TRAINING OF SPEAKER-CLUSTERED DISCRIMINATIVE ACOUSTIC MODELS FOR USE IN REAL-TIME RECOGNIZERS

J. Vaněk, J.V.Psutka, J. Zelinka and J. Trmal

Department of Cybernetics, Faculty of Applied Sciences, Univerzity of West Bohemia
Univerzitní 22, 301 00 Pilsen, Czech Republic
vanekyj, psutka_j, zelinka, jtrmal@kky.zcu.cz

Abstract
It is well known that gender-dependent (male/female) acoustic models are more acoustically homogeneous and therefore give better recognition performance than single gender-independent model in the case where the gender is successfully detected or a priory known. Speakers do not need to be split to two groups only. An algorithm to make higher number of speaker clusters is described in this paper. Further, the paper deals with a problem how to use these gender-based or speaker-clustered acoustic models in a real-time LVCSR where information from an automatic cluster detector is often delayed or incorrect. Moreover, various ways, how to incorporate discriminative training methods into training of the speaker-clustered acoustic models, are discussed in the paper.

1 Introduction

One of the most important problems of speaker-independent LVCSR systems is their worse ability to get over the inter-speaker variability. This problem becomes serious if the recognizer works in real time and in tasks where speakers change frequently. Such task is e.g. the on-line closed captioning of Parliament meetings - the task which is experimentally tested by the Czech TV since 11/2008 (experimental broadcasting). One of the ways how to handle this problem is the incremental speaker adaptation or using gender-dependent acoustic models (AMs) or even models obtained from more detailed clustered voices.

This paper describes our experiments with unsupervised speaker clustering followed by discriminative adaptation (DT) that seeks such solution (such acoustic models) which yield on one hand favorable quality (increased accuracy) of discriminative training, on the other hand obtained DT models should not be overly sensitive to imperfect function of a cluster-detector.

Additional goal of the work is to minimize an impact of delayed or incorrect response of cluster detector to the changes of speakers. Such situation is very frequent just in on-line subtitling of TV discussions. Therefore, we propose a fusion method for speaker-clustered AMs which give better results in real-time tasks than switching methods.

Let us mention that the clustering algorithm is described in Section 2, the Discriminative Adaptation is described in Section 3. Description of tested switching and fusion methods is in Section 4, and results of experiments are described in Section 6.

2 Automatic clustering

Training of gender-dependent models is the most popular method how to split training data into two more acoustically homogeneous classes [1]. But for particular corpora, it should be verified that the gender-based clusters are the optimal way, i.e. the criterion $L = \prod_u P\{u \mid M(u))$, where u is an utterance in a corpus and $M(u)$ is a relevant acoustic model

of its reference transcription, is maximal. Because of some male/female "mishmash" voices contained in corpora we proposed an unsupervised clustering algorithm which can reclassify training voices into more acoustically homogeneous classes. The clustering procedure can start from gender-dependent splitting and it finishes in somewhat refine distribution which yields higher accuracy score [2]. In addition, we can use the algorithm to find out more than only two acoustically homogeneous clusters. Thereafter, two ways of clustering procedure are possible. The first approach is just to split randomly initial training data into *n* clusters and run the algorithm. The second way is to prepare clusters hierarchically. It means to split data via the algorithm into two clusters and after that to continue in the same way with the both sub-clusters. The number of final clusters can naturally be the power of two only. This way produces more size-balanced clusters and it does not need as much computation time as the first direct way. But the final clusters do not need to be so compact.

2.1 Algorithm description

The algorithm is based on similar criterion like the main training algorithm - maximize likelihood *L* of the training data with reference transcription and models. The result of the algorithm is a set of trained acoustic models and a set of lists where all utterances are assigned to exactly one cluster. Number of clusters (classes) *n* has to be set in advance and for gender-dependent modeling or for hierarchical splitting is naturally $n = 2$. The process is modification of the Expectation-Maximization (EM) algorithm. The unmodified EM algorithm is applied for estimation of acoustic model parameters.

The clustering algorithm goes as follows:

1. Random splitting of training utterances into *n* clusters. The clusters should have similar size. In case of two initial classes there is reasonable to start the algorithm from gender-based clusters.

2. Train (retrain) acoustic models for all clusters.

3. Posterior probability density $P(u|M)$ of each utterance u with its reference transcription is computed for all models *M* (so-called forced-alignment).

4. Each utterance is assorted to the cluster with the maximal evaluation $P(u|M)$ computed in the previous step:

$$M_{t+1}(u) = \arg\max_{M} P(u \mid M).$$

5. If clusters changed than go back to step 2. Otherwise the algorithm is terminated.

Optimality of results of the clustering algorithm is not guaranteed. Besides, the algorithm depends on initial clustering. Furthermore, even convergence of the algorithm is not guaranteed, because there can be a few utterances which are reassigned all the time. Therefore, it is suitable to apply a little threshold as a final stopping condition or to use fixed number of iterations.

Thus, if we would like to verify that the gender-dependent splitting is "optimal" so we use this male/female distribution as initial and start algorithm. The intention is to complete the algorithm with more refined clusters, in which "masculine" female and "feminine" male

voices and also errors in manual male/female annotations will be reclassified. This should improve a performance of the recognizer.

3 Discriminative adaptation

Discriminative training (DT) was developed in a recent decade and provides better recognition results than classical training based on Maximum Likelihood criterion (ML) [3,4]. In principle, ML based training is a machine learning method from positive examples only. We used Frame-Discriminative MMI variant where denominator lattices generation and its forward-backward processing is not needed. The denominator posterior probability is calculated from a set of all states in HMM. This very general denominator model leads to good generalization to test data. This feature is especially valuable in the case where the training data are limited or, in our cases, the data are split into smaller clusters. Furthermore, statistics of only few major Gaussians are needed to be updated and its probability has to be exactly calculated in each time. It can tend to very time-efficient algorithm [5].

Additionally, for better training stability, discriminative variant of a maximum a posteriori probability method (DT-MAP) [6] has been used. It works in the same manner as the standard MAP, only the input HMM has to be discriminatively trained with the same objective function. For discriminative adaptation it is strongly recommended to use I-smoothing method to boost stability of new estimates [7].

4 Fusion instead switching

In the case where more than one AM is available, some of the switching or fusion methods have to be applied. In our recent paper [8], we investigate these methods. Three switching and four fusion methods were proposed, described and tested on gender-dependent LVCSR real-time system. Some of them gave significantly better results than the gender-independent modeling. The lowest WER has been obtained with weighted sum of the HMM state probabilities of all acoustic and its relative WER reduction was 2% absolutely and more than 11% relatively. The best method should be described in more detail:

4.1 Weighted sum with exponential forgetting

The method is relatively simple fusion method. It generates final output probabilities of HMM models as a weighted sum of all fused AMs:

$$\hat{P}(s_i \mid o_t) = \sum_{k=1}^{M} w_t^k P_k(s_i \mid o_t), \qquad (1)$$

where $\hat{P}(s_i \mid o_t)$ is the final output probability of HMM state *i* in time *t*, $P_k(s_i \mid o_t)$ is output probability of *i*-th state from *k*-th AM. Weighting coefficient w_t^k is computed according to

$$w_t^k = \frac{P_t(\lambda_k)}{\sum_{l=1}^{M} P_t(\lambda_l)}, \qquad (2)$$

where $P_t(\lambda_k)$ is smoothed total probability of AM *k* in time *t*. An exponential forgetting is used to the smoothing

$$P_t(\lambda_k) = \alpha P_{t-1}(\lambda_k) + (1-\alpha) P(\lambda_k \mid o_t), \qquad (3)$$

where α is smoothing constant and it was set to 0.95 which is in the middle of the best-performing region (0.9 to 0.99 from preliminary experiments).

5 Experiment description

5.1 Training data and acoustic processing

The corpus for training of the acoustic models contains 100 hours of parliament speech records. All data were manually annotated. The digitization of an analogue signal is provided at 44.1 kHz sample rate and 16-bit resolution format. The aim of the front-end processor is to convert continuous speech into a sequence of feature vectors. Several tests were performed in order to determine the best parameterization settings of the acoustic data (see [9] for methodology). The best recognition results were achieved using PLP parameterization [10] with 27 filters and 12 PLP cepstral coefficients with both delta and delta-delta sub-features (see [11] for details). Therefore one feature vector contains 36 coefficients. Feature vectors are computed each 10 milliseconds (100 frames per second).

5.2 Acoustic model

The individual basic speech unit in all our experiments was represented by a three-state HMM with a continuous output probability density function assigned to each state. As the number of Czech triphones is too large, phonetic decision trees were used to tie states of Czech triphones. Several experiments were performed to determine the best recognition results according to the number of clustered states and also to the number of mixtures. In all presented experiments, we used 8 mixtures of multivariate Gaussians for each of 5385 states. The baseline acoustic model was speaker and gender independent (there was no additional information about speaker available) and it was trained by discriminative training method described therein before.

5.3 Training data clustering

The whole training corpus was split into several (two or more) acoustically homogeneous classes via the algorithm introduced in the Subsection 2.1. In all cases the initial splitting was achieved randomly due to no additional speaker/sentence information available. The whole set of sentences (46k) was split into various numbers of clusters. Examples of shifts between clusters (sentences, which were moved from the one cluster to another) during individual iterations for two-cluster case can be seen in Table 1.

Step (i)	number of sentences [%]	
	$Cl(x)_{i-1} \rightarrow Cl(x)_i$	$Cl(x)_{i-1} \rightarrow Cl(y)_i$
1	83.26	16.73
2	87.30	12.70
3	92.05	7.95
4	97.10	2.90
5	98.44	1.56
6	98.81	1.18
7	99.29	0.71
8	99.32	0.67

Table 1: The shift between two clusters during individual iterations.

Where $Cl(x)_{i-1} \rightarrow Cl(x)_i$ means no-shift between cluster x and $Cl(x)_{i-1} \rightarrow Cl(y)_i$ means shift between cluster x to any other cluster y ($y \neq x$) in two following iteration steps (i-1,i).

5.4 Training of clustered models

In recent work [12], we explored explore a suitable way of a discriminative training procedure for clustered acoustic models. This procedure should hold favorable characteristics of DT models on one hand, but on the other hand developed acoustic models should not be overly sensitive to imperfect function of a cluster-detector, e.g. a negative impact of wrong-selected acoustic model. Such situation could happen for instance in real-time recognition tasks. In these cases, the reaction of the cluster-detector to a change of speaker is not immediate and/or the detector evaluates the change incorrectly. We had performed a set of experiments in which an impact of speaker-independent and speaker-clustered acoustic models both in combination with maximum likelihood and frame-based discriminative training were tested. The best method was discriminative adaptation. The baseline single-cluster discriminative model was adapted to cluster-model set via two iterations of DT-MAP.

5.5 Tests description

The test set consists of 100 utterances from 100 different speakers (64 male and 36 female speakers), which were not included in training data. There were no cross talking or speaker changes during each utterance. This portion of utterances was randomly separated to 10 sets so that each set contains at least one male and one female speaker. These multi-utterances were created in order to simulate real-time speaker changes. All recognition experiments were performed with a bigram back-off language model with Good-Turing discounting. The language model was trained on about 10M tokens of normalized Czech Parliament transcriptions. The SRI Language Modeling Toolkit (SRILM) [13] was used for training. The model contains 186k words and the perplexity of the recognition task was 12.36 and OOV was 2.4% (see [14] and [15] for details).

6 Results

For preliminary experiments, we follow up our last year paper [16], the same three acoustic models were used: gender-independent (GI), male and female. At first, all these models were tested stand alone. At second, the described fusion method was evaluated. The results are in table 2.

Stand alone models	WER [%]
Gender-independent	16.92
Male	22.08
Female	30.07
Fusion	14.96

Table 2: The results of recognition experiments. In fusion case, all three above models were used.

From the table 2, it is clear that fusion method gave significantly lower WER than GI model and its relative WER reduction is 2% absolutely and more than 11% relatively.

Proper setting of α parameter is needed for fusion with the exponential forgetting. The advisable α region is between 0.9 and 0.99. The relation between α value and word error rate is depicted on figure 1.

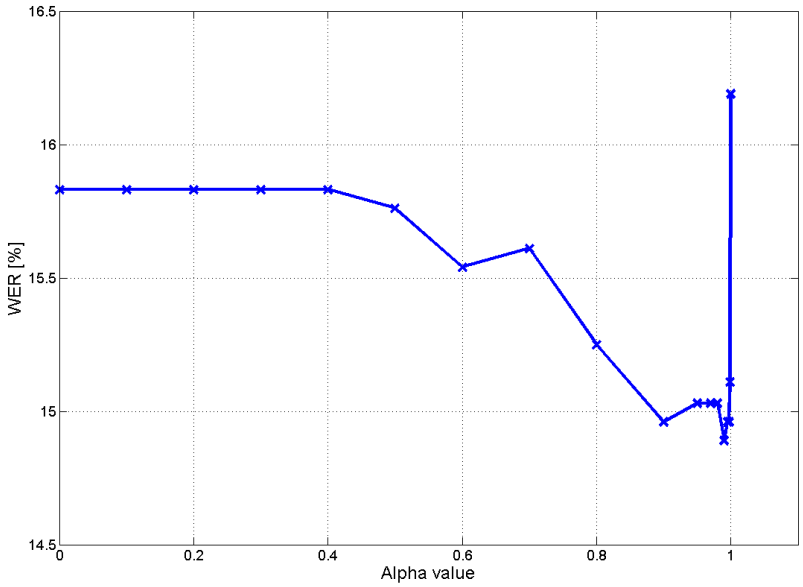

Figure 1: Relation of α value and WER.

Training of higher number of clusters is an extremely time consuming task. Therefore, these experiments are still under progress and results will be presented in near future. In this time, there are finished experiments with up to four clusters [12] where recognition was performed offline with ideal and anti-ideal speaker cluster detector. Finished online experiments for three clusters are described above.

7 Conclusion

This paper discusses using of higher number of AMs at ones. All the three key problems are analyzed; splitting of the training data, training of the models, and fusion of the AMs which can be used for real-time speech recognition. For three-cluster case, the described fusion method gave significantly lower WER than GI model and its relative WER reduction is 2% absolutely and more than 11% relatively.

8 Acknowledgements

This research was supported by Grant Agency of the Czech Republic, No. 102/08/0707 and by the Ministry of Education of the Czech Republic, project No. 2C06020.

References

[1] Stolcke A., et al.: The SRI March 2000 Hub-5 Conversational Speech Transcription System. Proc. NIST Speech Transcription Workshop, College Park, MD, May 2000.

[2] Zelinka J.: Audio-visual speech recognition. Ph.D. thesis, West Bohemia University, Department of Cybernetics, 2009. (in Czech)

[3] Povey D.: Discriminative Training for Large Vocabulary Speech Recognition. Ph.D. thesis, Cambridge University, Department of Engineering, 2003.

[4] McDermott E., et al: Discriminative training for large vocabulary speech recognition using minimum classification error. IEEE Trans. Speech and Audio Proc, Vol. 14. No. 2, 2006.

[5] Povey, D., et al.: Frame discrimination training for HMMs for large vocabulary speech recognition. In: Proceedings of the ICASSP, 1999.
[6] Povey, D., Gales M.J.F., Kim, D.Y., Woodland, P.C: MMI-MAP and MPE-MAP for acoustic model adaptation. In: EUROSPEECH, pp. 1981-1984, 2003
[7] Povey, D., Woodland, P.: Minimum phone error and I-smoothing for improved discriminative training. In: Proceedings of the ICASSP, Orlando, USA, 2002.
[8] Vaněk, J. and Psutka, J.V.: Gender-dependent acoustic models fusion developed for automatic subtitling of Parliament meetings broadcasted by the Czech TV. Lecture Notes in Artificial Intelligence, vol. 6231, p. 431-438, Springer, Berlin, 2010.
[9] Psutka, J., Müller, L., Psutka, J. V.: Comparison of MFCC and PLP Parameterization in the Speaker Independent Continuous Speech Recognition Task. In: EUROSPEECH, Aalborg, Denmark, 2001.
[10] Hermansky H.: Perceptual linear predictive (PLP) analysis of speech. J. Acoustic. Soc. Am.87, 1990.
[11] Psutka J.: Robust PLP-Based Parameterization for ASR Systems. In SPECOM 2007 Proceedings, Moscow State Linguistic University, 2007.
[12] Vaněk, J., et al.: Training of Speaker-Clustered Acoustic Models for Use in Real-Time Recognizers. In Proceedings of the International Conference on Signal Processing and Multimedia Application, p. 131-135, INSTICC, Setubal, 2009.
[13] Stolcke A.: SRILM - An Extensible Language Modeling Toolkit. In: International Conference on Spoken Language Processing (ICSLP 2002), Denver, USA, 2002.
[14] Pražák, A., Ircing, P., Švec, J, et al.: Efficient Combination of N-gram Language Models and Recognition Grammars in Real-Time LVCSR Decoder. In: 9th International Conference on Signal Processing, page 587-591, Beijing, CHINA, 2008.
[15] Pražák A., Müller, L., Šmídl, L. : Real-time decoder for LVCSR system. In: 8th World Multi-Conference on Systemics, Cybernetics and Informatics, Orlando FL, USA, 2004.
[16] Vaněk J., Psutka J.V., Zelinka J., Pražák A., and Psutka, J.: Discriminative training of gender-dependent acoustic models. Lecture Notes in Artificial Intelligence, vol. 5729, p. 331-338, Springer, Berlin, 2009.