



# Unsupervised Acoustic Model Adaptation for Multi-Origin Non Native ASR

Sethserey Sam<sup>1,2</sup>, Eric Castelli<sup>2</sup>, Laurent Besacier<sup>1</sup>

<sup>1</sup>LIG Laboratory, UMR CNRS 5524 BP 53, 38041 Grenoble Cedex 9, France

<sup>2</sup>MICA research center, UMI CNRS 2954, HUT, Hanoi, Vietnam

{sethserey.sam, laurent.besacier}@imag.fr, eric.castelli@mica.edu.vn

## Abstract

To date, the performance of speech and language recognition systems is poor on non-native speech. The challenge for non-native speech recognition is to maximize the accuracy of a speech recognition system when only a small amount of non-native data is available. We report on the acoustic model adaptation for improving the recognition of non-native speech in English, French and Vietnamese, spoken by speakers of different origins. Using online unsupervised adaptation acoustic modeling without any additional data for adapting purposes, we investigate how an unsupervised multilingual acoustic model interpolation method can help to improve the phone accuracy of the system. Results improvement of 7% of absolute phone level accuracy (PLA) obtained from the experiments demonstrate the feasibility of the method.

**Index Terms:** non-native ASR, language recognition, unsupervised adaptation, interpolation, multilingual acoustic modeling

## 1. Introduction

The accuracy of speech recognition usually drops when dealing with non-native speech utterances. Moreover, due to insufficient training data, it is difficult or unfeasible to bootstrap a non-native 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).

Recently, different acoustic model adaptation techniques have been proposed to improve the non-native 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 non-native speech (i.e. the speaker origin) was known in advance.

In this paper, we propose an online unsupervised approach to adapt the multilingual acoustic model in order to improve the accuracy of the acoustic-phonetic recognition system. "Online unsupervised" means that the adaptation is made at decoding time of each utterance (online) without any knowledge of the language and speaker origin of the spoken utterance (unsupervised). To achieve the goal, we propose a two-pass decoding architecture. The first-pass decoding captures the information by using a module called language observer (LO). Then the interpolation among multiple and multilingual acoustic models is made by using language posterior scores as interpolation language weights. Finally, we use the interpolated acoustic model in the second-pass decoding.

In fact, the technique used by our language observer (LO) to generate the language posterior scores is the same as that used for language identification (LID). But we prefer to use

the term "language observer" (LO) since, as opposed to LID, no hard decision is taken to identify a particular language spoken. The fact is that both LO and LID assign a set of language posterior scores, but LO considers all these scores, while LID selects the most likely language without really using the other language posterior scores. We believe that such a language observer gives not only the information about the spoken language in the speech segment but also on the native language of the speaker. For example, if the language observer gives  $P(EN) = 0.5$ ,  $P(FR) = 0.4$  and  $P(VN) = 0.1$ , then the speech segment may be in English spoken by a French speaker (or vice-versa).

This paper mainly presents the online unsupervised acoustic model adaptation to deal with three non-native speech utterances of 3 different origins: English, French and Vietnamese. It is organized as follows. Section 2 presents the multilingual meeting corpus setup from which we extract the test data. Section 3 and 4 present the multilingual acoustic-phonetic recognizer (baseline) and the language observer respectively. The acoustic model adaptation techniques are presented and evaluated in Section 5. We finish with a brief conclusion in Section 6.

## 2. Test corpora

We extract the test data from the MICA meeting speech corpus that was recorded in the professional meeting room of MICA research center [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 non-native 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 non-native speech).

Table 1: *Duration coverage matrix (in second) of languages spoken by different speaker origins.*

Speaker	Lang-KH	Lang-VN	Lang-FR	Lang-EN
Spk-KH	570	452	1822	3452
Spk-VN	0	1747	1177	675
Spk-FR	0	1550	2797	1370
Spk-EN	0	590	584	911

Due to the lack of non-native speakers of Khmer, we extract only the native and non-native speech data of EN, FR and VN

<sup>1</sup> EN, FR and VN mean English, French and Vietnamese, respectively.

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 non-native but no code-switching). Table 2 presents the quantity of test data that has been used in the experiments.

Table 2: *Quantity of testing data (value in seconds) used in our experiments.*

Language	Native/Non-native speech	TOTAL
EN	ENen=239; ENfr=715; ENvn=241	1195
FR	FRfr=241; FRen=253; FRvn=486	980
VN	VNvn=235; VNen=215; VNfr=483	933
TOTAL		3108

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).

Finally, we have a test data set of around 52 minutes where the non-native speech represents 69% of the total. We would like to emphasize that the native speech ENen, FRfr and VNvn presented in Table 2 will only be used to evaluate the quality of language observer (Section 4.2) but they will not be used in the experiment of the acoustic model adaptation because, in this paper, we study mainly how to improve acoustic models for non-native ASR.

### 3. 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 made 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 HMM with 3 states and 16 Gaussian components per state.

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.

## 4. The use of LID as language observer

### 4.1. Multilingual acoustic-phonetic recognizer followed by vector space modeling (MultPR-VSM)

To generate the posterior score of the involved three 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].

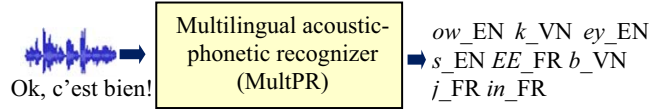


Figure 1: *Example of a MultPR output*

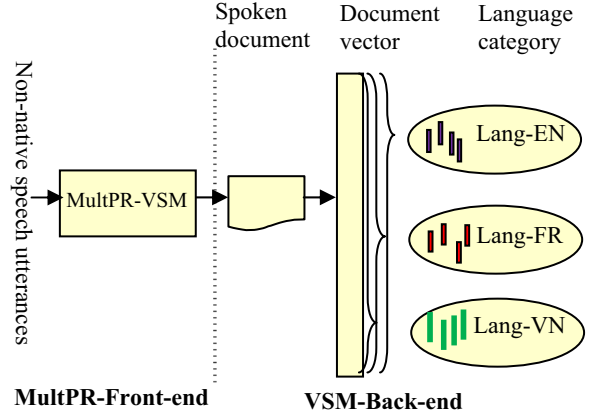


Figure 2: *Block diagram of the PR-VSM: MultPR front-end followed by VSM back-end*

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 any classifier learning techniques, such as support vector machines [11] or artificial neural networks [12], 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 [13].

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

$$P(L_i) = \text{Log}P(T | VSM(L_i)) \quad (1)$$

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.

### 4.2. 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.

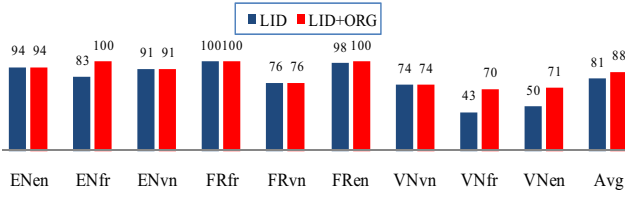


Figure 3: Evaluation of MultPR-VSM based on LID / LID+ORG accuracy (%)

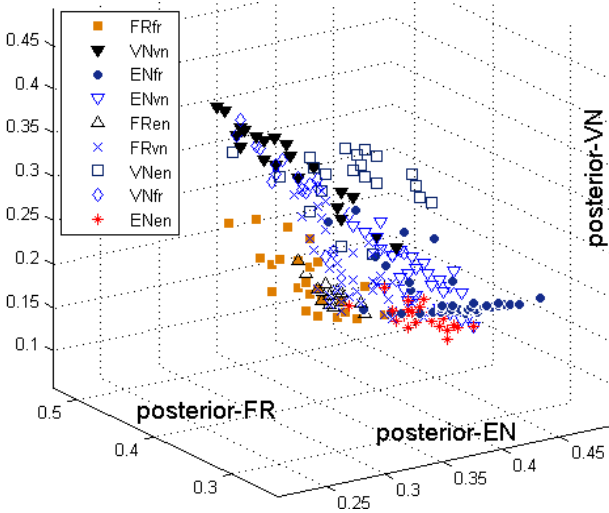


Figure 4: Localization in 3-D of different groups of native/non-native speech utterances

As shown in Figure 3, MultPR-VSM gives generally poor accuracy for non-native speech, compared to native speech in term of LID evaluation, but the native and non-native accuracy are very competitive in our LID+ORG evaluation.

According to [14] and [15], speakers borrow acoustic and phonetic features from their native languages in their non-native speech. So we also observe in a 3-dimensions space the language posterior scores generated by MultPR-VSM for all testing utterances. The purpose of this analysis is to study whether LO (MultPR-VSM) can determine the relationship between native and non-native speech utterances of the three languages (EN, FR and VN). In Figure 4, each point represents the 3 different language posterior scores of an utterance generated by the MultPR-VSM.

On the basis of the 3-D plan (Figure 4), we can say that:

- native English (ENen), French (FRfr) and Vietnamese (VNvn) utterance groups remain clearly separated in the

plan. But the ENen and FRfr groups seem a bit closer compared to the VNvn group. With that reserve, it seems that MultPR-VSM is able to separate native utterances;

- each group of non-native utterances is generally located closer to two groups of the native utterances: the one corresponding to the spoken language and the one corresponding to the origin of the speaker (mother tongue). By example, the non-native English spoken by French speakers (ENfr) and non-native French spoken by English speakers (FRen) are located closer to the ENen and FRfr utterances, compared to VNvn. It proves that the language posterior scores generated by MultPR-VSM could also be used to help in the unsupervised acoustic model adaptation for different groups of non-native speech.

## 5. Multilingual acoustic model adaptation

### 5.1. Adaptation techniques

We introduce now a new unsupervised acoustic model interpolation and compare it with the baseline unsupervised adaptation method MLLR, which is widely used in speaker adaptation.

- Baseline maximum likelihood linear regression (MLLR): the first-pass hypothesis from the unadapted multilingual system is used to generate the transformation matrices so that it can produce a new adapted mean (global mean) of the original multilingual acoustic model. In this case, no language observer (LO) is used during adaptation;
- Acoustic model interpolation (INTER): to date, most of the proposed interpolation techniques for non-native ASR use a fixed language weight which is not always matched for different native languages spoken by speakers of various origins. In our online unsupervised adaptation approach, the language weights to interpolate the acoustic models vary from utterance to utterance. The interpolation weights for interpolating acoustic models are based on the three language posterior scores (EN, FR, VN) generated by MultPR-VSM. The most likely language posterior score generated by MultPR-VSM is used to determine the target language ( $L_{target}$ ). Other languages are called source languages. The interpolation is made by:

$$A_{adapted} = \sum_{i=1}^{m-1} [(1 - P(L_i)) \cdot A_{L_{target}} + P(L_i) \cdot A_{L_i}] \quad (2)$$

where  $A$  is the acoustic model,  $P$  is the language posterior score (obtained by Equation 1),  $L_i$  is one of the source languages,  $m$  is the total number of languages to interpolate and  $\sum$  represents the combination of the interpolated acoustic models based on ML-sep [7].

- 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). In this case, the interpolation is always made based on the language posteriors generated by PR-VSM except that the target language is

considered as known in advance (not detected by the language observer). For example, suppose the utterance contains English language spoken by a French speaker, but PR-VSM produces language posterior scores such as  $P(\text{FR}) = 0.5$ ,  $P(\text{EN}) = 0.4$  and  $P(\text{VN}) = 0.1$ . In the oracle case, English is considered as the target language with a weight of 0.4, while the others (FR, VN) are considered as source languages with the remaining weights (0.5 and 0.1 respectively).

## 5.2. Experimental results

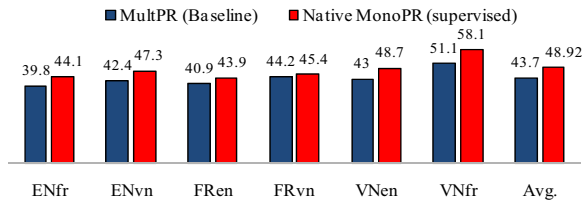


Figure 5: Phone level accuracy (PLA) [%] of baseline vs. monolingual native phone recognizer (MonoPR)

Table 3. PLA [%] of various adaptation techniques

Non-native	MultPR (Baseline)	MLLR	Using PRVSM		Oracle Case	
			INTER	MLLR	INTER	MLLR
ENfr	39.8	39.6	45.7	45.7	47.1	47.6
ENvn	42.4	42.4	47.3	47.2	50	50.4
FRen	40.9	41.5	48.3	48.7	49.5	50.1
FRvn	44.2	44.2	45	45	50.3	51.2
VNen	43	43.5	46.6	46.8	51.4	51.7
VNfr	51.1	51.8	54.1	54.2	56.1	56.3
Avg.	43.7	44.8	48.6	50.7	52.5	52.8

Figure 5 compares the Phone Level Accuracy (PLA) on the non-native speech parts for the baseline multilingual acoustic-phonetic recognizer (MultPR, baseline system) and the other three monolingual acoustic-phonetic recognizers (MonoPR) applied on the correct corresponding language. For MonoPR, we assume a perfect spoken language identification result is available in the case of monolingual phone recognizers (all non-native English, French and Vietnamese are decoded with the native phone recognizers of English, French and Vietnamese respectively).

It is important to recall that the test data involved is non-native speech; moreover, we experiment only acoustic-phonetic systems in this paper, which explains why lower PLAs are achieved in the monolingual phone recognizers as well as the baseline system. However, the overall difference in PLA between MultPR (baseline system) and MonoPRs is small (~5%).

With the result of various types of unsupervised acoustic model adaptation presented in Table 3, we can conclude that:

- MLLR improves slightly the phone level accuracy rate of non-native speech while the interpolation techniques based on the language posterior scores of MultPR-VSM outperform significantly the baseline system (by 7% higher performance than the baseline system).
- The performance of the adaptation really depends on the performance accuracy of MultPR-VSM. For instance, the adaptation performance of non-native Vietnamese is only slightly increased compared to the baseline system because of the poor performance of MultPR-VSM (Figure

3). But in the oracle case (perfect LID), the adaptation accuracy of non-native Vietnamese is as high as the performance for non-native English or French.

Finally, comparing the results of figures 5 and Table 3, we see that the unsupervised INTER-MLLR adaptation provides better average performance than the monolingual systems, in which perfect spoken language identification is considered for all utterances before decoding (50.7% vs. 48.92%). This confirms that for non-native speech, making a hard decision (LID: consider only the highest posterior score) on the language of the utterance, in order to choose the corresponding acoustic model, is not the best strategy. Alternatively, the online unsupervised method we proposed suggests that soft decisions (consider all language posterior scores) based on language observer outputs are useful for multilingual acoustic model adaptation.

## 6. Conclusion

In this paper, we explored and evaluated our concept of language observer and demonstrated its usefulness for online unsupervised adaptation of multilingual acoustic models for 3 different non-native languages spoken by speakers of different origins. Experiments have shown that, by considering all language posterior scores of MultPR-VSM in the adaptation process, online unsupervised interpolation followed by MLLR (INTER+MLLR) increases by 7% (absolute) the phone level accuracy, compared to our baseline system. In the future, it would be interesting to see experiments with multilingual acoustic models that share data across languages.

## 7. References

- [1] Wang, Z., Schultz, T., and Waibel, A., "Comparison of Acoustic Model Adaptation Techniques on Non-Native Speech," in Proc. ICASSP, 2003, pp. 540-543.
- [2] "http://www.mica.edu.vn."
- [3] "http://cmusphinx.sourceforge.net."
- [4] Paul, D. and Baker, J., "The Design for the Wall Street Journal-based CSR Corpus," in DARPA SLS Workshop, Pacific Grove, California, USA, 1992.
- [5] Lamel, L. F., Gauvain, J-L., and Eskénazi, M., "BREF, A Large Vocabulary Spoken Corpus for French," in Eurospeech, 1991, pp. 505-508.
- [6] Le, V.B., Tran, D.D., Castelli, E., Besacier, L., and J-F.Serignat, "Spoken and written language resources for Vietnamese," in LREC Lisbon, 2004.
- [7] Schultz, T. and Kirchhoff, K., Multilingual Speech Processing: Academic Press, 2006.
- [8] "http://www.phon.ucl.ac.uk/home/sampa".
- [9] Li, H., Ma, B., and Lee, C.H., "A Vector Space Modeling Approach to Spoken Language Identification," in IEEE Trans. on Audio, Speech and Language Processing, 2007, pp. 271-84.
- [10] Ma, B. and Li, H., "A Phonotactic-Semantic Paradigm for Automatic Spoken Document Classification," in SIGIR Salvador, Brazil, 2005.
- [11] Sebastiani, F., "Machine learning in automated text categorization," in ACM Comp. Surv., vol. 34, 2002, pp. 1-47.
- [12] Haykin, S., "Neural Networks: A Comprehensive Foundation", New York: MacMillan," N. Y. MacMillan, Ed.: Prentice Hall, 1994.
- [13] Joachims, T., "Learning to Classify Text Using Support Vector Machines," Kluwer, 2002.
- [14] Flege, J., Frieda, E., and Nozawa, T., "Amount of Native-Language (L1) Use Affects the Pronunciation of an L2," Journal of Phonetics, vol. 25, pp. 169-186, 1997.
- [15] Tan, T.P. and Besacier, L., "Modeling context and language variation for non-native speech recognition," in Interspeech, 2007, pp. 1429-1432.