (Equation 4) or for those target Gaussians without any associated source Gaussian (Equation 5). The threshold can be calculated for example by measuring the average distance among the Gaussians, and then multiplying it with a constant. In merging cases, their mixture weights will be reduced by the interpolation weight. Finally, the hybrid interpolation of two acoustic models have been formulated as follows:

\[
\text{Euclidean Distance} = \sqrt{\sum (\mu_t - \mu_s)^2} \tag{2}
\]

\[
g_{\text{new, sn}} = (1-w)g_{\text{tg, sn}} + w g_{\text{sc, sn}} \neq d, g_{\text{tg, sn}}, g_{\text{sc, sn}} \leq \text{dist} \tag{3}
\]

\[
g_{\text{new, sn}} = g_{\text{sc, sn}}, \omega_{\text{new, sn}} = w, \omega_{\text{sc, sn}} \neq \phi, d(g_{\text{tg, sn}}, g_{\text{sc, sn}}) > \text{dist} \tag{4}
\]

\[
g_{\text{new, sn}} = g_{\text{tg, sn}}, \omega_{\text{new, sn}} = (1-w), \omega_{\text{tg, sn}}, g_{\text{sc, sn}} = \phi \tag{5}
\]

where \(\mu_t, \mu_s\) represents the target language means and the source language means respectively. \(g_{\text{new, sn}}\) represents the interpolated/merged Gaussian, \(g_{\text{tg, sn}}\) is the target Gaussian, and \(g_{\text{sc, sn}}\) is the source Gaussian. \(w\) is the interpolation weight; \(\omega\) is the mixture weight for the Gaussian. \(d(.)\) is a distance function and \(\text{dist}\) is a threshold distance [16].

Regarding to the above equations (3, 4, 5), the hybrid interpolation of two acoustic models (AMinter) depending on three principal parameters: 1) acoustic model of target language (AMtarget); 2) acoustic model of source language (AMsource); 3) the interpolation weights (w). Equation 6 resumes the dependent parameters of equation 3, 4 and 5.

\[
\text{AMinter} = \{\text{AM}_{\text{target}}, \text{AM}_{\text{source}}, \omega\} \tag{6}
\]

To our knowledge, the recent studies on acoustic model interpolation between target and source models [1][16] are performed in the supervised adaptation context (the target and source models are known in advance, moreover, the interpolation weights are fixed).

In our unsupervised adaptation context, the three parameters (Equation 6) are unknown in advance. Furthermore, we have three languages (French, English and Vietnamese) to interpolate instead of only two languages (source and target) that are often studied in previous works. To determine the three parameters of Equation 6 from an unknown speech segment, we propose the interpolation process of the three acoustic models (EN, FR, VN) as follows:

- define the target and source languages: the language which has the best score among the three subsequent scores provided by the language observer (LO) is considered the target language (its model is AMtarget) and the other two languages are considered source languages (AMsource1, AMsource2 for languages that have, respectively, a score in the 2nd and 3rd rank)
- define the interpolation weights (w): the posterior scores of two sources languages are considered as the interpolation weight, so Equation 6 becomes:

\[
\text{AMinter} = \{\text{AM}_{\text{target}}, \text{AM}_{\text{source}}, P(L_{\text{source}})\} \tag{7}
\]

Because we have two source models and only one target model, we propose to do the hybrid interpolation in two times successively where two acoustic models (the target AM and one of the source AMs) are interpolated at each time. Finally, the adapted multilingual acoustic model is made by combining the two interpolated acoustic models based on the ML-sep combination method [7] as explained in Section 3.
Figure 5 illustrates our unsupervised interpolation process if the language observer gave the language posterior scores $P(EN)=0.5$, $P(FR)=0.3$ and $P(VN)=0.2$.

We also compute the performance of the acoustic model interpolation followed by MLLR (INTER-MLLR) by simply applying the MLLR adaptation to the adapted Mult-AM based on the above interpolation approach.

In order to evaluate more deeply the performance of the unsupervised interpolation adaptation based on MultPR-VSM, we also investigate the adaptation performance by using perfect language identification (oracle case: we suppose that the spoken languages are known in advance). In this case, the interpolation is made based on the language likelihood generated by the MultPR-VSM except that the target language is not always the most likely language identified by the language observation module. For example, if the utterance is English language and we suppose that MultPR-VSM produces the error language classification as $P(FR)=0.5$, $P(EN)=0.4$ and $P(VN)=0.1$; in the oracle case, English is the target language (not French language) while the others are considered as source languages. Finally the interpolation is made by using the source language likelihood ($P(FR)$ and $P(VN)$) as the source language weights in the acoustic model adaptation process (Equation 7).

Figure 5. An example of unsupervised interpolation process between three acoustic models

Figure 6. Confidence Interval (CI) of different adaptation techniques (Application on Nonnative Speech only)
Experiment Result

Table 3 presents experiment result dealing with both native and nonnative speech. Table 4 and present the experiment result applying only on nonnative speech of the involved languages. It is important to mention that the PERs values presented in Table 4 are exactly the nonnative speech PER values in Table 3. Figure 6 illustrates the confidence interval (CI) of PERs of different systems (baseline and adaptation).

With the experiment result shown in Table 3, we observe that:

- the online MLLR adaptation reduces the error rates of baseline system marginally because it uses a single speech utterance (the segment being decoded) in its adaptation process
- when both native and nonnative speech utterances were tested together, online adaptation INTER reduces PERs only by 2.5% absolute in unsupervised case (using MultPR-VSM to capture the information of L1 and L2) and by 4.7% absolute in oracle case (the spoken languages of the test utterances are supposed to know in advance). The degradation of PERs would be lower if more native speech presents in the testing utterances. The same observation for INTER-MLLR; it means that INTER and INTER-MLLR generally degrades baseline system performance if the segments contain native speech even the language observer module gives correct language information

If we consider only the result of the nonnative speech utterances (Table 4 and Figure 6), we observe that:

- the performance of INTER and INTER-MLLR adaptations (Table 4) depend on the performance of language information provided by language observer (MultPR-VSM). In the oracle case of language observer, INTER and INTER-MLLR reduce significantly the nonnative speech PERs of the baseline system; but, when the language observer gives poor language information (By example VNfr (Figure 3)) the INTER and INTER-MLLR slightly reduce PERs of the baseline system
- INTER and INTER-MLLR adaptations (Table 4) generally reduces PERs of the baseline system for all nonnative utterances, despite the fact that performance of the observer of languages is not good in some cases (e.g. VNen and VNfr speech utterances (Figure 3))
- based on PER results (Table 4), INTER and INTER-MLLR adaptation reduces the PER of the baseline system by around 5% and 7% absolute respectively while MLLR reduces the baseline’s PER by around 1% absolute only
- based on the confidence interval (CI) measurement (Figure 6), there is no significant different between baseline and MLLR performance while INTER and INTER-MLLR performances are significantly better than the previous two; however, the difference between INTER and INTER-MLLR performances are not significant while INTER-MLLR adaptation process takes more significant adaptation times than INTER technique

With these observations, we can conclude that the online unsupervised multilingual acoustic model adaptation based on the MultiPR-VSM language observer and acoustic models interpolation (INTER) is a promising option to decode several nonnative speech of different speaker origins. Unfortunately, this technique degrades baseline system performance if the decoding utterances are native speech.
Table 3: Phone Error Rate (PER [%]) of Online Unsupervised Adaptation Techniques (Application on Both Native and Nonnative Speech)

<table>
<thead>
<tr>
<th>Native/Nonnative Speech</th>
<th>Baseline</th>
<th>MLLR</th>
<th>Using MultPR-VSM</th>
<th>Oracle Case</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>INTER</td>
<td>INTER-MLLR</td>
</tr>
<tr>
<td>ENen</td>
<td>57.5</td>
<td>56.2</td>
<td>59.2</td>
<td>59.2</td>
</tr>
<tr>
<td>ENfr</td>
<td>60.2</td>
<td>60.4</td>
<td>54.3</td>
<td>54.3</td>
</tr>
<tr>
<td>ENvn</td>
<td>57.6</td>
<td>57.6</td>
<td>52.7</td>
<td>52.8</td>
</tr>
<tr>
<td>FRfr</td>
<td>56.7</td>
<td>56.5</td>
<td>58.2</td>
<td>58.2</td>
</tr>
<tr>
<td>FRen</td>
<td>59.1</td>
<td>58.5</td>
<td>51.7</td>
<td>51.3</td>
</tr>
<tr>
<td>FRvn</td>
<td>55.8</td>
<td>55.9</td>
<td>55.0</td>
<td>55.0</td>
</tr>
<tr>
<td>VNvn</td>
<td>47.3</td>
<td>45.7</td>
<td>53.2</td>
<td>53.1</td>
</tr>
<tr>
<td>VNen</td>
<td>57.0</td>
<td>56.5</td>
<td>53.4</td>
<td>53.2</td>
</tr>
<tr>
<td>VNfr</td>
<td>48.9</td>
<td>49.0</td>
<td>45.9</td>
<td>45.8</td>
</tr>
<tr>
<td>Avg.</td>
<td>55.8</td>
<td>55.4</td>
<td>53.3</td>
<td>53.4</td>
</tr>
</tbody>
</table>

Table 4: Phone Error Rate (PER [%]) of Online Unsupervised Adaptation Techniques (Application on Nonnative Speech)

<table>
<thead>
<tr>
<th>Nonnative Speech</th>
<th>Baseline</th>
<th>MLLR</th>
<th>Using PRVSM</th>
<th>Oracle Case</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>INTER</td>
<td>INTER-MLLR</td>
</tr>
<tr>
<td>ENfr</td>
<td>60.2</td>
<td>60.4</td>
<td>54.3</td>
<td>54.3</td>
</tr>
<tr>
<td>ENvn</td>
<td>57.6</td>
<td>57.6</td>
<td>52.7</td>
<td>52.8</td>
</tr>
<tr>
<td>FRen</td>
<td>59.1</td>
<td>58.5</td>
<td>51.7</td>
<td>51.3</td>
</tr>
<tr>
<td>FRvn</td>
<td>55.8</td>
<td>55.8</td>
<td>55.0</td>
<td>55.0</td>
</tr>
<tr>
<td>VNen</td>
<td>57.0</td>
<td>56.5</td>
<td>53.4</td>
<td>53.2</td>
</tr>
<tr>
<td>VNfr</td>
<td>48.9</td>
<td>48.2</td>
<td>45.9</td>
<td>45.8</td>
</tr>
<tr>
<td>Avg.</td>
<td>56.3</td>
<td>55.2</td>
<td>51.4</td>
<td>49.3</td>
</tr>
</tbody>
</table>

Conclusion

In this paper, we explored an online-unsupervised approach for acoustic model adaptation to better transcribe unknown speech utterances extract from the multilingual meeting corpus in which three spoken languages are involved: English, French and Vietnamese. The advantage of this approach is that it automatically readapts the multilingual acoustic models at the same time of decoding and without using any external data. According to the experimental results, our adaptation approach called Interpolation followed by MLLR based on MultPR-VSM reduces the PERs by 7% absolute compared to the baseline system when decoding unknown nonnative speech utterances.
To make the recognition process more robust for both native and nonnative speech, we currently study on using the native and nonnative discrimination module at the beginning of the adaptation process so that the adaptation will be used only with the nonnative speech while native recognizers will be used to recognize native speech.

References


