System output combination for improved speaker diarization

Simon Bozonnet [1], *Nicholas Evans* [1], *Xavier Anguera* [2],
Oriol Vinyals [3], *Gerald Friedland* [3] *and Corinne Fredouille* [4]

[1]EURECOM, Sophia Antipolis, France – [2]Telefonica Research, Barcelona, Spain
[3]ICSI, University of California at Berkeley, USA – [4]LIA, University of Avignon, France
{bozonnet,evans}@eurecom.fr, xanguera@tid.es, vinyals@eecs.berkeley.edu
fractor@icsi.berkeley.edu, corinne.fredouille@univ-avignon.fr

Abstract

System combination or fusion is a popular, successful and sometimes straightforward means of improving performance in many fields of statistical pattern classification, including speech and speaker recognition. Whilst there is significant work in the literature which aims to improve speaker diarization performance by combining multiple feature streams, there is little work which aims to combine the outputs of multiple systems. This paper reports our first attempts to combine the outputs of two state-of-the-art speaker diarization systems, namely ICSI's bottom-up and LIA-EURECOM's top-down systems. We show that a cluster matching procedure reliably identifies corresponding speaker clusters in the two system outputs and that, when they are used in a new realignment and resegmentation stage, the combination leads to relative improvements of 13% and 7% DER on independent development and evaluation sets.
Index Terms: speaker diarization, system combination, fusion

1. Introduction

Commonly referred to as the 'who spoke when?' task, speaker diarization has emerged as an increasingly important domain of speech research. More formally the task requires the unsupervised identification of each speaker within an audio stream and the intervals during which each is active. Speaker diarization has utility in any application where multiple speakers may be expected. Examples include audio and speaker indexing, content structuring, information retrieval, speaker verification (in the presence of multiple or competing speakers), to assist with speech-to-text transcription (with speaker-specific model adaptation or speaker-attributed speech-to-text) or, more generally, Rich Transcription (RT).

Speaker diarization performance tends to vary greatly between one application domain and another (i.e. between telephony, broadcast news and conference meetings) and even within domains. In addition to overall stability, the reliable estimation of the number of speakers and the modeling of speakers with widely varying floor times remain problematic. Furthermore, speaker overlap has emerged as a primary source of error for which there is currently no effective solution. Whilst the state-of-the-art in speaker diarization has advanced significantly since its early beginnings, performance seems to have reached a plateau of late, leaving the community in search of new ideas.

In other fields, for example in speaker recognition, combination or fusion strategies have led to significant leaps in performance. Approaches to fusion operate at the feature, score or decision levels and can produce systems which are more robust than the individual systems alone. The introduction of multiple microphones to the NIST RT evaluations led to the development of feature fusion strategies which combine acoustic features with inter-channel delay features and produced notable improvements in performance [1]. However, with the exception of some early work to combine system outputs through piped and hybridization strategies, there are very few efforts to combine speaker diarization systems at the equivalent score or decision levels. Given their entirely unsupervised nature the combination of diarization systems is troublesome.

Whilst other approaches have been proposed very recently, two standard approaches to speaker diarization have emerged through the NIST RT evaluations: the most popular bottom-up, hierarchical agglomerative clustering approach and the top-down approach. The hypothesis under investigation here is that, as such distinctly different approaches to speaker diarization, there is potential for their complementary outputs to be combined for improved speaker diarization performance. Hence in this paper we present our efforts to combine two state-of-the-art speaker diarization systems, one bottom-up, and one top-down, at the output levels.

The rest of this paper is organized as follows. In Section 2 we describe previous related work and the difficulties in combining speaker diarization systems. Section 3 describes the two speaker diarization systems used in this work and some of their characteristics which demonstrate the potential for combination. Section 4 describes our experimental work and results before our conclusions presented in Section 5.

2. System combination

In speech processing, the combination of decisions yielded by different systems/recognizers is a common means of improving performance. The combination of speaker recognition systems based on different parameterizations, different statistical classifiers and/or different data normalization techniques has proved extremely popular, e.g. [2]. For speech recognition, the ROVER (Recognizer Output Voting Error Reduction) technique is classically used to combine the outputs of different automatic speech recognizers [3].

The combination of speaker diarization outputs, however, is an extremely difficult task and must address numerous specific issues. First, the output of speaker diarization systems is not standardized in terms of labeling (as are words in speech recognition) i.e. there is no natural correspondence between system output labels. A preliminary matching algorithm is therefore necessary to identify speaker label pairs between two segmentation hypotheses. Second, the number of speakers detected may differ from one system to another one. Depending on the contribution, or floor time, of missed or falsely detected speakers, these differences alone can lead to significant variations in diarization performance. Combination strategies must therefore somehow unify outputs with differing numbers of speak-

ers. Finally, different segmentation outputs are generally not time-synchronized. This is particularly true if different Speech Activity Detection (SAD) algorithms are used. In this case, whilst one system might produce a speaker label, another may classify it as non-speech. Differences in SAD outputs and further down-stream dependent processes, such as speaker modeling and more general differences in the particular approach to speaker diarization, will all contribute to differences in the number of speaker boundaries, or turns, and different turn locations.

For these reasons, most of the approaches to system output combination reported thus far in the literature generally rely on the so-called 'piped' approach [4] where different algorithms or components are applied sequentially, based on the segmentation outputs of the previous steps, thereby refining both the number of detected speakers and speaker boundaries [5, 6]. Nevertheless, some attempts have been made to combine segmentation outputs. In [4], speaker labels produced by different diarization systems are merged at the frame level. Then, a resegmentation process is used to remove redundant speakers and to refine speaker turn points. In [7], a cluster voting approach is proposed to combine the outputs of two different speaker diarization systems operating on Broadcast News (BN) data. In [8], two systems uniquely differentiated by the input features (parameterizations based on Gaussianized and non-Gaussianized MFCCs) are successfully combined for the speaker diarization of telephone conversations. This approach relies on the identification of the most relevant, common clusters in the two system outputs. All the segments which are not identified as belonging to the common clusters are labeled as misclassified and are reassigned through a new resegmentation step based on the GMM modeling of the common clusters and a maximum likelihood-based decision.

3. Diarization systems and characteristics

Most of today's state-of-the-art speaker diarization systems fit into one of two categories: either bottom-up or top-down. Both approaches are generally based on Hidden Markov Models (HMMs) where states are Gaussian Mixture Models (GMMs) corresponding to individual speakers, and where transitions between states correspond to speaker turns.

The bottom-up, or agglomerative hierarchical clustering approach is the most popular. The audio stream is over-segmented into a number of segments which exceeds the anticipated number of speakers. Closely matching clusters are then iteratively merged, hence reducing the number of clusters by one upon each iteration. A reassignment of frames to clusters is usually performed after each cluster merging, via Viterbi realignment, for example, and the whole process is repeated iteratively, until some stopping criterion is reached, upon which there should remain only one cluster for each detected speaker.

In contrast, the top-down approach first models the entire audio stream with a single, root speaker model, and upon its successive splitting, new speaker models are added one-by-one with interleaved Viterbi realignment and adaptation. Stopping criteria similar to those employed in bottom-up systems may be used to terminate the process or it can continue until there remain no more unlabeled segments with which to train new speaker models. Top-down approaches are less popular than their bottom-up counterparts but nonetheless perform respectably well.

As two distinctly contrasting approaches to speaker diarization there may be scope for effective combination or fusion. This paper presents our initial efforts to combine two such state-of-the-art speaker diarization systems, at the output level.

	Av. no. spkrs		Av. Err	
Source	RT'07	RT'09	RT'07s	RT'09s
Ground Truth	4.37	5.42	-	-
ICSI	6.62	5.28	2.25	1.86/1.33
LIA-EURECOM	4.75	5.28	0.87	1.28/0.66
Combined	4.62	5.28	0.65	1.28/0.66

Table 1: Average number of speakers and average error for the ground-truth reference, the two individual systems and their combination, for RT'07 and RT'09 datasets. Results in column 5 illustrated with/without the inclusion of the *NIST_20080307-0955* show which is an outlier and biases results.

They are ICSI's bottom-up, agglomerative clustering system [9] and LIA-EURECOM's top-down system with purification [10]. Both systems have achieved competitive results in official NIST RT evaluations and in the following we present a comparison of each system output on the NIST RT'07 and RT'09 datasets in order to highlight the potential for improved speaker diarization performance through their combination.

3.1. Number of speakers

Reliably estimating the number of speakers is both extremely challenging and crucial to the overall performance of any diarization system. Table 1 shows the number of speakers per show, averaged across the full RT'07 and RT'09 datasets in columns 2 and 3 respectively, for the ground-truth reference (row 1) and the segmentation hypotheses obtained from the ICSI and LIA-EURECOM systems (rows 2 and 3 respectively). Details of the datasets are given later in Section 4.

In addition, shown in columns 4 and 5 of Table 1, is the error in the number of speakers detected by each system, also averaged across the full datasets. This is computed by averaging the absolute value of the difference between the real number of speakers (i.e. that in the reference) and the number hypothesized by each system for each meeting. For the RT'07 dataset both systems are shown to under-cluster, i.e. they produce more than a single cluster per speaker (results of 6.62 and 4.75 speakers cf. 4.37). For the RT'09 dataset, however, both systems over-cluster, i.e. some clusters correspond to more than a single speaker (results of 5.28 for both systems cf. 5.42). In both cases, the average error is lower for the LIA-EURECOM system than for the ICSI system.

Where both systems under-cluster the robust matching of clusters identified by the two systems may give improved performance when their outputs are combined. Where both systems over-cluster improvements may only be obtained if the clusters in each system which correspond to more than a single speaker do not overlap, i.e. we can find clusters in one system output that do not correspond to clusters in the other system output and hence introduce 'new' clusters into the combined output. This is likely to be more difficult.

3.2. Segment sizes

Table 2 shows the average number of segments and segment length in seconds, for the ground-truth data (row 1) and for both system outputs (rows 2 and 3). The ICSI system estimates the number of segments more reliably than the LIA system (617 and 307 cf. 676). Similar results are obtained for the RT'09 dataset. The ICSI system also better reflects the average segment length (2.2s and 4.5s cf. 2.0s) and once again similar results are obtained for the RT'09 dataset.

Thus, whilst one system better estimates the true number of speakers with a smaller average error, the other system better

Source	No. segments RT'07	No. segments RT'09	Av. seg. length (s) RT'07	Av. seg. length (s) RT'09
Ground Truth	676	882	2.0	1.8
ICSI	617	694	2.2	2.2
LIA-EURECOM	307	313	4.5	6.3
Combined	353	315	3.9	6.2

Table 2: Average number of segments and average segment length in seconds for the ground-truth reference, each individual system and their combination for the RT'07 and RT'09 datasets.

reflects the true number of segments and their average length. Should it be possible to exploit the beneficial characteristics of each system then this observation supports the hypothesis that a combined system has the potential to deliver better results.

4. Experimental Work

Here we first describe the datasets for all work reported here and then two approaches to combine the system outputs: one artificial combination using the ground-truth reference to illustrate the potential and a practical combination without the reference.

4.1. Datasets

Experiments were performed on meeting recordings from the NIST RT'07 and RT'09 evaluation datasets. Each meeting recording contains 4-11 speakers and each dataset is about 3 hours long, though scoring is performed on periods of approximately 20 minutes only, as defined by NIST. We decided to concentrate on the single distant microphone (SDM) condition since it has the highest potential for improvement through system combination. In all cases performance is reported in terms of the Diarization Error Rate (DER) as defined by NIST[1]. All results **include** the scoring of overlapping speech.

4.2. System combination

Here we report two experiments. First, we aim to assess the potential of combination by establishing a lower bound on performance with an artificial experiment using the ground-truth reference. Second, we report a practical experiment where the reference is not used.

For the artificial experiment the two system outputs are combined in an optimal manner using the ground-truth reference. Segment boundaries (i.e. speaker turns) from both systems are merged and virtual clusters are defined by taking the product space of the clusters for each of the two systems. For example, if, for a given segment, system 1 outputs label c_i^1 and system 2 outputs label c_j^2, then we attribute the virtual cluster assignment $c_{(i,j)}^V$. Thus, the resulting cardinality for our virtual cluster space becomes $N_1 N_2$ where N_i refers to the total number of clusters output by system i.

The virtual clusters are then merged in an optimal manner in order to minimise the DER, without violating cluster groupings nor changing the segment boundaries. This is achieved with a dynamic programming search making use of the ground truth data to find optimal many to one mappings. Results are illustrated in Tables 3 and 4 for RT'07 and RT'09 datasets respectively. In both cases columns 2 and 3 show results for the individual systems whereas results for the optimally combined system are shown in column 4. Average results (last row) show that a relative improvement of almost 50% over the best single system are achieved on the RT'07 dataset (individual system result of 18% cf. 10% when optimally combined). For the RT'09 dataset the maximum relative improvement is 25% and thus there appears to be less scope for improvement on this dataset. In the following we present an approach to perform this combination in a practical scenario without the ground-truth reference.

For practical system combination without the ground-truth reference we performed cluster alignment using a cluster confusion matrix obtained from the output of both systems. The elements of the matrix contain the total speech time assigned to speaker x in system 1 and speaker y in system 2. Then, for each cluster in the output of system 1 we seek to identify a matching cluster in the output of system 2. This is done according to two criteria for each cluster in the output of system 1. First, we calculate the information change rate (ICR) [11] for all possible cluster alignments and select the cluster in the output of system 2 which has the highest ICR as a candidate cluster pair. Second, among all of the other clusters in the output of system 2, we verify that the candidate cluster is that with the highest value in that column of the confusion matrix. Note that in some cases the cluster pairing with the highest ICR is not the same as the pairing with the highest value in the confusion matrix and thus some clusters in the outputs of each system are not aligned through this process.

The two system outputs are then aligned frame-by-frame to produce a new segmentation hypothesis according to the following procedure. First, all frames that have labels in the outputs of systems 1 and 2 which match according to the determined cluster alignment are assigned the label from system 1. All frames which have mis-matching labels are rejected during this stage. Then, for each cluster in system 1 which does not have a paired cluster in system 2 we retrain only a percentage α of frames which best match the cluster in system 1, according to those which have the highest likelihood. α is the only parameter which requires optimization. This new hypothesis is then used to perform a resegmentation of the data using the MAP adaptation of a 128-component Gaussian mixture model (GMM) previously trained on external data. Several iterations of realignment and adaptation are performed until a stable hypothesis is obtained. At each stage clusters with less than 8 seconds of assigned speech are removed. Finally the features are normalized segment-by-segment to fit a zero-mean and unity-variance distribution before a second resegmentation is applied using the normalized features.

4.3. Results

The combination algorithm described above was optimized on the RT'07 dataset and then applied to the RT'09 dataset without modification, in both cases using the LIA-EURECOM output as system 1 and the ICSI output as system 2 with $\alpha = 20\%$. Results are illustrated in column 5 of Tables 3 and 4 for each dataset. In all but two cases for both the RT'07 development set and RT'09 evaluation set, illustrated in bold in Tables 3 and 4 respectively, results for the combined system are as good as, or better than the best results for either of the single systems. For the RT'07 dataset single system results of 21% and 18% fall to 15% when combined, a relative improvement of 13% over the best single system. For the RT'09 evaluation set single system results of 31% and 21% fall to 20% which corresponds to a relative improvement of 7% over the best single system. It should be noted that, in order to combine the systems, some of ICSI's standard optimizations had to be turned off for different technical reasons, i.e. ICSI's system did not include a prosodic feature stream [9] and no adaptive initialization [12].

[1] http://www.itl.nist.gov/iad/mig/tests/rt/2009/index.html

RT07	ICSI	LIA-EUR	Optimal	Combined
CMU_20061115-1030	36.08	21.88	16.82	21.62
CMU_20061115-1530	19.65	35.15	9.65	19.87
EDI_20061113-1500	32.39	20.30	16.51	19.14
EDI_20061114-1500	22.73	29.96	12.72	**28.85**
NIST_20051104-1515	7.56	10.88	6.76	**11.09**
NIST_20060216-1347	9.34	9.72	6.81	10.31
VT_20050408-1500	16.92	4.60	4.26	4.53
VT_20050425-1000	27.31	11.34	9.14	9.84
Average	21.30	17.72	10.23	15.48

Table 3: Speaker diarization performance in DER for the RT'07 dataset. Results illustrated for the two individual systems, and optimally (with reference) and practically combined (without reference) systems.

RT09	ICSI	LIA-EUR	Optimal	Combined
EDI_20071128-1000	20.34	10.00	9.38	10.01
EDI_20071128-1500	18.12	25.24	15.56	16.63
IDI_20090128-1600	18.94	11.64	6.03	10.40
IDI_20090129-1000	23.69	15.29	13.15	**17.49**
NIST_20080227-1501	45.09	17.69	13.46	**18.31**
NIST_20080307-0955	47.11	31.85	21.58	31.59
NIST_20080201-1405	65.79	51.66	45.06	46.89
Average	31.15	21.06	15.70	19.61

Table 4: As for Table 3 except for the RT'09 dataset

Comparative speaker statistics for the combined system are also illustrated in Table 1. We note that, even though *both* systems over-estimate the number of speakers for the RT'07 dataset, the combined system gives a more accurate estimate. For the RT'09 dataset both single systems estimate the same number of speakers and no improvement is obtained with the combined system. Similar improvements are observed with the error in the number of detected speakers: an improvement for RT'07 but no difference for RT'09. When we compare the number of segments and their average length, as illustrated in Table 2, we notice consistent improvements over the LIA-EURECOM system only. This behavior is to be expected since it is the LIA-EURECOM system that is used as system 1: when the speaker labels in the two individual system outputs do not match according to the cluster alignment procedure outlined above we revert to those assigned by system 1.

The comparison of columns 4 and 5 in Tables 3 and 4 shows how well the combination performs with respect to the optimum combination. We see that in many cases the combined system achieves performance very close to the optimum but also that there are plenty of examples where the combined system gives results which are far off and thus more work is required to improve practical combination performance.

Finally, we performed a cross validation by optimizing the combination system using the RT'09 dataset and evaluating it using the RT'07 dataset. In this case we obtained results of 16% and 19% for the two datasets respectively cf. 15% and 20% before. Here the optimised value of $\alpha = 60\%$ differs significantly but the resulting DER was in any case observed to be quite stable with α in the range of 20 to 60%.

5. Conclusions

This paper presents our first efforts to combine two state-of-the-art speaker diarization systems at the output level. Average relative improvements of 13% and 7% DER are achieved for the RT'07 and RT'09 datasets. Although modest, these improvements remain consistent in a cross validation experiment and demonstrate the potential of system output combination. Results obtained with an artificial, optimal combination experiment show that more work is required to fully exploit this potential.

Future work should consider the combination of speech activity detection systems and improved combination of final speaker diarization system outputs making better use of system likelihoods or confidence measures. Another possible direction of future work involves the use of an HMM model with emissions being two independent multinomial observations from the system outputs. This work could consider models that use non-parametric Bayes such as the Hierarchical Dirichlet Process HMM, where the number of clusters is directly inferred from the observed data.

6. Acknowledgments

This research is partly supported by Microsoft (Award #024263) and Intel (Award #024894) funding and by matched funding from U.C. Discovery (Award #DIG07-10227). X. Anguera was partially funded by the Torres Quevedo Spanish program.

7. References

[1] J. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Transaction on Computers*, vol. 56, no. 9, pp. 1212–1224, 2007.

[2] L. Burget *et al.*, "BUT system for NIST 2008 speaker recognition evaluation," in *Proceedings of Interspeech*, 2009, pp. 2335–2338.

[3] J. M. Fiscus, "A post processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *IEEE ASRU Workshop*, 1997, pp. 347–352.

[4] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," in *CSL, selected papers from the Speaker and Language Recognition Workshop (Odyssey'04)*, 2006, pp. 303–330.

[5] D. Vijayasenan, F. Valente, and H. Bourlard, "Combination of agglomerative and sequential clustering for speaker diarization," in *Proc. ICASSP*, Las Vegas, USA, 2008, pp. 4361–4364.

[6] E. El-Khoury, C. Senac, and S. Meignier, "Speaker diarization: combination of the LIUM and IRIT systems," in *Internal report*, 2008.

[7] S. E. Tranter, "Two-way cluster voting to improve speaker diarisation performance," in *IEEE ASRU Workshop*, 1997, pp. 347–352.

[8] V. Gupta, P. Kenny, P. Ouellet, G. Boulianne, and P. Dumouchel, "Combining gaussianized/non-gaussianized features to improve speaker diarization of telephone conversations," in *Signal Processing letters, IEEE*, 2007, pp. 1040–1043.

[9] G. Friedland, O. Vinyals, Y. Huang, and C. Muller, "Prosodic and other long-term features for speaker diarization," *IEEE TASLP*, vol. 17, no. 5, pp. 985–993, July 2009.

[10] S. Bozonnet, N. W. D. Evans, and C. Fredouille, "The LIA-EURECOM RT'09 speaker diarization system: enhancements in speaker modelling and cluster purification," in *Proc. ICASSP*, March 2010.

[11] K. J. Han, S. Kim, and S. S. Narayanan, "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1590–1601, November 2008.

[12] D. Imseng and G. Friedland, "Robust speaker diarization for short speech recordings," in *Proc. of the IEEE workshop on Automatic Speech Recognition and Understanding*, December 2009, pp. 432–437.