Component Structuring and Trajectory Modeling for Speech Recognition

Arseniy Gorin, Denis Jouvet

Component Structuring and Trajectory Modeling for Speech Recognition

Arseniy Gorin[1,2,3], *Denis Jouvet*[1,2,3]

Speech Group, LORIA
[1]Inria, 615 rue du Jardin Botanique, F-54600, Villers-ls-Nancy, France
[2]Universit de Lorraine, LORIA, UMR 7503, Villers-ls-Nancy, F-54600, France
[3]CNRS, LORIA, UMR 7503, Villers-ls-Nancy, F-54600, France
{arseniy.gorin, denis.jouvet}@inria.fr

Abstract

When the speech data are produced by speakers of different age and gender, the acoustic variability of any given phonetic unit becomes large, which degrades speech recognition performance. A way to go beyond the conventional Hidden Markov Model is to explicitly include speaker class information in the modeling. Speaker classes can be obtained by unsupervised clustering of the speech utterances.

This paper introduces a structuring of the Gaussian components of the GMM densities with respect to speaker classes. In a first approach, the structuring of the Gaussian components is combined with speaker class-dependent mixture weights. In a second approach, the structuring is used with mixture transition matrices, which add dependencies between Gaussian components of mixture densities (as in stranded GMMs). The different approaches are evaluated and compared in detail on the TIDIGITS task. Significant improvements are obtained using the proposed approaches based on structured components. Additional results are reported for phonetic decoding on the NEOLOGOS database, a large corpus of French telephone data.

Index Terms: speech recognition, unsupervised clustering, speaker class modeling, stochastic trajectory modeling

1. Introduction

Speaker variability is a well-known problem of state-of-the-art Automatic Speech Recognition (ASR) systems. The variability of the acoustic features across different speakers makes it difficult to build accurate Speaker-Independent (SI) systems. Main sources of variability include speaker gender, age and accent [1, 2]. Hidden Markov Models with Gaussian Mixture observation densities (HMM-GMM) are not able to accurately represent highly heterogeneous feature distributions because of the strong independence assumptions. Model adaptation and feature normalization techniques (i.e. MLLR [3], MAP [4], VTLN [5]) are widely used in state-of-the-art ASR systems to reduce speaker variability. The adaptation data are usually associated with a speaker, or some class of speakers (i.e. gender, age, or accent).

This work focuses on unsupervised clustering of the speech utterances, assuming that the speaker class is not changing within the sentence [6]. With respect to the training process, increasing the number of classes decreases the number of utterances associated with each class. This problem can be partially handled by soft clustering techniques, such as eigenvoice approach, where the parameters of an unknown speaker are determined as a combination of class models [7], or by explicitly enlarging the class-associated data by allowing one utterance to belong to several classes [8, 9].

In this paper, it is proposed to include the speaker class information into the model structure, instead of building separate class-based models. To do this, the components of GMM are initialized from GMMs of smaller dimensionality trained on class-associated data. In conventional HMM, GMM components are trained independently. In contrast, speaker class structuring leads to GMM, in which each k^{th} component of the density (or a subset of components) is associated with a given class.

First, to efficiently use class-structured HMM-GMM, Speaker class-dependent Weights can be used (SWGMM). This model was originally investigated in a radio broadcast transcription system [10]. In this model, the mixture weights are class-dependent and the Gaussian means and variances are class-independent, but class-structured. Another way of using class-structured GMM is to replace class-dependent mixture weights by Mixture Transition Matrices (MTMs) of the Stranded Gaussian Mixture Model (SGMM). SGMM is similar to conditional Gaussian model [11], which was recently extended, reformulated and investigated for robust ASR [12] and investigated for child data and non-stationary noise conditions [13]. MTMs explicitly define the dependencies between the components of the adjacent Gaussian mixture observation densities.

It was originally proposed in [12] to initialize SGMM from the conventional HMM-GMM. Instead, for a class-Structured SGMM (SSGMM), the SGMM is initialized from SWGMM and each GMM component (or each set of components) mainly represents a different speaker class. MTMs in SSGMM are used to model the probabilities of either staying in the same component (speaker class) over time, or dynamically switching between dominating components (classes). Explicit trajectory modeling improves the recognition accuracy. Moreover, it does not require an additional classification step to determine the class of the utterance in decoding.

The paper is organized as follows. Section 2 describes the ASR system for TIDIGITS task and formulates the problem. Section 3 discusses the unsupervised class-based approach. Section 4 introduces the class-structured SWGMM. Section 5 recaps SGMM framework and introduces the class-Structured SGMM. Sections 2 to 5 contain theoretical explanations and experimental verification of the concept on TIDIGITS task. Finally, Section 6 discusses additional phonetic decoding experiments on the NEOLOGOS data, a large corpus of French telephone data. The paper ends with conclusions and future work.

2. Baseline ASR system for TIDIGITS task

The main part of the paper is supported by the experiments conducted on TIDIGITS connected digits task that contains data from speakers of different age and gender [14]. The full training data set consists of 41224 digits (28329 for adult and 12895 for child speech). The test set consists of 41087 digits (28554 for adult and 12533 for child). The Sphinx3 toolkit [15] was modified to handle SGMM. The digits are modeled as sequences of word-dependent phones. Each phone is modeled by a 3-state HMM without skips. Each state density is modeled by 32 Gaussian components. The front-end is the same in all experiments described in the paper, and it consists of 13 standard MFCC (12 cepstral + log energy) with the first and second derivatives. Similar to other work with TIDIGITS [16], the signal is downsampled to 8 kHz. Word Error Rates (WERs) of the baseline systems are reported in Table 1. Two SI models are trained, one from the adult subset only and another from the full training set. For the last two lines, the Age and Gender-Age dependent models are built with MLLR+MAP adaptation.

Model description	Adult	Child
Training on adult data	0.64	9.92
Training on adult+child data	**1.66**	**1.88**
+Age adaptation (age is known in decoding)	1.42	1.56
+Gender-Age adaptation (gender and age are known in decoding)	1.31	1.31

TABLE 1 – TIDIGITS baseline WERs for SI, Age and Gender-Age adapted models with known speaker classes in decoding

Training on adult data provides the best results for adult speakers, but shows a weak performance on child speech. When child data are included in the training set, the performance improves on child, but degrades on adult subset. Using class-adapted models (whether Age only, or Gender-Age) further improves the baseline performance. In further experiments with TIDIGITS only full training set (i.e., adult and child data) will be considered with no class information available.

3. Unsupervised class-based ASR

Let us consider a set of training utterances without any knowledge about the speaker identity, or class. The objective is to split the training set into classes of acoustically similar data. In this case a GMM-based utterance clustering algorithm can be applied [8]. The resulting class data are used for adapting the SI model parameters. The classification GMMs are also used in decoding to identify the class for selecting the best class-model for each utterance of the test set (i.e., 2 pass decoding).

Experiments with class-based ASR on TIDIGITS data. Let us apply the described unsupervised clustering on the TIDIGITS data. The classification GMM consists of 256 components. The first clustering step (2 classes) mainly splits male speakers from female and child. The second split (4 classes) allows to separate female from child speakers. It seems impossible to distinguish boys from girls, even with more classes.

After classification, the SI acoustic models are adapted on each class data using MLLR+MAP. The bars "*CA-GMM*" in Figure 4 illustrate WERs with the associated 95% confidence intervals. The average best result is achieved with 4 classes, for which the WER is similar to the supervised Gender-Age adaptation (compare the line "*4 cl. CA-GMM*" of Table 2 with Table 1). After 4 classes, the performance degrades, because there is not enough data to adapt the class-based models.

4. Class-Structured HMM with Class-Dependent Mixture Weights

Instead of adapting all HMM parameters for each class of data, a more compact parameterization was investigated: HMM with Speaker class-dependent Weights (SWGMM) [10]. GMM components of this model are shared and structured with respect to speaker classes and only mixture weights are class-dependent. The SWGMM density for a state j and a given speaker class c has the following form:

$$b_j^{(c)}(o_t) = \sum_{k=1}^{M} w_{jk}^{(c)} \mathcal{N}(o_t, \mu_{jk}, U_{jk}) \qquad (1)$$

where M is the number of components per mixture, o_t is the observation vector at time t, $w_{jk}^{(c)}$ is the mixture weight and $\mathcal{N}(o_t, \mu_{jk}, U_{jk})$ is the Gaussian density with the mean vector μ_{jk} and the covariance matrix U_{jk}.

In decoding, each utterance to be recognized is firstly automatically assigned to some class c. Then, Viterbi decoding with the corresponding set of mixture weights is performed.

The class-structured GMMs are initialized by concatenating the components of GMMs of smaller dimensionality, separately trained from different class-data. For example, to train a target model with mixtures of M Gaussian components from Z classes, first Z models with $L = M/Z$ components per density are trained. Then, these components are concatenated into a single mixture (Figure 1).

$$\left[\mu_{j1}^{(c_1)}, \ldots, \mu_{jL}^{(c_1)}\right] \cdots \left[\mu_{j1}^{(c_Z)}, \ldots, \mu_{jL}^{(c_Z)}\right] \Rightarrow \left[\mu_{j1}, \ldots, \mu_{jL}, \ldots, \mu_{M-L+1}, \ldots, \mu_M\right]$$

FIGURE 1 – Initializing SWGMM from class-dependent models

For the combined (structured) model, mixture weights are also concatenated, copied and re-normalized. Finally, the means, variances and mixture weights are re-estimated in the iterative Expectation-Maximization manner. The class-specific data are used for updating the class-dependent mixture weights, whereas the whole data set is used for re-estimating the shared means and variances:

$$\omega_{jk}^{(c_i)} = \frac{\sum_{t=1}^{T} \gamma_{jk}^{(c_i)}(t)}{\sum_{t=1}^{T}\sum_{l=1}^{M} \gamma_{jl}^{(c_i)}(t)}; \quad \mu_{jk} = \frac{\sum_{i=1}^{Z}\sum_{t=1}^{T} \gamma_{jk}^{(c_i)}(t) \cdot o_t}{\sum_{i=1}^{Z}\sum_{t=1}^{T} \gamma_{jk}^{(c_i)}(t)} \qquad (2)$$

where $\gamma_{jk}^{(c_i)}(t)$ is the Baum-Welch count [17] for the k^{th} component of the state j with respect to the observation o_t from the class c_i. Summation over t means summation over all frames of all training utterances of the class. Variances are re-estimated in a similar way as means. Means can also be estimated in a Bayesian way (MAP) to take into account the prior distribution.

After such re-estimation the class-dependent mixture weights are larger for the components, that are associated with the corresponding classes of data (Figure 2 shows the examples of class-dependent mixture weights, averaged over all HMM densities, for classes c_7, c_{17} and c_{27}).

Experiments with SWGMM on TIDIGITS. The previous GMM-based unsupervised clustered data of TIDIGITS were used to build the proposed SWGMM. In order to build models with 32 Gaussians per density, smaller class-dependent models are combined, for example, 2 classes modeled with 16 Gaussians per density, 4 classes - with 8 Gaussians per density, and so on up to 32 classes.

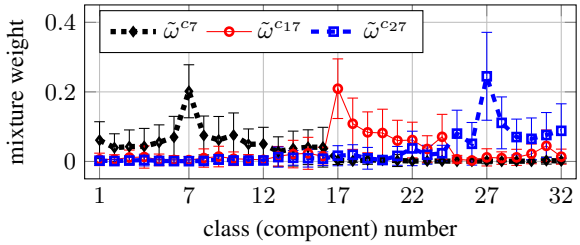

FIGURE 2 – Class-dependent weights after joint re-estimation. Here mixture weights are averaged over HMM states with corresponding standard deviation in bars (here Z=32, M=32)

Once the SWGMM is initialized, the model is re-estimated. ML estimation (MLE) is used for mixture weights and MAP for means and variances. The corresponding results are described by the bars "*SWGMM*" in Figure 4.

This parameterization allows to use the information from all classes for a robust estimation of the means and variances, and significantly reduces the WER with a limited number of parameters, due to the sharing of the Gaussian parameters. This model achieves the best result of 0.80% for adult and 1.05% for child data (see "*32 cl. SWGMM*" row in Table 2).

5. Class-Structured Stranded GMM

Stranded GMM (SGMM) was proposed in the robust ASR framework [12]. This model expands the observation densities of HMM-GMM and explicitly adds dependencies between GMM components of the adjacent states. Conventional SGMM is initialized from an HMM-GMM. In this section a class-Structured SGMM (SSGMM) is proposed.

5.1. Conventional Stranded GMM

The conventional SGMM consists of the state sequence $\mathcal{Q} = \{q_1, ..., q_T\}$, the observation sequence $\mathcal{O} = \{o_1, ..., o_T\}$, and the sequence of components of the observation density $\mathcal{M} = \{m_1, ..., m_T\}$, where every $m_t \in \{1, ..., M\}$ is the component of the observation density at the time t, and M denotes the number of such components in the mixture.

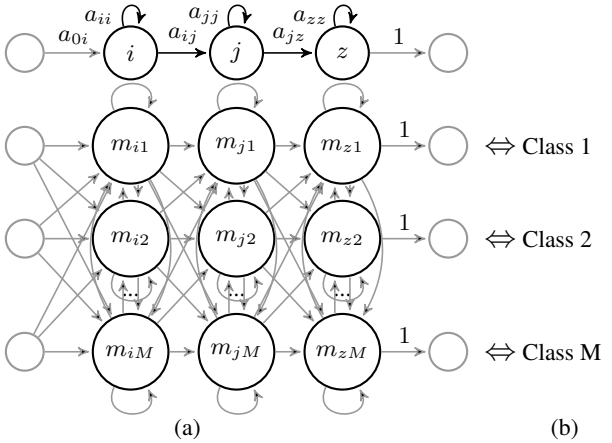

FIGURE 3 – (a) Stranded GMM with schematic representation of the component dependencies. (b) the idea of Structured SGMM, i.e., associating each k^{th} component with a data class

The difference of SGMM with respect to HMM-GMM is that an additional dependency between the components of the mixture at the current frame m_t and at the previous frame m_{t-1} is introduced (Figure 3-a). The joint likelihood of the observation, state and component sequences is defined by:

$$P(\mathcal{O}, \mathcal{Q}, \mathcal{M}|\lambda) = \prod_{t=1}^{T} P(o_t|m_t, q_t) P(m_t|m_{t-1}, q_t, q_{t-1}) P(q_t|q_{t-1}) \quad (3)$$

where $P(q_t = j|q_{t-1} = i) = a_{ij}$ is the state transition probability, $P(o_t|m_t = l, q_t = j) = b_{jl}(o_t)$ is the probability of the observation o_t with respect to the single density component $m_t = l$ in the state $q_t = j$ and $P(m_t = l|m_{t-1} = k, q_t = j, q_{t-1} = i) = c_{kl}^{(ij)}$ is the component transition probability, and $\sum_{l=1}^{M} c_{kl}^{(ij)} = 1, \forall i, j, k$. For each connected pair of states i and j a mixture transition matrix (MTM) is defined as $C^{(ij)} = \{c_{kl}^{(ij)}\}$.

Experiments with conventional SGMM on TIDIGITS. In conventional SGMM, MTM rows are initialized from the mixture weights of conventional HMM-GMM, and the model parameters are re-estimated with MLE. Such initialization and training are repeated in this section for TIDIGITS data. In addition, to reduce the number of parameters, only 2 MTMs are used for each state (i.e., cross-phone MTMs are shared). The WERs for SGMM are shown in the bar "*SGMM*" in Figure 4 and in the corresponding row of Table 2.

Compared to the HMM-GMM, SGMM improves from 1.66% to 1.11% on adult and from 1.88% to 1.27% on child speech. Both improvements are statistically significant with respect to 95% confidence interval. The SGMM performance is even better than the Gender-Age adapted baseline, but it does not outperform SWGMM, proposed in the previous section.

5.2. Class-Structured Stranded GMM

The idea of class-Structured SGMM (SSGMM) is to structure the components of SGMM, such that initially the k^{th} component of each density corresponds to a class of data (Figure 3-b). To do this, the SSGMM is initialized from the SWGMM, described in Section 4. The means and variances are taken from SWGMM and MTMs are defined with uniform probabilities. Class-dependent mixture weights of the SWGMM are not used.

When the initialization of SWGMM is done from class-models with 1 Gaussian per density, each component corresponds to a class. After EM re-estimation of all parameters, the diagonal elements of MTMs are dominating, which leads to the consistency of the class within utterance decoding. At the same time, non-diagonal elements allow other Gaussian components to contribute to the acoustic score computation.

The advantage of SSGMM is that it explicitly parameterizes speech trajectories and also allows to switch between different components (speaker classes). Therefore, the first pass classification algorithm is no more needed in decoding.

Experiments with Structured SGMM on TIDIGITS. In the experimental study with TIDIGITS data, the SSGMM is initialized from SWGMM, which was constructed using 32 classes with 1 Gaussian per class and re-estimated with ML for mixture weights and MAP for Gaussian means and variances (corresponds to the result "*32 cl. SWGMM*" in Table 2). Two MTMs per state are initialized with uniform probabilities. Then, the parameters of SSGMM are re-estimated with MLE.

The WERs for such SSGMM are described with the bars "*SSGMM* " in Figure 4 and in the corresponding rows of Table 2. Initializing SSGMM from SWGMM with different number of classes (2, 4, 8 and 16) was always leading to an accuracy improvement, compared to SGMM. Only the result, corresponding

to 32 classes, is reported. Class-Structured SGMM (*SSGMM*) further improves and achieves 0.52% WER on adult and 0.86% on child data.

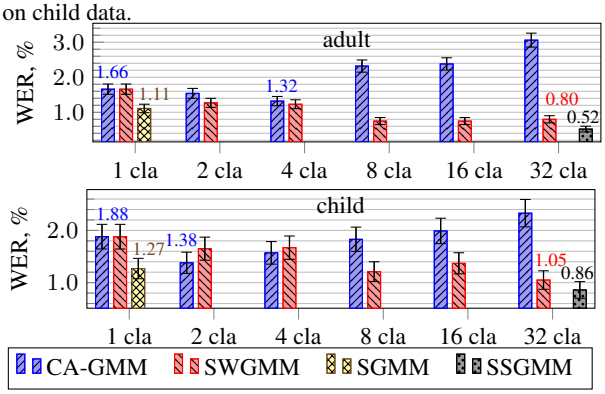

FIGURE 4 – WER for adult (top) and child (bottom) sets of TIDIGITS, computed with Class-Adapted models (CA-GMM), HMM with class-dependent weights (SWGMM), SGMM, and a Structured SGMM(SSGMM) initialized from 32 classes

Model	Decoding	Parameters/state	Adult	Child
SI GMM	1 pass	$78{\cdot}32{+}32{=}2528$	1.66	1.88
4 cl. CA-GMM	2 pass	$4{\cdot}(78{\cdot}32{+}32){=}10112$	1.32	1.57
8 cl. SWGMM	2 pass	$78{\cdot}32{+}8{\cdot}32{=}2752$	0.75	1.21
32 cl. SWGMM	2 pass	$78{\cdot}32{+}32{\cdot}32{=}3520$	**0.80**	**1.05**
SGMM	1 pass	$78{\cdot}32{+}2{\cdot}32{\cdot}32{=}4544$	1.11	1.27
SSGMM	1 pass	$78{\cdot}32{+}2{\cdot}32{\cdot}32{=}4544$	**0.52**	**0.86**

TABLE 2 – Detailed summary of best results and the number of model parameters per state for the TIDIGITS task

6. Experiments on NEOLOGOS database

The French database NEOLOGOS consists of 3 databases, that were recorded over the fixed telephone network [18]. First, IDIOLOGOS1 contains 1000 adult speakers of different gender and accent. Each speaker produces 50 sentences. Second, 200 selected speakers from IDIOLOGOS1 additionally produce 450 sentences for IDIOLOGOS2 database. Finally, PAIDIOLOGOS contains 1000 children speakers, each producing 37 sentences. Some sentences are repeated by different speakers and the total vocabulary consists of 3.3k running words. This is not crucial, as the experiments are done with phonetic decoding.

There are not many publications that deal with this database. Part of NEOLOGOS was used for phonetic decoding in the context of rapid speaker adaptation with Reference Model Interpolation (RMI) research [19]. The authors used another large database for SI model training and NEOLOGOS data for RMI adaptation and decoding (up to 3 utterances per speaker for adaptation and the remaining utterances for decoding). The back-end was also different, as they used 2-gram phone language model in decoding and 4-gram language model for the lattice re-scoring. With RMI speaker adaptation approach, they achieved 37.0% Phone Error Rate (PER) on 50 adult speakers and 61.8% on 100 child speakers. High PER shows that the task is far from being solved. The main reasons are the diversity of speakers and recording conditions and variety of non-speech acoustic events and fillers (as speakers were doing the recordings mainly from home).

6.1. NEOLOGOS phonetic decoding experiment setup

For the reported experiments, the training data includes 200 speakers that appear in both IDIOLOGOS1 and IDIOLOGOS2 (adult), plus random 200 speakers from IDIOLOGOS1 (adult) and 700 speakers from PAIDIOLOGOS (child). Development set includes 3 random phrases from 100 speakers of IDIOLOGOS1 and from 100 speakers of PAIDIOLOGOS. The remaining 500 speakers of IDIOLOGOS1 and 200 speakers of PAIDIOLOGOS are used for the test set.

After removing the sentences, that contain non-intelligible words, the training set consists of about 5M running phones (4.4M phones for adult and 0.6M for child speech); the development set contains 13594 phones (10708 phones for adult and 2886 for child). The test set contains 781011 phones (712773 phones for adult and 68238 for child).

A set of 30 phonemes is used for both training and evaluation. In the chosen phoneme set, the apertures of the vowels are not considered; i.e., the open and the close /o/ are merged in a single unit, same for the open and the close /e/, as well as for the open and the close /ø/. In addition, 6 fillers, a silence and a short pause units are used in modeling. Each context-independent unit is modeled by 3 states, and each density has 32 Gaussian components. A 3-gram phone language model is derived from training data. For evaluation purposes the development and test data were forced-aligned with the model that was trained on large vocabulary radio broadcast data and adapted on NEOLOGOS data. The word insertion penalty and the language model weight were optimized on development set.

SWGMM is initialized with 32 class models, 1 Gaussian per class-model. SSGMM is initialized from SWGMM. Means and variances are re-estimated with MAP and MTMs are updated with ML. The corresponding Phone Error Rates are summarized in Table 3.

Model	Decoding	All	Adult	Child
SI GMM	1 pass	42.42	41.16	55.55
32 cl. SWGMM	2 pass	**41.36**	**40.20**	**53.43**
SSGMM	1 pass	**41.14**	**40.03**	**52.75**

TABLE 3 – Phone Error Rate on NEOLOGOS test data

The 95% confidence interval is about $\pm 0.11\%$ for adult and $\pm 0.38\%$ for child test sets. Structured SGMM outperforms SWGMM and does not require an additional classification pass in decoding (1 pass decoding).

7. Conclusion and future work

In this paper, an efficient class-structured parameterization of HMM was proposed. The structuring consists in associating the subsets of Gaussian components with given speaker classes. Two class-structured models were investigated and demonstrated significant improvements of the ASR accuracy.

The first model uses Speaker class-dependent Weights (SWGMM). Unlike full class model adaptation and because of parameter sharing, the performance does not degrade, when the number of classes increases (which decreases the amount of training data for each class). Similar idea of class structuring was investigated for Stranded GMM, an explicit trajectory model with additional dependencies between the components of the observation densities. Structured SGMM is initialized from SWGMM. SSGMM provides promising results and, importantly, does not require the classification algorithm before the utterance decoding.

Future work must be done to improve SSGMM for dealing with real-world data (like NEOLOGOS). In particular, a specific processing should be applied for the fillers. It might be useful to exclude filler units from the clustering procedure and from SGMM component score propagation.

8. References

[1] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

[2] R. M. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 34–43, 2012.

[3] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[4] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *Speech and audio processing, IEEE transactions on*, vol. 2, no. 2, pp. 291–298, 1994.

[5] P. Zhan and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," in *Technical report*, DTIC Document, 1997.

[6] F. Beaufays, V. Vanhoucke, and B. Strope, "Unsupervised Discovery and Training of Maximally Dissimilar Cluster Models," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 66–69.

[7] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation." in *Proc. ICSLP*, vol. 98, 1998, pp. 1774–1777.

[8] D. Jouvet, A. Gorin, and N. Vinuesa, "Exploitation d'une marge de tolérance de classification pour améliorer l'apprentissage de modèles acoustiques de classes en reconnaissance de la parole," in *JEP-TALN-RECITAL*, 2012, pp. 763–770.

[9] A. Gorin and D. Jouvet, "Class-based speech recognition using a maximum dissimilarity criterion and a tolerance classification margin," in *Proc. Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 91–96.

[10] ——, "Efficient constrained parametrization of GMM with class-based mixture weights for Automatic Speech Recognition," in *Proc. LTC-6th Language & Technologies Conference*, 2013, pp. 550–554.

[11] C. J. Wellekens, "Explicit time correlation in hidden Markov models for speech recognition," in *Proc. ICASSP*, 1987, pp. 384–386.

[12] Y. Zhao and B.-H. Juang, "Stranded Gaussian mixture hidden Markov models for robust speech recognition," in *Proc. ICASSP*, 2012, p. 4301–4304.

[13] A. Gorin, D. Jouvet, E. Vincent, and D. Tran, "Investigating Stranded GMM for Improving Automatic Speech Recognition," in *HSCMA (to appear)*, 2014.

[14] R. G. Leonard and G. Doddington, "Tidigits speech corpus," *Texas Instruments, Inc*, 1993.

[15] CMU, "Sphinx toolkit http://cmusphinx.sourceforge.net," 2014.

[16] D. C. Burnett and M. Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in *Proc. ICSLP*, vol. 2. IEEE, 1996, pp. 1145–1148.

[17] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[18] D. Charlet, S. Krstulovic, F. Bimbot, O. Boëffard, D. Fohr, O. Mella, F. Korkmazsky, D. Mostefa, K. Choukri, A. Vallée *et al.*, "Neologos: an optimized database for the development of new speech processing algorithms." in *INTERSPEECH*, 2005, pp. 1549–1552.

[19] T. Wenxuan, G. Gravier, F. Bimbot, and F. Soufflet, "Rapid speaker adaptation by reference model interpolation," in *INTERSPEECH*, 2007, pp. 258–261.