Subspace Gaussian Mixture Models for Vectorial HMM-states Representation

Mohamed Bouallegue, Driss Matrouf, Mickael Rouvier, Georges Linarès

University of Avignon, LIA, France

{mohamed.bouallegue, driss.matrouf, mickael.rouvier ,georges.linares}@univ-avignon.fr

***Abstract*—In this paper we present a vectorial representation of the HMM states that is inspired by the Subspace Gaussian Mixture Models paradigm (SGMM). This vectorial representation of states will make possible a large number of applications, such as HMM-states clustering and graphical visualization. Thanks to this representation, the Hidden Markov Model (HMM) states can be seen as sets of points in multi-dimensional space and then can be studied using statistical data analysis techniques. In this paper, we show how this representation can be obtained and used for tying states of an HHM-based automatic speech recognition system without any use of linguistic or phonetic knowledge. In experiments, this approach achieves significant and stable gain, while conserving the classical approach based on decision trees. We also show how it can be used for graphical visualization, which can be useful in other domains like phonetics or clinical phonetics.**

Index Terms: HMM-state vector representation, Speech recognition, Acoustic Modelling, HMM states clustering, Subspace Gaussian mixture.

I. Introduction

Hidden Markov Models are the dominant technology in speech recognition systems; almost all automatic recognition systems simulate acoustic-phonetic phenomena using HMM. However, progress in HMM-based recognition systems has slowed, and thus research in alternative speech recognition methods has become more critical. Phonemic treatment and classification may be one such supplemental approach.

Phoneme classification in continuous speech is a particular pattern classification problem. Each phoneme in the phoneme set is represented by a three-state left-to-right HMM. The most common method uses a mixture of weighted Gaussian probability density functions which characterize the distribution of observations within each state. In the phonemes classification, the scientific communities are tried to define of a similarity distance between phonemes or allophones. This distance can be used to reduce the number of models have in a given context, to find the optimal set of phonemes in a language or to establish a common phonetic set to several languages. Young [1] defined the similarity between allophones by expressing the difference of two Gaussians according to their mean and their variance. However, this approach is only applicable to Markov models with one state and one Gaussian. Replacing the concept of divergence from the Bhattacharyya distance, Mak [2] proposed an alternative expression of the similarity which depends on model parameters. This approach is applicable to Markov models with Gaussian mixtures of a single state. The comparison of two Markov models in several states and Gaussian mixtures requires a measure of similarity based on the likelihood of the acoustic data (cepstral coefficients) compared to Markov models. In the same context, Juang [3] et Khler [4] proposed a similarity distance defined as the difference from acoustic likelihoods, this measure is derived from divergence concept between Markov models.

In the previously cited works, the definition of similarity distance between phonemes is based on Gaussian Mixture Model parameters (means and variances) and the likelihood of the acoustic data. However, the estimation of similarity distance with proposed methods are complicated and time-consuming, and requires good estimation of large parameters (means, weights and covariance matrix).

In this paper we propose a new vectorial representation of HMM states. A state is represented by a vector of low dimension, called a factor state, which is obtained from recent proposals based on SGMM [5]. We also propose using a simple method to calculate the similarity distance, based on factor states, in the procedure of states-tying [6]. This method involves putting together all data frames corresponding to states which are acoustically similar, and then estimating a unique GMM to reduce the total number of states in the HMM. The most popular approach used in literature is based on the use of a decision tree algorithm, associated with phonetic knowledge together with the training data to decide which contexts are acoustically similar [5].

With the vectorial representation, we propose in this work, clustering states becomes easier and faster, and no longer requires phonetic or linguistic knowledge. Likelihood computation is replaced by simple distance and state clustering may be formulated as a classical classification problem in R^d. This representation will also allow the use of scientific results obtained during several years of research in pattern classification and data analysis. Moreover, the state vector representation could facilitate treatment of nuisance acoustic variability, as it did in the speaker recognition domain [7].

The rest of the paper is organized as follows : Section 2 describes the SGMM approach and its ability to represent a state as a low dimension vector. In Section 3, we present the strategy to estimate the factor states. Section 4 details our new approach for state clustering based on factor states and some experimental results. Finally, discussion and conclusions are provided in Sections 6 and 7 respectively.

II. SGMM FOR A SIMPLE VECTORIAL REPRESENTATION OF STATES

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 [8]. 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 mean and the covariance matrix of the g^{th} Gaussian.
Let m be the mean super-vector obtained by concatenating all Gaussian means. In the SGMM the mean 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 mean 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 number of Gaussians in the UBM and D the is cepstral feature size. $\mathbf{x}_s$ are the state factors (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 using all training data and the MAP point estimate $x_{(s)}$ of $\mathbf{x}_s$ is obtained for each state [9].

In the SGMM modeling [8], the specific Gaussian weights are obtained from the UBM as follow :

$$w_g^s = \frac{exp\ \mathbf{w}_g^T \mathbf{x}_s}{\sum_{g'=1}^{M} exp\ \mathbf{w}_{g'}^T \mathbf{x}_s} \tag{2}$$

where $\mathbf{x}_s \in \Re^R$ is the "state vector" with R the subspace dimention. $\mathbf{w}_g$ is weight vector depending on the Gaussian but not on the state.
This state specific weights estimation is difficult to perform. Its derivation is based on a combination of Jensen-type inequalities, local second-order Taylor-series expansions, and a modification to the resulting quadratic auxiliary function which ensures stability while maintaining the same local gradient [8]. As an alternative, we proposed in [10] to re-estimate the gaussian weights by simple iteration EM (Expectation-maximization algorithm). Its derivation is further more simplified and allows us to gain computing time. Let w_g be the weight of the Gaussian g in the UBM. The weight w_g^s for that Gaussian in the state s is calculated as follows:

$$w_g^s = \frac{\sum_{x \in s} P(g|x)}{N_s} \tag{3}$$

where,

$$P(g|x) = \frac{w_g * f(x|g)}{\sum_{g'} w_g * f(x|g')} \tag{4}$$

$N_{(s)}$ is the number of frames of the to HMM-state (s), $P(g|x)$ is the a *posteriori* probability of the Gaussian g given the frame x and $f(x|g)$ is the likelihood for the frame x given the Gaussian g.

In a previous work [10], we demonstrated that acoustic models using GMM states (which are estimated by Equation 1) result in similar performance when compared to a baseline system (Sect IV.A.2). These results show that the state factors $\mathbf{x}_s$, with their limited number of parameters, are sufficient to allow us to characterize their states. Based on these results, we propose the adoption of the factor state $\mathbf{x}_s$ as a vectorial representation of HMM states. The simplicity of vector processing makes possible many new applications based on factor states. In the rest of this article we present how we calculate factor states. We also present our new approach for clustering HMM-states based on their factor states without any use of other linguistic information or phonetic knowledge. This approach will be compared with the standard approach based on decision trees.

III. ESTMATION OF U AND LATENT VARIABLE xs

In this section, we show a strategy to estimate the matrix $\mathbf{U}$ and the state factors $\mathbf{x}_s$. The $\mathbf{U}$ matrix is estimated using training data; the MAP point $x_{(s)}$ of $\mathbf{x}_{(s)}$ is estimated using state-dependent data frames.

Let $\mathbf{N}_{(s)}$ and $\mathbf{X}_{(s)}$ be two vectors containing the zero order and first order state statistics respectively. The statics are estimated against the UBM:

$$\mathbf{N}_{(s)}[g] = \sum_{t \in s} \gamma_g(t);\ \{\mathbf{X}_{(s)}\}_{[g]} = \sum_{t \in (s)} \gamma_g(t) \cdot t \tag{5}$$

where $\gamma_g(t)$ is the *a posteriori* probability of Gaussian g for the observation t. In the equation, $\sum_{t \in s}$ represents the sum over all frames belonging to the state s.

Let $\overline{\mathbf{X}}_{(s)}$ be the state dependent statistics defined as follows:

$$\{\overline{\mathbf{X}}_{(s)}\}_{[g]} = \{\mathbf{X}_{(s)}\}_{[g]} - \mathbf{m}_{[g]} \cdot \sum_{s} \mathbf{N}_{(s)}[g] \tag{6}$$

Let $\mathbf{L}_{(s)}$ be a $R \times R$ matrix, and $\mathbf{B}_{(s)}$ a vector of dimension R, both defined as:

$$\begin{aligned} \mathbf{L}_{(s)} &= \mathbf{I} + \sum_{g \in \text{UBM}} \mathbf{N}_{(s)}[g] \cdot \{\mathbf{U}\}_{[g]}^t \cdot \mathbf{\Sigma}_{[g]}^{-1} \cdot \{\mathbf{U}\}_{[g]} \\ \mathbf{B}_{(s)} &= \sum_{g \in \text{UBM}} \{\mathbf{U}\}_{[g]}^t \cdot \mathbf{\Sigma}_g^{-1} \cdot \{\overline{\mathbf{X}}_{(s)}\}_{[g]}, \end{aligned} \tag{7}$$

By using $\mathbf{L}_{(s)}$ and $\mathbf{B}_{(s)}$, $x_{(s)}$ can be obtained using the following equation:

$$x_{(s)} = \mathbf{L}_{(s)}^{-1} \cdot \mathbf{B}_{(s)} \tag{8}$$

The matrix $\mathbf{U}$ can be estimated line by line, with $\{\mathbf{U}\}_{[g]}^i$ being the i^{th} line of $\{\mathbf{U}\}_{[g]}$ then:

$$\mathbf{U}_{[g]}^i = \mathbf{L}\mathbf{U}_g^{-1} \cdot \mathbf{R}\mathbf{U}_g^i, \tag{9}$$

Algorithm 1: Estimation algorithm of $\mathbf{U}$ and latent variable $\mathbf{x}$.

For each state s : $x_{(s)} \leftarrow 0$, $\mathbf{U} \leftarrow random$;
Estimate statistics: $\mathbf{N}_{(s)}$, $\mathbf{X}_{(s)}$ (eq.5);
for $i = 1$ *to* $nb\_iterations$ **do**
    for $all\ s$ **do**
        Center statistics: $\overline{\mathbf{X}}_{(s)}$ (eq.6);
        Estimate $\mathbf{L}_{(s)}$ and $\mathbf{B}_{(s)}$ (eq.7);
        Estimate $x_{(s)}$ (eq.8);
    end
    Estimate matrix $\mathbf{U}$ (eq. 9 and 10) ;
end

where $\mathbf{RU}_g^i$ and $\mathbf{LU}_g$ are given by:

$$\begin{aligned} \mathbf{LU}_g &= \sum_s \mathbf{L}_{(s)}^{-1} + x_{(s)} x_{(s)}^t \cdot \mathbf{N}_{(s)}[g] \\ \mathbf{RU}_g^i &= \sum_s \{\overline{\mathbf{X}}_{(s)}\}_{[g]}^{[i]} \cdot x_{(s)} \end{aligned} \tag{10}$$

Algorithm 1 presents the method adopted to estimate the state variability matrix with the above developments where the standard likelihood function can be used to assess the convergence.

IV. State tying with state factors

The most popular technique used for states-tying is clustering using a decision-tree [11]. 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 using top-down [12] or bottom-up [13] procedures. However, top-down procedures suffer from two major drawbacks: they require linguistic knowledge, which may be unavailable for certain languages, and they have slow runtime due to the exhaustive evaluation of probabilities for each question at each node in the tree structure. In bottom-up approaches, a large number of context-dependent GMMs are estimated and then they are 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) according to the limited amount of state-dependent training data. Reported results are relatively close to the one obtained with the decision tree approach.

Here we present our new approach of states-tying to reduce the number of states in the acoustic model. The clustering is only based on the information carried by the state factor $\mathbf{x}_s$ 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 ESTER1&2 [15]. The context-independent HMM models are utilized during this procedure. Then, we expand the segmentation to a context-dependent phoneme segmentation: a translation between context-independent and context-dependent phonemes. The number of context dependent phonemes that have been found in the corpus is 21,889, corresponding to 65,667 states before clustering.

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, thus we process only the states having more than 50 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 non-hierarchical clustering by minimizing intra-cluster variance based on Euclidean distance; this process is a simple way to classify a given data set into a certain number of clusters (denoted k) fixed beforehand.

With this vectorial representation we can visualize the results of our Clustering algorithm in two dimensional space by using the mathematical procedure called Principal Component Analysis (PCA). This procedure allows for the size reduction of the initial space into a smaller dimension with minimal loss of information. In our case we project the factor states of one hundred parameters into two space. We determine the principal components using PCA procedure. figure 1 shows the projection of nine different groups of factor-states.

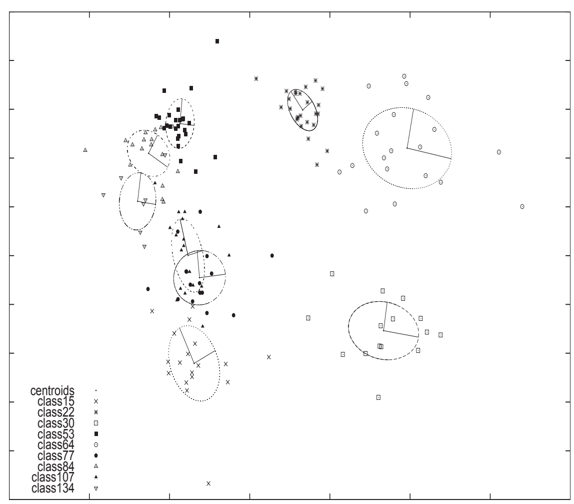

Fig. 1. *Projection showing different classes of factor-states*

After Clustering, we model the states belonging to the same class as a GMM. The class-state GMMs are derived from the GMM-UBM. This derivation is obtained by using a MAP adaptation on data belonging to the single class-states. We associate the unclassified states to the class-state which appears to have the maximum association with the unclassified state based on the already obtained GMMs. 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).

In order to further reduce the number of parameters in a GMM for a state s, we select the N-best Gaussians with the largest weights and ignore the others. N is chosen in such a way that the sum of the weights of the selected Gaussians reaches a predefined threshold (defined 0.9).

The class-state GMMs become the GMMs for the HMM-state of our acoustic model. Finally, to improve the performance of our HMM the standard recursion of re-alignment and parameter re-estimation is performed.

Through this classification, we can reduce the total number of modeled states and solve the problem of infrequent context modeling. The advantages of this clustering are its simplicity, low complexity and speed, it is based only on the state factors that do not use linguistic or phonetic knowledge as a decision-tree.

A. Experimental framework and Results

1) The LIA broadcast system: Experiments are carried out using the Laboratoire Informatique d'Avignon (LIA) broadcast news (BN) system which was used in the ESTER evaluation campaign. This system relies on the HMM-based decoder developed at the LIA, Speeral [14], and is based on an A* decoder using a state-dependent HMM for acoustic modeling. The language models are classical trigrams estimated using about 200M words from the French newspaper Le Monde and about 1M words from the ESTER1&2 broadcast news corpus. The lexicon contains 67K words. In these experiments, only one decoding pass is performed in 3x Real Time. The training parts of the data set are based on the training corpus provided for the ESTER1&2 evaluation campaign with 190 hours. The ESTER1&2 corpus consists of French radio broadcasts from the Radio-France group . We test our approach on 9.5 hours of speech extracted from the ESTER2 test set.

2) Baseline : decision tree based system: The baseline system adopted in this work is based on Hidden Markov Models. The state-level probabilities are estimated by using Gaussian Mixture Models. Our baseline uses a left-to-right HMM architecture of 13,316 context-dependent phonemes. To reduce the size of the acoustic models, we used a states-tying technique based on decision-tree top-down context clustering approache. This technique allows one to model 13,316 context-dependent phonemes with only 5080 states instead of 39,948, with 64 Gaussians per state and 39 PLP (Perceptual Linear Predictive) coefficients per frame (13 static with first and second order derivatives).

3) Results: In our experiment, we created acoustic models using the new proposed approach. The performance of our new models are rated by comparison with models which accomplish states-tying using the standard technique based on a decision-tree. We evaluate the system's performance according to the full state number (i.e. the number of states in the HMM set), and the number of parameters in the factors state. The obtained results are presented in the following tables. Table I shows the results in terms of Word Error Rates (WER) on our test corpus. The first column contains the titles of the radio stations from which the samples were collected. The second column shows the WER of the baseline system. In the last three columns, we find the WER of three different models, each with a different number of states: 2672, 3730 and 4932. The number of parameters in a state factor $x_{(s)}$ is 60. The last line indicates the percentage of absolute gain compared to the baseline system.

Station	Baseline (5079)	2672	3730	4932
Inter (5h40mn)	34.88	33.32	32.85	32.49
RFI (1h10mn)	18.62	18.06	17.39	17.29
tvme (1h)	28.74	26.45	26.02	26.03
africa(1h30mn)	32.42	31.16	30.90	30.69
Total (9h30)	30.92	29.46	29.03	28.8
absolute gain	-	1.46	1.89	2.12

TABLE I
Word Error Rates in % for different radio stations.

The results obtained show an improvement in performance even though the new models use a smaller number of states than does the baseline. We obtained an absolute gain of 2.12 when running the new model with almost the same number of states as the baseline. These results confirm that the clustering of states based on factors $x_{(s)}$ allows the model to be more independent of the linguistics in the acoustic modeling.

Table II reports the results obtained by three new models with the same number of states (4932). For each one, the classification is based on state factors of 40, 60 and 100 parameters, respectively.

Stations	4932-40	4932-60	4932-100
Inter (5h40mn)	32.74	32.49	32.85
RFI (1h10mn)	17.31	17.29	17.63
tvme (1h)	26.11	26.03	25.98
africa(1h30mn)	30.75	30.69	30.43
Total (9h30)	28.89	28.80	28.91

TABLE II
Word Error Rates in % for different stations.

Word Error Rates for the three models are very close. This demonstrates that a vector of 40 parameters is enough to characterize the state. This result encourages the use of this vector representation of states for other applications, such as analysis of variability that can affect the phoneme and graphic visualization.

V. Discussion

The new characterization of HMM states using vector representations improves the treatment of states. We have shown in this article an interesting use of factor states and the clustering of HMM states in the tying procedure. In our approach from states-tying we do not need to linguistic knowledge, That this facilitates the processing of languages which have no such knowledge

The results confirm that the factor states representations are good characterizations of HMM states, which supports the choice to base other works on factor states.

One interesting possible application of the vectorial representation is graphical visualization; using factor states the graphical visualization of phonemes becomes much simpler. To do this, ones needs only to move from the dimension of the N factor states to the dimension of a given visual medium (3D or 2D) by using the mathematical procedure Principal Component Analysis. This visualization allows us to observe several phenomena such as the influence of speaker variability on phonemes, the variability of pronunciation of a single speaker and the other variable may be affected the phoneme. To illustrate this we choose, at random, three states pronounced by several speakers. We calculate the factors states x_{s_p} for each state s pronounced by the speaker p, and after the calculation of the PCA, we project the factor states onto two dimensional space. We observe three clouds of points for the three different states. The dispersion of the points of each state around their center is due to the pronunciation variability of the different speakers.

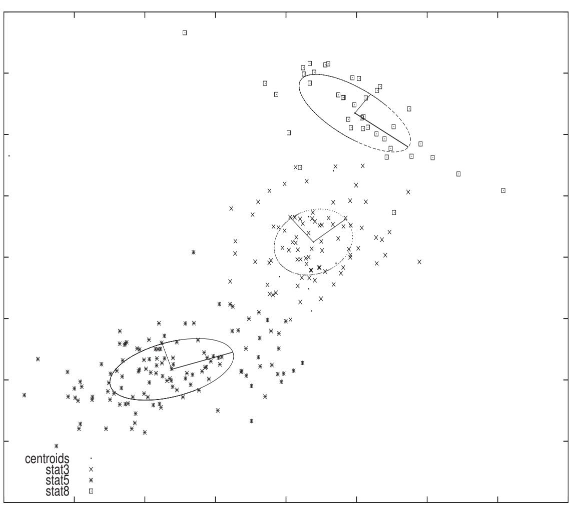

Fig. 2. *Projection showing the speaker variability of the three different states*

VI. Conclusions

In this paper, we highlighted the value of representing states as vectors in a low-dimensionality subspace. This representation is based on the subspace GMM paradigm. Once the states are transformed into vectors, analysis, visualization and comparison of states becomes easier. The application of state clustering demonstrates the accuracy and efficiency of the proposed vectorial representation. Without any use of linguistic or phonetic knowledge, the vectorial state representation achieves significant and stable gains in comparison to the standard decision tree based technique. Future works will analyze the different kinds of variability that may affect the states. We also plan to evaluate new methods of integrating state factors into the decoding procedure.

References

[1] S.J. Young, P.C. Woodland, The use of state tying in continuous speech recognition, in Proc. of European Conference on Speech Technology, EuroSpeech, pp. 2203-2206, Berlin 1993.

[2] B. Mak, E. Barnard, Phone Clustering using the Bhattacharyya Distance, in Proc. of International Conference on Speech and Language Processing,Philadelphie, 1996.

[3] B.H. Juang, L.R. Rabiner, A Probabilistic Measure for Hidden Markov Models, AT&T Technical Journal 64(2), 1985.

[4] J. Khler, Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds, in Proc. of International Conference on Speech and Language Processing, Philadelphie, 1996.

[5] D. Povey, and al., Subspace Gaussian mixture models for speech recognition, In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 10), March 14-19, 2010, Dallas, Texas, U.S.A, pp. 1144-1160, 2010.

[6] G. Boulianne, P. Kenny, Optimal Tying of HMM Mixture Densities Using Decision Trees, Proceedings of the Fourth International Conference on Spoken Language Processing, pp. 350-353, Philadelphia, Pennsylvania, USA, October 1996.

[7] Dehak Najim, Dehak Rèda, Kenny Patrick, Brümmer Niko, Ouellet Pierre, Dumouchel Pierre (2009): Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification, In INTERSPEECH, pp. 1559-1562, 2009.

[8] D. Povey, and al., A Tutoriel Introduction to Subspace Gaussien Mixture Models for Speech Recognition, Tech. Rep. MSR-TR-2009-111, Microsoft Research,2009.

[9] Driss Matrouf, Nicolas Scheffer, Benoit Fauve, Jean-Franccois Bonastre, A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification, INTERSPEECH Conference 2007, Antewerp, Belgium, pp. 1242-1245, 2007.

[10] Mohamed Bouallegue, Driss Matrouf, Georges Linarès Simplified Subspace Gaussian Mixture To Compact Acoustic Models For Speech Recognition (ICASSP 11), May 22-27, 2011, Prague, Czech Republic.

[11] W. Reichl and W. Chou, Decision Tree State Tying Based Segmental Clustering for Acoustic Modeling, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 801-804, Seattle, Washington, USA, May 1998.

[12] K. Beulen and H. Ney, Automatic Question Generation for Decision Tree Based State Tying, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 805-808, Seattle, Washington, USA, May 1998.

[13] X. Aubert and P. Beyerlein, A Bottom-Up Approach for Handling Unseen Triphones in Large Vocabulary Continuers Speech Recognition, Proceedings of the Fourth International Conference on Spoken Language Processing, pp. 14-17, Philadelphia, Pennsylvania, USA, October 1996.

[14] P. Nocera, C. Fredouille, G. Linarès, D. Matrouf, S. Meignier, J.-F. Bonastre, D. Massonié, and F. Béchet, The LIA's french broadcast news transcription system, in SWIM: Lectures by Masters in Speech Processing, 2004.

[15] S.GALLIANO, G.GRAVIER AND L.CHAUBARD (2009). The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts, 10th Annual Conference of the International Speech Communication Association , InterSpeech 2009, Brighton UK.