Speaker-clustered Acoustic Models Evaluated on GPU for on-line Subtitling of Parliament Meetings

Mr. Blind

Blinded.
email@email,
WWW home page: http://www

Abstract. This paper describes the effort with building speaker-clustered acoustic models as a part of the real-time LVCSR system that is used more than one year by the Czech TV for automatic subtitling of parliament meetings broadcasted on the channel ČT24. Speaker-clustered acoustic models are more acoustically homogeneous and therefore give better recognition performance than single gender-independent model or even gender-dependent models. Frequent changes of speakers and a direct connection of the LVCSR system to the audio channel require an automatic switching/fusion of models as quickly as possible. An important part of the solution is real time likelihood evaluations of all clustered acoustic models, taking advantage of a fast GPU(Graphic Processing Unit). The proposed method achieved a WER reduction to the baseline gender-independent model over 2.34% relatively with more than 2M Gaussian mixtures evaluated in real-time.

1 Introduction

Recently, we introduced the system for automatic subtitling of the 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 [9], [12] and [11]). An unpleasant problem that accompanies the automatic speech recognition of deputies is frequent and sometimes very rapid changes of speakers. It complains about the use of the on-line speaker adaptation techniques, which need relatively a longer part of speech for adaptation. To avoid using common speaker adaptation techniques we suggested a method, which operates simultaneously with several speaker clustered acoustic models. The fast model switching/fusion makes possible to ”adapt” the ASR system on-line to a new voice within a few frames. An increase in computational demands during calculations of output likelihoods was eliminated by using the GPU(Graphic Processing Unit).

2 Methods

2.1 Unsupervised clustering

Recently, we describe [13] an automatic clustering algorithm that divides speakers by voice into homogeneous classes. It is based on iterative k-means-like approach composed from three main steps. The algorithm starts with partitioning of initial training

data into a predefined number of clusters (randomly or using any relevant prior information). In the first step, the acoustic models for the individual clusters are trained or adapted. The second step consists of a criterion calculation for all training data across all models. The last step of the iteration is the data (uttered sentences in our case) reassignment to the cluster with the best criterion value. If the percentage of cross-assigned utterances in all the clusters decreases bellow some predefined value, the iteration process is stopped.

For more than two clusters, the hierarchical or direct variant of the algorithm can be applied. In the hierarchical case, the initial data are split into two clusters and the actual cluster with the largest criterion is than split into another two clusters. This process continues until the target number of clusters is achieved. The main advantage of this approach is the speed. Because only a small part of the data is processed in after-initial splitting steps. The disadvantage of this approach is a possibility to create "gaps" between the binary-tree branches. In contrast, the direct approach splits the whole training set directly into target number of clusters. It is very computationally intensive. The computational demands grow linearly per iteration with the number of clusters. Moreover, the required number of iterations usually grows as well. However, the quality of the models set produced by direct approach is slightly higher. Disadvantage of the direct approach next to the computational intensity is a higher sensitivity to the initial data partitioning.

Various criteria that are used for acoustic model training can be also used for the clustering. It could be the traditional Maximal Likelihood criterion (ML) but also discriminative criteria, e.g. Maximum Mutual Information (MMI) or Minimum Phone Error (MPE). After obtaining the training data from the clustering algorithm (list of sentences in our case), the final models set has to be trained. Appropriate training techniques were examined in [10] and [13]. The best performance was achieved using the discriminative criterion. If the final number of clusters is small and the clusters are large enough, the full discriminative training procedure is appropriate. In the case of the higher number of clusters with some relatively empty clusters, the discriminative adaptation techniques are helpful. The models do not differ so much but their parameters are estimated robustly.

2.2 Acoustic models fusion

Recently, various techniques for acoustic models switching/fusion were proposed (see [11] for details). All presented techniques were designed for the real-time applications therefore only a small history for actual processed frames is needed. Results of an extensive experimental work suggest that the best solution is a weighted sum with exponential forgetting. This method can be written in the form of

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

where $P_k(s_i|\mathbf{o}_t)$ is an output probability of the state s_i of the k-th acoustic model, $\hat{P}(s_i|\mathbf{o}_t)$ is the new evaluated state's probability and M is number of acoustic models. The weights in the time t are computed as

$$w_t^k = \frac{\alpha P_{t-1}(\lambda_k) + (1-\alpha) P(\lambda_k|\mathbf{o}_t)}{\sum_{l=1}^{M} \alpha P_{t-1}(\lambda_l) + (1-\alpha) P(\lambda_l|\mathbf{o}_t)}. \quad (2)$$

where α parameter is set to 0.95 and where $P_0(\lambda_k)$ are set to zero for all acoustic models k.

2.3 GPU accelerated acoustic model evaluation

The computation of acoustic model likelihoods accounts for the largest processing part in automatic speech recognition systems. We use GPU cards to offload this computation-intensive part from CPU. Our optimized algorithm efficiently exploits the GPU [14]. The efficiency together with the high GPU performance enable us to evaluate very large acoustic models much faster than in real-time. Evaluation of several acoustic models together is now possible even on low-end or laptop GPUs.

Our implementation splits the acoustic model evaluation into the data-parallel blocks. The individual blocks evaluate 8 or 16 feature-vectors together with 64 tied-states which are based on Gaussian mixture models with diagonal covariance matrix. The number of evaluated feature-vectors is a trade-off between the recognition system delay and the system performance. In the case of a small number of feature vectors, the overhead of CPU-GPU communication together with GPU-kernel management is significant. Fully asynchronous GPU handling is used for the maximum total system performance. Two likelihood-buffers are prepared. One is used by the decoder and the other is asynchronously filled by the results of evaluation of the next feature-vector window. In the most cases, the GPU evaluation is faster than decoder part. Therefore, CPU-implemented decoder is a bottleneck of the entire system and total recognition speed depends on the decoder performance.

3 Train data description

3.1 Annotated data

The training corpus consists of two different parts. The first one contains 100 hours of parliament speech records collected till 2010. All this data has been manually annotated and carefully revised. However, after the parliament election in 2010, more than 50% of the parliament members were changed.

3.2 Unsupervised data

The second part of the corpus contains 300 hours of speech. This huge part of parliamentary speeches was collected from deputies elected in 2010. There were no manual transcriptions available for this data. But the shorthand records of all Parliament meetings must be (by law) available for public use on the Internet. Unfortunately, these shorthand records are amended to avoid slips of the tongue and to meet the grammatical rules, so they do not meet the demands for exact transcriptions suitable for acoustic model training. Anyway, we can use these transcriptions of the individual meetings to create the meeting-specific language models by combination of the language

model trained from the meeting transcriptions (dynamic LM) with the standard language model (static LM). Static language model was trained on about 27M tokens of the normalized Czech Parliament meeting transcriptions (Chamber of Deputies only) from different electoral periods. Dynamic language models were trained on meeting transcriptions containing from 3k to 100K tokens. Resulted trigram language model with modified Kneser-Ney smoothing was trained by SRI Language Modeling Toolkit [8]. This approach reduced the word error rate of the recognized transcriptions to about 50 %.

Since the recognition process was not error-free, some technique for confidence tagging of the recognized words was used to choose only well-recognized segments of the speech that were taken into the acoustic model training process. We used the posterior word probabilities computed on the word graph as a confidence measure [15]. To use only the trustworthy segments of the speech, we applied a quite strict criterion for the word selection - only the words, which had confidence greater than 0.99 and their neighboring words with the same confidence (greater than 0.99), were selected. This ensures that the word boundaries of selected words are correctly assigned for retraining of acoustic model.

4 Experimental setup

4.1 Acoustic processing

The digitization of an analogue signal was provided at 44.1 kHz sample rate and 16-bit resolution format. The front-end worked with PLP parameterization [5] with 27 filters and 12 PLP cepstral coefficients with both delta and delta-delta sub-features (see [6] for details). Therefore one feature vector contains 36 coefficients. Feature vectors are computed each 10 milliseconds (100 frames per second). Cepstral mean normalization (CMN) was used in order to reduce the effect of constant channel characteristics.

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 the Czech triphones is large, phonetic decision trees were used to tie the states of the 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 48 mixtures of multivariate Gaussians for each of 5385 states. A silence model was trained by borrowing the most relevant Gaussians from all non-speech HMMs in proportion to their state and mixture occupancies. Thus the resulting silence model contained 253 mixtures on average per state. The prime 48 Gaussians triphone acoustic model trained by Maximum Likelihood (ML) criterion was made using HTK-Toolkit v.3.4 [7]. At second, final models were obtained via two iterations of MMI-FD discriminative training [10] or [2].

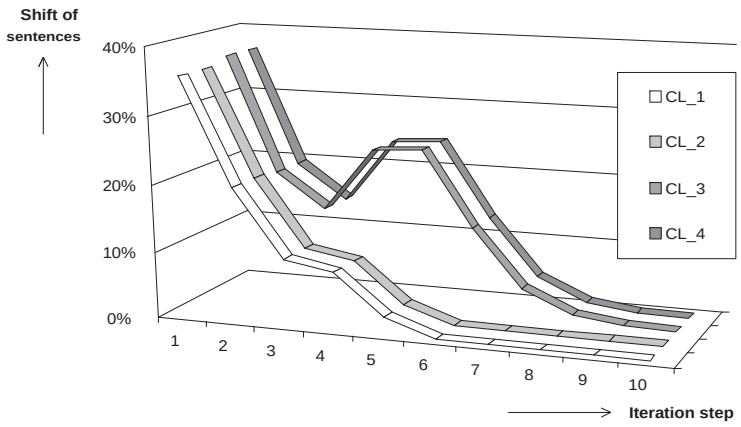

Fig. 1. *An example of behaviour of stopping criterion ($4Cl_h$)*

4.3 Unsupervised speaker clustering

As was presented in [11], by splitting via manual male/female markers and creating the gender-based acoustic models we achieved a significant gain in terms of the recognition accuracy than a simple speaker-independent acoustic model. This method is the most popular method how to split training data into two more acoustically homogeneous classes [4], [1]. As can be seen in [13] the increasing number of the speaker-clusters brings even better recognition results in comparison with the gender-dependent acoustic models. The whole training corpus was split hierarchically into two, four and eight acoustically homogeneous classes via the algorithm introduced in the Subsection 2.1. However, in all cases, the initial splitting was achieved randomly because additional speaker/sentence information was unavailable (especially for the second part of the training corpus). The stopping criterion in all presented experiments was based on the shift between clusters (sentences, which were moved from the one cluster to another in the consecutive steps) less than 1% of sentences. An expample of such criterion can be seen on the Figure 1.

4.4 Tests description

The test set consists of 1 hour of the special part of the parliament meetings - interpellation. Interpellation is the right of a parliament to submit questions (oral or formal) to the government. This part of the parliamentary speeches was chosen because of the limited time for one speaker (2 minutes per question). This portion of speech contains 15 different speakers (11 male and 4 female) since each speaker has the right to submit several questions. All recognition experiments were performed with a trigram language models with modified Kneser-Ney smoothing that was trained by SRI Language Modeling Toolkit [8]. The language model was trained on about 27M tokens of normalized Czech

parliament transcriptions. The model contained 192k words with OOV amounting 4%. The perplexity of the recognition task was 547.

5 Results

In all our experiments, the word error rate (WER) as well as clustering criterion (Maximal Likelihood criterion and Maximum Mutual Information) were evaluated. We tried only the hierarchical division methods to split all the training sentences (175k) into two, four, or eight clusters. This type of division method was used according to former experiments (see [13] for details).

The hierarchical division method means that we divided the training set into two classes ($2Cl_h$) and than each class was split again into another two classes ($4Cl_h$) and so on (finally we had eight clusters $8Cl_h$). The recognition results as well as some parameters which describe the clustering criterion are shown in the Table 1.

Table 1. Recognition results

	WER [%]		ML Criterion	MMI
	Ideal	Real		
baseline	13.70		-	-
$2Cl_h$	13.18	13.53	68.49	-3.2860
$4Cl_h$	12.86	13.47	68.68	-3.2231
$8Cl_h$	12.60	13.38	68.83	-3.1809

The column *Ideal WER* shows the recognition results for the ideal off-line recognition, where the tests were performed on the list of single speaker sentences across all clustered acoustic models. On the other hand the column *Real WER* shows the recognition results for the real-time recognition, where the fusion of all clustered acoustic models was applied. From the obtained results we can see that the best recognition result was achieved for eight clusters ($8Cl_h$). The number of on-line evaluated Gaussians in this case is more than 2M.

6 Conclusion

The goal of this paper was to describe our work with building speaker-clustered acoustic models in a real-time LVCSR system for automatic subtitling of Parliament meetings that are broadcasted on the TV channel ČT24. To be able to use the speaker-clustered acoustic models in the task of on-line recognition of parliamentary speeches with the frequent changes of speakers, we suggested the fast switching/fusing method of acoustic models evaluation. This approach works with the support of a GPU unit and is able to enumerate 2M Gaussian mixtures in real-time. The proposed method achieved a WER reduction by more than 2% relatively compared to the baseline gender-indendent model.

7 Acknowledgments

This research was supported by some grant.

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. 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.
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. 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.
5. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. J. Acoustic. Soc. Am.87, 1990.
6. Self-citation.
7. S. Young et al.: The HTK Book (for HTK Version 3.4), Cambridge, 2006.
8. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: International Conference on Spoken Language Processing (ICSLP 2002), Denver, USA, 2002.
9. Self-citation.
10. Self-citation.
11. Self-citation.
12. Self-citation.
13. Self-citation.
14. Self-citation.
15. Wessel F, et al. : Confidence measures for large vocabulary continuous speech recognition. IEEE Transactions on Speech and Audio Processing, 2001.