Discriminative Training for Hierarchical Clustering in Speaker Diarization

Oriol Vinyals[1,2], *Gerald Friedland*[2], *Nelson Morgan*[1,2]

[1]University of California, Berkeley, CA, USA
[2]International Computer Science Institute, Berkeley, CA, USA

vinyals@eecs.berkeley.edu, fractor@icsi.berkeley.edu, morgan@icsi.berkeley.edu

Abstract

In this paper, we propose a discriminative extension to agglomerative hierarchical clustering, a typical technique for speaker diarization, that fits seamlessly with most state-of-the art diarization algorithms. We propose to use maximum mutual information using bootstrapping i.e., initial predictions are used as input for retraining of models in an unsupervised fashion. This article describes this new approach, analyzes its behavior, and presents results on the official NIST Rich Transcription datasets. We show an absolute improvement of 4 % DER with respect to the generative approach baseline. We also observe a strong correlation between the original error and the amount of improvement, that is, the better our predicted labels are, the more gain we obtain from discriminative training, which we interpret as a strong indication for the high potential of the extension.

Index Terms: Discriminative learning, Maximum Mutual Information, Speaker Diarization

1. Introduction

The goal of Speaker Diarization is to segment audio into speaker-homogeneous regions trying to answer the question “who spoke when?”. Speaker diarization has utility in any application where multiple speakers may be expected. Examples include audio and speaker indexing, information retrieval, speaker verification (in the presence of multiple or competing speakers), to assist with speech-to-text transcription (via speaker-dependent modeling) and, more generally, rich transcription (RT).

Most state-of-the-art systems use a combination of agglomerative hierarchical clustering with Bayesian Information Criterion (BIC) and Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs), usually done in a top-down approach, that is, to overcluster our observation $x_1^T = X$, merge clusters and stop based on some model selection criterion. The training of such systems is usually performed using the Expectation Maximization (EM) algorithm, where we update of the parameters of our model θ as to maximize the conditional probability of the observed data X given the hypothesized labels S, $p(X|S;\theta)$ (i.e. likelihood). This model is generative as, given a set of labels, it tries to explain or generate the observations X as accurately as possible.

In this paper, we propose to change that paradigm and to, instead, optimize adding discriminative methods to our inference procedure. The rest of the article is organized as follows: Section 2 presents the agglomerative clustering approach used as baseline as well as related work, Section 3 provides some necessary background, and motivates the use of a particular discriminative objective function, in Section 4 we provide results on a large dataset, and give a justification on the observed performance gain/loss. We conclude with Section 5, in which we express our conclusions on the work.

2. Speaker Diarization Overview

As already explained in Section 1, the goal of speaker diarization is to segment audio into speaker-homogeneous regions with the ultimate goal of answering the question “who spoke when?” [1]. In contrast to speaker recognition or identification, speaker diarization attempts to use no prior knowledge of any kind; in particular, usually no specific speaker models are trained for the speakers that are to be identified in the recording. This means a speaker diarization system has to answer the following questions in an unsupervised manner:

- What are the speech regions?
- How many speakers are in the recording?
- Which speech regions belong to the same speaker?

The speaker diarization engine that we developed uses an agglomerative clustering approach to perform both segmentation of the audio track into speaker-homogeneous time segments, and the grouping of these segments into speaker-homogeneous clusters in one step, as detailed in [2].

The algorithm is initialized using a higher amount of clusters than speakers assumed in the audio track. An initial segmentation is generated by partitioning the audio track into k segments of the same length. Using the initial segmentation, Gaussian Mixture Models (GMMs) for each clusters are trained. A minimum duration of 2.5 seconds is assumed for each speech segment. Viterbi alignment is then used to combine the individual decisions via an ergodic HMM. The algorithm then performs the following loop:

- Re-Segmentation: Compute the optimal segmentation for the given models (e.g. Viterbi path).
- Re-Training: Given the new segmentation of the audio track, compute new Gaussian Mixture Models for each of them using maximum likelihood training. Transitions are not retrained, and are fixed to provide uniform probabilities to jump to any state in the HMM.
- Cluster Merging: Given the new Gaussian Mixture Models, find the two models that most likely represent the same speaker. This is done by computing the BIC score (Bayesian Information Criterion) of each of the models and the BIC score of a new GMM trained on the merged segments for two clusters. If the BIC score of the merged Gaussian Mixture Model is smaller than or equal to the

Copyright © 2010 ISCA

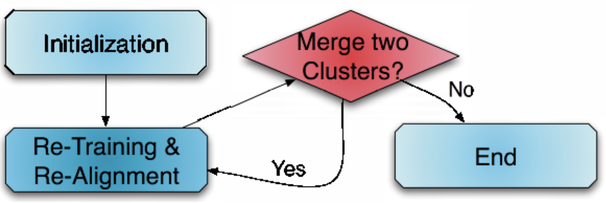

Figure 1: The agglomerative clustering approach of the ICSI Speaker Diarization Engine as explained in Section 2 and in [2]. Retraining and re-segmentation ends when no more models can be merged as of the BIC score. At the end, the number of clusters is hoped to be equal to the number of speakers.

sum of the individual BIC scores, the two models are merged and the algorithm loops at the re-segmentation using the merged Gaussian Mixture Model. If no pair is found, the algorithm stops. Note that when merging two mixture models, the number of parameters (i.e. complexity) is maintained. Thus, the ΔBIC score is just the likelihood ratio (or log likelihood difference) in this case.

Figure 1 illustrates the steps of the algorithm.

The output consists of meta-data describing speech segments in terms of starting time, ending time, and speaker cluster name. This output is usually evaluated against manually annotated ground truth segments. A dynamic programming procedure is used to find the optimal one-to-one mapping between the hypothesis and the ground truth segments so that the total overlap between the reference speaker and the corresponding mapped hypothesized speaker cluster is maximized. The difference is expressed as Diarization Error Rate which is defined by NIST[1]. The Diarization Error Rate (DER) can be decomposed into three components: misses (speaker in reference, but not in hypothesis), false alarms (speaker in hypothesis, but not in reference), and speaker-errors (mapped reference is not the same as hypothesized speaker).

3. Maximum Mutual Information in Speaker Diarization

Maximum Mutual Information (MMI) techniques have been used successfully in the speech community. Such techniques have been applied both on speech recognition [3, 4] and speaker recognition [5] for GMM training using MMI instead of ML, and in speaker diarization via the information bottleneck approach [6], where mutual information is used to directly cluster the data. Our approach involves a different objective function in place of ML, namely the mutual information between the observations X and true labels S. However, in diarization, the true labels are not known, and because of that a bootstrapping technique is used where, at each iteration, we assume that the current segmentation is the correct one. A major advantage of the method proposed here is that is can be easily used to extend current speaker diarization algorithm that are based on agglomerative hierarchical clustering.

As stated in the previous section, the algorithm proposed here ensures the update of the model to provide better likelihood in each of the steps: GMM training via ML, Viterbi alignment with the new models, and BIC based decision that ensures that, by merging two models with positive ΔBIC score, the likelihood grows (as we keep the number of parameters fixed throughout the iterative procedure).

Let x_1^T denote the observations of our sequence, and let s_1^T be the hidden state sequence. Note that the cardinality of s is unknown, and thus the agglomerative clustering will keep merging clusters until the delta BIC score of any two pairs is no longer positive. Our algorithm is finding a set of parameters θ_{ML} such that:

$$\theta_{ML} = \arg\max_{\theta} p(x_1^T|s_1^T;\theta)$$

is maximized.

This maximum likelihood approach will fit the models so that they represent the data with high fidelity. However, in the case of speaker recognition, or speaker diarization, many similarities will be observed between speakers (given, for examples, that both of them are males), and thus many degrees of freedom of our GMM model will be used to capture information that is redundant in regards of identifying and separating speakers by their acoustic dissimilarities. Thus, a discriminative approach similar that the one proposed in [4] will help enhance the performance. The objective function then becomes:

$$\theta_{MMI} = \arg\max_{\theta} p(s_1^T|x_1^T;\theta) = \arg\max_{\theta} \frac{p(x_1^T|s_1^T;\theta)}{p(x_1^T|\theta)^f}$$

Note that, in the MMI formulation there is a normalization factor that appears in the denominator in comparison with ML (which is controlled by a parameter $0 \leq f \leq 1$ such that $f = 0$ corresponds to ML and $f = 1$ corresponds to full MMI). In order to compute $p(x_1^T|\theta)$ we would need to use the total probability theorem, that is, to sum over all possible state sequences s_1^T. The sum is not tractable for the typical size of our data (T is on the order of tens of thousands), therefore we chose to use the approximation made in [3], which assumes that each time frame x_i is independent given θ. This is not true in our case as our HMM model enforces a minimum state duration that does not allow for any transition to happen at any given time. With this simplification, we obtain (dropping θ in all terms):

$$E_{\theta,S|X}[\log p(x_1^T)] = \sum_t E_{\theta,s_t|x_t}[\log p(x_t)] \equiv$$

$$\equiv \sum_{t=1}^{T}\sum_{j=1}^{S} p(s_t=j|x_t)\log(p(x_t|s_t=j)p(s_t=j)) \quad (1)$$

This reformulation of our baseline algorithm to maximize the MMI function instead of ML also makes use of the findings on the usage of MMI for speech recognition in [4]. For purposes of space, we will not derive the complete formulation discussed there, but the modified version of the algorithm to update the gaussian mean of state j and mixture m becomes:

$$\mu_j' = \frac{\theta_j^{num} - f\theta_j^{den} + D\mu_j}{\gamma_j^{num} - f\gamma_j^{den} + D}$$

$$\theta_j^{num} = \sum_t p(s_t=j|x_1^T)x_t$$

$$\theta_j^{den} = \sum_t p(s_t=j|x_t)x_t$$

$$\gamma_j^{num} = \sum_t p(s_t=j|x_1^T)$$

[1] http://nist.gov/speech/tests/rt/rt2004/fall

$$\gamma_j^{den} = \sum_t p(s_t = j|x_t)$$

Our starting point are the ML estimates, based solely on numerator statistics, proceeding with the update of the parameters using modified Baum Welch [4], while if we set $D = 0$ the maximization step becomes the standard gradient descend for MMI as in [3]. θ_{jm}^{num} refers to the sum of the observations, weighted by occupancy $p(s_t = j|x_1^T)$, for mixture component m of state j, and γ_{jm}^{num} are the Gaussian occupancies summed over time. Similarly for the denominator, except that the occupancies here become $p(s_t = j|x_t)$. D is a parameter that selects how much we want to diverge from the initial ML estimate, and is empirically set as in [4] as 5 times the maximum value of D that would be needed for all variances to be positive (this is found by solving a quadratic equation). Lastly, f is a parameter that we tuned to select how much we want to keep from the initial ML estimate as the objective function (i.e. the objective function becomes an interpolation between ML and MMI).

Note that we can interpret the MMI updates to be the same as ML, except that we discount observations for which $p(s_t = j|x_t)$ is high (that is, the posterior probability is already good for the given set of parameters, or a sample that should not align well with the model does (thus, we are adding discriminative power to the model)), and we give more importance to data samples that are hard to classify locally, and negative weight to those points that should not align with the current state but do. Furthermore, there is a heuristic that we used, which is to discount only if the assignment given by $\max_j p(s_t = j|x_t)$ does not correspond to the assignment given by $\max_j p(s_t = j|x_1^T)$ (i.e., the Viterbi output does not match the local decision taken by the models).

Lastly, in the original formulation BIC score is used to decide whether we should merge two clusters, and what two clusters should be merged. BIC computes an approximation of the marginal likelihood, that is, to integrate out nuisance parameters:

$$BIC \approx p(x_1^T|s_1^T) = \int_\theta p(x_1^T|s_1^T, \theta)p(\theta)d\theta$$

As we discussed in the previous Section, this ensures that the ML objective function is going up at every merging point, since the number of parameters is maintained fixed across iterations (note that in this case, the ΔBIC score becomes just the likelihood function). However, in the new scenario, we need to merge based on whether the MMI function is increased or not, and select the cluster pair that increases the objective function the most. Thus, we propose a modified BIC criterion as follows:

$$BIC_{ML}(H_0)(i,j) \propto$$
$$\propto -\log p(x_i|s_i) - \log p(x_j|s_j) + \lambda(k_i + k_j)\log N$$
$$BIC_{ML}(H_1)(i,j) \propto$$
$$\propto -\log p(x_{i\bigcup j}|s_{i\bigcup j}) + \lambda(k_i + k_j)\log N$$
$$\Delta BIC_{ML}(i,j) = BIC_{ML}(H_1)(i,j) - BIC_{ML}(H_0)(i,j) =$$
$$= \log p(x_i|s_i) + \log p(x_j|s_j) - \log p(x_{i\bigcup j}|s_{i\bigcup j})$$

where $p(x_{i\bigcup j}|s_{i\bigcup j})$ is the likelihood of a jointly trained model with the union of samples from clusters i and j, k_i is the number of parameters for model i, N is the total number of samples, H_0 correspond to the non merging hypothesis, and H_1 correspond to the merging hypothesis. Note that since we keep the number of parameters constant, the second term of the BIC criterion vanishes.

For the MMI case, we obtain:

$$\Delta BIC_{MMI}(i,j) = BIC_{MMI}(H_1)(i,j) - BIC_{MMI}(H_0)(i,j) =$$
$$= \log \frac{p(x_i|s_i)}{p(x_i)^f} + \log \frac{p(x_j|s_j)}{p(x_j)^f} - \log \frac{p(x_{i\bigcup j}|s_{i\bigcup j})}{p(x_{i\bigcup j})^f}$$

With this modifications, the algorithm is guaranteed to find a maximum of the mutual information between X and S, rather than the ML of X given S. In the next section, we report how various parameters affect the Diarization Error Rate (DER) of the system on various datasets.

System	Diarization Error Rate
Baseline	17.58%
MMI	15.40%
MMI+mod. BaumWelch	13.72%

Table 1: Results on the Dev07 NIST RT development set. Comparison of two main approaches to train MMI, previously used in speech recognition.

System	Diarization Error Rate
Baseline 07	19.11%
MMI+mod. BaumWelch 07	15.8%
Baseline 09	29.13%
MMI+mod. BaumWelch 09	24.31%

Table 2: Results on the Eval07/09 NIST RT development set.

4. Experimental Results

In this section, we discuss some of the choices made to select an optimal set of parameters, as well as what data was used for development and what data was used for test. In particular, we did not tune the D parameter, and we set it to be 5 times the maximum D needed for all variances to be positive (which works well for speech recognition), we kept the number of iterations done to update means and variances with starting point the ML estimates to 6. We tuned the f parameter, and we also used the heuristic described in the previous section (i.e. just to discount if the local alignment does not match the Viterbi alignment) as well as the update function using modified Baum Welch [4], or just plain MMI [3] (D=0).

The data used was what we define as Dev07 data, which contains several past NIST evaluations (2004/2005/2006), and a total of 21 meetings. Results can be seen in Table 1 for the development set, where the best parameter f was selected for with and without using the most recent approach with modified Baum-Welch [4]. We tested our proposed modification in the recent 2007 and 2009 Evaluation set (there was none for 2008), as can be seen in Table 2. Note that there is a significant improvement by using this technique in the Dev07 dataset, and that it also outperforms other methods such as a non parametric Bayesian approach like the HDP-HMM. In Figure 2 we see the error rate for three systems (baseline, MMI, and MMI with modified Baum Welch). All of those have the heuristic turned on, as it consistently performed better.

Furthermore, on the evaluation set there is also improvement. However, the Eval09 set was difficult because of presence of overlapped speech (for which our algorithm simply fails),

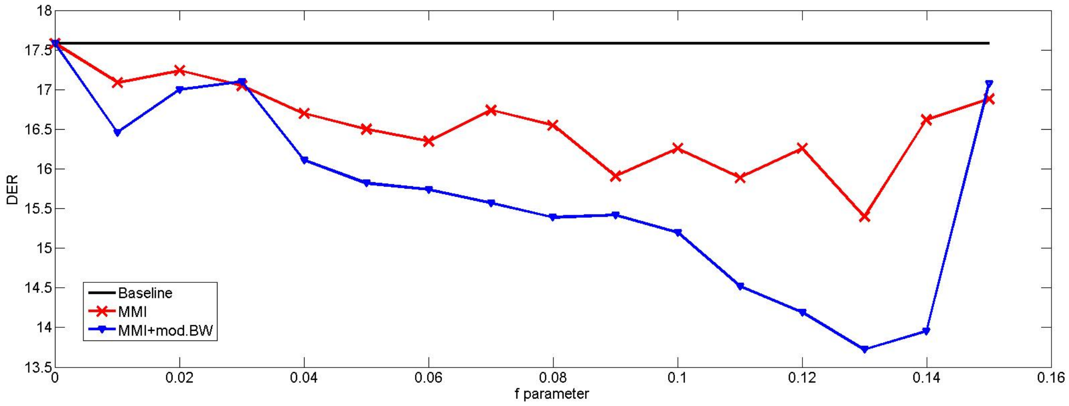

Figure 2: DER as a function of the parameter f for three systems: baseline, MMI, and MMI with extended Baum-Welch. Observe that, in general, the behavior is as expected, with an optimal parameter f around 0.13.

Baseline Performance	Contrib. to DER improvement
Good quality (DER $< 25\%$)	38%
Bad quality (DER $> 25\%$)	62%

Table 3: Contribution to DER improvement based on baseline performance (i.e. bootstrapping quality).

and thus the error due to speech activity detection is already 10%. Also, some of the meetings had a very poor performance as some contained many speakers and fast speaker turns. Therefore, as we use bootstrapping to add discrimination to diarization, applying discrimination on noisy labels may not only not be beneficial but be harmful. Surprisingly, this is not the case for this task, as can be seen in Table 3, where the decomposition of errors is shown for meetings in which our approach does better than 25% (and thus we can expect bootstrapping to be robust), and for those that the baseline performance is worse than 25%. Although the relative improvements are generally higher for meeting with lower absolute DER, the contribution to the total DER gain comes mostly from meetings that were inherently hard (i.e. high baseline DER). It must be noted that the number of meetings is small, and thus the results shown in Table 3 may be affected by statistical noise.

5. Conclusion and Future Work

In this paper we explored a method for extending current agglomerative hierarchical clustering-based speaker diarization systems using discriminative training. In particular, we changed the agglomerative clustering objective function from ML (generative) to MMI (discriminative), which has been applied successfully to other speech tasks. Initial experiments and results show significant improvement and the method seems to complement the current algorithms well. A limit of the approach is that some smoothing parameters need to be tuned in order for the model to provide better results, usually through training with a development set or knowledge transfer from other machine learning tasks. It must also be noted that MMI techniques are computationally more expensive than ML, and thus clustering time was roughly three times slower than ML baseline.

One line of future research includes the tuning of parameters that were fixed due to time limitations, such as the D parameter, or the number of training iterations for MMI. It would also be useful to explore other discriminative functions instead of MMI, and to extend this work to other unsupervised clustering techniques that could make use of bootstrapping and discriminative training to enhance their prediction accuracy.

6. Acknowledgements

This work was partially supported by the Swiss National Science Foundation's IM2 Research network project.

7. References

[1] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proceedings of the IEEE ICASSP*, 2005.

[2] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Proceedings of the RT07 Meeting Recognition Evaluation Workshop*, 2007.

[3] L.R. Bahl, M. Padmanabhan, D. Nahamoo, and P. S. Gopalakrishnan, "Discriminative training of gaussian mixture models for large vocabulary speech recognition systems," in *In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing*, 1996, pp. 613–616.

[4] P.C. Woodland and D. Povey, "Large scale discriminative training for speech recognition," 2000.

[5] F. Korkmazski and B.-H. Juang, "Discriminative adaptation for speaker verification," 1996.

[6] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard, "Agglomerative information bottleneck for speaker diarization of meetings data," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2007, IDIAP-RR 07-31.