SPEAKER DIARIZATION OF FRENCH BROADCAST NEWS

Vishwa Gupta, Gilles Boulianne, Patrick Kenny, Pierre Ouellet, and Pierre Dumouchel

Centre de recherche informatique de Montréal (CRIM)

{Vishwa.Gupta, Gilles.Boulianne, Patrick.Kenny, Pierre.Ouellet and Pierre.Dumouchel}@crim.ca

ABSTRACT

We report results on speaker diarization of French broadcast news and talk shows on current affairs. This speaker diarization process is a multistage segmentation and clustering system. One of the stages is agglomerative clustering using state-of-the-art speaker identification methods (SID). For the GMMs used in this stage, we tried many different feature parameters, including MFCCs, Gaussianized MFCCs, Gaussianized MFCCs with cepstral mean subtraction, and Gaussianized MFCCs with cepstral mean substraction containing only frames with high energy. We found that this last set of feature parameters gave the best results. Compared to Gaussianized MFCCs, these features reduced the diarization error rate (DER) by 12% on a development set and by 19% on a test set. We also combined clusters resulting from Gaussianized and non-Gaussianized feature sets. This cluster combination resulted in another 4% reduction in DER for both the development and the test sets. The best DER we have achieved is 15.4% on the development set, and 14.5% on the test set.

Index Terms— speaker diarization, speaker segmentation and clustering, BIC clustering, SID clustering.

1. INTRODUCTION

Speaker diarization is the task of automatically partitioning an input audio stream into homogeneous segments and assigning these segments to sources. These sources generally include particular speakers, music, or background noise. The speaker diarization task is relative to a given show or audio file and there is no prior knowledge of the number of speakers involved. The speaker labels produced show which audio segments were spoken by the same speaker, but do not indicate the true identity of the speaker.

Speaker diarization has many applications. Some well-known applications include tracking speakers through various recordings, speaker-based indexing of data, speaker adaptation in speech recognition, etc. This paper focuses on speaker diarization of French broadcast news in Quebec. The speaker diarization is part of rich transcription for an *Assisted Indexation* project. The *Assisted Indexation* project is part of the E-Inclusion network research program. The goal of this research program is to create audio-video tools that will allow multi-media content producers to improve the richness of the multi-media experience for the blind, the deaf, the hard of hearing, and the hard of seeing, by automating key aspects of the multi-media production and post-production processes.

Recent work on speaker diarization for NIST Rich Transcription has primarily focused on broadcast news. Tranter and Reynolds [1] give a good overview of speaker diarization for broadcast news. Barras et al. [2] got good results on the French ESTER radio broadcast news evaluation data. This data consists of various radio broadcast news shows in France from 10 minutes to 1 hour in length. Here, we are working with French TV broadcasts on news, weather, finance, and talk shows on current affairs in Quebec. These shows vary from 45 minutes to 2 hours in length. The two-hour long talk shows on current affairs contain many speakers and a lot of background music.

In speaker recognition, Gaussianized MFCCs (also known as feature-warped MFCCs) [3] give lower error rates than MFCCs. These Gaussianized MFCCs have been successfully used for speaker diarization of broadcast news [2] [4]. Gaussianization normalizes the mean and variance in a 3 sec window. For this reason, it eliminates the need for cepstral mean subtraction before Gaussianization. That is why MFCCs are Gaussianized without any cepstral mean subtraction [2]. We have found that Gaussianization after real-time cepstral mean subtraction significantly reduces the diarization error rate (DER). This real-time cepstral mean is computed from a much larger window as a weighted average of prior frames, with the frames in the distant past getting an exponentially lower weight. The combination of this cepstral mean subtraction followed by Gaussianization reduced the DER by 8% for the development set, and by 11% for the test set.

We noticed that many errors were due to loud background music. Many speakers in each show are split into two clusters: one without music background and one with loud music background. To minimize these errors, we removed low energy frames, and only retained features from high energy frames. The rationale is that using only the high energy segments will reduce the masking effect due to music, and these segments are generally voiced, thus carrying more speaker-specific information. Using only the high energy frames for clustering reduced the DER by 5% for the development set and by 9% for the test set.

The paper is organized as follows: Sec. 2 gives the overview of the system, Sec. 3 describes the data used for the French broadcast news, Sec. 4 discusses the effect of relevant modules and the experiments carried out to optimize the modules. Sec. 5 gives the conclusions.

2. SPEAKER DIARIZATION SYSTEM OVERVIEW

A flowchart of our speaker diarization system is shown in Fig. 1. We first remove the silence, noise, and music segments from the audio. Silence is removed by an energy-based voice activity detector. The music and noise segments are removed with the help of GMMs for noise, music, speech, music+speech. The remaining speech and speech+music segments go through an acoustic change point detection step (CPD) that uses a symmetric Kullback-Leibler (KL2) metric, and a 13-dimensional feature vector (12 MFCCs + energy) with diagonal covariance matrix [5]. This is followed by an

This project is made possible with the support of the Canadian Heritage New Media Research Networks Fund, which promotes innovation in new media or interactive digital content that pertain to the Canadian cultural sector.

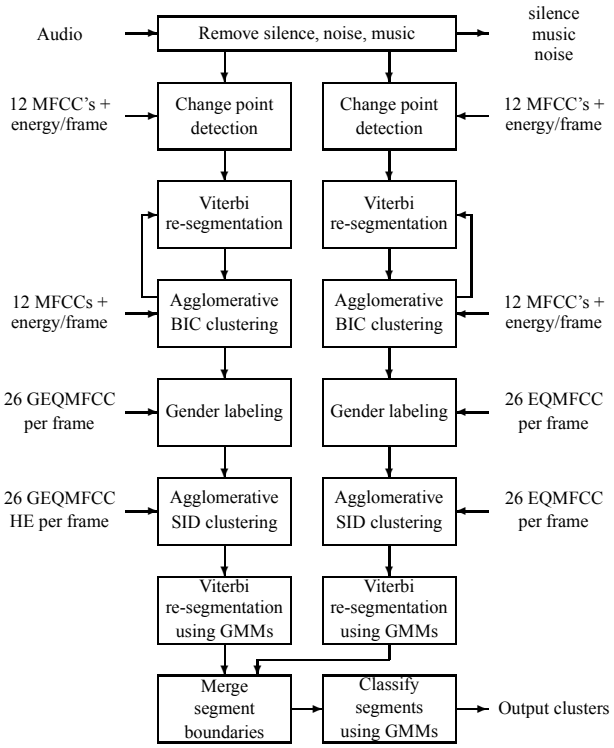

Fig. 1. *Multistage speaker diarization algorithm combining clusters from Gaussianized and non-Gaussianized features.*

iterative Viterbi re-segmentation stage that models each segment by its mean and variance and finds the optimal boundaries between segments. The resulting segments are clustered using BIC agglomerative clustering that uses a 13-dimensional feature vector (12 MFCCs + energy) with full covariance matrix [2]. In this step, the clustering threshold is set so as to under-cluster the segments. The Viterbi re-segmentation and BIC-clustering steps are iterated twice. The next stage is gender determination, which labels each cluster from the previous step as male or female. The next step is separate male/female speaker identification-style (SID) clustering that uses more complex models of the clusters for final clustering. For this step, we tried many different feature parameters in order to minimize the DER. This is followed by iterated Viterbi re-segmentation using adapted GMMs for each cluster. The final step merges the clusters from the Gaussianized and the non-Gaussianized features.

3. DATA SET FOR FRENCH BROADCAST NEWS

We recorded general news, weather news, news on finance, and talk shows on current affairs from French TV stations in Quebec. The development set consisted of six 45-minute news shows, one two-hour show on finance, two one-hour weather reports from a weather channel, and two two-hour talk shows on current news, for a total of 13 hours in the development set. The number of speakers in the audio files varied from 9 to 71, and the average number of speakers per file was 30. The two talk shows on current affairs had 60 and 71 speakers. We will refer to this development set as DEV. For the test set, we kept a similar mix for a total of 13 hours of audio from 11 shows. The number of speakers in the audio files varied from 5 to 63, and the average number was 30. For the test set also, the two talk shows on current affairs had 58 and 63 speakers. We will refer to this set as TEST.

Each audio file was segmented into speaker, music, noise, or advertisement segments by a transcriber. The audio sections corresponding to advertisements were skipped during evaluation. The audio portions with more than one person talking were marked as *don't care* (non-lexical) regions.

For training the GMMs used in the universal background models (UBMs), we took a total of 2.5 hours of speech from 925 different male segments (from several shows) to train GMMs for male speakers, and 1.5 hours of speech from 570 different audio segments to train GMMs for female speakers. The shows for the development and test set were telecast chronologically later than those used for training. Both the male and female GMMs contained 256 diagonal Gaussians.

4. EXPERIMENTS AND RESULTS

We carried out many experiments to measure DER on both the DEV and TEST data sets. The philosophy was to measure the effect on overall performance of the system when we perturb the parameters for one single module. In the text, we refer to the flowchart on the left as the Gaussianized system, and the flowchart on the right as the non-Gaussianized system.

4.1. Diarization Error Rate

The main metric of performance is the diarization error rate (DER) as defined by NIST in the RT-04 Fall evaluation [6]. The DER is the sum of three errors: missed speech (speech in the reference but not in the hypothesis), false alarm speech (speech in the hypothesis but not in the reference), and speaker match error (reference and hypothesized speakers differ). We used the md-eval-v17.pl Perl script from the NIST website to estimate this DER.

4.2. Gaussianized and non-Gaussianized Systems

Here, we outline in detail the features pertinent to this paper. As outlined in Sec. 2, the CPD algorithm [5] looks for a maximum in overlapping n second windows, and classifies this maximum as a change point if the KL2 metric exceeds a distance threshold. This scanning window length n is important, as it has a significant effect on the overall DER.

The GMMs used in SID agglomerative clustering and in Viterbi re-segmentation using GMMs are generated by adapting universal background models (UBMs) with the corresponding cluster data. For adaptation, we used variable-prior MAP adaptation (VP-MAP) [4] since this adaptation gave us the best results.

In agglomerative BIC clustering, the overall DER is sensitive to the λ used to compute the Bayesian Information Criterion (ΔBIC) [2] [4]. The optimal value of λ chosen was 3.5 in order to under-cluster the data. In SID agglomerative clustering, the DER was sensitive to the threshold δ [2] used for stopping the clustering process (optimal $\delta = 0.2$ for Gaussianized system, 1.8 for MFCCs and 0.6 for MFCCs after cepstral mean subtraction). With the optimized parameters for DEV, we got 16.2% DER for the best Gaussianized system, and 19.8% DER for the best non-Gaussianized system.

4.3. Feature Parameters

The features used in SID agglomerative clustering have a significant effect on the overall diarization error rate. We compared the overall

DER for the following features:

1. MFCC: 12 MFCCs plus the normalized energy plus their first differences (26 parameters every 10 msec).

2. GMFCC: Gaussianized MFCCs (26 parameters) using a moving window of 3 seconds. The entire audio file is Gaussianized.

3. GMFCC seg: Gaussianized MFCCs per segment (after BIC clustering stage) using a moving window of 3 seconds. Frames beyond the segment boundaries are not used for Gaussianization.

4. EQMFCC: MFCCs after real-time cepstral mean subtraction [7]. In real-time cepstral mean subtraction, we take μ_0 as the initial estimate of the cepstral mean vector, and update it as successive frames become available. This leads to the following cepstral mean subtraction procedure which we implemented with $\alpha = 0.005$:

$$\begin{aligned} \mu_t &= (1-\alpha)\mu_{t-1} + \alpha Y_t \\ X_t &= Y_t - \mu_t \end{aligned} \quad (1)$$

Here, Y_t represents the cepstral feature vector at time t and X_t the cepstral feature vector after subtracting the cepstral mean μ_t. Explicitly,

$$\mu_t = (1-\alpha)^t \mu_0 + \alpha \sum_{\tau=1}^{t} (1-\alpha)^{t-\tau} Y_\tau \quad (2)$$

so that cepstral mean μ_t is a weighted average of the initial estimate μ_0 and the observations up to time t, with the contribution of μ_0 decaying exponentially over time.

5. GEQMFCC: Gaussianization of EQMFCC using a moving 3 second window. The entire audio file is Gaussianized.

6. GEQMFCC HE: Gaussianization of EQMFCC frames with high energy only. An energy-based voice activity detector is used with a noise floor of -35 dB instead of -60 dB. Only the frames classified as speech by this voice activity detector are then used for Gaussianization using a 3-sec moving window. This process removed 19% of the speech frames in the DEV set, and 21% of the speech frames in the TEST set. (Note that Gaussianization after throwing away low energy frames is important. Throwing away frames after Gaussianization does not reduce the DER.) Since we are throwing away frames, we need to elaborate how the clustering is done and how we measure the DER. The male/female classification module before SID clustering uses for comparison all the frames in each segment. However, only the high energy frames for each segment take part in agglomerative SID clustering. After clustering, the resulting cluster segments are mapped back to the original segments (containing all the frames) for estimating the DER. It can happen that some segments do not have any high energy frames. These segments do not take part in the SID agglomerative clustering. However, they are associated with the same cluster as before SID clustering. If it happens that the entire cluster did not take part in SID clustering, then this cluster is given a separate speaker name. The module following SID clustering is Viterbi re-segmentation using GMMs. In this step, we use GEQMFCC features for realignment. The GEQMFCC HE features are only used during SID clustering.

The diarization error rate for the DEV and TEST sets corresponding to the various feature parameters is shown in Table 1. For the TEST set, we used the same values of λ and δ as those used for the DEV set. This Table shows the error rates for scanning window lengths of 1.3 and 1.7 secs. The LIMSI paper [2] and the Cambridge paper [4] use the Gaussianized features per segment (GMFCC seg). We got slightly better results when we Gaussianize the entire file (GMFCC). Compared to GMFCC, the Gaussianized cepstra after real-time mean subtraction (GEQMFCC) reduce the DER by 8% for the DEV set and by 11% for the TEST set.

Table 1. *DER for DEV and TEST sets using various feature parameters for scanning window lengths (SCW) of 1.3 and 1.7 secs. Note that the DER includes 0.3% missed speaker time and 1.9% false alarm speaker time for DEV set. For the TEST, the DER includes 1.0% missed speaker time and 1.7% false alarm speaker time.*

Feature	DEV set SCW 1.3s	DEV set SCW 1.7s	TEST set SCW 1.3s	TEST set SCW 1.7s
MFCC	23.7	24.2	24.8	29.2
GMFCC	18.7	18.5	18.6	19.7
GMFCC seg	19.4			
EQMFCC	20.5	19.8	18.0	17.8
GEQMFCC	17.7	17.1	16.6	17.4
GEQMFCC HE	**17.2**	**16.2**	**15.1**	**15.8**

The GEQMFCC HE parameters reduce the DER by 5% for the DEV set and by 9% for the TEST set compared to the GEQMFCC features. The reason for this is the significant amount of background music in all the shows. Many speakers in each show are split into two clusters: one without music background and one with loud music background. Gaussianizing only the loud portions for clustering reduces such errors. The real-time cepstral mean subtraction also reduces such errors.

4.4. Viterbi Re-segmentation using GMMs

For telephone conversations, we showed in [8] that Viterbi re-segmentation following SID clustering reduces the DER by 10%. We experimented with a similar module for broadcast news diarization. We used the adapted GMMs for each cluster to perform Viterbi re-segmentation again. We carried out iterative re-segmentation until convergence or for a maximum of 6 iterations. After each iteration, we re-computed the adapted GMMs using the new segment boundaries. The number of segments and their association to clusters was not changed. We also imposed a 1 second minimum duration for segment boundaries between any two consecutive segments. For the GMMs, we used Gaussianized EQMFCC (GEQMFCC) parameters. (We could not use the GEQMFCC HE parameters as they are computed for only high energy frames). For the DEV set, this module reduced the DER from 17.1% to 16.2% (5% reduction in DER). For the TEST set, this module reduced the DER from 16.9% to 15.1% (11% reduction in DER).

4.5. Merging Clusters from Gaussianized and non-Gaussianized Systems

For telephone conversations [9] we showed that by combining the clusters from the Gaussianized and non-Gaussianized systems, we could reduce the DER by 10% to 20%. We tried similar combination for broadcast news. The overriding principle in combining clusters from the two diarization systems is to keep the clusters common to both systems, since we have more confidence in the correct assignment of these common clusters. We generate VP-MAP adapted GMMs for these clusters. These GMMs are then used to re-classify the remaining segments. The remaining segments are the segments not common to the two systems. The cluster combination algorithm is explained in detail in [9]. A simple example of cluster merging is shown in Fig. 2. There are a few differences between cluster combination for telephone speech and that for broadcast news. For broadcast news, silence and music segments have been removed, so the clusters contain only speech segments. For

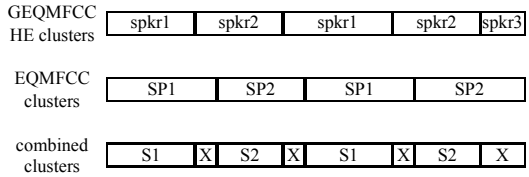

Fig. 2. *Example showing combination of clusters from Gaussianized and non-Gaussianized (EQMFCC) systems. Segments marked X in the combined cluster are reclassified using adapted GMMs for clusters S1 and S2.*

telephone conversations, we had merged clusters using the features MFCCs and Gaussianized MFCCs (GMFCC). For broadcast news, we get lower DER for GEQMFCC HE. Therefore, we tried cluster combination with GEQMFCC HE (Gaussianized features) and the EQMFCC (non-Gaussianized features). For rescoring, we used the GEQMFCC HE features. Since we can use GEQMFCC HE features only for the high energy regions, segments without any frames in high energy region were scored using GEQMFCC features (and the corresponding adapted GMMs).

Table 2 shows the results for various scanning window length combinations for the DEV set, and Table 3 shows the results for the TEST set. The reason for using different scanning window lengths for the Gaussianized and non-Gaussianized features is that we found that this strategy worked well for the telephone conversations [9]. From the Tables, we can see that the combined system gives DER roughly 4% lower than the DER for the best single Gaussianized system. For the DEV set, the DER goes down from 16.2% to 15.5%. For the TEST set, the DER goes down from 15.1% to 14.5%. For both the DEV and TEST sets, scanning window length of 1.3 secs for the Gaussianized system and 1.7 secs for the non-Gaussianized system gave the best results. The speaker match error reduced by 6% for both the DEV and TEST sets (2.2% of the errors in the DEV set and 2.8% of the errors in the TEST set are due to missed speech + false alarm speech).

For telephone conversations, we were able to reduce the DER by 10 to 20 percent by this cluster combination. However, for French broadcast news, the reduction in DER is only 4%. The reason may be the background music in broadcast news. Significant portions of the errors seem to be due to the loud background music. It results in a speaker split in two: one with background music, and one without background music. These errors cannot be corrected by cluster combination using two different features. Only soft errors due to random noise, etc., seem to be corrected by cluster merging using two different features. For this reason, the DER reduction by cluster merging is much higher for telephone conversations than for broadcast news.

Table 2. *Scanning window lengths (SWL) versus DER for GEQMFCC HE (G), EQMFCC (NG), and combined systems for DEV set.*

SWL G	SWL NG	DER G	DER NG	DER combined
1.7	1.3	16.2%	20.5%	15.8%
1.3	**1.7**	**17.2%**	**19.8%**	**15.5%**

5. CONCLUSIONS

We have applied state-of-the-art speaker diarization algorithms on French broadcast news and talk shows on current affairs. These algorithms are similar to the multistage segmentation and clustering systems [2] [4] used successfully in broadcast news. We added a Viterbi re-segmentation stage using GMMs that reduced the DER by 5% for DEV set and 11% for TEST set. We show that the choice of feature parameters has a significant impact on DER. Switching from Gaussianized MFCCs to Gaussianized MFCCs that have gone through real-time cepstral mean subtraction reduces the DER by 8% for the DEV set and by 11% for the TEST set. Using only the frames with high energy results in another 5% reduction in DER for the DEV set, and by 9% for the TEST set. Combining the clustering results from two independent speaker diarization systems: one using Gaussianized feature parameters (GEQMFCC HE) and the other using non-Gaussianized feature parameters (EQMFCC), results in another reduction of 4% in DER for both the DEV and TEST sets.

Table 3. *Scanning window lengths (SWL) versus DER for GEQMFCC HE (G), EQMFCC (NG), and combined systems for TEST set.*

SWL G	SWL NG	DER G	DER NG	DER combined
1.7	1.3	15.8%	18.0%	15.4%
1.3	**1.7**	**15.1%**	**17.8%**	**14.5%**

6. REFERENCES

[1] S. E. Tranter, and D. A. Reynolds, "An Overview of Automatic Speaker Diarization Systems", IEEE Trans. ASLP, vol. 14, no. 5, 1557–1565, 2006.

[2] C. Barras, X. Zhu, S. Meignier and J. Gauvain, "Multistage Speaker Diarization of Broadcast News", IEEE Trans. ASLP, vol. 14, no. 5, 1505–1512, 2006.

[3] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification", Proc. Odyssey Spkr Lang. Recog. Workshop, Crete, Greece, 2001, pp. 213–218.

[4] R. Sinha, S. E. Tranter, M. J. F. Gales and P. C. Woodland, "The Cambridge University March 2005 Speaker Diarisation System", Interspeech 2005, pp. 2437–2440.

[5] M. Siegler, B. Jain and R. Stern, "Automatic segmentation and clustering of broadcast news audio", Proc. DARPA Speech Recognition Workshop, Feb. 1997, pp. 97–99.

[6] NIST. Fall 2004 Rich Transcription (RT-04F) evaluation plan. Online: www.nist.gov/speech/tests/rt/rt2004/fall/docs/rto4f-eval-plan-v14.pdf

[7] P. Kenny, V. Gupta, G. Boulianne, P. Ouellet, and P. Dumouchel, "Feature normalization using smoothed mixture transformations", Interspeech 2006, pp. 25–28.

[8] V. Gupta, P. Kenny, P. Ouellet, G. Boulianne, and P. Dumouchel, "Multiple Feature Combination to Improve Speaker Diarization of Telephone Conversations", IEEE ASRU workshop Dec. 2007, pp. 705–710.

[9] V. Gupta, P. Kenny, P. Ouellet, G. Boulianne, and P. Dumouchel, "Combining Gaussianized/non-Gaussianized Features to Improve Speaker Diarization of Telephone Conversations", IEEE Sig. Proc. Letters, vol. 14, no. 12, Dec. 2007, pp. 1040–1043.