Two-Step Unsupervised Speaker Adaptation Based on Speaker and Gender Recognition and HMM Combination

Petr Cerva, Jan Nouza and Jan Silovsky

SpechLab, Department of Electronics and Signal Processing
Technical University of Liberec, Halkova 6
461 17 Liberec, Czech Republic
{petr.cerva, jan.nouza, jan.silovsky}@vslib.cz

Abstract

In this paper, we present a new strategy for unsupervised speaker adaptation. In our approach, the adaptation is performed in two steps for each test utterance. In the first online step, we utilize speaker and gender identification, a set of speaker dependent (SD) hidden Markov models (HMMs) and our own fast linear model combination approach to create a proper model for the first speech recognition pass. After that the recognized phonetic transcription of the utterance is used for maximum likelihood (ML) estimation of more accurate weights for the final model combination step. Our experimental results on different types of broadcast programs show that the proposed method is capable to reduce the word error rate (WER) relatively by more than 17 %.

Index Terms: unsupervised speaker adaptation, HMM combination, speaker and gender recognition

1. Introduction

In recent years, a speaker adaptation method known as speaker selection training (SST) [1] has been developed. The main advantage of this framework consists in the fact that only a small amount of adaptation data is necessary to reduce the error rate of the system significantly. The SST is based on the usage of a set of SD models, which are created offline for a group of training speakers. The main idea of this method is to find a cohort of training speakers, who are close in the acoustic space to the test speaker, and to build the adapted model from models belonging to cohort speakers. These models can be created by speaker dependent training, when huge amount of training data is available, but speaker adaptation is often used for this purpose in practice.

The SST can be performed in several ways. The SD models of all speakers in the training set can be used for calculating of the likelihood of the testing utterance to find the cohort of the nearest speakers [1]. Alternatively, speaker identification based on Gaussian mixture models (GMMs) can be utilized for this purpose too [2]. After the cohorts are found, there exist also several possibilities how to create the adapted model for speech recognition. When at least some transcribed training data for cohort speakers are available, they can be transformed to better map the test speaker's acoustic space. The adapted model is then created by re-estimation [1] on these transformed data. Another possibility in this case is to determine the weighting coefficients for combination by MAP or ML estimation [3]. The model combination can be also performed online and in unsupervised way when sufficient HMM statistics are stored during the phase of SD models training [2].

In this paper, we propose a new unsupervised approach based on utilization and modification of several mentioned variants of the SST. The goal is to propose an effective method, which could be used mainly (but not only) in systems for transcription of various spoken data streams (e.g. broadcast news, parliament debates or commented sport games) because one of the most challenging problems in these systems is the acoustic variability across various speakers appearing in each stream and the fact that the speaker in each segment of the stream is unknown.

This paper is structured as follows: The next section is focused on the detailed description of the proposed method. In section 3 we evaluate it in experiments performed on the database of parliament debates and broadcast news. In the last section we discuss the results.

2. Description of the proposed speaker adaptation method

The proposed speaker adaptation method is illustrated in Fig.1. Its input is one utterance or single speaker speech segment that was automatically cut off from the given audio stream. The output is the acoustic model (a set of phoneme HMMs) which should fit best to the given test speaker. The whole framework works in two main steps.

First speaker and gender identification is performed to select a cohort of speakers (with the same gender as it was identified) from the training set, who are acoustically close to the given test speaker. After that the models of cohort speakers are combined to create the adapted model for the first speech recognition pass. In the second adaptation step, the recognized phonetic transcription of the utterance is used together with ML estimation to calculate more accurate weights for the next combination step. The final adapted model is then used in the second speech recognition pass.

The individual parts and steps of our approach are explained in detail in the following subsections.

2.1. Models of key speakers

For the set of training (key) speakers, SD HMMs and GMMs are prepared offline. The GMMs are trained in several iterations of standard ML estimation while the HMMs are created by MLLR [4] or MAP [5] based adaptation of mean vectors. For this adaptation, we use gender dependent (GD) models - rather than the general SI models – as prior sources. These GD models are trained offline in several iterations of the standard EM algorithm and they may have different numbers of Gaussian components (mixtures) due to this reason. This number depends for each acoustic model on the available amount of gender specific training data.

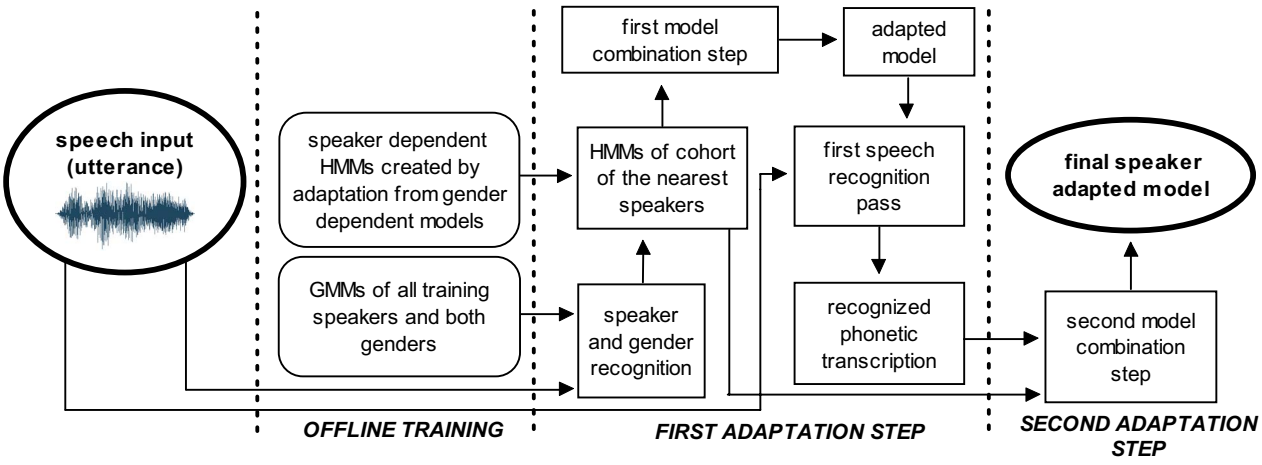

Figure 1: The goal of the proposed adaptation scheme is to compute the optimal acoustic model for each test utterance

2.2. Speaker and gender identification

The initial part of our framework is speaker and gender identification. For the given speech segment, a likelihood score is calculated for each of the key speakers represented by their GMMs (speaker identification step). The same is also done for GMMs of both genders. The gender with the higher value of likelihood then determines if the unknown test speaker is a male or a female.

2.3. Forming of cohorts of the nearest speakers

In the second step of the proposed scheme, we utilize the scores from speaker identification for forming the cohorts of nearest speakers. We chose N speakers with the highest scores and the same gender as it was identified in the previous step. The constraint on the same gender is not only natural but also practical because the two genders differ in the number of Gaussian components as explained in section 2.1.

2.4. The first model combination

In this step, there is not available any information about the phonetic transcription of the utterance so we can not use any classical estimation (like ML or MAP) to maximize some criteria. The adaptation can only be based on knowledge of prior information. Due to this reason, we use only simple but fast adaptation method based on model combination.

In the last few years, several methods have been proposed to estimate the weighting coefficients for model combination when any transcription of the data is not available. For example, one often used approach relies on occupation likelihoods of individual mixtures, which are collected for all training speakers during the phase of SD model training [2]. In our case, the deployment of this method is problematic due to two main reasons. First the SD models are created by adaptation on different amount of data, so the values of occupation likelihoods, which depend on the amount of used data too, are not comparable for different speakers. The second reason is that we perform only adaptation of means during the SD model building (because of the small amount of available speaker specific data), so any variances and weights of mixtures can not be combined because they are the same for all speakers.

In this paper, we propose another approach, which weights the models of key speakers according to their similarity in the acoustic space to the test speaker. It is only based on linear combination of mean vectors belonging to the N key speakers, who form the cohort. Variances and mixture weights are copied from the corresponding GD model (determined during gender identification) without any modification. This fact has two main advantages: a) the robustness of the adapted model is improved in comparison with other possible types of adaptation while b) the variances from the given GD model still perform significantly better than the original ones from the SI model.

Generally, the combined mean vector $\boldsymbol{\mu}_m$ of the m-th Gaussian component can be expressed as

$$\boldsymbol{\mu}_m = \mathbf{M}_m \boldsymbol{\lambda}, \tag{1}$$

where $\mathbf{M}_m = [\boldsymbol{\mu}_m^1, \boldsymbol{\mu}_m^1, ..., \boldsymbol{\mu}_m^N]$ is the matrix of N cohort speakers, $\boldsymbol{\mu}_m^n$ is the m-th mean vector belonging to the n-th cohort speaker and $\boldsymbol{\lambda}$ is the estimated vector of weights.

In our approach, first the speakers in the cohort are sorted in ascendant order according to their likelihood scores obtained in speaker identification. Then only one global weight is calculated for the n-th speaker as

$$\lambda^n = n / \sum_{j=1}^{N} j \tag{2}$$

The equation (2) ensures that the mean vectors of the nearest speaker will have N-times higher weight than those of the most distant speaker and also that $\sum_{n=1}^{N} \lambda^n = 1$ and $\lambda^n \geq 0$ for all n. After this online model combination step, the adapted model is used in the first speech recognition pass to create the phonetic transcription of the given utterance.

2.5. The second model combination step

In the second adaptation step, we utilize the recognized phonetic transcription to estimate more accurate weights for final combination of means. Variances and mixture weights are again copied from the corresponding GD model, because the adaptation is again unsupervised and we do not have

enough accurate transcription to update them. This time, we do not estimate only one global weigh, but all Gaussian components of all acoustic models are split into a binary regression tree. During adaptation, the tree is searched down from the root towards the leaves while calculating the vector of weights only for those nodes where sufficient amount of the adaptation data is available. In this case, we use ML estimation to find $\boldsymbol{\lambda}$. Given the sequence of adaptation observations $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T]$, where T is the number of frames, and assuming that all observations are independent, the goal is to find $\boldsymbol{\lambda}$ according to

$$\arg\max_{\boldsymbol{\lambda}} \{\sum_{m=1}^{M}\sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \boldsymbol{\lambda})\} \quad (3)$$

where M is the number of all Gaussian components associated with observation $\mathbf{x}_t$. The complex solution of (3) can be found in [3] for example.

3. Experimental evaluation

3.1. Testing database

The evaluation was done on several types of broadcast programs. First we used 2 hours long stream of parliament debates (13,624 words) as the development set to perform detailed experiments focused on the first and second adaptation step (section 3.4 and 3.5). These debates were split manually into 225 segments, each containing a single speaker's utterance. Then we used broadcast news data prepared within the European COST278 project [6] to show the total performance of our approach on other types of data too (section 3.6). The news shows were recorded from radio (16,677 words, 2 hours) and TV (29,887 words, 3 hours) and they were again manually segmented into parts belonging mostly to a single speaker.

3.2. Used speech recognition system

In all experiments, we employed our own transcription system [7]. Its core is formed by a LVCSR system operating with a vocabulary containing up to several hundred thousands of Czech words. For broadcast news, the size of the vocabulary was 312,490 items and the language model was based on smoothed bigrams estimated on a corpus compiled from about 2.6 GB of Czech (mainly newspaper) texts.

For parliament debates, this language model was modified by mixing of the previous corpora with a parliament specific data (0.18 GB of texts) and the most of words missing in the parliament corpus were removed from the general vocabulary (see [8] for details). Its size was then 154,463 items. The purpose of this step was to create the best possible language model and to provide similar baseline recognition accuracy for all types of testing data.

3.3. Acoustic models

We used models of 41 Czech phonemes and 7 types of noise. They were three-state context independent CDHMM with up to 100 mixtures per state. The feature vector was composed of 39 MFCC parameters (13 static coefficients and their first and second derivatives). For speaker recognition, we used GMM with 256 mixtures and just 12'th order MFCC (excluding the c_0 coefficient). The speech database used for acoustic model training contained 49 hours of speech recordings. These were a mix of microphone and broadcast signal. From some 1000 speakers occurring in the database, 190 women and 310 men were selected as training speakers. For these speakers, a varying amount of adaptation data (from 18 seconds to 25 minutes) was available.

We performed all experiments maximally objectively: when the speaking person in the given test segment was randomly one of those 500 key speakers, it was removed temporarily (just for adaptation of this segment) from the database of training speakers.

3.4. Results after the first adaptation step

The results from testing of several approaches in the first adaptation step are summarized in Tab. 1. They show WER (Word Error Rate) values vs. the number of cohort speakers. The three rows represent the following three strategies:

1. The framework described in this paper, i.e. gender identification and gender specific cohorts together with the model mixing scheme according to eq. 2.
2. Similar framework like above except the model combination part. Here, the mean combination is based on HMM statistic collected during the training phase as proposed in [2].
3. Cohorts formed from male and female speakers, their models created by adaptation of SI models, HMM statistics used for combination of means.

Table 1: Values of WER [%] for different approaches in the first adaptation step

speaker independent models : WER = 26.80							
gender dependent models: WER = 24.75 (just gender ident. was done without any adaptation)							
value of *N*	**5**	**25**	**50**	**75**	**100**	**150**	**190**
proposed method	24.45	23.77	23.83	24.10	24.02	24.19	24.39
2. strategy	24.78	24.39	24.56	24.67	24.74	24.48	24.61
3. strategy	25.61	25.51	25.50	25.81	26.13	26.02	26.09

From Table 1., it is evident that the biggest improvement of the WER was reached by the use of GD models (from 26.80% to 24.75%). The combination of mean vectors added smaller, but still statistically significant improvement – mainly for values of N about 25. The above results also demonstrate that the recognition accuracy can only be improved significantly in the case when cohorts are formed from SA models created by adaptation from GD models (the first and second row). In the third scheme, the WER reduction was only negligible. When comparing the first and second row we can observe that the proposed combination of means leads to slightly better improvement than the combination based on HMM statistics. The advantage of our approach is not only this small improvement but namely in the simplicity and speed of the adaptation. Only one global weight is used for all mixtures of each speaker and no HMM statistics must be stored. We also tried to use some other simple approaches for combination of means (like the same weight for all

speakers), but the results were always worse than the proposed approach based on the similarity in the acoustic space between the test speaker and cohort speakers.

3.5. Results after the second adaptation step

The next experiment (Tab. 2.) is focused on the second adaptation step. The recognized phonetic transcription of each segment was used a) to calculate more accurate weights for model combination as described in section 2.5 and b) for unsupervised adaptation based on MLLR. In the former case, the value of *N* was always the same as in the first step. In the latter one, the adapted model from the first step was used as a prior source for MLLR. After both types of adaptation, the final adapted model was used for the second speech recognition pass.

Table 2: Values of WER [%] after the second adaptation step

value of *N*	5	25	50	75	100	150	190
MLLR method	23.81	23.10	23.03	23.26	23.16	23.29	23.33
ML based model combination	23.39	22.45	21.78	21.58	21.49	21.02	20.83

The results of this experiment show that the model combination method performs for unsupervised adaptation better than the MLLR method. While the improvement against the first pass (the first row of Tab. 1.) reached by MLLR is logically near the same for all values of *N*, the improvement reached by model combination raises with increasing value of *N*. It is because in this case, a lot of cohort speakers are selected from the database while their weights are set dynamically by ML estimation. The model combination approach is also faster than MLLR due to its lower computation complexity.

3.6. Total results on different types of data

In the previous two subsections we focused on the detailed evaluation of both adaptation steps. Now, let us consider the performance of the complete framework. In Tab. 3., we present the total results achieved not only on the development set (parliament debates), but also on other types of broadcast programs (as described in section 3.1).

We used the best settings obtained on the development set: the value of *N* in the first step and second step was set on 25 and 190 respectively. Yet, in all tasks we can see a significant improvement. The WER values were reduced relatively by more than 17 % in all tasks.

Table 3: The WER [%] for different tasks after the application of the whole framework

program	SI models	SA models	rel. reduction of WER [%]
radio news	19.45	15.03	22.7
TV news	22.96	19.04	17.0
parliament debates	26.80	20.74	22.6

4. Conclusions

In this paper, we proposed a new two step unsupervised adaptation strategy that is suitable for speech recognition tasks where speakers change frequently. This happens namely in broadcast news, parliament debates, talk-shows, etc. The performance of our framework was tested in several different tasks. In all of them we could report significantly improved results. On the contrary the total computation time needed for our whole two-step framework was more than two times higher than the time needed by the one pass baseline speaker independent system (approximately 3x real time). It is because in our approach, a little time is necessary for the speaker recognition, estimation of parameters from the recognized phonetic transcription and ML based combination of means too. In the recent version of our transcription system [9], this problem is solved easily by the use of parallel system architecture: the system employs several recognition servers to transcribe each segment of the given stream separately. Our future work will be also focused on further improvement of the computation speed of individual steps.

In should be also noted that all the described experiments were done with data streams that were manually split into acoustically homogenous segments. This manual segmentation was necessary if we wanted to evaluate each of the framework steps separately. For practical implementation in the complete transcription system, an automatic segmentation algorithm is used.

5. Acknowledgements

This work was partly supported by project 1QS108040569 of the Grant Agency of the Czech Academy of Science.

6. References

[1] M. Padmanabhan, L. Bahl, D. Nahamoo and M. Picheny, "Speaker Clustering and Transformation for Speaker Adaptation in Speech Recognition Systems," *IEEE Transactions on Speech and Audio Processing*, vol. 6, n1, pp. 71-77, 1998.

[2] S. Yoshizawa, A. Baba, K. Matsunami et al, "Evaluation on Unsupervised Speaker Adaptation Based on Sufficient HMM Statistics of Selected Speakers," *Proc. of Eurospeech2001*, vol. 2, pp.1219-1222, 2001.

[3] C. Huang, T. Chen and E. Chang, "Adaptive model Combination for dynamic speaker selection training," *in Proc. ICSLP2002*, vol. 1, pp. 65-68, 2002.

[4] Leggetter, C.J. & Woodland, P.C., "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression", *Proc. ARPA Spoken Language Technology Workshop*, pp. 104-109, 1995, Morgan Kaufmann.

[5] Gauvain, J.L., Lee, C.H., "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Trans. SAP*, Vol. 2, pp. 291-298, 1994.

[6] Vandecatseye A. et al, "The COST278 pan-European Broadcast News Database", *Proc. of LREC 2004*, Lisbon.

[7] Nouza, J., Zdansky, J., Cerva, P., Kolorenc, J., "Continual On-line Monitoring of Czech Spoken Broadcast Programs", *Proc. of Interspeech2001.*

[8] Nouza, J., Cerva, P., Zdansky, J., Kolorenc, J., David, P.: Towards automatic transcription of parliament speech. In: Electronic Speech Signal Processing 2005, pp. 237-244, Prague, Czech Republic,