Discriminative training of gender-dependent acoustic models

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,
WWW home page: http://www.kky.zcu.cz

Abstract. The main goal of this paper is to explore the methods of gender-dependent acoustic modeling that would take the possibly of imperfect function of a gender detector into consideration. Such methods will be beneficial in real-time recognition tasks (eg. real-time subtitling of meetings) when the automatic gender detection is delayed or incorrect. The goal is to minimize an impact to the correct function of the recognizer. The paper also describes a technique of unsupervised splitting of training data, which can improve gender-dependent acoustic models trained on the basis of manual markers (male/female). The idea of this approach is grounded on the fact that a significant amount of "masculine" female and "feminine" male voices occurring in training corpora and also on frequent errors in manual markers.

1 Introduction

The gender-dependent acoustic modeling is a very efficient way how to increase the accuracy in LVCSR systems. The training of acoustic models is usually based on manual markers connected with each utterance stored in a corpus. Such training of male/female acoustic models usually ignores diametrically different types of voices, e.g. "masculine" female and "feminine" male voices, whose occurrence in the corpus is not negligible. Also a problem with frequent errors in manual markers (male/female) connected with individual utterances is not solved. We proposed an unsupervised clustering algorithm which can reclassify training voices into more acoustically homogeneous classes. The clustering procedure starts from gender-dependent splitting and finishes in somewhat refined distribution which yields higher accuracy score. This approach is discussed in more detail in Section 2.1.

In the following part of the paper we discuss discriminative training (DT) of gender-dependent acoustic models. All the discussed methods come from frame-based discriminative training that seeks such solution (such acoustic models) which yield on one hand favorable quality (increased accuracy) of DT models, on the other hand these DT models should not be overly sensitive to imperfect function of a gender detector. A profit of this solution can be observed in real-time recognition tasks (e.g. real-time subtitling of meetings) when a reaction of the gender detector to the changes of speakers is not immediate or the detector evaluates changes incorrectly. The goal is to minimize an impact to the correct function of the recognizer. Let us mention that the Discriminative Training (DT) or Frame-Discriminative training (FD) are described in Section 2.2 and

incorporating DT to a gender-dependent training procedure is discussed in Section 4.4. Obtained results are presented in Section 5.

2 Methods

2.1 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 $L = \prod_u P(u|M(u))$, 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 starts from gender-dependent splitting and finishes in somewhat refined distribution which yields higher accuracy score. [2].

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 naturally $n = 2$. The process is a 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. Initial splitting of training utterances into n clusters. The clusters should have similar size. In case of two initial classes it is reasonable to start the algorithm from gender-based clusters/lists. In general case it should be a random splitting.
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:

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

5. If clusters changed then 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 a fixed number of iterations. Thus, if we want to verify that the gender-dependent splitting is "optimal" we use this initial male/female distribution and start the algorithm. The intention is that the algorithm finishes with more refined clusters, in which "masculine" female and "feminine" male voices and also errors in manual male/female annotations will be reclassified which should ensure better performance of a recognizer.

2.2 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) [3–6]. 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) [7], Minimum Classification Error (MCE) [5], Minimum Word/Phone Error (MWE/MPE) [3]. 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 (FD), which is independent on a used dictionary and language model [8]. In addition, this approach is much faster.

2.3 Frame-Discriminative training

In lattice based method with MMI objective function the training algorithm seeks to maximize the posterior probability of the correct utterance given the used models [7]:

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

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 [8]:

$$\hat{\mu}_{jm} = \frac{\Theta_{jm}^{num}(O) - \Theta_{jm}^{den}(O) + D_{jm}\mu'_{jm}}{\gamma_{jm}^{num} - \gamma_{jm}^{den} + D_{jm}} \tag{3}$$

$$\hat{\sigma}_{jm}^2 = \frac{\Theta_{jm}^{num}(O^2) - \Theta_{jm}^{den}(O^2) + D_{jm}(\sigma'^2_{jm} + \mu'^2_{jm})}{\gamma_{jm}^{num} - \gamma_{jm}^{den} + D_{jm}} - \mu_{jm}^2, \tag{4}$$

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 $E\gamma_{jm}^{den}$, 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 [9]. 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 a few major Gaussians are need to be updated and their probability has to be exactly calculated in each time. It can lead to very time-efficient algorithm [10]. Optimization of the model parameters uses the same two equations (3) and (4), the computation of $\Theta_{jm}^{den}(O)$ and γ_{jm}^{den} is modified only.

2.4 Frame-Discriminative adaptation

In case that only limited data are available, maximum a posteriori probability method (MAP) [11] can be used even for discriminative training [12]. 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 [13].

3 Train data description

For training of acoustic models a microphone-based high-quality speech corpus was used. The high-quality speech corpus of read-speech consists of the speech of 800 speakers (384 males and 416 females). Each speaker read 170 sentences. The database of text prompts from which the sentences were selected was obtained in an electronic form from the web pages of Czech newspaper publishers[14]. Special consideration was given to the sentences selection, since they provide a representative distribution of the more frequent triphone sequences (reflecting their relative occurrences in natural speech). The corpus was recorded in the office where only the speaker was present. Sentences were recorded by a close-talking microphone (Sennheisser HMD410-6). The recording sessions yielded totally about 220 hours of speech.

4 Experimental setup

4.1 Acoustic processing

The digitization of an analogue signal is provided at 22.05 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 [15] for methodology). The best results were achieved using PLP parameterization [16] with 27 filters and 12 PLP cepstral coefficients with both delta and delta-delta sub-features (see [17] 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 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 16 mixtures of multivariate Gaussians for each of 4922 states. The baseline acoustic model was made using HTK-Toolkit v.3.4 [18].

4.3 Two class splitting

As was presented above, the whole training corpus was split into two acoustically homogeneous classes (gender-based). Initial splitting was achieved via manual markers. However, due to several "masculine" female and "feminine" male voices occurring in the training corpora and also because of possible errors in manual annotations we applied algorithm introduced in Subsection 2.1 to refine initial "gender-based" training subcorpora. The whole set of sentences (109.5k) was split into male (52.4k) and female (57.1k) parts based on manually assigned markers. The percentage of sentences which were moved from the male to female ($M_{i-1} \rightarrow F_i$) cluster as well as from the female to male ($F_{i-1} \rightarrow M_i$) cluster in two following iteration steps (i,$i-1$) is given in Table 1.

Table 1. The shift between male and female clusters

Iteration step (i)	[%]			
	$M_{i-1} \rightarrow M_i$	$M_{i-1} \rightarrow F_i$	$F_{i-1} \rightarrow F_i$	$F_{i-1} \rightarrow M_i$
1	96.81	4.81	95.76	2.63
2	99.37	0.90	99.21	0.52
3	99.34	0.37	99.93	0.36
4	99.75	0.25	99.69	0.31
5	99.65	0.25	99.78	0.32
6	99.90	0.06	99.94	0.10
7	99.97	0.01	99.99	0.03

4.4 Discriminative training of two-class models

Our next attention was to explore a suitable way of discriminative training of gender-dependent acoustic models which would yield on one hand favorable characteristics of DT models but on the other hand developed models should not be overly sensitive to imperfect function of a gender detector, e.g. a negative impact of reversely selected (male/female) acoustic model. Such situation could happen for instance in real-time recognition tasks in case that the reaction of a gender detector to the 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 gender-dependent acoustic models were tested in combination with a technique of frame-based discriminative training. 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 SI (Speaker Independent) and SI_DT for ML and DT, respectively. The DT model was developed from SI via two training iterations based on FD-MMI objective function. 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 is split into more than one class, 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 is concluded by a set of independent models. For a real application this approach is not a good option because final models have different topology which is generated during tied-states clustering 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 as $ClusterGD$ and $ClusterGD\_DT$ for ML and DT, respectively. Secondly, the ML or DT adaptation can be applied. In our experiments the adaptation starts from SI or SI_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 as $SI\_MLAdapt$ and $SI\_DTAdapt$.

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 a quality of developed acoustic models. In all experiments the perplexity of the task was 2190, there were no OOV words.

5 Results

As can be seen from Table 2 we achieved a significant gain in terms of recognition results for all gender-dependent acoustic models ($ClusterGD$, $ClusterGD\_DT$, $SI\_MLAdapt$ and $SI\_DTAdapt$) against speaker independent acoustic models (SI and SI_DT). Moreover the automatic clustering procedure decreases the word error rate more than 1.7% relatively, see the row GD with recognition results for manually marked training data and $ClusterGD$ with recognition results after the automatic re-clustering procedure of training data was performed. This gain is more than 11% relatively for $ClusterGD\_DT$ (discriminative re-training of gender dependent ML model) when the information on speaker gender is correct. But on the other hand the recognition results are considerably worse when the speaker gender information

Table 2. The results of recognition experiments

	WER [%]	
SI	40.19	
SI_DT	39.02	
Gender identification	correct	non correct
GD	37.50	64.08
$ClusterGD$	36.89	63.57
$ClusterGD\_DT$	35.81	61.92
$SI\_MLAdapt$	38.08	52.18
$SI\_DTAdapt$	36.99	46.60

is not correct. From this point of view the best tradeoff between recognition results of gender-dependent acoustic model with correct and non-correct gender information is $SI\_DTAdapt$ ($SI\_DTAdapt$ is SI_DT after two iterations via DT-MAP). In this case the recognition results are slightly worse (improvement 8% relatively to SI) than in case of $ClusterGD\_DT$ but the non-correct gender information decreases the recognition results only slightly comparing with the original SI acoustic model.

6 Conclusion

The goal of our work was to build the gender-dependent acoustic model which is more robust to the incorrect decisions of gender detector. We tried several methods based on combination of gender-based data and discriminative training procedures. In all experiments a zero-gram language model was applied in order to better judge the quality of developed acoustic model. The best gender-dependent training procedure depends on the performance of gender detection. If the gender detector works perfectly the GD_DT model is the best solution. But if the gender detector works incorrectly, e.g. a change of speaker is not detected in time or is evaluated sometimes wrongly then $SI\_DTAdapt$ acoustic model seems to be a good trade off.

7 Acknowledgements

This research was supported by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1QS101470516.

References

1. A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauche, C. Rickey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng: 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. D. Yu, L. Deng, X. He, and A. Acero: Use of incrementally regulated discriminative margins in MCE training for speech recognition. Proc. Interspeech 2006.
5. E. McDermott, T. Hazen, J. L. Roux, A. Nakamura, and S. Katagiri: Discriminative training for large vocabulary speech recognition using minimum classification error. IEEE Trans. Speech and Audio Proc, Vol. 14. No. 2, 2006.
6. W. Reichl, G. Ruske: Discriminative Training for Continuous Speech Recognition. Proc. 1995 Europ. Conf. on Speech Communication and Technology, Vol. 1, pp. 537-540, Madrid, September 1995.
7. Bahl L.R., Brown P.F, de Souza P.V., and Mercer L.R.: Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition. In ICASSP, 1986.
8. Kapadia S.: Discriminative Training of Hidden Markov Models. Ph.D. thesis, Cambridge University, Department of Engineering, 1998.
9. Povey, D. and Woodland, P.C.: Improved discriminative training techniques for large vocabulary continuous speech recognition. In: IEEE international Conference on Acoustics Speech and Signal Processing, 7-11 May 2001, Salt Lake City, Utah.
10. Povey D., Woodland P.C.: Frame discrimination training for HMMs for large vocabulary speechrecognition. In: Proceedings of the ICASSP, Phoenix, USA, 1999.
11. Gauvain, L., Lee, C.H.: Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. In: IEEE Transactions SAP, 1994.
12. 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
13. Povey, D., Woodland, P.: Minimum phone error and I-smoothing for improved discriminative training. In: Proceedings of the ICASSP, Orlando, USA, 2002.
14. V. Radov, J. Psutka, J., UWB-S01 Corpus: A Czech Read-Speech Corpus, Proceedings of the 6th International Conference on Spoken Language Processing ICSLP2000, Beijing 2000, China.
15. Psutka, J., Müller, L., Psutka, J. V.: Comparison of MFCC and PLP Parameterization in the Speaker Independent Continuous Speech Recognition Task. In: 7th European Conference on Speech Communication and Technology (EUROSPEECH 2001), Aalborg, Denmark, 2001.
16. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. J. Acoustic. Soc. Am.87, 1990.
17. Psutka, J. Robust PLP-Based Parameterization for ASR Systems. In SPECOM 2007 Proceedings. Moscow : Moscow State Linguistic University, 2007.
18. S. Young et al.: The HTK Book (for HTK Version 3.4), Cambridge, 2006.
19. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: International Conference on Spoken Language Processing (ICSLP 2002), Denver, USA, 2002.