INTERSPEECH 2009 BRIGHTON

A Study of New Approaches to Speaker Diarization

Douglas Reynolds[1] *, Patrick Kenny*[2]*, Fabio Castaldo*[3]

[1]MIT Lincoln Laboratory, USA
[2]CRIM, Canada
[3]Politecnico di Torino, Italy

dar@mit.ll.edu, patrick.kenny@crim.ca, fabio.castaldo@polito.it

Abstract

This paper reports on work carried out at the 2008 JHU Summer Workshop examining new approaches to speaker diarization. Four different systems were developed and experiments were conducted using summed-channel telephone data from the 2008 NIST SRE. The systems are a baseline agglomerative clustering system, a new Variational Bayes system using eigenvoice speaker models, a streaming system using a mix of low dimensional speaker factors and classic segmentation and clustering, and a new hybrid system combining the baseline system with a new cosine-distance speaker factor clustering. Results are presented using the Diarization Error Rate as well as by the EER when using diarization outputs for a speaker detection task. The best configurations of the diarization system produced DERs of 3.5-4.6% and we demonstrate a weak correlation of EER and DER,

Index Terms: Diarization, speaker recognition, segmentation, clustering, factor analysis, eigenvoice, variational Bayes

1. Introduction

[1] Audio diarization is the process of annotating an input audio channel with information that attributes (possibly overlapping) temporal regions of signal energy to their specific sources. These sources can include particular speakers, music, background noise sources and other signal source/channel characteristics. Diarization systems are typically used as a preprocessing stage for other downstream applications, such as providing speaker and non-speech annotations to text transcripts or for adaptation of speech recognition systems. In this work we are interested in improving diarization to aid in speaker recognition tasks where the training and/or the test data consists of speech from more than one speaker. In particular we focus on two speaker telephone conversations and multi-microphone recorded interviews as used in the latest NIST Speaker Recognition Evaluation (SRE)[2].

This paper reports on work carried out at the 2008 JHU Summer Workshop examining new approaches to speaker diarization. Four different systems were developed and experiments were conducted using data from the 2008 NIST SRE. Results are presented using a direct measure of diarization error (Diarization Error Rate) as well as the effect of using diarization outputs for a speaker detection task (Equal Error Rate). Finally we conclude showing the relation of DER to EER and summarize the effective components common to all systems.

[1]This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

[2]See http://www.nist.gov/speech/tests/sre/2008/ for more details.

2. Diarization Systems

Four systems were developed for the 2008 JHU Summer Workshop. The systems range from a baseline agglomerative clustering system, to a new system based on variation Bayes theory, to a streaming audio clustering system and a new hybrid system using elements of the baseline system and newly developed speaker factor distances.

2.1. Agglomerative Clustering System (Baseline)

The baseline system represents the framework of most widely used diarization systems [1]. It consists of three main stages.

In the first stage speaker change points are detected using a Bayesian Information Criterion (BIC) based distance between abutting windows of feature vectors. The features for the baseline system consist of 13 cepstral coefficient (including c0) with no channel normalization. This technique searches for change points within a window using a penalized likelihood ratio test of whether the data in the window is better modeled by a single distribution (no change point) or two different distributions (change point). If a change is found, the window is reset to the change point and the search restarted. If no change point is found, the window is increased and the search is redone. Full covariance Gaussians are used as distribution models.

The purpose of the second stage is to associate or cluster segments from the same speaker together. The clustering ideally produces one cluster for each speaker in the audio with all segments from a given speaker in a single cluster. Hierarchical, agglomerative clustering with a BIC based stopping criterion is used consisting of the following steps:

0. Initialize leaf clusters of tree with speech segments.
1. Compute pair-wise distances between each cluster.
2. Merge closest clusters.
3. Update distances of remaining clusters to new cluster.
4. Iterate steps 1-3 until stopping criterion is met.

The clusters are represented by a single full covariance Gaussian. Since we have prior knowledge of two speakers present in the audio, we stop when we reach two clusters.

The last stage is iterative re-segmentation with GMM Viterbi decoding to refine change points and clustering decisions. Additionally, a form of Baum-Welch re-training of speaker GMMs using segment posterior-weighted statistics can be used before a final Viterbi segmentation. This step was inspired by the Variation Bayes approach and is also referred to as ”soft-clustering.”

2.2. Variational Bayes System

This is one of the new systems developed during the workshop and is based on the Variational Bayes method of speaker diarization described by Valente [2]. Work on this system was motivated by the desire to build on the success of factor analysis methods in speaker recognition and to capitalize on some of the advantages a Bayesian approach may bring to the diarization problem (e.g., EM-like convergence guarantees, avoiding premature hard decisions, automatic regularization).

To build on the factor analysis work, we begin by using an eigenvoice model to represent the speakers. The assumption in eigenvoice modeling is that supervectors[3] have the form

$$\boldsymbol{s} \quad = \quad \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}.$$

Here $\boldsymbol{s}$ is a randomly chosen speaker dependent supervector; $\boldsymbol{m}$ is a speaker independent supervector (i.e., UBM); $\boldsymbol{V}$ is a rectangular matrix of low rank whose columns are referred to as eigenvoices; the vector $\boldsymbol{y}$ has a standard normal distribution; and the entries of $\boldsymbol{y}$ are the speaker factors. From the point of view of Bayesian statistics, this is a highly informative prior distribution as it imposes severe constraints on speaker supervectors. Although supervectors typically have tens of thousands of dimensions, this representation constrains all supervectors to lie in an affine subspace of the supervector space whose dimension is typically at most a few hundred. The subspace in question is the affine subspace containing $\boldsymbol{m}$ which is spanned by the columns of $\boldsymbol{V}$.

In the Variational Bayes diarization algorithm, we start with audio file in which we assume there are just two speakers and a partition of the file into short *segments*, each containing the speech of just one of the speakers. This partitioning need not be very accurate. A uniform partition into one second intervals can be used to begin with; this assumption can be relaxed in a second pass.

We define two types of posterior distribution which we refer to as *speaker posteriors* and *segment posteriors*. For each of the two speakers, the speaker posterior is a Gaussian distribution on the vector of speaker factors which models the location of the speaker in the speaker factor space. The mean of this distribution can be thought of as a point estimate of the speaker factors and the covariance matrix as a measure of the uncertainty in the point estimate. For each segment, there are two segment posteriors q_1 and q_2; q_1 is the posterior probability of the event that the speaker in the segment is speaker 1 and similarly for speaker 2.

The Variational Bayes algorithm consists in estimating these two types of posterior distribution alternately as explained in detail in [3]. At convergence, it is normally the case that q_1 and q_2 takes values of 0 or 1 for each segment but q_1 and q_2 are initialized randomly so that the Variational Bayes algorithm can be thought of as performing a type of soft speaker clustering, as distinct from the hard decision making in the agglomerative clustering phase of the baseline system.

The Variational Bayes algorithm can be summarized as follows:

Begin:

- Partition the file into 1 second segments and extract Baum Welch statistics from each segment
- Initialize the segment posteriors randomly
- No initialization is needed for the speaker posteriors

On each iteration of Variational Bayes:

- For each speaker s:
  - Synthesize Baum Welch statistics for the speaker by weighting the Baum Welch statistics for each segment by the corresponding segment posterior q_s
  - Use the synthetic Baum Welch statistics to update the speaker posterior
- For each segment:
  - Update the segment posteriors for each speaker

End:

- Baum Welch estimation of speaker GMM's together with iterative Viterbi re-segmentation (as in the baseline system)

In the Variational Bayes system, 39 dimensional feature vectors derived from HLDA transforms of 13 cepstra (including c0) plus single, double and triple deltas are used. The cepstra were processed with short-term (300 frame) Gaussianization. For the re-segmentation, 13 un-normalized cepstra (c0-c12) were used. The eigenvoice analysis used a 512 mixture GMMs and 200 speaker factors.

2.3. Streaming Systems

In this section we describe another way to combine speaker diarization and join factor analysis. Speaker diarization using factor analysis was first introduced in [4] using a stream-based approach. This technique performs an on-line diarization where a conversation is seen as a stream of fixed duration time slices. The system operates in a causal fashion by producing segmentation and clustering for a given slice without requiring the following slices. Speakers detected in the current slice are compared with previously detected speakers to determine if a new speaker has been detected or previous models should be updated.

Given an audio slice, a stream of cepstral coefficients and their first derivatives are extracted. With a small sliding window (about one second) a new stream of speaker factors (as described in the previous section) is computed and used to perform the slice segmentation. The dimension of speaker factor space is quite small (about twenty) with respect to the number used for speaker recognition (about three hundred) due to the short estimation window.

In this new space, a clustering of the speaker factors stream is done obtaining a single multivariate Gaussian for each speaker. A BIC criterion is used to determine how many speaker there are in the slice. A Hidden Markov Model (HMM) using the Gaussian for each state associated to a speaker is built and through the Viterbi algorithm a slice segmentation is obtained.

In addition to the segmentation, a Gaussian Mixture Model (GMM) in the acoustic space is created for each speaker found in the audio slice. These models are used in the last step, slice clustering, where we determine if a speaker in the current audio slice was present in previous slices, or is a new one. Using an approximation to the Kullback-Leibler divergence, we find the closest speaker model built in previous slices to each speaker model in the current slide. If the divergence is below a threshold

[3]The term *supervector* is used to refer to the concatenation of the mean vectors in a Gaussian mixture model.

the previous model is adapted using the model created in the current slice, otherwise the current model is added to the set of speaker models found in the audio.

The final segmentation and speakers found from the on-line processing can further be refined using Viterbi re-segmentation over the entire file.

2.4. Hybrid System

The last system was motivated by other work at the 2008 JHU Workshop [5] that showed good speaker recognition performance could be obtained using a simple cosine-distance between speaker factor vectors. The idea is to use the baseline agglomerative clustering system to build a tree up to a certain level where the nodes contain sufficient data, then to extract speaker factors for these nodes and continue the clustering process using the cosine distance.

The critical factor in determining when to stop the initial clustering is the amount of speech in each node since we wished to work with 200 speaker factors. Two approaches were used for stopping the initial clustering: level cutting and tree searching. Level cutting consists of merely running the clustering until a preset number of nodes exist - typically 5-15. The tree searching consists of building the entire tree, then searching the tree from the top-down to select the set of nodes that had at least a preset amount of speech.

As with the other systems, a final Viterbi re-segmentation is applied to refine the diarization.

3. Experiment Design

3.1. Data

Summed channel telephone data from the 2008 SRE was used for diarization experiments with the above systems. This data was selected since we could derive reference diarization, needed for measuring DER, by using time marks from ASR transcripts produced on each channel separately. In addition the data corresponded to one of the speaker detection tasks in the 2008 SRE, so we could measure the effect of diarization on EER. The test set consists of 2215 files of approximately five minutes duration each ($\approx$ 200 hours total). To avoid confounding effects of mismatched speech/non-speech detection on the error measures, all diarization systems used a common set of reference speech activity marks for processing.

3.2. Measures of Performance

As mentioned earlier, we used two measures of performance for the diarization systems. The diarization error rate (DER) is the more direct measure which aligns a reference diarization output with a system diarization output and computes a time weighted combination of miss, false alarm and speaker error [4]. Since all systems use reference speech activity marks, miss and false alarm, which only are affected by speech activity detection, are not used. Speaker error measures the percent of time a system incorrectly associates speech from different speakers as being from a single speaker. In these results we report the average and standard deviation DER computed over the test set to show both the average as well as the variation in performance for a given system.

To measure the effect of diarization on a speaker detection task, we used the diarization output in the recognition phase of one of the summed-channel telephone tasks from the 2008 SRE. In the *3conv-summed* task, the speaker models are trained with three single channel conversations and tested with a summed channel conversation. The diarization output is used to split the test conversation into two speech files (presumably each from a single speaker) which are scored separately and the maximum score of the two is the final detection score. A state-of-the-art Joint Factor Analysis (JFA) speaker detection system developed by Loquendo [6] is used for all diarization systems. Results are reported in terms of the equal error rate (EER).

[4] DER scoring code available at www.nist.gov/speech/tests/rt/2006-spring/code/md-eval-v21.pl

4. Results

In Table 1 we present DER results for some key configurations of the diarization systems. Overall we see that the final Viterbi re-segmentation significantly helps all diarization systems. For the baseline system, it was further seen that the soft-clustering, inspired by the Variational Bayes system, reduces the DER by almost 50%. The Variational Bayes system achieves similarly low DER when a second pass is added that relaxes the first pass assumption of fixed one second segmentation. The streaming system had the best performance out of the box, with some further gains with the non-causal Viterbi re-segmentation. Disappointingly, the hybrid system did not achieve performance better then the original baseline. This may be due to the first stage baseline clustering biasing the clusters too much or the inability to reliably extract 200 speaker factors from the small amounts of speech in the selected clusters.

Table 1: *Mean and standard deviation of diarization error rates (DER) on the NIST 2008 summed channel telephone data for various configurations of diarization systems.*

	mean DER (%)	σ (%)
Baseline + Viterbi	6.8	12.3
Baseline + soft-cluster + Viterbi	3.5	8.0
Var. Bayes	9.1	11.9
Var. Bayes + Viterbi	4.5	8.5
Var. Bayes + Viterbi + 2-pass	3.8	7.6
Stream	5.8	11.1
Stream + Viterbi	4.6	8.8
Hybrid + Viterbi (level cut)	14.6	17.1
Hybrid + Viterbi (tree search)	6.8	13.6

Lastly, in Figure 1 we show EER for the 3conv-summed task for different configurations of the above diarization systems. The end point DER values of 0% and 35% represent using reference diarization and no diarization, respectively. We see that there is some correlation of EER to DER, but this is relatively weak. It appears that systems with a DER $< 10\%$ produce EERs within about 1% of the "perfect" diarization. To sweep out more point with higher DER, we ran the baseline system with no Viterbi re-segmentation (DER=20%). While the EER did increase to 10.5% it was still better than the no-diarization result of EER=14.1%.

5. Conclusions

In this paper we have reported on a study of several diarization system developed during the 2008 JHU Summer Workshop. While each of the systems had a different approach to speaker diarization, we found that ideas and techniques proved out in one system were also able to be successfully applied to other

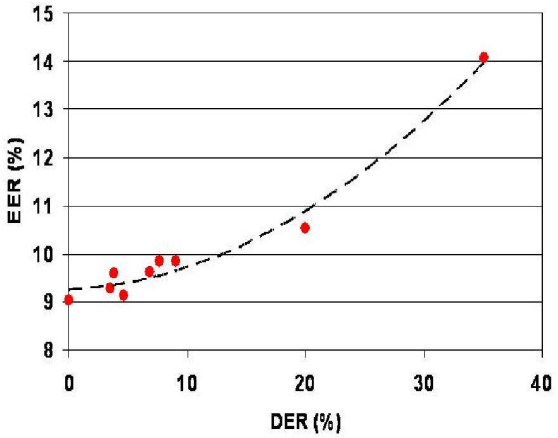

Figure 1: *EER vs DER for several diarization systems.*

systems. The Viterbi re-segmentation used in the baseline system, was a very useful stage for the other systems. Also the idea of soft-clustering from the Variation Bayes approach was incorporated into the agglomerative clustering baseline system to reduce the DER by almost 50%. The best configurations of the diarization systems produced DERs of 3.5-4.6% on summed-channel conversational telephone speech. We further examined the impact of using different diarization system with varying DERs on a speaker recognition task. While there was some weak correlation of EER to DER, it was not as direct as one would like in order to optimize diarization systems using DER independent of the recognition systems using the their output. In future work we plan on applying these diarization systems to the interview recordings in the 2008 SRE. This new domain will present several new challenges, including variable acoustics due to microphone type and placement as well as different speech styles and dynamics between face to face interviewer and interviewee.

6. References

[1] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, Sept. 2006.

[2] Fabio Valente, *Variational Bayesian methods for audio indexing*, Ph.D. thesis, Eurecom, Sep 2005.

[3] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," 2008.

[4] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *Proc. ICASSP*, Las Vegas, Nevada, Mar. 2008.

[5] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, and F. Castaldo, "Support vector machines and joint factor analysis for speaker verification," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009.

[6] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, , and P. Laface, "Loquendo - politecnico di torino's 2006 nist speaker recognition eva luation system," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007.