ACOUSTIC MODELING FOR UNDER-RESOURCED LANGUAGES BASED ON VECTORIAL HMM-STATES REPRESENTATION USING SUBSPACE GAUSSIAN MIXTURE MODELS

Mohamed Bouallegue, Emmanuel Ferreira, Driss Matrouf, Georges Linarès, Maria Goudi, Pascal Nocera

University of Avignon, LIA, France
{mohamed.bouallegue, emmanuel.ferreira, driss.matrouf,
georges.linares, maria.goudi, pascal.nocera }@univ-avignon.fr

ABSTRACT

This paper explores a novel method for context-dependent models in automatic speech recognition (ASR), in the context of under-resourced languages. We present a simple way to realize a tying states approach, based on a new vectorial representation of the HMM states. This vectorial representation is considered as a vector of a low number of parameters obtained by the Subspace Gaussian Mixture Models paradigm (SGMM). The proposed method does not require phonetic knowledge or a large amount of data, which represent the major problems of acoustic modeling for under-resourced languages. This paper shows how this representation can be obtained and used for tying states. Our experiments, applied on Vietnamese, show that this approach achieves a stable gain compared to the classical approach which is based on decision trees. Furthermore, this method appears to be portable to other languages, as shown in the preliminary study conducted on Berber.

***Index Terms*—** Acoustic Modelling, under-resourced languages, HMM-state vector representation, state-tying, Subspace Gaussian Mixture Models

1. INTRODUCTION

Among the approximately 7000 languages of the world, only a small number of languages possess the resources (i.e. electronic text, transcribed speech corpora, etc.) required for advanced human technologies such as speech recognition. Attempting to counterbalance this problem, research in the field of speech recognition often focuses on finding strategies which deal with the lack of training data.

In this article, we focus on adapting the acoustic model estimation in such a way that it could overstep the limited available resources. Numerous approaches have been explored in order to deal with data scarcity in acoustic model estimation. Data combination strategies constitute the most popular solution that has been proposed in the literature for this purpose. In the case of the bootstrapping strategies [1], which are the most commonly used, the acoustic models are initialized, based on models from one or many source languages and are then re-estimated by using target language data only. According to [2], multilingual models that were used in the bootstrapping process proved to outperform approaches based on only one language. Other strategies of data combination have also been proposed: i.e. data pooling [3], cross-language adaptation and cross-language transfer [4]. In order to apply the data combination strategies just mentioned, we should proceed to a source/target acoustic model mapping, which can be either knowledge-based or data-driven. The first type of mapping uses the phoneme inventory of each language and the International Phonetic Alphabet (IPA) information in order to build a phoneme mapping table [5]. Another way of mapping, called data-driven, consists in calculating the distance between models [6].

The method previously described is used in building context-independent models. In order to create context-dependent models, [5] proposes to calculate the distance between any two source/target clustered context-dependent phoneme models. These clusters are obtained by using the decision tree which is a standard method of clustering. Calculating this distance allows us to determine the most similar model of context-dependent-cluster in the source language, for each model of context-dependent-cluster in the target language. This model is then copied into the acoustic model of the target language and adapted by using a small amount of its training data.

As previously shown, building a context-dependent model for under-resourced languages seems to involve multiple inter-related steps, which are not always easy to accomplish. In the first place, the clustering process of the target language is usually based on a decision tree that is time consuming due to the exhaustive evaluation of probabilities at each node in the tree structure. The questions used in order to build the decision tree could be created either manually or automatically. In the first case, the required phonetic knowledge is not always available for many languages. In the case of an automatic approach, many works like [7][8] show that the data-driven systems do not present a good performance in the cases of data scarcity, of insufficient phonetic description or of noisy training conditions. Furthermore, the fact that we don't possess of a sufficient amount of data influences the accuracy of the estimation of the parameters related to the context-dependent phoneme models. Consequently, the evaluation of the distance between source and target clusters of the context-dependent phoneme models could also be affected.

In this work, we try to overcome some of these constraints in the creation of a context-dependent acoustic model for under-resourced languages. Thus, we propose a method of context-dependent acoustic modeling with a simple technique of clustering in the procedure of states-tying [9]. The similarity distance used in the clustering is based on a vector of low dimension, which is called *factor state* . This vector, that characterizes the states, is obtained from recent proposals based on SGMM [10]. With this *factor state*, clustering becomes easier and faster, and it no longer requires phonetic knowledge. The parameter estimation of the *factor state* does not require a large amount of data, which represents an advantage for under-resourced languages. Also the likelihood computation in the tree decision approach is replaced by simple distance; moreover, state clustering may be formulated as a classical classification problem in R^d.

Our paper is organized as follows : in section 2, we expose the principal works related to context-dependent acoustic models, based on the state tying approach. In Section 3, we describe the SGMM approach and its ability to represent a state as a low dimension vector; in this section we also detail our new approach for state clustering. Section 4, demonstrates the application of our method on the Vietnamese language. Finally, conclusions are provided in Sections 5.

2. CONTEXT-DEPENDENT ACOUSTIC MODEL : STATE TYING

Most speech recognition systems are based on context dependent Hidden Markov models (HMM). In such a system, the context to the left/right of each phoneme is also taken into account ; thus, an HMM is used for each context-dependent phoneme. This type of modeling requires a large amount of data because of the large number of context-dependent phonemes to be trained. Usually, the estimation process runs into data insufficiency problems. The state tying approach appears as a common solution for this problem. It consists of clustering the models in acoustically similar groups.

The most popular technique used for states-tying is clustering using a decision-tree [9]. This technique is a computationally heavy process where competing linguistic questions are extensively evaluated. Moreover, it is based on phonetic knowledge. Clustering can be performed either by using top-down [7] or bottom-up [11] procedures. However, top-down procedures suffer from two major drawbacks: first of all, they are time consuming due to the exhaustive evaluation of probabilities for each question at each node in the tree structure. Nevertheless, they require linguistic knowledge for generating the questions for the decision tree. In case of a lack of linguistic knowledge, the automatic generation of questions constitutes a necessary additional step. In bottom-up approaches, a large number of context-dependent GMMs are estimated and they are afterwards iteratively merged according to a minimum likelihood-loss criterion. Only small mixtures are used at the leaf-level (typically from 1 or 4 gaussian components) because of the limited amount of context-dependent training data. Reported results of this approach are relatively close to the ones obtained with the decision tree approach.

Besides the complexity of the tying-state approach based on the decision tree, in the case of certain under-resourced languages, we dont possess the necessary phonetic knowledge in order to generate the questions concerning this approach. Usually, these questions are based on human expert judgment concerning similarities between contextual phonemes. Currently, a common solution for this problem is to generate these questions automatically. Several approaches are proposed. For instance, Beulen et al [7] introduces a data-based method which uses the statistical similarity as a clustering metric. [8] also attempts a data-driven approach employing local similarities between the probability density functions of hidden Markov models. [12] proposes a clustering technique which is a mixture or hybrid of the top-down and bottom-up clustering procedures. Nevertheless, even if many of these works improve the performance of ASR systems of large vocabulary languages, they seem to underperform when data is insufficient or when the phonetic units are poorly described for automatic treatment purposes, as for the case of under-resourced languages.

In our work, we present an innovative approach of states tying. In this approach, the clustering is only based on the information carried by the *factor state* $\mathbf{x}_s$ obtained using the SGMM paradigm. Our method was successively applied to French in a previous work [13] for which, on one hand, we disposed a large amount of training data and on the other, the decision tree was created by an expert. Nevertheless, the proposed method presents an advantage especially for under-resourced languages, since it does not require a large amount of data, any phonetic knowledge or an automatic generation of questions. In the next section, we explain our strategies used to estimate the *factor state* $\mathbf{x}_s$ and the manner in which they can be used to realize the state-tying procedure.

3. STATE TYING BASED ON SUBSPACE GAUSSIAN MIXTURE

3.1. SGMM for vectorial HMM-state representation : factor state

The Subspace Gaussian Mixture Model is a modeling approach based on the Gaussian Mixture Model. In SGMM, the HMM states share a common Gaussian Mixture Model, called the Universal Background Model (UBM). The means

and mixture weights are allowed to vary in a subspace of the full parameter space. Indeed, the means and the weights for each state are derived from the GMM-UBM [10]. The global GMM-UBM is defined as follows: UBM=$(\alpha_g, m_g, \Sigma_g)$, where α_g, m_g and Σ_g are respectively the weight, the means and the covariance matrix of the g^{th} Gaussian.
Let m be the means super-vector obtained by concatenating all Gaussian means. In the SGMM the means super-vector random variable of the state s is written as follows:

$$\mathbf{m}_s = m + \mathbf{U}\mathbf{x}_s \tag{1}$$

where $\mathbf{m}_s$ is the state dependent means super-vector (random vector variable) and $\mathbf{U}$ represents the inter-state variability matrix (a $MD \times R$ matrix) of low rank R. M is the Gaussian components in the UBM and D is the cepstral feature size. $\mathbf{x}_s$ are the *factor states* (an R vector). The vector $\mathbf{x}_s$ is assumed to be normally distributed among $\mathcal{N}(0, I)$. In the training phase, the $\mathbf{U}$ matrix is estimated to use all training data and the MAP point estimate $x_{(s)}$ of $\mathbf{x}_s$ is obtained for each state [14].

In [13], we detail how we calculate the *factor states* with the algorithm that presents the adopted strategy to estimate the $\mathbf{U}$ matrix.

In a previous work [15], we demonstrated that acoustic models using GMM states (which are estimated by Equation 1) result in a similar performance when compared to a baseline system. These results show that the *factor states* $\mathbf{x}_s$, with their limited number of parameters, are sufficient to allow us to characterize their states. The simplicity of the vector processing makes possible many new applications based on *factor states*. In this work, we used this *factor states* in order to accomplish a state tying procedure for the building of a context-dependent acoustic model for under-resourced languages.

In the following sections, we present our new approach for clustering HMM-states. This approach will be compared with the standard approach based on decision trees.

3.2. State tying based on factor states

In this part, we present our approach of states-tying which reduces the number of states in the acoustic model. The clustering is only based on the information carried by the *factor state* $\mathbf{x}_s$, obtained using the SGMM paradigm, without any use of phonetic rules related to the context. We first estimate the factor vectors of all states of the context-dependent phonemes that exist in the training corpus. Then, we search for the states that are acoustically similar by using a standard clustering algorithm k-means.

We start by conducting a context-independent phoneme segmentation from our training corpus. The context-independent HMM models are used during this procedure. Then, we expand the segmentation to a context-dependent phoneme segmentation: a correspondence between context-independent and context-dependent phonemes.

Afterwards, we estimate the *factor states* $\mathbf{x}_s$ for each state s using the method described in the previous section. A good estimation of $\mathbf{x}_s$ requires a minimum number of frames. It is preferable to process the states having more than 100 frames but if the amount of data is not sufficient, we can process the states having at least 40 frames.

In the classification step, which involves grouping the states of the context-dependent HMM phonemes into acoustically homogeneous classes (called class-states), we use the well known unsupervised classification algorithm $k-means$. This algorithm achieves a non-hierarchical clustering by minimizing intra-cluster variance based on Euclidean distance. This process constitutes a simple way to classify a given data set into a certain number of clusters (denoted k) fixed beforehand.

After the clustering step, we use a GMM to model the states belonging to the same class. The class-state GMMs are derived from the GMM-UBM. This derivation is obtained by using a MAP adaptation technique on data belonging to each class-state. Based on the previously obtained GMMs, we associate the unclassified states to the class-state which appears to have the maximum association with the unclassified state. To obtain the final class-state GMMs, we adapt the former class-state GMMs from data that belongs to these class-states (included the new states classified).

The class-state GMMs become the GMMs for the HMM-state of our acoustic model. Finally, in order to improve the performance of our HMM, the standard recursion of re-alignment and parameter re-estimation is carried out. Through this classification, we can reduce the total number of modeled states and solve the problem of infrequent context modeling. As we have already pointed out, the advantages of this clustering are its simplicity, its low complexity and its speed.

In the next paragraphs, we present the results of this method when applied to Vietnamese, an under-resourced language.

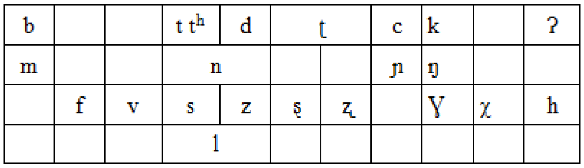
4. APPLICATION FOR VIETNAMIEN LANGUAGE

4.1. linguistic aspects

Vietnamese language belongs to the Viet-Muong group which is a member of the Mon-Khmer branch of the Austroasiatic language family. It is spoken by about 82 million people mainly in Vietnam. Vietnamese was originally written with a Siniform script but a Latin-based writing system was introduced during the 17th century and has been widely used ever since.

Vietnamese is characterized by the scholars as a typical case of a syllabic and isolating language where the syllable, the morpheme and the word seem to coincide [16]. However, even if it is traditionally considered as a monosyllabic language, the fact that compound words in Vietnamese can be formed by multiple syllables, that don't necessarily exist as independent morphemes, should also be considered [17]. But even in the case of multisyllabic words, in the written form of the language, syllables are systematically separated by a space.

Concerning the phonological system of the Vietnamese language and given the rather fixed structure of its syllable, we could firstly list the following 22 phonemes that can be realized at the onset of the syllable [17] :



b			t tʰ	d	ʈ		c	k		ʔ
m			n				ɲ	ŋ		
	f	v	s	z	ʂ	ʐ		ɣ	χ	h
			l							

Fig. 1. *Vietnamese consonants figuring at the onset of a syllable.*

At the nucleus can be realized 9 vowels (/i e ɛ a ɯ ɣ u o ɔ/), 4 short vowels (/ɛ̆ ă ɤ̆ ɔ̆/) and 3 diphthongs (/ie ɯɤ uo/). Only a small number of consonants and two semi-vowels can be found at the coda (final) of the syllable : i.e. /p t k m n ŋ/ and /j w/ respectively [17]. Vietnamese is a tonal language possessing 6 tones. The majority of the syllables can be realized with six different tones except for the syllables whose coda is composed of one of the three consonants /p t k/. Syllables of this type can only carry two of the whole set of tones. The role of the tone in Vietnamese is distinctive, that is, two different tones differentiate the meaning of a syllable [17]. It is worth noting at this point that compared to the French phonemic system that we use as the basis of the bootstraping process , the Vietnamese one presents some differences regarding, for example, the opposition of the vowel length (e.g. /ă/,/a/), the presence of diphthongs (e.g. [ie]) as well as some consonants and vowels (e.g. the retroflex consonants [ʈ], [ʂ], [ʐ], the glottal stop [ʔ], and the posterior vowels [ɯ], [ɤ]), not present in the French system.

4.2. Experimental framework and Results

4.2.1. available corpus

Speech corpus : The VNSPEECHCORPUS corpus used in our tests is a read speech corpus, recorded in a quiet studio [18]. The corresponding files are digitized in Wave format, with 16 KHz sampling rate and Analog-to-Digital precision of 16 bits. There are two types of read text: paragraph (80%) and conversation (20%). We only use the records of standard dialect speakers (North of Vietnam). In all, approximately 9 hours of speech are recorded, corresponding to eighteen speakers : 10 men and 8 women. For our experiment, we used 7 hours for the training corpus (8 men and 6 women) and 2 hours for the test corpus (2 men and 2 women).

Text corpus : The Vietnamese text corpus, used to estimate the language model, was exclusively collected from web site pages and numeric newspapers. Our corpus is composed of 2,7 millions of sentences, that is, with 45 millions syllables.

4.2.2. Baseline : decision tree based system

To build the contex-independent acoustic model, the correspondence between the already estimated French phoneme states and the Vietnamese ones was achieved by using the IPA proximities criterion (mainly based on an expert's linguistic knowledge). The bootstrapping method was employed according to [5].

Concerning the context-dependent model, we performed automatic question generation for decision tree based state tying, as described in [7]. This technique is based on the bottom-up clustering algorithm.

Firstly, the bottom-up clustering, based on the Cross Likelihood Ratio (CLR) distance measure, is processed on each triphone state sharing the same central phoneme. The triphone states are modeled with a GMM of three gaussians using the Expectation Maximization (EM) algorithm.

Secondly, after a pruning procedure based on the log-likelihood gain, the intermediate clusters obtained in each clustering constitute a specific questions set. Then we construct a decision tree by selecting the accurate sequence of questions, using the log-likelihood maximum gain.

Based on the method that is described above, we created several acoustic models of different size in order to find the optimal parameters. The parameters to be optimized consist of the number of states in the HMM as well as the number of gaussians which model the states. In the following table, we expose the results in terms of Word Error Rates (WER).

	200s	400s	600s	800s	1200s
64g	32.81	32.73	32.70	33.00	34.90
128g	33.83	33.80	34.40	35.30	36.40
256g	34.00	33.97	36.10	35.40	38.40

Table 1. *Result in WER of new model of different size.*

According to this table, the model, which is composed of 600 states, with 64 gaussians for each state, has the best results. We remark that we cannot create models with a large number of parameters otherwise we would have to face the constraint of overlearning. In Section 4.3.1, these results will be compared to the new models, obtained by the method proposed in this work.

4.3. Context-dependent models : States tying based on factor states

In this part, we apply our approach of state tying on the Vietnamien language. As we have indicated above, we only have 7 hours of speech as a training corps. Firstly, based on the independent-context HMM model, as described in 4.2.2, we obtained the context-independent phoneme segmentation. Then, we expanded this segmentation to a context-dependent phoneme segmentation. The number of context dependent phonemes that have been found in the corpus is 8624, corresponding to 25873 states before clustering.

Secondly, we estimate the *factor states* $\mathbf{x}_s$ for each state s based on the SGMM. Since a good estimation of $\mathbf{x}_s$ requires a minimum number of frames, we process only the states having more than 40 frames.

Thirdly, we use the unsupervised classification algorithm $k - means$ in order to group the states of the context-dependent HMM phonemes into acoustically homogeneous classes. After clustering, we model the states of each class by a GMM, derived from the GMM-UBM. In order to associate to a class the states that have less than 40 frames, we were based on the likelihood.

Finally, we adapt a second time the GMM models, already obtained from data (including the new states classified), that belong to these class-states. The class-state GMMs becomes the GMMs for the HMM-state of our context phoneme acoustic model.

To obtain the final class-state GMMs, we adapt the former class-state GMMs from data that belongs to these class-states (included the new states classified).

4.3.1. Results

In our experiments, we created several context-dependent acoustic models using the new proposed approach in the states-tying procedure. The performance of our new models are rated by comparison with models which, accomplish a states-tying procedure, based on a decision tree. The questions that were used for generating the decision trees are automatically generated, as mentioned in 3.2.2. We evaluate the system's performance according to three parameters: the full state number (i.e. the number of states in the HMM set), the gaussian number of the GMM which models the HMM-state and the number of parameters in the *factor states*. The obtained results, in terms of WER, are presented in the following tables. These results will be compared with the best Baseline of a 32.70% WER, as described in 4.2.2.

In the first experiment, we tested a package of models in order to find the optimal parameters. We estimated four *factor states* which have respectively, 20, 40, 80 and 120 parameters. Subsequently, we were based on each one of the *factor states* in order to cluster the states of the context-dependent phonemes in respectively, 1200, 1400, 1600, 1800 and 2000 classes. The number of classes is the number of states in our acoustic model. The gaussian number of GMMs, which model the states, are fixed in 128 gaussians. In Table 1, we show the WER for these models. Each line of the table corresponds to the number of states in the acoustic models. In the columns, we find the number of parameters of the *factor states* that we used in the clustering.

GMM-UBM of 128 gausssion

	20	40	80	120
1000	28.80	29.50	38.50	57.10
1200	28.00	28.80	38.40	60.70
1400	28.10	28.20	35.70	52.30
1600	27.20	27.80	35.10	49.90
1800	26.80	27.20	34.20	50.50
2000	26.60	27.50	33.80	49.00

Table 2. *Result in WER of new model of different size.*

The results obtained show that the models with the biggest number of states perform better. In addition, we observe that, after 1800 states, the gain becomes stationary. This result shows that, because of a small amount of training data, we can not model the acoustic model with a big number of states. For the same reason, we couldn't estimate a satisfactory *factor state* of a large number of parameters. Therefore, we obtained the best clustering results, based on a *factor state* of the smallest number of parameters, that is 20 parameters.

In the second experiment, we tried to find the optimal number of gaussians which model the states. We fixed the number of the *factor states* parameters in 20. We modeled several models with respectively, 1600, 1800 and 2000 states. For each one of these states, we modeled the GMM states by respectively, 64, 128 and 256 gaussians.

Table 3 shows that the best model has 256 gaussians. Unlike the standard clustering approach, these results show the possibility to build an acoustic model with an important number of states. The good clustering, obtained based on factor states, allows us to well estimate the states by using an important umber of gaussians without having to face the over-learning problem. The best model shows an absolute gain of 7.8%.

		64g	128g	256g
x_s of 20 parameters	1600	30.60	27.20	26.00
	1800	29.80	27.80	25.50
	2000	29.30	26.6 0	24.90

Table 3. *Results in WER of new models of different number of gaussians.*

In the last experiment, we tested a new approach of clustering. Firstly, we grouped the HMM-states per phoneme. Then, based on the *factor states*,we classified the HMM-states of each group. We fixed the number of gaussian in 256 and the number of parameters of the *factor states*, in 20. The results presented in Table 4 show that in this way, we obtain an absolute gain of 0.7% compared to the former method.

GMM-UBM of 256 gausssion, x_s of 20 parameters

	1600s	1800s	2000s	2200s	2400s
WER	25.00	24.80	24.90	24.30	24.2

Table 4. *Results in WER of guided classification.*

5. CONCLUSIONS

This work proposes a new approach of modeling context-dependent phonemes for acoustic models of under-resourced languages. In this approach, the tying-state procedure is based on *factor states*, obtained with an SGMM paradigm. The application of our method on the Vietnamese language showed an improvement compared to the baseline system, which was obtained by the standard decision tree technique. Moreover, our approach is efficient in the case of a lack of large amounts of training data. Another important aspect of our technique consists in the fact that it replaced the generation of a decision tree, necessary for the context-dependent acoustic models. In a supplementary experiment, we applied our method on Berber, another under-resourced language. The results showed an absolute gain of 4%, compared to the baseline system. Consequently, these results are encouraging for testing this method to other under-resourced languages, in order to verify its portability feature.

6. REFERENCES

[1] L. Osterholtz, C. Augustine, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, and M. Woszczyna, “Testing generality in janus: A multilingual speech to speech translation system,” 1992, vol. Vol. 1, pp. 209–212.

[2] T. Schultz and A. Waibel, “Language independent and language adaptive acoustic modeling for speech recognition. in speech communication,” 2001, vol. Vol. 35, p. 3151.

[3] T. Schultz and A. Waibel, “Multilingual and crosslingual speech recognition,” 1998, pp. 259–252.

[4] A. Constantinescu and G. chollet, “On cross-language experiments and data-driven units for alisp,” 1997, pp. 606–613.

[5] V. Bac and L. Besacier, “Automatic speech recognition for under-resourced languages : Application to vietnamese language,” 2009, vol. 17, pp. 1471–1482.

[6] O. Anderson, P. Dalsgaard, and W. Barry and, “On the use of data-driven clustering techniques for language identification of poly and mono-phonemes for four european languages,” 1994, pp. 121–124.

[7] K. Beulen and H. Ney, “Automatic question generation for decision tree based state tying,” 1998, vol. vol 2, pp. 805–809.

[8] R. Singh, B. Raj, and R. M.Stern, “Automatic clustering and generation of contextual questions for tied states in hidden markov models,” 1999, vol. vol. 1, pp. 117–120.

[9] Gilles Boulianne and Patrick Kenny, “Optimal tying of hmm mixture densities using decision trees,” in *ICSLP*, 1996.

[10] Daniel Povey, Lukas Burget, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondrej Glembek, Nagendra K. Goel, Martin Karafiát, Ariya Rastrow, Richard C. Rose, Petr Schwarz, and Samuel Thomas, “Subspace gaussian mixture models for speech recognition,” in *ICASSP*, 2010, pp. 4330–4333.

[11] Xavier Aubert, Peter Beyerlein, and Meinhard Ullrich, “A bottom-up approach for handling unseen triphones in large vocabulary continuous speech recognition,” in *CLEO collaboration), Cornell preprint CLNS 94/1306, CLEO 94/24*, 1996, pp. 14–17.

[12] F. Diehl and A. Moreno, “Acoustic phonetic modelling using local codebook features,” 2004.

[13] Mohamed Bouallegue, Driss Matrouf, Mickael Rouvier, and Georges Linarès, “Subspace gaussian mixture models for vectorial hmm-states representation,” in *ASRU*, 2011, pp. 512–516.

[14] Driss Matrouf, Nicolas Scheffer, Benoit G. B. Fauve, and Jean-Franccois Bonastre, “A straightforward and efficient implementation of the factor analysis model for speaker verification,” in *INTERSPEECH*, 2007, pp. 1242–1245.

[15] Mohamed Bouallegue, Driss Matrouf, and Georges Linares, “A simplified subspace gaussian mixture to compact acoustic models for speech recognition,” in *ICASSP*, 2011, pp. 4896–4899.

[16] O. Cao Xuân-Ha, “Phonologie et linéarité. réflexions critiques sur les postulats de la phonologie contemporaine,” 1985, vol. SELAF, Paris.

[17] Thi-Thuy-Hien Tran, “Processus d’acquisition des clusters et autres séquences de consonnes en langue seconde : de l’analyse acoustico-perceptive des séquences consonantiques du vietnamien à l’analyse de la perception et production des clusters du francais par des apprenants vietnamiens du fle,” 2012.

[18] Viet Bac Le, Do Dat Tran, Eric Castelli, Laurent Besacier, and Serignat Jean-Francois, “Spoken and written language resources for vietnamese,” in *LREC*. 2004, European Language Resources Association.