MODEL COMPLEXITY SELECTION AND CROSS-VALIDATION EM TRAINING FOR ROBUST SPEAKER DIARIZATION

Xavier Anguera[1,2] *, Takahiro Shinozaki*[1,3,4]*, Chuck Wooters*[1] *and Javier Hernando*[2]

[1] International Computer Science Institute, Berkeley, CA 94704, U.S.A.
[2] Technical University of Catalonia (UPC), 08034 Barcelona, Spain
[3] Department of Electrical Engineering, University of Washington, Washington, U.S.A
[4] Kyoto University, Kyoto, Japan

staka@u.washington.edu, {xanguera, wooters}@icsi.berkeley.edu

ABSTRACT

Accurate modeling of speaker clusters is important in the task of speaker diarization. Creating accurate models involves both selection of the model complexity and optimum training given the data. Using models with fixed complexity and trained using the standard EM algorithm poses a risk of overfitting, which can lead to a reduction in diarization performance. In this paper a technique proposed by the author to estimate the complexity of a model is combined with a novel training algorithm called "Cross-Validation EM" to control the number of training iterations. This combination leads to more robust speaker modeling and results in an increase in speaker diarization performance. Tests on the NIST RT (MDM) datasets for meetings show a relative improvement of 10.6% relative on the test set.

Index Terms— Speaker Diarization, speaker segmentation and clustering, complexity selection, cross-validation EM training.

1. INTRODUCTION

The task of speaker diarization involves the automatic segmentation and clustering of acoustic data into speakers, attempting to answer the question "who spoke when?" in an audio recording. It is usually done so without any prior information about the number of speakers or their identities. Agglomerative clustering is the most commonly used technique and is used in the system presented in this paper [1]. The system starts by creating many clusters from the input data, which are modelled using Gaussian Mixture Models (GMM), and then iteratively merges the closest pair of clusters (according to some defined metric) until a stopping criterion indicates that the optimum number of clusters has been reached. The GMM is trained using the Expectation Maximization (EM) algorithm [2]. In standard implementations of this algorithm ([3], [4]) the model complexity is chosen independently of the amount or nature of the acoustic data to be modeled and using a fixed number of EM iterations. Doing so, it is easy to cause the models to overfit to the data resulting in errors when comparing pairs of clusters. This is particularly the case when little data is available for training.

Several methods have been proposed to solve both of these problems. Methods such as the Bayesian Information Criterion (BIC) [5] or the Minimum Description Length (MDL) [6] for model complexity selection Data driven techniques such as cross-validation and bootstrapping [7] have been used for training. These methods are not necessarily the best solutions for speaker diarization. On one hand, BIC and MDL usually carry a large computational cost. On the other hand, the use of these algorithms for training is prone to instabilities in the mixtures placement when little training data is available.

In this paper we apply a recently proposed iterative training algorithm called Cross-Validation EM (CV-EM) to the model training of speaker diarization for meetings. CV-EM is introduced by T. Shinozaki in [8] for robust model training and applied to the task of large vocabulary speech recognition. CV-EM performs an iterative EM-like training where multiple models are trained on subsets of data. The portion of the data that is held out in estimating the cross validation model is used to determine whether to continue or stop training. This method is combined with a cluster complexity algorithm [9] to obtain robust models.

In section 2 we review the speaker diarization system used in this work. Then, in section 3 we describe the model complexity algorithm and in 4 we explain the CV-EM training algorithm in detail. Section 5 describes experiments to show the performance of these algorithms, and section 6 concludes.

Xavier Anguera was visiting ICSI within the Spanish visitors program at the time of this work.

Takahiro Shinozaki was at Department of Electrical Engineering, University of Washington and at ICSI during most of the time of this work

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

2. AGGLOMERATIVE SPEAKER DIARIZATION SYSTEM

As explained in [1], the speaker diarization system is based on an agglomerative clustering technique. It initially splits the data into K clusters (where K must be greater than the number of speakers and is chosen using the algorithm presented in [9]), and then iteratively merges the clusters (according to the ΔBIC metric described by [5] and modified by [3]) until a stopping criterion is met. Each cluster is modeled via a Gaussian Mixture Model (GMM) of variable complexity, chosen automatically.

The system modified for this paper works as follows:

1. When multiple channels are available, acoustic beamforming is used to combine the channels into a single "enhanced" channel.
2. Run speech/non-speech detection to eliminate non-speech regions and then extract acoustic features.
3. Estimate the number of initial clusters K and create cluster models. The complexity of the models is determined by the algorithm explained in section 3.
   (a) Run a Viterbi decode to resegment the data and retrain the models using the CV-EM algorithm presented in 4. Iterate between segmentation and training until the segmentation stabilizes.
   (b) Select the cluster pair with the largest merge score (based on ΔBIC) that is > 0.0.
   (c) If no such pair of clusters is found, stop and output the current segmentation.
   (d) Merge the pair of clusters found in step (b). The models for the individual clusters in the pair are replaced by a single, combined model and its complexity is recomputed.
   (e) Go to step (a).

This system does not require any external training data and has been developed with the goal of robustness to changes in the acoustics of the data, thus allowing it to easily port to new acoustic domains.

3. MODEL COMPLEXITY SELECTION

The acoustic models used to represent each cluster are a key part of the agglomerative clustering process. On the one hand, comparing the models is how it is decided whether two models belong to the same cluster, while on the other hand, the models are used in the decoding process to redistribute the acoustic data into the different clusters. When comparing two models via ΔBIC, if the models are too general, they tend to over-merge. If the models are too specific they under-merge. Therefore it is important to find the optimal number of mixtures to use, i.e. the model complexity.

We presented an algorithm in [9] that selects the number of mixtures based on the number of data frames assigned to the cluster. In the current work, we combine this approach with a variation of EM training in an attempt to obtain the optimum cluster models. The algorithm works as follows: whenever there is a change in the amount of data assigned to a cluster (normally due to a segmentation step), the number of acoustic frames that are assigned to the model is used to determine the new number of mixtures in the GMM using:

$$M_i^j = round(\frac{N_i^j}{CCR}) \quad (1)$$

where the number of Gaussian mixtures to model cluster i at iteration j (M_i^j) is determined by the number of frames belonging to that cluster at that time (N_i^j) divided by the Cluster Complexity Ratio (CCR), which is a constant value across all meetings.

When the desired model complexity changes, the new gaussians are created by either splitting the mixtures with largest weight (when $M_i^j > M_i^{j-1}$) or forgetting the current model and training it from "scratch".

4. CROSS-VALIDATION SPEAKER MODEL EM TRAINING

The EM algorithm is the most used iterative training method for training models that includes hidden variables. It has the disadvantage that there is no mechanism to avoid overfitting of the model to the data. As a result, the algorithm is sometimes unstable and can lead to overtraining depending on the structure of the models, especially for the cases when small training data is available.

This often happens when training Gaussian Mixture Models (GMM) as they are prone to instability. For example, a two-mixture Gaussian distribution gives large likelihoods for training data if one of the Gaussians covers only a single data point (and has a very small variance) and the other Gaussian spans the rest of the data points. The same phenomenon can occur, when the model trains too much to the data, losing generality. Therefore, it is important for the EM training to find the optimal number of iterations. However, the optimal number of iterations depends on the data and it is difficult to predict.

For these reasons we are experimenting with a new training algorithm called cross-validation EM (CV-EM), which is presented by T. Shinozaki in [8]. CV-EM uses cross-validation in the iterative process of EM, addressing the problems of overfitting and potential local maxima.

Figure 1 shows the CV-EM procedure. The system starts from an initial single model to be trained and finishes also with a single model. On the initial E-step of the EM processing the training data is split into N partitions as evenly as possible (in the speaker diarization using GMM models, each consecutive frame is assigned to a different partition sequentially until all frames have been assigned). Then the con-

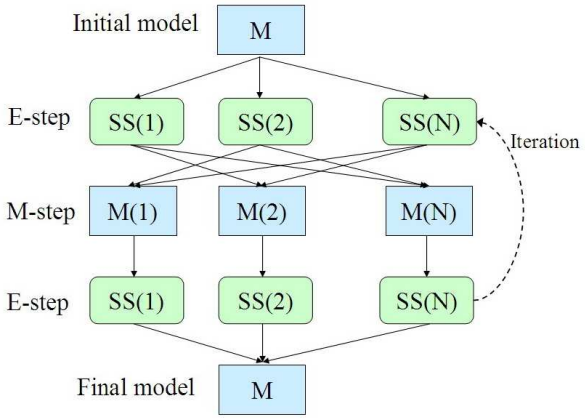

Fig. 1. *Cross-validation EM training algorithm*

ditional probability of each frame to each Gaussian mixture in the initial model is computed. This process is identical to the initial E-step in the technique called parallel EM training [10].

In the following M-step, each model M_i is reestimated using the sufficient statistics computed for all partitions except for SS_i, which is kept as cross-validation data (differing from the parallel-EM procedure). In the CV-EM algorithm, once all the N models have been reestimated, new conditional probabilities are computed for the frames in each partition SS_i using model M_i. As data in partition SS_i was not involved in the reestimation of the parameters in M_i, the accumulated likelihood from all partitions can be used as a check for convergence, avoiding overfitting to the data. Once convergence is reached, the current sufficient statistics computed for each of the subsets are used to derive a single output model.

In [8] CV-EM is applied to speech recognition using a fixed number of (five) iterations. In speaker diarization, convergence based on an increase in the likelihood is preferred in order to bound the likelihood variation between iterations of all models and therefore make them more comparable. In the implementation here, a likelihood increase of $\Delta\mathcal{L}_{inc} = 0.1\%$ is used.

The advantages of the CV-EM algorithm are threefold. While the EM algorithm iterates the E-step and the M-step using the same data, CV-EM uses different data subsets so that there is no overlap. In this way, CV-EM is more stable than EM with respect to overtraining since distributions highly specialized to a particular data point cannot produce a large likelihood during training. Another advantage of CV-EM is that the likelihood obtained in the E-step is more reliable than the optimistic likelihood in the EM training and can be used as a termination criterion for the training iteration. Because the likelihood is estimated using cross-validation, it decreases when the model looses generality. Therefore, a good termination criterion is to stop iterating when the likelihood decreases. Finally, the increase in computational cost of the CV-EM is small as only the sufficient statistics accumulation needs to be repeated for each of the cross-validation models.

5. EXPERIMENTS

To test the effectiveness of the proposed algorithms we use the ICSI speaker diarization system as described in section 2. As a baseline system we refer to the submission used in the RT06s evaluation, without the use of any purification [9], using linear cluster initialization and only acoustic MFCC-19 features. In this baseline system 5 iterations of standard EM was performed and model complexity was fixed to 5 initial Gaussians per cluster, with complexity accumulated when two clusters merge.

The development data is composed of the NIST RT02s, RT04s and RT05s [11] conference room datasets (26 meeting excerpts) and the test set is the RT06s eval data (8 meetings). Experiments have been run on the Single Distant Microphone (SDM) and Multiple Distant Microphone (MDM) tasks. The metric used in all cases is the Diarization Error Rate, defined by NIST as the percentage of misassigned time.

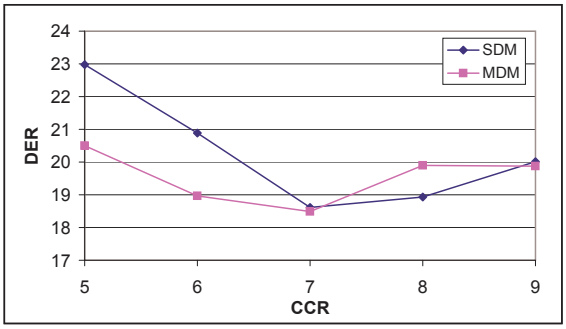

Fig. 2. *Model complexity selection DER changing the CCR parameter*

Figure 2 shows the effect of the Cluster Complexity Ratio (CCR) when performing model complexity selection on the development set. We can see that for both SDM and MDM cases the optimum value is located at $CCR = 7$. Using this optimum complexity setting we now substitute the standard EM training by the CV-EM algorithm. In order to determine the optimum number of cross-validation models used in CV-EM, we plot the DER for the range from 15 to 45 cross-validation models in figure 3.

The best average DER is obtained with 25 cross-validation models, but between 20 and 45 cross-validation models the differences in the average are less than 1%, which shows the robustness of the algorithm. With 15 models (or fewer), the errors grow as the data differs too much between models and therefore the training of the cross-validation models doesn't converge. In fact, each model differs from the others in 1/(N-1) parts of the total number of frames, which becomes impor-

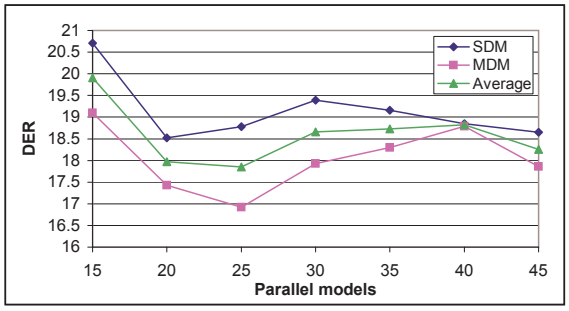

Fig. 3. *DER for different number of cross-validation models*

tant as N decreases below 15. This increase can also be due to the fragmentation of the training data being used.

Compared to the standard EM training (doing a fixed 5 iterations per training) now the training applies more iterations to models whose complexity has changed (for example after a merge) and does the minimum training for models containing almost the same data as in the previous iteration. Given that only a few models change every iteration, this brings a general speed-up to the system.

System	DER Devel		DER Test	
	SDM	MDM	SDM	MDM
Baseline	20.60%	19.04%	24.54%	26.50%
Complex select	18.61%	18.49%	28.84%	23.68%
CV-EM	19.58%	19.08%	22.70%	26.74%
Complex + CV-EM	18.78%	**16.92%**	25.00%	**23.68%**

Table 1. Comparison of the DER for the different techniques presented

Table 1 summarizes the results obtained on the development and test sets. Compared to the baseline both methods individually perform in mixed ways in the different tasks. Complexity selection outperforms the baseline in all except on test SDM, where the decrease in performance is mostly due to two shows from the same meeting room with an increase in error of over 80%. The CV-EM alone works well in general for SDM and is comparable to the baseline for MDM. When combining both methods the biggest improvements are in the MDM case, improving by 11.3% relative on the dev set and 10.6% on the test set. The SDM task also improves by 8.8% relative in the dev set but is slightly worse for the eval set.

6. CONCLUSIONS

In this paper two newly proposed techniques are presented to obtain robust speaker models for the task of speaker diarization. When doing speaker diarization via agglomerative clustering we need to robustly model the speakers given varying amounts of data, thus the complexity of each speaker model and its optimum training become important decisions. A simple (yet effective) cluster complexity algorithm based on the data size is combined with a recently proposed cross-validation EM training algorithm that does an ML training of the data while avoiding overfitting. We tested these two techniques using an extensive development set composed of 26 meeting excerpts from the NIST RT evaluations and a test set with 8 excerpts and find a relative 11.3% improvement on the dev set and 10.6% on the eval set, for the multiple distant microphones (MDM) case.

7. REFERENCES

[1] Xavier Anguera, Chuck Wooters, and Jose M. Pardo, "Robust speaker diarization for meetings: ICSI RT06s meetings evaluation system," in *RT06s Meetings Recognition Evaluation*, Washington DC, USA, May 2006.

[2] A. P. Dempster, N. M. Laird, , and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[3] Jitendra Ajmera and Chuck Wooters, "A robust speaker clustering algorithm," in *Proc. ASRU*, US Virgin Islands, USA, Dec. 2003.

[4] Xuan Zhu, Claude Barras, Sylvain Meignier, and Jean-Luc Gauvain, "Combining speaker identification and bic for speaker diarization," in *Proc. ICSLP*, Lisbon, Portugal, September 2005.

[5] Shaobing S. Chen and P.S. Gopalakrishnan, "Clustering via the bayesian information criterion with applications in speech recognition," in *Proc. ICASSP*, Seattle, USA, 1998, vol. 2, pp. 645–648.

[6] K. Shinoda and T.Watanabe, "Acoustic modeling based on the mdl criterion for speech recognition," in *Proc. Eurospeech*, 1997, vol. 1, pp. 99–102.

[7] T. Hastie, R. Tibshirani, , and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, 2001.

[8] Takahiro Shinozaki and Mari Ostendorf, "Cross-validation EM training for robust parameter estimation," *Proc. ICASSP*, 2007, submitted.

[9] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Automatic cluster complexity and quantity selection: Towards robust speaker diarization," in *MLMI'06*, Washington DC, USA, May 2006.

[10] Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland, *The HTK Book*, Cambridge University Engineering Department, 2005.

[11] "NIST spring rich transcription evaluation in meetings website, http://www.nist.gov/speech/tests/rt/rt2005/spring," 2006.