Gender-dependent acoustic models fusion developed for automatic subtitling of Parliament meetings broadcasted by the Czech TV

Jan Vaněk and Josef V. Psutka

Department of Cybernetics, West Bohemia University, Pilsen, Czech Republic.
{vanekyj,psutka_j,}@kky.zcu.cz,
WWW home page: http://www.kky.zcu.cz

Abstract. Gender-dependent (male/female) acoustic models are more acoustically homogeneous and therefore give better recognition performance than single gender-independent model. This paper deals with a problem how to use these gender-based acoustic models in a real-time LVCSR (Large Vocabulary Continuous Speech Recognition) system that is for more than one year used by the Czech TV for automatic subtitling of Parliament meetings that are broadcasted on the channel ČT24. Frequent changes of speakers and the direct connection of the LVCSR system to the TV audio stream require switching/fusion of models automatically and as soon as possible. The paper presents various techniques based on using the output probabilities for quick selection of a better model or their combinations. The best proposed method achieved over 11% relative WER reduction in comparision with the GI model.

1 Introduction

In recent years, there appeared some projects for hearing-impaired people to help them to access to the information contained in acoustic signal especially of mass media. One of those projects is automatic subtitling of live broadcasted teleview. Recently, we introduced the system for automatic subtitling of Parliament meetings that are broadcasted by the Czech Television (ČT). This system is now used for more than one year by the ČT on the channel ČT24 (see details in [1]).

Frequent changes of speakers and the direct connection of the LVCSR system to the TV audio stream brings interesting challenges. This paper describes our effort to build and use gender dependent acoustic models. The gender-dependent acoustic modeling is a very efficient way how to increase the accuracy over a gender independent modeling in LVCSR and has been previously considered in the literature [2]. The most typical applications work in two-passes where in the first pass a gender-detection method is used (based on GMMs or on multilayer perceptrons-MLP) and in the second pass the speech is recognized with the corresponding gender-specific acoustic model [3].

In this paper we proposed a new combination methods for fusion of the acoustic models. These methods were applied on the level of acoustic models output probabilities. In recent years, the huge amount of computations related to acoustic model become negligible due to the increasing computer speed and capacity of computer memory.

From that point of view it is possible to compute several acoustic models simultaneously and switch or even combine their output probabilities in real-time applications. We would like to discuss and compare such methods with methods commonly used.

2 Methods

Various techniques for acoustic models switching/fusion were proposed. All techniques were designed for the real-time applications therefore only a small history for actual processed frame is needed. The first two methods are based on pure switching of individual acoustic models. The third method switches output probabilities for each time/state independently through all acoustic models. The other methods are based on evaluated total probability of the actual frame for all acoustic models. Some of the proposed methods use exponential forgetting to smooth probability volatility. The detailed description of the methods follows.

2.1 Frame arg max

This method marked as *Frame_argmax* chooses for the actual frame the acoustic model that maximizes given criterion. This criterion can be defined in several ways. The commonly used criterion is output probability from GMM or MLP. Because it was necessary to compute the output probabilities for all states in all acoustic models for other switching and fusion methods anyway, we used the total probability of all states of the acoustic model for the actual frame as our criterion:

$$P(\lambda_k|\boldsymbol{o}_t) = \sum_{i=1}^{I} P_k(s_i|\boldsymbol{o}_t), \tag{1}$$

where the total probability is the sum of the all I states s_i of the acoustical model λ_k and $P_k(s_i|\boldsymbol{o}_t)$ is an output probability of the state s_i of the k-th acoustical model and the feature vector $\boldsymbol{o}_t$ in time t. This criterion has, according to our experiments, similar results as the commonly used criterion based on GMMs. Method *Frame_argmax* chooses for actual frame model with the highest total probability. It means that at first the k_{max} is evaluated as

$$k_{max} = \underset{k \in 1 \ldots M}{\operatorname{argmax}} P(\lambda_k|\boldsymbol{o}_t) \tag{2}$$

and thus the new probabilities are

$$\hat{P}(s_i|\boldsymbol{o}_t) = P_{k\_max}(s_i|\boldsymbol{o}_t). \tag{3}$$

where M is number of acoustic models and $\hat{P}(s_i|\boldsymbol{o}_t)$ is the new evaluated state's probability.

2.2 Frame arg max with exponential forgetting

Because the time behavior of the total probability is volatile, some kind of smoothing should be used. The exponential forgetting is a good choice for the real-time applications. The total probabilities for all models are computed as

$$P_t(\lambda_k) = \alpha P_{t-1}(\lambda_k) + (1-\alpha) P(\lambda_k|\boldsymbol{o}_t), \tag{4}$$

where α parameter was set to 0.95. This value was in the center of optimal region in preliminary experiments. Relation between α value and word error rate were examined and results are shown in section 5. This method marked as *Frame_argmax_exp* is practically the same as the previous method except for using smoothed total probability $P_t(\lambda_k)$ instead of $P(\lambda_k|\boldsymbol{o}_t)$.

2.3 Independent maximum

The method marked as *Maximum* puts as the new probability of the state s_i the highest value of all M acoustic models.

$$\hat{P}(s_i|\boldsymbol{o}_t) = \max_{k \in 1 \ldots M} P_k(s_i|\boldsymbol{o}_t). \tag{5}$$

It means that the highest output probabilities are searched for each state s_i though all M acoustic models at every time t.

2.4 Independent multiplication

The following methods, contrary to the previous ones, are fusion of the output probabilities for states across all available acoustic models. The first method marked as *Multiply* is a simple multiplication of M acoustic models likelihoods for individual state:

$$\hat{P}(s_i|\boldsymbol{o}_t) = \sqrt[M]{\prod_{k=1}^{M} P_k(s_i|\boldsymbol{o}_t)}, \tag{6}$$

where $P_k(s_i|\boldsymbol{o}_t)$ is an output probability of the state s_i of the k-th acoustical model. The M-th root is there used to normalize probability back into original range. This approach is implemented internally as an average in log-likelihood domain.

2.5 Independent average

The second fusion method marked as *Average* is a simple average of M acoustic models likelihoods for individual state:

$$\hat{P}(s_i|\boldsymbol{o}_t) = \frac{1}{M} \sum_{k=1}^{M} P_k(s_i|\boldsymbol{o}_t). \tag{7}$$

2.6 Weighted multiplication with exponential forgetting

Similar to the the switching methods some kind of smoothing should be used. The last two methods use smoothing via weighted sum or multiplication of all probabilities. The weights in time t are computed as

$$w_t^k = \frac{P_t(\lambda_k)}{\sum_{l=1}^{M} P_t(\lambda_l)}. \tag{8}$$

The method marked as *W_mult_exp* evaluates new probabilities as

$$\hat{P}(s_i|\boldsymbol{o}_t) = \prod_{k=1}^{M} P_k(s_i|\boldsymbol{o}_t)^{w_t^k}. \tag{9}$$

In log-likelihood domain this approach can be implemented more simple as weighted sum of the log-likelihoods with precomputed weights w_t^k.

2.7 Weighted sum with exponential forgetting

The method with exponential forgetting is the last fusion method which is proposed in this paper. It is marked as *W_sum_exp* and it evaluates new probabilities as weighted sum

$$\hat{P}(s_i|\boldsymbol{o}_t) = \sum_{k=1}^{M} w_t^k P_k(s_i|\boldsymbol{o}_t). \tag{10}$$

In summary, the three switching and four fusion methods were proposed. All of them are fitted to real-time processing and do not pose any restriction to the number of acoustic models being used.

There's no need to compute all probabilities of all acoustic models for the first two switching methods. It is necessary to compute only one model in actual time if we have some estimate of total probability of individual models. This estimate can be done via much smaller GMM or with some algorithm using Gaussians pruning of the evaluated HMM model.

For fusion methods all state's probabilities of all models need to be evaluated but pruning or other fast HMM evaluation technique can be used. In addition, in the first stage just single acoustic model can be evaluated and, in the second stage, only small number of relevant states can be evaluated for other acoustic models. By using this scenario the computation burden increases over single-model only slightly.

3 Train data description

For acoustic model training a microphone-based high-quality speech corpus was used. This 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[4]. Special consideration was given to the sentences selection, since they provide a representative distribution of the more frequent triphone sequences (reflecting their relative occurrence in natural speech). The corpus was recorded in the office where only the speaker was present. The recording sessions yielded totally about 220 hours of speech.

4 Experimental setup

4.1 Acoustic processing

The digitization of an analogue signal was provided at 22.05 kHz sample rate and 16-bit resolution format. The aim of the front-end processor was 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 [5] for methodology). The best results were achieved using PLP parameterization [6] with 27 filters and 12 PLP cepstral coefficients with both delta and delta-delta sub-features (see [7] for details). Therefore one feature vector contained 36 coefficients. Feature vectors were computed each 10 milliseconds (100 frames per second).

4.2 Acoustic modeling

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 the 4922 states. The prime single mixture triphone acoustic model trained by Maximum Likelihood (ML) criterion was made using HTK-Toolkit v.3.4 [8]. Further, three 16 mixtures models were trained from the prime model: gender-independent, male and female. The training procedure has two stages. At first, 16 mixtures models were trained with HTK using ML criterion. At second, final models were obtained via two iterations of MMI-FD discriminative training [9, 10].

4.3 Gender based splitting

As was presented in [9], the splitting via manual male/female markers need not to be optimal due to several ”masculine” female and ”feminine” male voices occurring in the training corpora and also because of possible errors in manual annotations. Therefore, an initial splitting (achieved via manual markers) was realigned via automatic clustering algorithm. After this process, two more acoustically homogeneous classes were available for gender-dependent acoustic modeling which was described in previous subsection.

4.4 Test conditions

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. This 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)[11] was used for training. The model contains 186k words and the perplexity of the recognition task was 12362 and OOV was 2.4% (see [12] and [13] for details).

5 Results

To follow up our last year paper [9], the same three acoustic models were used: gender-independent (GI), male and female. At first, all these models were tested stand alone. At second, all switching and fusion method were evaluated. All the results are in table 1.

Table 1. The results of recognition experiments

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

From the table 1 it is clear that *Multiply* and *Frame_argmax* methods gave even higher WER than GI model. On the other hand, some methods gave significantly lower WER than GI. The lowest WER has been obtained via *W_sum_exp* method and its relative WER reduction is 2% absolutely and more than 11% relatively.

Proper setting of α parameter is needed for methods with the exponential forgetting. For all these methods the optimal value range was very similar. 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.

6 Conclusion

Various methods of employing gender-dependent acoustic models to the LVCSR system were tested in this paper. The methods had to be designed for real-time automatic subtitling task which is connected to the live TV audio stream. Three switching and four fusion methods were proposed, described and tested. 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 models (method marked as *W_sum_exp*) and its relative WER reduction is 2% absolutely and more than 11% relatively. All proposed methods are able to combine even higher number of acoustic models than they were tested with.

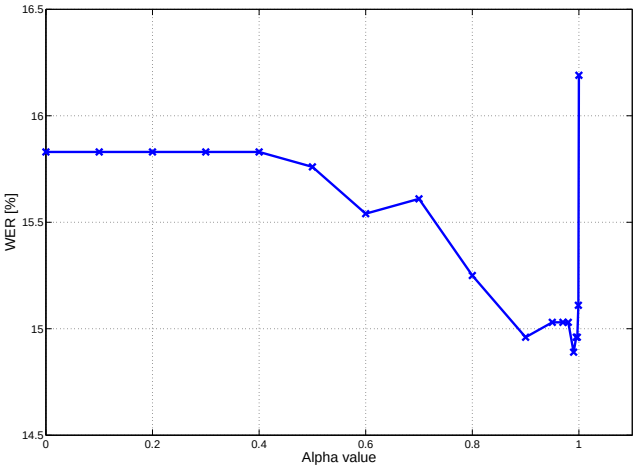

Fig. 1. *Relation of α value and WER.*

7 Acknowledgements

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

References

1. Pražák, A. and Psutka, J. and Hoidekr, J. and Kanis, J. and Müller, L. and Psutka, J. : Automatic online subtitling of the Czech parliament meetings . Lecture Notes in Artificial Intelligence, Lecture notes in artificial intelligence. 0302-9743 ; 4188, 4188, p. 501-508, Springer, Berlin, 2006.
2. Olsen, P.A., Dharanipragada, S. : An efficient integrated gender detection scheme and time mediated averaging of gender dependent acoustic models, In: 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003),Geneva, Switzerland, 2003.
3. Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D. : Broadcast News Subtitling System In Portuguese. In: Proceedings of the ICASSP, Las Vegas, USA, 2008.
4. Radová, V. and Psutka, J., UWB-S01 Corpus: A Czech Read-Speech Corpus, Proceedings of the 6th International Conference on Spoken Language Processing ICSLP2000, Beijing 2000, China.
5. 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.
6. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. J. Acoustic. Soc. Am.87, 1990.
7. Psutka, J. Robust PLP-Based Parameterization for ASR Systems. In SPECOM 2007 Proceedings. Moscow : Moscow State Linguistic University, 2007.
8. S. Young et al.: The HTK Book (for HTK Version 3.4), Cambridge, 2006.

9. Vaněk, J. and Psutka, J.V. and Zelinka, J. and Pražák, A. and Psutka, J. : Discriminative training of gender-dependent acoustic models. Lecture Notes in Artificial Intelligence, Lecture Notes in Artificial Intelligence, 5729, p. 331-338, Springer, Berlin , 2009.
10. Vanek J.: Discriminative training of acoustic models. Ph.D. thesis, West Bohemia University, Department of Cybernetics, 2009. (in Czech)
11. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: International Conference on Spoken Language Processing (ICSLP 2002), Denver, USA, 2002.
12. 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.
13. 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.