

Structured GMM Based on Unsupervised Clustering for Recognizing Adult and Child Speech

Arseniy Gorin, Denis Jouvet

▶ **To cite this version:**

Arseniy Gorin, Denis Jouvet. Structured GMM Based on Unsupervised Clustering for Recognizing Adult and Child Speech. SLSP 2014, 2nd International Conference on Statistical Language and Speech Processing, Oct 2014, Grenoble, France. pp.108 - 119, 10.1007/978-3-319-11397-5_8 . hal-01090472

HAL Id: hal-01090472

https://inria.hal.science/hal-01090472

Submitted on 3 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Structured GMM based on unsupervised clustering for recognizing adult and child speech

Arseniy Gorin and Denis Jouvet

Speech Group, LORIA
Inria, 615 rue du Jardin Botanique, F-54600, Villers-lès-Nancy, France
Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
{arseniy.gorin, denis.jouvet}@inria.fr

Abstract. Speaker variability is a well-known problem of state-of-the-art Automatic Speech Recognition (ASR) systems. In particular, handling children speech is challenging because of substantial differences in pronunciation of the speech units between adult and child speakers. To build accurate ASR systems for all types of speakers Hidden Markov Models with Gaussian Mixture Densities were intensively used in combination with model adaptation techniques.
This paper compares different ways to improve the recognition of children speech and describes a novel approach relying on Class-Structured Gaussian Mixture Model (GMM).
A common solution for reducing the speaker variability relies on gender and age adaptation. First, it is proposed to replace gender and age by unsupervised clustering. Speaker classes are first used for adaptation of the conventional HMM. Second, speaker classes are used for initializing structured GMM, where the components of Gaussian densities are structured with respect to the speaker classes. In a first approach mixture weights of the structured GMM are set dependent on the speaker class. In a second approach the mixture weights are replaced by explicit dependencies between Gaussian components of mixture densities (as in stranded GMMs, but here the GMMs are class-structured).
The different approaches are evaluated and compared on the TIDIGITS task. The best improvement is achieved when structured GMM is combined with feature adaptation.

Keywords: speech recognition, unsupervised clustering, speaker class modeling, stochastic trajectory modeling

1 Introduction

Hidden Markov Models with Gaussian Mixture observation densities (HMM-GMM) are successfully applied in automatic speech recognition systems, despite their inability to accurately model the dynamic properties of speech coming from different speakers and recording conditions. The accuracy is usually improved by applying various tuning techniques and more advanced feature processing.

Children speech is a good example of the data that is hard to recognize with conventional HMM-GMM because of the variability of the acoustic features of the same phonetic units spoken by adult and child speakers. Such variability comes from the differences in the size of the vocal tract and mispronunciation of certain phones by children [2, 14]. For example, children have shorter vocal tract, than adults, which leads to higher F0 (fundamental frequency) [12].

The task becomes more complicated as the amount of available annotated children speech is not large enough for training separate models for children data. Also, frequently, the information about speaker age is available neither for test, nor for training data.

An effective strategy for handling child speech (or speaker variability in general) consists in adapting the ASR systems. These techniques either modify the acoustic features (VTLN [17], fMLLR [5]), or the model parameters (MLLR, MAP [6]) to maximize the likelihood of the adaptation data. A review paper [13] discusses various improvements and applications of VTLN-based algorithms for improving automatic recognition of children speech.

The conventional approach for handling speaker variability assumes age and gender known at least for the training data. In this case separate models are constructed for different age and gender classes by adapting the Speaker-Independent (SI) model trained on the full training dataset. In decoding the corresponding model is selected for each utterance based on knowledge of the speaker age and gender (if available), or on an automatic classification. A different approach relying on interpolation of several models was proposed in [16] and demonstrated significant improvements also on children speech data.

The main part of this work focuses on the general situation, when the dataset contains speakers of different age and gender, but the speaker age and gender are known neither for testing, nor for training. In such case unsupervised clustering is applied at the utterance level, assuming that the speaker class is not changing within the sentence [1]. Increasing the number of classes decreases the number of available training 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 [10], or by explicitly enlarging the class data by allowing one utterance to belong to several classes [9, 7].

Furthermore, a novel approach is proposed in this work for using speaker classes to structure an HMM-GMM. The idea is to include the speaker class information into the structure of a single HMM-GMM instead of building separate models for each class. To do this, the components of GMMs are composed from GMMs with a smaller number of components per density and trained (or adapted) on class data. 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 in contrast to conventional GMM, where the components are trained independently.

When the components are structured, the speaker class is represented as a subspace of the structured GMM (k^{th} component, or subset of components

of each GMM corresponds to k^{th} speaker class). To select the corresponding subspace, additional modifications are proposed in the form of dependencies added on weights of the Gaussian components.

Class-structured GMM was first used with mixture Weights dependent on the speaker class in addition to the associated HMM state. Such a model with Speaker class-dependent Weights (SWGMM) was originally investigated in a radio broadcast transcription system [8]. 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 state and class-dependent mixture weights by only state-dependent Mixture Transition Matrices (MTMs) of Stranded Gaussian Mixture Model (SGMM). SGMM is similar to conditional Gaussian model [15], which was recently extended, re-formulated and investigated for robust ASR [18]. In SGMM the Mixture Transition Matrix (MTM) defines the dependencies between the components of adjacent Gaussian mixture observation densities.

In [18] it was originally proposed to initialize SGMM from the conventional HMM-GMM. Instead, here, 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. MTM in SSGMM is used to model the probabilities of either keeping the same component (speaker class) over time, or to dynamically switch between dominating components (classes).

The advantage of using explicit component dependencies over class-dependent mixture weights is that the weights are no more fixed at the utterance level (determined by the speaker class), but rather change depending on the observation from the previous frame. As a result, 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 system and discusses the conventional adaptation-based approach. Section 3 discusses unsupervised class-based-adaptation approach for ASR (CA-GMM). Section 4 introduces class-structured GMM with Speaker class-dependent Weights (SWGMM) and describes the corresponding experiments. Section 5 recaps Stranded GMM (SGMM) framework, describes the initialization of the class-Structured SGMM (SSGMM) from SWGMM and explains the corresponding experiments. The paper ends with conclusion and future work.

2 Adaptation for handling age and gender variability

The section describes conventional approaches based on gender and age adaptation with MLLR, MAP and VTLN. Unlike the main objective of the work (use no prior information about speakers), within this section the speaker classes (adult/child and male/female) are assumed to be known for the training data.

2.1 TIDIGITS baselines

The experiments in this paper are conducted on the TIDIGITS connected digits task [11]. 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). Similarly to other work with TIDIGITS [3] the signal is down sampled to 8 kHz in order to roughly model the telephone-quality data.

The Sphinx3 toolkit [4] is used for modeling. 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 computes 13 standard MFCC (12 cepstral + log energy) plus the first and second derivatives and a cepstral mean normalization (CMN) is applied.

Two speaker-independent (SI) models are trained from the adult subset only and from the full training set. The corresponding Word Error Rates (WER) for baseline models are shown in Table 1

	Adult	**Child**
Training on adult data	**0.64**	**9.92**
Training on adult+child data	**1.66**	**1.88**

Table 1: Baseline WERs on TIDIGITS data

Training on adult data provides the best results for adult speakers, but shows a weak performance on the child subset. When child data are included in the training set, the conventional HMM-GMM improves on child, but degrades on adult subset.

2.2 Model adaptation

Better baselines are achieved when age-gender classes are used for adapting the SI baselines with MLLR for GMM mean values followed by MAP for all model parameters.

With class-based modeling, decoding is usually done in 2 passes. In the 1st pass, for each utterance, the corresponding class is determined using a GMM classifier trained on age-gender labels of the training data. In 2nd pass the standard decoding is done with the corresponding class-based model.

In addition, the recognition hypothesis can be used for applying rapid adaptation of the features (VTLN) using only the utterance data. After such VTLN-based feature transformation a 3rd pass decoding is done.

Word Error Rates for baselines, 2-pass and 3-pass decoding of TIDIGITS data are summarized in Table 2.

Although for SI baseline using all data in training provides better results, the adaptation is more efficient when initial SI model is trained on adult data. In all cases additional VTLN pass in decoding further improves the model accuracy.

	Decoding	Adaptation in decoding	WER	
			Adult	**Child**
Training on adult data	1 pass	–	**0.64**	**9.92**
+Gender-Age adaptation	2 pass	–	0.54	1.08
+Utterance Rapid adaptation	3 pass	VTLN	0.54	0.97
Training on adult+child data	1 pass	–	**1.66**	**1.88**
+Gender-Age adaptation	2 pass	–	1.34	1.45
+Utterance Rapid adaptation	3 pass	VTLN	1.29	1.41

Table 2: Baseline WERs for SI and Gender-Age adapted models

3 Unsupervised clustering for multi-model ASR

Let us consider a set of training utterances without any knowledge about the speaker identity or class (age, gender, etc.). The objective is to automatically group the training data into classes of acoustically similar data.

A GMM-based utterance clustering algorithm is applied [9]. In this approach, a single GMM with a large number of components is first trained on the full dataset. Then, the GMM is duplicated and the mean values are perturbed. Next, the data are classified with Maximum Likelihood criterion and the GMMs are trained from the corresponding classes. The classification and training steps are repeated until convergence. This split-classification-training process is repeated until the desired number of classes is achieved. The class data are then used for adapting the SI HMM-GMM model parameters. The same classification GMMs are used in decoding to identify the class for selecting the best model for each utterance of the test set.

Although clustering of the utterances is not exactly equivalent to speaker clustering, here and later we assume that the main source of variability comes from the speaker and we will refer to the described process as speaker clustering and to the resulting classes of utterances as speaker classes.

Analyzing data clustering for mixed adult-child data

This unsupervised clustering is applied on the TIDIGITS train data. The classification GMMs consist of 256 components. The corresponding distributions of Age-Gender over these classes are summarized in Figure 1.

The first clustering step (2 classes) mainly splits male speakers from female and child speakers. The second split (4 classes) allows to separate female speakers from child speakers. It seems impossible to distinguish boys from girls, even with more classes.

After clustering, the SI acoustic model (32 Gaussian per density) trained on full train data (adult and child) is adapted using each class data with MLLR+MAP. The bars "*CA-GMM*" in Figure 4 illustrate WERs with the associated 95% confidence intervals. The best result is achieved with 4 classes, for which the WER (see details in the "*4 classes CA-GMM*" row of the table 3) is similar to the supervised Gender-Age adaptation of the mixed Adult-Child SI model results (see Table 1). After 4 classes, the performance degrades, because there is not enough data to adapt the class-based models.

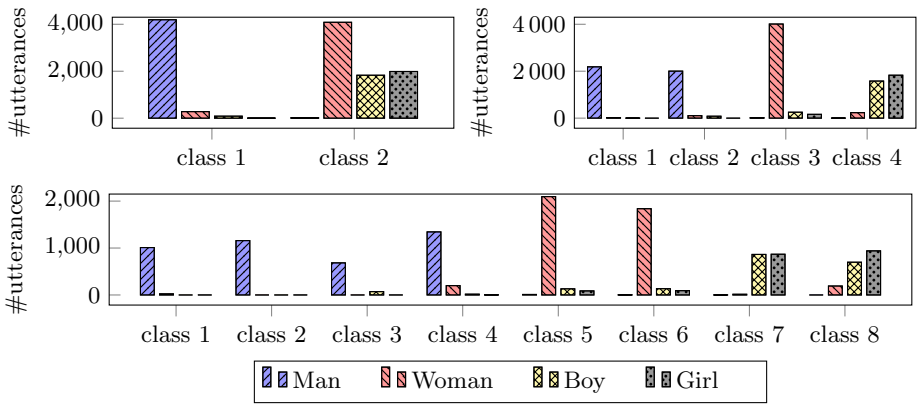


Fig. 1: Number of training utterances for each Age-Gender in the resulting 2, 4 and 8 classes

4 Class-structured GMM with Class-Dependent Weights

Instead of adapting all GMM parameters for each class of data, a more efficient and compact parameterization was investigated: structured GMM with Speaker class-dependent Weights (SWGMM) [8]. GMM components of this model are shared and structured with respect to speaker classes and only the mixture weights are class-dependent.

The SWGMM pdf for an HMM state j and a given speaker class c has the following form:

$$b_j^{(c)}(\boldsymbol{o}_t) = \sum_{k=1}^{M} w_{jk}^{(c)} \mathcal{N}(\boldsymbol{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{U}_{jk}) \tag{1}$$

where M is the number of components per mixture, $\boldsymbol{o}_t$ is the observation vector at time t and $\mathcal{N}(\boldsymbol{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{U}_{jk})$ is the Gaussian pdf with the mean vector $\boldsymbol{\mu}_{jk}$ and the covariance matrix $\boldsymbol{U}_{jk}$.

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

The class structuring consists in concatenating the components of GMMs of smaller dimensionality, separately trained from different classes. 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 merged into a single mixture as follows:

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

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 means and variances:

$$\omega_{jk}^{(c_i)} = \frac{\sum_{t=1}^{T} \gamma_{jk}^{(c_i)}(t)}{\sum_{t=1}^{T} \sum_{l=1}^{M} \gamma_{jl}^{(c_i)}(t)} \qquad \boldsymbol{\mu}_{jk} = \frac{\sum_{i=1}^{Z} \sum_{t=1}^{T} \gamma_{jk}^{(c_i)}(t) \boldsymbol{o}_t}{\sum_{i=1}^{Z} \sum_{t=1}^{T} \gamma_{jk}^{(c_i)}(t)} \tag{2}$$

where $\gamma_{jk}^{(c_i)}(t)$ is the Baum-Welch count of the k^{th} component of the state j, generating the observation $\boldsymbol{o}_t$ from the class c_i. Summation over t means summation over all frames of all training utterances of the class. The 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 of structured GMM, averaged over HMM states, for classes c_7, c_{17} and c_{27}).

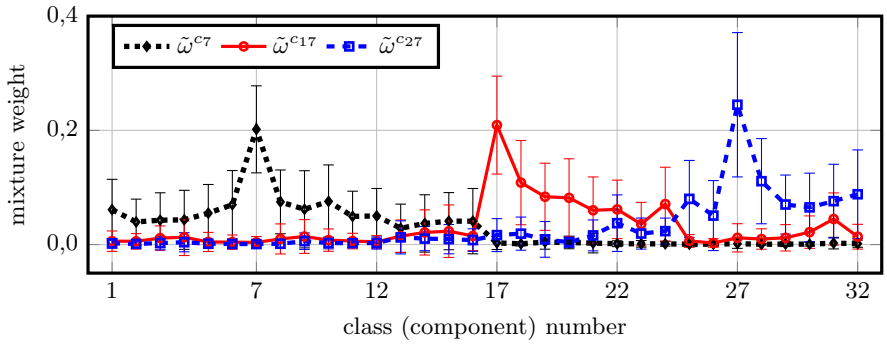


Fig. 2: Example of class-dependent mixture weights of structured GMM after joint re-estimation. Here mixture weights are averaged over HMM states with corresponding standard deviation in bars (here Z=32, M=32)

Experiments with class-structured SWGMM

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

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 *8* and *32 classes SWGMM* rows in Table 3).

5 Class-Structured Stranded Gaussian Mixture Model

Stranded GMM was proposed [18] in the robust ASR framework. The corresponding extended training and decoding algorithms were also introduced in the original paper. This model expands the observation densities of HMM-GMM and explicitly adds dependencies between GMM components of the adjacent states.

Originally, an SGMM is initialized from an HMM-GMM. In this section after briefly recalling the conventional Stranded GMM approach, a class-Structured SGMM (SSGMM) is proposed.

5.1 Conventional Stranded GMM

The conventional SGMM consists of the state sequence $\boldsymbol{\mathcal{Q}} = \{q_1, ..., q_T\}$, the observation sequence $\boldsymbol{\mathcal{O}} = \{\boldsymbol{o}_1, ..., \boldsymbol{o}_T\}$, and the sequence of components of the observation density $\boldsymbol{\mathcal{M}} = \{m_1, ..., m_T\}$, 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.

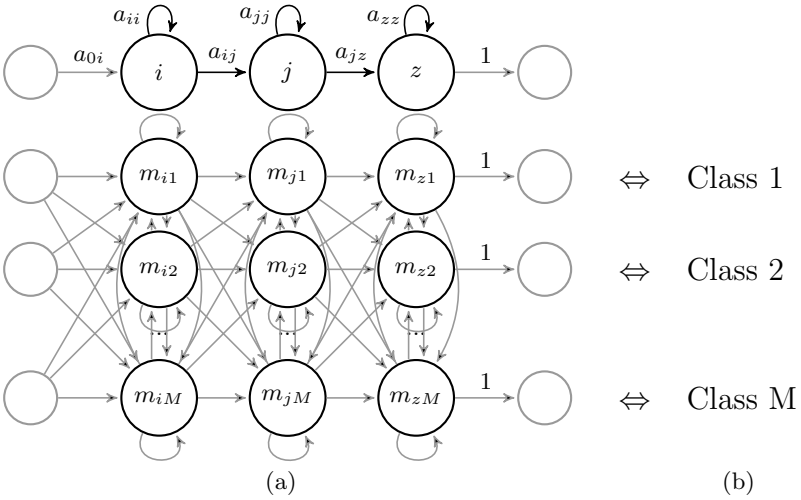


Fig. 3: (a) Stranded GMM with schematic representation of the component dependencies; (b) the idea of Structured SGMM, i.e., associating each k^{th} component with some class of data

The difference of SGMM from 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:

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

where $P(q_t = j|q_{t-1} = i) = a_{ij}$ is the state transition probability, $P(\boldsymbol{o}_t|m_t = l, q_t = j) = b_{jl}(\boldsymbol{o}_t)$ is the the probability of the observation $\boldsymbol{o}_t$ 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 mixture transition probability.

The set of component transition probabilities corresponds to the mixture transition matrices (MTMs) $C^{(ij)} = \{c_{kl}^{(ij)}\}$, where $\sum_{l=1}^{M} c_{kl}^{(ij)} = 1, \forall i, j, k$.

Experiments with conventional SGMM

In conventional SGMM, MTM rows are initialized from the mixture weights of convention HMM-GMM, and the model parameters are re-estimated with MLE. Such initialization and training processes are applied in this section. 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 3.

Compared to the conventional HMM-GMM trained on all data (adult+child), 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 re-estimated SWGMM, described in Section 4. The means and variances are taken from SWGMM and MTMs are defined with uniform probabilities. The 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 allows to automatically switch between different components (speaker classes). Therefore, the classification algorithm is no more needed in decoding.

Experiments with class-Structured Stranded GMM

In the experimental study, 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 classes SWGMM* in Table 3). Two MTMs per states are defined 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 3. Initializing SSGMM from SWGMM

with different number of classes (2, 4, 8 and 16) was always leading to accuracy improvement, compared to SGMM. Only the best result, corresponding to 32 classes, is reported.

While conventional SGMM improves from 1.66% to 1.11% on adult and from 1.88% to 1,27% on child data, compared to the SI GMM trained on full train data (adult+child), the proposed Class-Structured SGMM (*SSGMM*) further improves by achieving 0.52% WER on adult and 0.86% on child data.

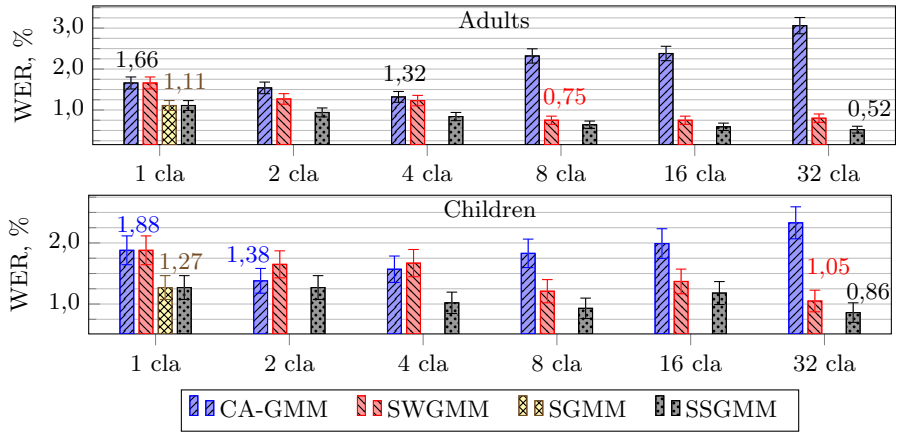


Fig. 4: WER for adult (top) and child (bottom) sets, computed with full Class-Adapted model (CA-GMM), class-structured GMM with Speaker-class dependent Weights (SWGMM), conventional Stranded GMM and class-Structured Stranded GMM built from 32 classes (SSGMM)

The key improvements from all proposed techniques are summarized in Table 3. Notice, that SSGMM can be further combined with rapid feature adaptation to further slightly improve the recognition result on child data (see row *SSGMM+VTLN*).

Model	Decoding	Parameters/state	**Adult**	**Child**
SI GMM	1 pass	78*32+32=2528	**1.66**	**1.88**
4 classes CA-GMM	2 pass	4*(78*32+32)=10112	1.32	1.57
8 classes SWGMM	2 pass	78*32+8*32=2752	0.75	1.21
32 classes SWGMM	2 pass	78*32+32*32=3520	0.80	1.05
SGMM	1 pass	78*32+2*32*32=4544	1.11	1.27
SSGMM	1 pass	78*32+2*32*32=4544	**0.52**	**0.86**
SSGMM+VTLN	2 pass	78*32+2*32*32=4544	**0.52**	**0.81**

Table 3: Summary of the best results and the number of model parameters. Compared the baseline (SI GMM), 4 full Class-Adapted model (CA-GMM), 8 and 32 class-structured GMM with Speaker-class dependent Weights (SWGMM), conventional Stranded GMM and class-Structured Stranded GMM built from 32 classes (SSGMM) without and with additional VTLN pass in decoding)

6 Conclusion and future work

This paper investigated an efficient unsupervised approach for handling heterogeneous speech data without prior knowledge about speaker age and gender. Unsupervised clustering does not allow to build many speaker class models, when the amount of training data is limited. To address this problem, an efficient class-structured parameterization of GMM components has been proposed.

The structuring consists in associating subsets of Gaussian components with given speaker classes. Two models, which include this class-structured parameterization, have been investigated and lead to significant improvements of the ASR accuracy.

The first model uses Speaker class-dependent Weights (SWGMM). Unlike standard class model adaptation, the performance does not degrade, when the number of classes increases and when the number of class-associated data decreases. The class structuring approach was also applied for Stranded GMM - an explicit trajectory model with additional dependencies between the components of the observation densities. Class-Structured SGMM is initialized from SWGMM, in which Gaussian components are structured with respect to speaker classes. Mixture Transition Matrices (MTMs) were then used to replace class-dependent mixture weights and to model dependencies between components (speaker classes). SSGMM provides very promising results for both child and adult data. Moreover, it does not require classification algorithm before utterance decoding. SSGMM combined with VTLN achieves overall best performance, outperforming even the strong 3-pass MLLR+MAP age-gender adapted baseline with VTLN pass in decoding.

In the future the proposed techniques should be applied for large vocabulary speech recognition task including adult and child speakers.

References

1. Beaufays, F., Vanhoucke, V., Strope, B.: Unsupervised Discovery and Training of Maximally Dissimilar Cluster Models. In: Proc. INTERSPEECH. pp. 66–69. Makuhari, Japan (2010), http://www.isca-speech.org/archive/interspeech_2004/i04_0377.html
2. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Tyagi, V., Wellekens, C.: Automatic speech recognition and speech variability: A review. Speech Communication 49(10), 763–786 (2007)
3. Burnett, D.C., Fanty, M.: Rapid unsupervised adaptation to children's speech on a connected-digit task. In: Proc. ICSLP. vol. 2, pp. 1145–1148. IEEE (1996)
4. CMU: Sphinx toolkit http://cmusphinx.sourceforge.net (2014)
5. Gales, M.J.: Maximum likelihood linear transformations for HMM-based speech recognition. Computer speech & language 12(2), 75–98 (1998)
6. Gauvain, J.L., Lee, C.H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. Speech and audio processing, IEEE transactions on 2(2), 291–298 (1994)

7. Gorin, A., Jouvet, D.: Class-based speech recognition using a maximum dissimilarity criterion and a tolerance classification margin. In: Proc. Spoken Language Technology Workshop (SLT), 2012 IEEE. pp. 91–96. IEEE (2012)
8. Gorin, A., Jouvet, D.: Efficient constrained parametrization of GMM with class-based mixture weights for Automatic Speech Recognition. In: Proc. LTC-6th Language & Technologies Conference. pp. 550–554 (2013)
9. Jouvet, D., Gorin, A., Vinuesa, N.: 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. pp. 763–770 (2012)
10. Kuhn, R., Nguyen, P., Junqua, J.C., Goldwasser, L., Niedzielski, N., Fincke, S., Field, K., Contolini, M.: Eigenvoices for speaker adaptation. In: Proc. ICSLP. vol. 98, pp. 1774–1777 (1998)
11. Leonard, R.G., Doddington, G.: Tidigits speech corpus. Texas Instruments, Inc (1993)
12. O'Shaughnessy, D.: Acoustic analysis for automatic speech recognition. Proceedings of the IEEE 101(5), 1038–1053 (2013)
13. Panchapagesan, S., Alwan, A.: Frequency warping for vtln and speaker adaptation by linear transformation of standard mfcc. Computer speech & language 23(1), 42–64 (2009)
14. Stern, R.M., Morgan, N.: Hearing is believing: Biologically inspired methods for robust automatic speech recognition. Signal Processing Magazine, IEEE 29(6), 34–43 (2012), http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6296528
15. Wellekens, C.J.: Explicit time correlation in hidden Markov models for speech recognition. In: Proc. ICASSP. pp. 384–386 (1987)
16. Wenxuan, T., Gravier, G., Bimbot, F., Soufflet, F.: Rapid speaker adaptation by reference model interpolation. In: Proc. INTERSPEECH. pp. 258–261 (2007)
17. Zhan, P., Waibel, A.: Vocal tract length normalization for large vocabulary continuous speech recognition. In: Technical report. DTIC Document (1997)
18. Zhao, Y., Juang, B.H.: Stranded Gaussian mixture hidden Markov models for robust speech recognition. In: Proc. ICASSP. p. 4301–4304 (2012)