MULTIPLE FEATURE COMBINATION TO IMPROVE SPEAKER DIARIZATION OF TELEPHONE CONVERSATIONS

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

Centre de recherche informatique de Montréal (CRIM)

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

ABSTRACT

We report results on speaker diarization of telephone conversations. This speaker diarization process is similar to the multistage segmentation and clustering system used in broadcast news. It consists of an initial acoustic change point detection algorithm, iterative Viterbi re-segmentation, gender labeling, agglomerative clustering using a Bayesian information criterion (BIC), followed by agglomerative clustering using state-of-the-art speaker identification methods (SID) and Viterbi re-segmentation using Gaussian mixture models (GMMs). The Viterbi re-segmentation using GMMs is new, and it reduces the diarization error rate (DER) by 10%. We repeat these multistage segmentation and clustering steps twice: once with MFCCs as feature parameters for the GMMs used in gender labeling, SID and Viterbi re-segmentation steps, and another time with Gaussianized MFCCs as feature parameters for the GMMs used in these three steps. The resulting clusters from the parallel runs are combined in a novel way that leads to a significant reduction in the DER. On a development set containing 30 telephone conversations, this combination step reduced the DER by 20%. On another test set containing 30 telephone conversations, this step reduced the DER by 13%. The best error rate we have achieved is 6.7% on the development set, and 9.0% 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. In speaker diarization, 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 are *relative* to the audio recording. They show which audio segments were spoken by the same speaker, but do not indicate the true identity of the speaker.

This work was partly funded by the Canadian Department of National Defence.

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 telephone conversations. A potential application of speaker diarization of telephone conversations is the automated recording of target speakers. In general, law enforcement officials can get permission to record calls when a particular person is involved in these conversations. To respect the court order that only calls containing this speaker be recorded, speaker diarization followed by speaker identification are necessary steps. This process can eventually be followed by automated transcription of the audio segments of the speaker of interest.

A constraint imposed by our funder was that the speaker diarization task on telephone conversations should not be restricted to two speakers. For this reason, we concatenated telephone conversations to generate recordings with more than two speakers, and pursued algorithms that do not assume a fixed number of speakers.

Some initial work on speaker segmentation of telephone conversations was done at AT&T [5] on cutomer care conversations. Recent work on speaker diarization for NIST Rich Transcription has primarily focused on broadcast news. Tranter and Reynolds [6] give a good overview of speaker diarization for broadcast news. Several methods of combining different diarization systems exist. One example is the *piped* system [7] [8] where the segmentation from the CLIPS system is piped to the LIA system for better initialization. Another example is the cluster voting scheme [9] that combines the clusters from two speaker diarization systems. Here, we have merged the outputs of our diarization system using two different feature parameters to lower the diarization error rate (DER).

To get good speaker diarization results on telephone conversations, we implemented multi-stage speaker diarization that gave good results for broadcast news audio [1] [2]. The philosophy is to first use a fast acoustic change point detection algorithm that over-segments the data, followed by an iterative Viterbi re-segmentation to refine the segment boundaries. The ensuing BIC agglomerative clustering combines the segments into bigger clusters. These bigger clusters can then be

modeled by more complex models for further clustering. We added a Viterbi re-segmentation stage using GMMs to this multi-stage system in order to improve this system even further. This new stage reduced the overall diarization error rate (DER) by 10%. Another contribution here is in combining speaker clusters using two different feature parameters to get even lower DER.

In speaker recognition, Gaussianized MFCCs (also known as feature-warped MFCCs) [4] give lower error rates than MFCCs. These Gaussianized MFCCs have been successfully used for speaker diarization of broadcast news [1] [2]. We used these Gaussianized MFCCs as feature parameters for the GMMs used in gender labeling, SID clustering and in Viterbi re-segmentation using GMMs. We repeat these steps in parallel using MFCCs instead of Gaussianized MFCCs in gender labeling, in SID clustering, and in Viterbi re-segmentation using GMMs, and combine the resulting clusters from the two systems. This combination reduces the DER by another 10% to 20%. The combination steps are shown in Figs. 1 and 2. For the two separate clusterings of the acoustic data, we first find the common clusters. The common clusters are the cluster segments where the corresponding cluster labels match. For each cluster in these resulting common clusters, we generate a MAP-adapted GMM. For the remaining segments, we use these MAP-adapted GMMs to classify each segment as belonging to the cluster giving the highest likelihood. Overall, this combination reduced the DER for the development set by 20%, and for the test set by 13%.

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

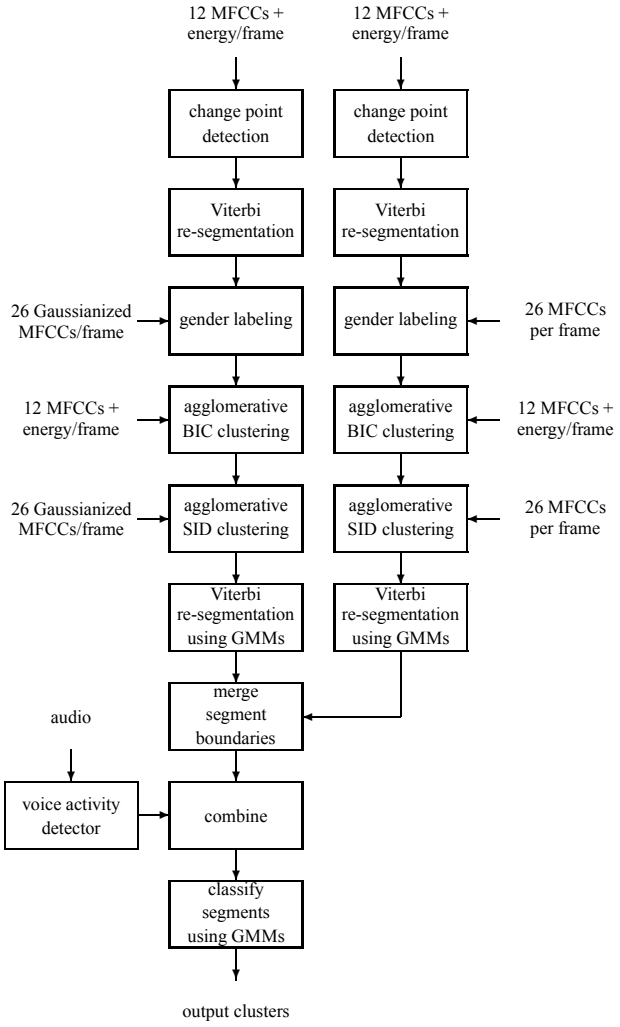

Fig. 1. *Multistage speaker diarization algorithm combining clusters from Gaussianized and non-Gaussianized features.*

2. SPEAKER DIARIZATION SYSTEM OVERVIEW

A flowchart of our speaker diarization system is shown in Fig. 1. It consists of 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 [3]. This is followed by an iterative Viterbi re-segmentation stage that models each segment by its mean and variance and finds the optimal boundaries between segments. The next stage is gender determination that labels each segment from the previous step as male or female. The resulting male/female segments are clustered separately using BIC agglomerative clustering that uses a 13-dimensional feature vector (12 MFCCs + energy) with full covariance matrix [1]. In this step, the clustering threshold is set so as to under-cluster the segments. The next step is separate male/female speaker identification-style (SID) clustering that uses more complex models of the clusters for final clustering. This is followed by iterated Viterbi re-segmentation using adapted GMM models for each cluster from the SID clustering stage.

The novelty here is the use of two different features in the GMMs used to carry out speaker diarization as shown in the left and right flowcharts of Fig. 1. The two separate features are 26 MFCCs (12 MFCCs + energy + their first differences), and their Gaussianized versions [4] using an incremental 3-sec window. The resulting clusters from the Gaussianized and non-Gaussianized features are then combined. The combination results in common clusters and audio segments that are marked for re-classification (see Fig. 2). We generate adapted GMMs (from male/female UBMs) for the common clusters, and classify the remaining segments using these cluster-adapted GMMs.

3. DATA SET FOR TELEPHONE CONVERSATIONS

For speaker diarization of telephone conversations, we need recorded telephone conversations with well-marked speaker segment boundaries. Such recordings are available from NIST RT-2004 conversational telephone speech (CTS) recordings. We only had access to the RT-2004 training set, not the development or the evaluation set. We took 30 conversations from the RT-2004 CTS training set and labeled them as a development set. We concatenated pairs of calls to create calls with four speakers per audio file. This was done in order to avoid tuning the algorithms to two speakers per audio file. We refer to this set as DEV2Calls. DEV2Calls contains 15 audio files of 20 minute duration each. We took another set of 30 calls from the RT-2004 CTS training set (disjoint from DEV2Calls) and created 15 audio files with 2 calls each. We call this set TEST2Calls. Two of these audio files had a common speaker in the two calls they contain. Therefore, 13 audio files have four speakers each, and two audio files have three speakers each. All the audio recordings use summed sides (a.k.a. two-wire).

We manually determined the gender of the speakers in another set of 25 calls from the RT-2004 CTS training set (disjoint from DEV2Calls and TEST2Calls), and used these as a training set for male/female Gaussian mixture models (GMM) used as universal background models (UBM). We call these audio files TRAIN. TRAIN contains 20 female and 30 male speakers, for a total of roughly 4 hours of speech.

4. EXPERIMENTS AND RESULTS

We carried out many experiments to measure DER on both the DEV2Calls and TEST2Calls 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 [10]. 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. Some of the details are also given in [11]. As outlined in Sec. 2, the CPD algorithm [3] 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 are generated by adapting universal background models (UBM) with the corresponding cluster data. The male/female UBMs with 256 diagonal Gaussians are trained on the TRAIN and the development data. For the development data, we used the segments labeled as male or female after the gender labeling step. For adaptation, we used variable-prior MAP adaptation (VP-MAP) [2] 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) [1] [2]. The optimal value of λ was 3.0 for the Gaussianized system, and 3.5 for non-Gaussianized system. In SID agglomerative clustering, the DER was sensitive to the threshold δ [1] used for stopping the clustering process (optimal δ = -0.05). With the optimized parameters for DEV2Calls, we got 8.4% DER for the Gaussianized system, and 8.3% DER for the non-Gaussianized system.

4.3. Viterbi Re-segmentation using GMMs

The initial change point detection algorithm is followed by Viterbi re-segmentation. The segment boundaries obtained from Viterbi re-segmentation step are carried over all the way to SID clustering. These segment boundaries were obtained with segment means and variances, while the SID clustering uses much more complex models. It seemed obvious that using the more complex models from the last stage of SID clustering should lead to better segment boundaries, so 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 Gaussianized system, this process reduced the overall DER for DEV2Calls from 9.5% to 8.4%. For the test set TEST2Calls also, Viterbi re-segmentation reduced the DER for the Gaussianized system from 11.5% to 10.4%.

4.4. Merging Clusters from Gaussianized and non-Gaussianized Systems

We combine the clusters from the Gaussianized and non-Gaussianized systems to reduce the DER even further. The overriding principle in combining clusters from the two diarization systems is to keep the clusters common to both

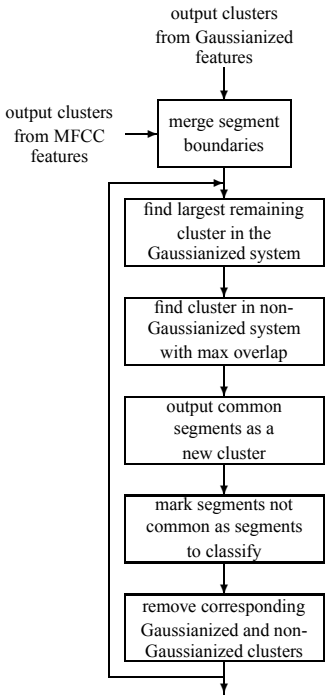

Fig. 2. *Flowchart of process for combining clusters resulting from Gaussianized and non-Gaussianized systems.*

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 used to re-classify the remaining segments. The remaining segments are the segments not common to the two systems. The flowchart for this cluster combination is shown in Fig. 2. We explain the algorithm for combining the clusters using a simple example.

Suppose that the Gaussianized system results in 2 clusters and 10 segments (11 segment boundaries). Assume that the non-Gaussianized system results in 3 clusters and 15 segments (16 segment boundaries). If all the segment boundaries are different (except for the first and the last), then when we pool the boundaries, there will be 25 segment boundaries or 24 segments altogether. If there are some common boundaries, then there will be between 15 and 24 segments after pooling. Each of these segments is labeled with the corresponding cluster IDs from both systems. This pooling of segments simplifies the implementation of the rest of the steps. For example, to compute the overlap of cluster 0 from the Gaussianized system and cluster 1 from the non-Gaussianized system, we simply go through all the segments and add the durations of the segments that belong to cluster 0 of the Gaussianized system and cluster 1 of the non-Gaussianized system.

As shown in Fig. 2, we start with the largest cluster in the Gaussianized system. We find the corresponding cluster in the non-Gaussianized system with the maximum number of frames in common with this cluster. All the common segments in the two corresponding clusters form the first output cluster. The segments that are not common between these two clusters are marked for re-classification. These two clusters are then removed from further consideration. We proceed similarly to find the largest remaining cluster in the Gaussianized system and find the corresponding cluster in the non-Gaussianized system with the maximum overlap. In the end, for the example given, we will probably end up with two output clusters and many segments that need re-classification.

Sometimes, near the end, some of the smaller Gaussianized clusters may have no matching non-Gaussianized cluster. This can happen if all the segments for this Gaussianized cluster correspond to the segments of the non-Gaussianized clusters that have already been matched to bigger Gaussianized clusters. In that case, these smaller Gaussianized clusters are lost (all the segments for these clusters have been marked for re-classification). This actually reduces diarization error rate, since in most cases, these clusters happen to be spurious clusters.

The re-classification of the segments is done as follows. We first remove all the silence frames from the output clusters and the segments to be reclassified. The silence frames are the frames that have been tagged as silence by the voice activity detector. These silence frames are assigned to a new cluster labeled as *silence*. We generate one VP-MAP-adapted GMM for each output cluster, and the silence cluster. If the cluster is male, we use the male UBM for adaptation, and we proceed similarly for female clusters. (For silence, we used the male UBM for adaptation.) We re-label each segment that has been tagged for re-classification using these GMMs: the segment is given the label of the cluster with the highest likelihood. A simple example of cluster merging is shown in Fig. 3.

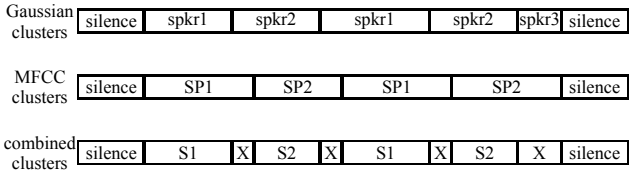

Fig. 3. *Example showing combination of clusters from Gaussianized and non-Gaussianized (MFCC) systems. Segments marked X in the combined cluster are reclassified using adapted GMMs for S1, S2 and silence.*

While optimizing the DER for DEV2Calls with the combination of the two systems as outlined above, we realized that the DER is sensitive to the scanning window length used for change point detection. We varied the scanning length for both the Gaussianized and the non-Gaussianized systems and measured the combined DER. Table 1 shows that for all combinations of scanning lengths, we get the lowest DER with the combined system. The lowest DER is 6.7% for a scanning length of 1.9 sec for the Gaussianized system and 1.3

sec for the non-Gaussianized system. Compared to the lowest DER for any scanning length for the single system (8.3%), this is a reduction of 20% in DER. For this combination, the missed speech is 0.8%, the false alarm speech is 1.7%, and the speaker match error is 4.2%. The primary difference for the combined system is the lowering of the speaker match error rate from 5.7% to 4.2%, a reduction of 26% in speaker match error rate.

The silence models seem to have little impact on the missed speech and the false alarm speech rates. What could have made a difference is if we used two different sensitivity levels for the voice activity detector. Then there would have been more silent segments (that were potentially speech segments) that needed reclassification.

Table 1. *Scanning window lengths (SWL) versus DER for Gaussianized (G), non-Gaussianized (NG), and combined systems for DEV2Calls.*

SWL G	SWL NG	DER G	DER NG	DER combined
1.3	1.3	8.6%	9.0%	7.7%
1.5	1.3	8.4%	9.0%	7.5%
1.7	1.3	8.6%	9.0%	7.4%
1.9	**1.3**	**8.6%**	**9.0%**	**6.7%**
2.1	1.3	8.9%	9.0%	7.9%
1.5	1.5	8.4%	8.3%	7.7%
1.7	1.5	8.6%	8.3%	7.6%
1.9	1.5	8.6%	8.3%	7.3%
2.1	1.5	8.9%	8.3%	8.0%

4.5. Results on the Test Set

We ran the TEST2Calls test set through the same algorithms using the same thresholds as for DEV2Calls. For this test set, we created separate male/female UBM models trained from the training set and the labeled male or female segments in the test set after the gender labeling step. The scanning window length was varied in the same fashion as for DEV2Calls. The DER for the Gaussianized, non-Gaussianized, and the combined system are shown in Table 2. As we can see, the best DER for any single system is 10.4%, while the best combined DER is 9.0%, a drop of 13% in DER. For every pair of scanning lengths for Gaussianized and non-Gaussianized systems, the DER of the combined system is the lowest. The boldface row in table 2 shows the results corresponding to the thresholds for best DEV2Calls results (boldface row in table 1).

We notice from table 2 that for TEST2Calls, the DER for the Gaussianized system (10.4%) is lower than that for the non-Gaussianized system (12.3%). For DEV2Calls, the DER for the Gaussianized system is close to the DER for the non-Gaussianized system (see table 1). The difference in DER between the Gaussianized and the non-Gaussianized systems is not as big as reported for broadcast news [2]. This is probably because it is the same call (same channel) and the background noise does not vary much. The difference is much more pronounced in the broadcast news probably due to the varying music noise in the background.

Table 2. *Scanning window lengths (SWL) versus DER for Gaussianized (G), non-Gaussianized (NG), and combined systems for TEST2Calls.*

SWL G	SWL NG	DER G	DER NG	DER combined
1.5	1.3	11.1%	12.5%	10.1%
1.7	1.3	10.4%	12.5%	9.0%
1.9	**1.3**	**11.7%**	**12.5%**	**10.0%**
1.7	1.1	10.4%	14.0%	9.2%
1.7	1.5	10.4%	12.3%	9.3%
1.7	1.7	10.4%	13.1%	9.5%
1.5	1.5	11.1%	12.3%	10.3%
1.5	1.7	11.1%	13.1%	10.3%

5. CONCLUSIONS

In this paper, we have applied state-of-the-art speaker diarization algorithms on telephone conversations. These algorithms are similar to the multistage segmentation and clustering systems [1] [2] used successfully in broadcast news. We added a Viterbi re-segmentation stage using GMMs that reduced the DER by 10%. We have further enhanced these algorithms by combining the clustering results from two independent speaker diarization systems: one using Gaussianized feature parameters and the other using non-Gaussianized feature parameters. These enhancements result in the reduction of DER from 8.3% to 6.7% for DEV2Calls, and from 10.4% to 9.0% for Test2Calls. This is approximately a 20% reduction in error rate for DEV2Calls and 13% for TEST2Calls. Also, combining the two systems using different scanning window lengths is more effective than using the same scanning window length for the two systems.

One issue is the choice of the two feature sets: the Gaussianized features are considered channel/noise-robust while the MFCCs are channel/noise-sensitive. This choice results in significantly different cluster assignments that probably lead to the improvements that we have observed. Whether other feature sets will lead to similar improvements can only be answered after extensive experimentation.

The other issue is how our system combination compares with other system combinations. As far as the ELISA *piped* system [7] is concerned, the two systems seem to be complementary. In theory, we could possibly pipe our segmentation using the Gaussian features to the HMM-based LIA system [7] and get clusters with lower DER. We could apply the same process to the non-Gaussianized system and get clusters with

lower DER. Combining the two output clusters using our approach would possibly result in even lower DER.

6. REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

[5] A. E. Rosenberg, A. Gorin, Z. Liu, and S. Parthasarathy, "Unsupervised speaker segmentation of telephone conversations", Proc. ICSLP 2002, pp. 565–568.

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

[7] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.-F. Bonastre, "The Elisa consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation", Proc. ICASSP 2004, pp. I-373–I-376.

[8] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization", Comput. Speech Lang., no. 20, pp. 303–330, 2006.

[9] S. E. Tranter, "Two-way cluster voting to improve speaker diarisation performance", Proc. ICASSP 2005, pp. I-753–I-756.

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

[11] 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. 15, no. 2, Feb. 2008.