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

Jan Vaněk, Josef V. Psutka, Jan Zelinka, Aleš Pražák and Josef Psutka
Department of Cybernetics, West Bohemia University, Pilsen, Czech Republic.
{*vanekyj,psutka_j,aprazak,zelinka,psutka*}*@kky.zcu.cz*

Keywords: Acoustic models training, discriminative training, clustering, gender-dependent models.

Abstract: The paper deals with training of speaker-clustered acoustic models. Various training techniques - Maximum Likelihood, Discriminative Training and two adaptation based on the MAP and Discriminative MAP were tested in order to minimize an impact of speaker changes to the correct function of the recognizer when a response of the automatic cluster detector is delayed or incorrect. Such situation is very frequent e.g. in on-line subtitling of TV discussions (Parliament meetings). In our experiments the best cluster-dependent training procedure was discriminative adaptation which provided the best trade-off between recognition results with correct and non-correct cluster detector information.

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 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 or even models obtained from more detailed clustered voices. This paper describes our experiments with unsupervised speaker clustering and following discriminative training of various initial acoustic models. The 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. All the discussed methods came from frame-based discriminative training (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.

Let us mention that the clustering algorithm is described in Section 2, the Discriminative Training or Frame-Discriminative training are described in Section 3. The incorporating DT to a cluster-dependent training procedure is discussed in Section 4.4 and results of completed experiments are described in Section 5.

2 Automatic clustering

Training of gender-dependent models is the most popular method how to split training data into two more acoustically homogeneous classes (Stolcke, 2000). But for particular corpora, it should be verified that the gender-based clusters are the optimal way, i.e. the criterion $L = \prod_u P(u|M(u))$, 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 (Zelinka, 2009). 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:

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

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 Training

Discriminative training (DT) was developed in a recent decade and provides better recognition results than classical training based on Maximum Likelihood criterion (ML) (Povey, 2003; McDermott, 2006). In principle, ML based training is a machine learning method from positive examples only. DT on the contrary uses both positive and negative examples in learning and can be based on various objective functions, e.g. Maximum Mutual Information (MMI) (Bahl at al., 1986), Minimum Classification Error (MCE) (McDermott, 2006), Minimum Word/Phone Error (MWE/MPE) (Povey, 2003). Most of them require generation of lattices or many-hypotheses recognition run with appropriate language model. The lattices generation is highly time consuming. Furthermore, these methods require good correspondence between training and testing dictionary and language model. If the correspondence is weak, e.g. there are many words which are only in the test dictionary then the results of these methods are not good. In this case, we can employ Frame-Discriminative training, which is independent on a used dictionary and language model (Kapadia, 1998). In addition, this approach is much faster. In lattice based method with the MMI objective function the training algorithm seeks to maximize the posterior probability of the correct utterance given the used models (Bahl at al., 1986):

$$\mathcal{F}_{MMI}(\lambda) = \sum_{r=1}^{R} \log \frac{P_\lambda(O_r|s_r)^\kappa P(s_r)^\kappa}{\sum_s P_\lambda(O_r|s)^\kappa P(s)^\kappa}, \qquad (2)$$

where λ represents the acoustic model parameters, O_r is the training utterance feature set, s_r is the correct transcription for the r'th utterance, κ is the acoustic scale which is used to amply confusions and herewith increases the test-set performance. $P(s)$ is a language model part. Optimization of the MMI objective function uses Extended Baum-Welch update equations and it requires two sets of statistics. The first set, corresponding to the numerator (num) of the equation (2), is the correct transcription. The second one corresponds to the denominator (den) and

it is a recognition/lattice model containing all possible words. An accumulation of statistics is done by forward-backward algorithm on reference transcriptions (numerator) as well as generated lattices (denominator). The Gaussian means and variances are updated as follows (Kapadia, 1998):

$$\hat{\mu}_{jm} = \frac{\Theta^{num}_{jm}(O) - \Theta^{den}_{jm}(O) + D_{jm}\mu'_{jm}}{\gamma^{num}_{jm} - \gamma^{den}_{jm} + D_{jm}} \quad (3)$$

$$\hat{\sigma}^2_{jm} = \frac{\Theta^{num}_{jm}(O^2) - \Theta^{den}_{jm}(O^2) + D_{jm}(\sigma'^2_{jm} + \mu'^2_{jm})}{\gamma^{num}_{jm} - \gamma^{den}_{jm} + D_{jm}} - \mu^2_{jm}, \quad (4)$$

where j and m are the HMM-state and Gaussian index, respectively, γ_{jm} is the accumulated occupancy of the Gaussian, $\Theta_{jm}(O)$ and $\Theta_{jm}(O^2)$ are a posteriori probability weighted by the first and the second order accumulated statistics, respectively. The Gaussian-specific stabilization constants D_{jm} are set to maximum of (i) double of the smallest value which ensures positive estimated variances, and (ii) value $E\gamma^{den}_{jm}$, where constant E determines the stability/learning-rate and it is a compromise between stability and number of iteration which is needed for well-trained models (Povey at al., 2001). In Frame-Discriminative case, the 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. 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 (Povey at al., 1999). Optimization of the model parameters uses the same two equations (3) and (4), the computation of $\Theta^{den}_{jm}(O)$ and γ^{den}_{jm} is modified only. In case that only limited data are available, maximum a posteriori probability method (MAP) (Gauvain at al., 1994) can be used even for discriminative training (Povey at al., 2003). 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 (Povey at al., 2002).

4 EXPERIMENTS DESCRIPTION

4.1 Train 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 (Psutka, 2001) for methodology). The best recognition results were achieved using PLP parameterization (Hermansky, 1990) with 27 filters and 12 PLP cepstral coefficients with both delta and delta-delta sub-features (see (Psutka, 2007) for details). Therefore one feature vector contains 36 coefficients. Feature vectors are computed each 10 milliseconds (100 frames per second).

4.2 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 triphones is too large, phonetic decision trees were used to tie states of 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 were no additional information about speaker available).

4.3 Training data clustering

The whole training corpus was split into several (two or more) acoustically homogeneous classes via algorithm introduced in the Subsection 2. 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 four classes in two different ways. Firstly, we used hierarchical division method. It means that we divided the training set into two classes ($2Cl$) and than each class was split again into another two classes (finally we had four clusters $4Cl_{Hi}$). Secondly, we split the whole training set into four classes directly ($4Cl_{Di}$). All splittings were done using algorithm presented above. Examples of shifts between clusters (sentences, which were moved from the one cluster to another) for hierarchical division can be seen in Table 1.

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)

Table 1: Example of the shift between clusters

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

4.4 Discriminative training of clustered models

Our next attention was to 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 case that the reaction of the cluster-detector to a change of speaker is not immediate and/or the detector evaluates the change incorrectly. We 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. In case when only single acoustic model is trained, the situation is simple. The model is trained from all data under ML approach or some DT objective function. Nevertheless some parameters could be tuned, for example a number of tied-states and a number of Gaussians per state. In DT case, the number of tuned parameters is higher but it is still an optimization task. In our experiments corresponding models are marked as *SC* (Single Cluster), precisely *SC_ML* and *SC_DT* for ML and DT, respectively. The DT model was developed from *SC* via two discriminative training iterations. The E constant was set to one. Furthermore, the I-smoothing was applied and smoothing constant τ^I was set to 100. If the training data are split into more than one cluster, the situation is a bit complicated because of more training strategies that we have in our disposal. Naturally the same training procedure can be used for each part of data. This would be concluded by a set of independent models. For a real application such approach is not a good option because final models have different topology which is generated during a tied-states clustering procedure and therefore obtained models cannot be simply switched/replaced in the recognizer. The better strategy is to split the training procedure just after state clustering. In our experiments such model sets are marked with suffix *_ML* and *_DT* for ML and DT, respectively. Secondly, the ML or DT adaptation can be applied. In our experiments the adaptation starts from *SC_ML* or *SC_DT* and two iterations were done via MAP or DT-MAP with parameter τ equal to 25. Two models developed by these techniques are marked with suffix $\_ML_{Adapt}$ and $\_DT_{Adapt}$.

4.5 Tests description

The test set consists of 100 minutes of speech from 10 male and 10 female speakers (5 minutes from each) which were not included in training data. In all recognition experiments a language model based on zerograms was applied in order to judge better a quality of developed acoustic models. In all experiments the perplexity of the task was 3828, there were no OOV words.

5 RESULTS

In all our experiments the recognition accuracy was evaluated. Obtained results are shown in Table 2, where *SC_ML* and *SC_DT* were trained from the whole training data via Maximum Likelihood and Discriminative Training, respectively. In the second part of Table two recognition results for each training procedure and clustering method are shown. The *best* recognition result was found for each utterance across appropriate acoustic models (the same level of clustering and the same training procedure). The *worst* result was found by analogical way. The difference between corresponding results illustrates the drop of recognition accuracy when the cluster-detector fails. As was described in Section 4.3, $2Cl$ is an acoustic model with two clusters. $4Cl_{Hi}$ and $4Cl_{Di}$ are the four-clusters acoustic models which were obtained by hierarchical and direct clustering, respectively. We achieved a significant gain in terms of recognition accuracy for all cluster-dependent acoustic models against standard (single-cluster) acoustic models (*SC_ML* and *SC_DT*). The two-clusters acoustic models gave only slightly better results than single-cluster models. But in the four-clusters case the achieved improvement is significant. The comparison between the hierarchical ($4Cl_{Hi}$) and direct ($4Cl_{Di}$) clustering method showed that the direct method gave clusters whose corresponding acoustic models yield bet-

Table 2: The results of recognition experiments

	Acc [%]	
SC_ML	71.37	
SC_DT	73.60	
Cluster identification	best	worst
$2Cl\_ML$	71.76	66.65
$2Cl\_DT$	74.01	69.29
$2Cl\_ML_{Adapt}$	71.64	67.53
$2Cl\_DT_{Adapt}$	74.03	71.36
$4Cl_{Hi}\_ML$	72.62	52.14
$4Cl_{Hi}\_DT$	75.17	55.83
$4Cl_{Hi}\_ML_{Adapt}$	72.83	62.78
$4Cl_{Hi}\_DT_{Adapt}$	74.39	69.48
$4Cl_{Di}\_ML$	74.69	54.18
$4Cl_{Di}\_DT$	74.01	56.69
$4Cl_{Di}\_ML_{Adapt}$	74.65	59.03
$4Cl_{Di}\_DT_{Adapt}$	76.66	67.28

ter recognition results. The maximum gain (improvement 5.29% absolutely for *SC_ML*) was achieved for $4Cl_{Di}\_DT_{Adapt}$ (Discriminatively trained directly clustered four-clusters acoustic models). In this case, the accuracy 76.66% was obtained if the cluster-detector works ideally. But on the other hand the recognition results are considerably worse when the cluster information is not correct. From this point of view the best tradeoff between recognition results of the cluster-based acoustic model with correct and non-correct cluster information are $2Cl\_DT_{Adapt}$ in the two-clusters case and $4Cl_{Di}\_DT_{Adapt}$ in the four-clusters case. In the two-clusters case the recognition results are slightly worse (improvement 2.66% absolutely for *SC_ML*) than for the four-clusters approach. But if the cluster detector information is wrong, the recognition results were almost the same in comparison with the original *SC_ML* acoustic model.

6 CONCLUSION

The goal of our work was to build the cluster-dependent acoustic model which yields higher recognition accuracy than non-clustered model and which is more robust when a response of the automatic cluster-detector is delayed or incorrect. This problem becomes serious if the recognizer works in real-time and in tasks where speakers change frequently. We tested several methods based on a combination of unsupervised clustered training data and discriminative/non-discriminative training procedures. If the cluster-detector works "almost" correctly then the best cluster-dependent training procedure is $4Cl_{Di}\_DT_{Adapt}$. But the question is what results we obtain if the splitting process continues, e.g. for levels 8, 16 ... In our next research we would like to concentrate on this problem and also on the question how to build a quick cluster-detector which will work correctly and really in real-time.

7 Acknowledgements

The work was supported by the Ministry of Education of the Czech Republic, project no. MŠMT 2C06020.

REFERENCES

Povey, D. at al. (1999). Frame discrimination training for hmms for large vocabulary speechrecognition. In *ICASSP*.

Povey, D. at al. (2001). Improved discriminative training techniques for large vocabulary continuous speech recognition. In *ICASSP*.

Povey, D. at al. (2002). Minimum phone error and i-smoothing for improved discriminative training. In *ICASSP*.

Povey, D. at al. (2003). Mmi-map and mpe-map for acoustic model adaptation. In *EUROSPEECH*.

Bahl, L.R. at al. (1986). Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *ICASSP*.

Gauvain, L. at al. (1994). Maximum a-posteriori estimation for multivariate gaussian mixture observations of markov chains. In *IEEE Transactions SAP*.

Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech. *Acoustic. Soc., Am.87*.

Kapadia, S. (1998). *Discriminative Training of Hidden Markov Models*. PhD thesis, Cambridge University, Department of Engineering.

McDermott, E. (2006). Discriminative training for large vocabulary speech recognition using minimum classification error. *IEEE Trans. Speech and Audio Processing, Vol. 14. No. 2*.

Povey, D. (2003). *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, Cambridge University, Department of Engineering.

Zelinka, J. (2009) *Audio-Visual Speech Recognition*. PhD thesis, West Bohemia University, Department of Cybernetics.

Psutka, J. (2001) Comparison of MFCC and PLP Parameterization in the Speaker Independent Continuous Speech Recognition Task. In *EUROSPEECH*.

Psutka, J. (2007) Robust PLP-Based Parameterization for ASR Systems. In *SPECOM*.

Stolcke, A. (2000). The sri march 2000 hub-5 conversational speech transcription system. In *NIST Speech Transcription Workshop*. College Park, MD.