WHERE ARE THE CHALLENGES IN SPEAKER DIARIZATION?

Mark Sinclair[*]*, Simon King*[†]

The Centre for Speech Technology Research, The University of Edinburgh, UK

M.Sinclair-7@sms.ed.ac.uk, Simon.King@ed.ac.uk

ABSTRACT

We present a study on the contributions to Diarization Error Rate by the various components of speaker diarization system. Following on from an earlier study by Huijbregts and Wooters, we extend into more areas and draw somewhat different conclusions. From a series of experiments combining real, oracle and ideal system components, we are able to conclude that the primary cause of error in diarization is the training of speaker models on impure data, something that is in fact done in every current system. We conclude by suggesting ways to improve future systems, including a focus on training the speaker models from smaller quantities of pure data instead of all the data, as is currently done.

Index Terms— speaker diarization, diarization error rate

1. INTRODUCTION

Speaker Diarization involves segmenting audio into speaker-homogenous regions and labelling regions from each individual speaker with a single label. Knowing both *who* spoke and *when* has useful applications and can form part of a rich transcription of speech. The task is challenging because it is generally performed without any *a priori* knowledge about the speakers present, not even how many speakers there are.

The NIST Rich Transcription (RT) evaluation campaign [1] ran annually between 2002 and 2009, focusing on promoting Metadata Extraction (MDE) for speech. For some of the years, the campaign included a dedicated speaker diarization task and the use of the associated datasets and evaluation tools have come to form the standard for developing and comparing most current systems. The NIST RT challenges have probably been the most significant driving force for community interest and support for speaker diarization.

The more recent campaigns (RT05/06/07/09) focused on diarization of meetings data. However, system performance on this task has been notoriously meeting-dependent and hyper-sensitive to system parameters [2]. Diarization systems based on agglomerative clustering generally involve an initialisation step, followed by interleaved iterations of re-segmenting the speech, re-estimating the speaker models, and merging models, to gradually converge on the correct number of speakers and the best segmentation and speaker assignments. This architecture means that the final system performance is a complex function of the performance of the individual parts, making it very difficult to identify the causes of error. The work we present here was motivated by the need for a better understanding of the system component factors that contribute to diarization error. Our ultimate goal is to identify where improvements are needed and, conversely, which parts of the system already work well.

Huijbregts and Wooters [3] conducted an investigation along similar lines in 2007. Our investigation is complementary to that work: we investigate several aspects of the system that they did not consider in detail, and we also reach different conclusions about where efforts should be focused in order to reduce diarization error rate. Our methodology is broadly similar to theirs: we start with a diarization system that is capable of good performance in the standard fully-unsupervised mode, and then conduct various 'oracle' experiments to isolate the effects of various components.

First, we describe the system in Section 2 and then introduce the methodology and experiments in Section 3, summarising our findings and making conclusions about where to focus effort in Sections 4 and 5.

2. SYSTEM DESCRIPTION

There are several diarization systems with competetive state-of-the-art performance such as ICSI [4], IDIAP [5], LIA-EURECOM [6] and I^2R [7]. We used our own modular speaker diarization system and chose parameters and methods that would closely emulate that of the ICSI system. The performance of our system is therefore comparable e.g. for single distant microphone (sdm) RT09 data, ICSI has an average of 31.3%DER [4] vs. our 31.8%.

Unlike many other systems (e.g., [8]) we choose not to use a beamformed signal from multiple channels of a microphone array and instead opt for single distant microphone data. A beamformed signal typically improves DER results for systems that ignore overlap [9], but could be a poor choice if we wish to detect a number of simultaneous speakers.

[*]Funded by an EPSRC studentship.

[†]Partially funded by EPSRC grant EP/I031022/1 (Natural Speech Technology) and from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287678 (Simple4All).

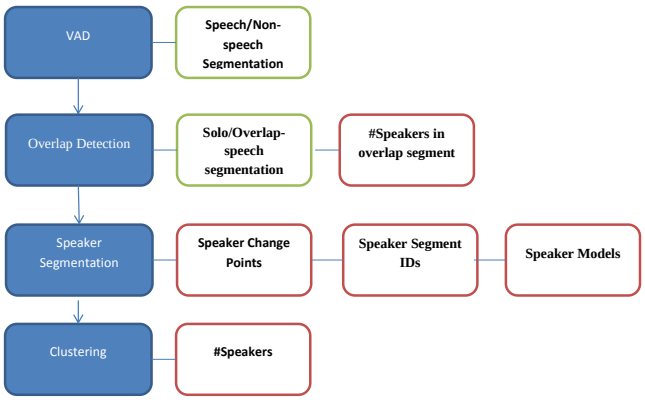

Fig. 1. [The basic speaker diarization system design, showing information at each stage that can be replaced with oracle knowledge]

Speech Activity Detection (SAD) is performed by the QIO-Aurora toolkit [10] as this has proven to work well with the RT datasets [11]. For the regions labelled as speech, feature extraction is performed using HTK [12]: we use the first 19 MFCCs computed from a bank of 26 Mel-scaled triangular filters with a pre-emphasis coefficient of 0.97 and cepstral lifting coefficient of 22. We used an analysis window of 30ms and a timeshift of 10ms.

The system uses a GMM-HMM framework whereby 16 clusters (states) are initialised with speech data by dividing the speech frames uniformly into 32 parts and using 2 parts (from different points in the data) to initialise each of the 16 GMMs. Given these models, the system then segments all speech using the Viterbi algorithm with a forced minimum duration constraint of 250ms. After segmentation, the models are retrained, and this is followed by a clustering step in which the most similar clusters are merged – the choice of which clusters to merge is based on the Bayesian Information Criterion. The putative merged model has a complexity (i.e., number of model parameters) equal to the sum of the complexity of the models being merged, which means that a penalty factor parameter is not required.

The process of segmentation and clustering is then iterated until a termination criteria is met: for example, all BIC scores for putative cluster merges are negative. Fig.1 shows an outline of the system design and also illustrates information at each stage that can be replaced by oracle knowledge.

3. EXPERIMENTS

3.1. Data

We used the data from RT06, RT07 and RT09 in a series of experiments designed to control for the influence of separate system components by replacing them with oracle or *ideal* equivalents. Often, in the literature, we see that results on the RT corpora are presented by campaign year. However there are no inherent differences in terms of task or conditions and, while inter-meeting variations are observed in results, no inter-campaign variations are. Therefore, results for all meetings are presented here together as a single set.

3.2. Diarization Error Rate

The main metric for system evaluation is the Diarization Error Rate (DER) which is a sum of three contributing factors as shown in Eq.1: speaker missclassification *SpkErr*, false alarm *FA* (speaker attributed when no speech exists) and missed speech *Miss* (speaker not attributed when speech exists).

$$DER = E_{Spkr} + E_{FA} + E_{Miss} \quad (1)$$

However there is some contention between how overlap should be considered. Some authors [3] choose to take the *FA_speech* and *Miss_speech* errors from the SAD which essentially ignores overlap and results in a lower overall DER. This error is referred to as *speech* time error in the results computed by NIST tools[1].

Others [13] choose to report the *FA_speaker* and *Miss_speaker* errors inclusive of overlap, e.g. a segment which contains two speakers that has been completely missed by the system will have twice the error. This error is referred to as *speaker* time error by the results of the NIST tools and is in fact the default formulation of the *overall speaker diarization error* of the output. This form is used for all results shown in this paper.

The difference between *Miss_speech* and *Miss_speaker* is attributed to overlap. For systems which do not consider overlap, there is no difference between *FA_speech* and *FA_speaker*.

3.3. System configurations

The system was configured to use various combinations of real, oracle and 'ideal' components.

3.3.1. End-to-End

This is the fully automatic unsupervised system. The system is not provided with any oracle knowledge. Apart from a few heuristically-selected parameters (as is the case for all diarization systems), it is completely unsupervised. SAD is done automatically using the QIO-Aurora toolkit. These are the standard conditions for speaker diarization.

3.3.2. Oracle Number of Speakers

One key problem in the clustering stages of diarization is knowing when to stop. Over-clustering will lead to the speech being labelled with too few speakers and typically this results in a sudden increase in DER. In this condition the clustering stops at precisely the known number of speakers per meeting.

[1] http://www.itl.nist.gov/iad/mig/tools

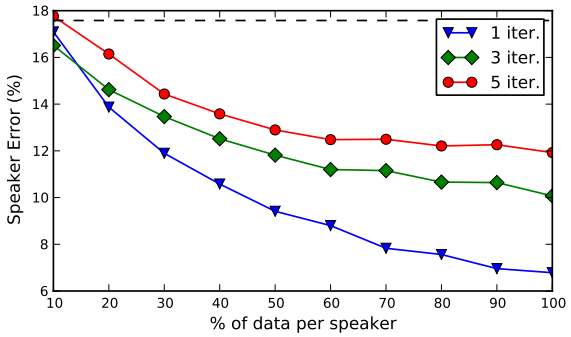

Fig. 2. The *SpkErr* obtained when creating ideal speaker models using varying amounts of data, up to and including all the data. The three lines show the results for segmentation using these models directly ('1 iter.') and when re-segmenting and re-training the models on all the data in the usual iterative fashion ('3 iter.' and '5 iter.'). The dashed line is the DER of the ??? system.

3.3.3. Oracle SAD

This condition is the same as End-to-End, except only that the initial SAD segmentation is provided by reference transcriptions. This is intended to give insight into how SAD-related errors made at the beginning of the process propagate to other parts of the system. All speaker IDs are relabelled as *speech* and are then collapsed into a standard speech/non-speech segmentation (i.e., overlap is not represented).

3.3.4. Ideal Cluster Initilisation

Normally the initial seed clusters to the algorithm are derived by uniformly dividing the data and attributing a portion of it to each cluster. In this condition the clusters are instead each initialized with homogenous data belonging to only one speaker. In order to maintain a similar amount of data in each cluster as in the End-to-End condition, each speaker's data is split across a number of clusters based on the proportion of his or her speaking time. The number of initial clusters is the same as in the End-to-End condition. Ideally, at each iteration, the algorithm should choose to merge clusters in such a way as to maximise cluster purity – that is, belonging to the same speaker. This condition allows us to check how sensitive the clustering process is to initialization.

3.3.5. Ideal Models

The speaker models are rather simple: Gaussian mixture models with simple duration modelling. It is reasonable to ask whether these are adequate for the task. The reference transcription is used to create optimal speaker models by creating a number of clusters equal to the known number of speakers and training each with data from one speaker only. This way, we can discover whether the models themselves and the associated acoustic feature set have sufficient speaker discrimination power for the task.

Table 1. Experimental Results for RT06/07/09

System Stage	***Miss_Spkr***	***FA_Spkr***	***SpkrErr***	***DER***
end2end	14.86	2.34	16.38	33.58
numspks	14.86	2.34	16.68	33.88
SAD	10.50	0.00	17.58	28.09
SAD_idealclust	10.50	0.00	15.27	25.78
SAD_numspks_idealclust	10.50	0.00	15.46	25.96
SAD_idealmodels	10.50	0.00	6.78	17.29
0OL	19.04	0.0	0.07	19.11
1OL	10.2	0.0	0.99	11.19
2OL	1.35	0.0	5.21	6.56
allOL	0.0	0.0	5.87	5.87

We also vary the amount of data used to train these ideal models, from 10% of the available data per speaker up to 100%. We examine the effect of further iterations of re-segmentation + re-training (no merging) too, from 1 iteration (i.e., segmentation with ideal models) up to 5 iterations of re-segmenting + re-training. These iterations *should* improve the models (or, in the 100% data case, do nothing).

3.3.6. Overlap Segmentation

SAD is used prior to diarization to classify the signal into speech and non-speech (e.g., silence, music, noise, etc.). We could also benefit from knowing if each speech segment contains one speaker (solo speech) or multiple (overlap speech). This condition employs such a three-class segmentation derived from reference transcriptions. We first use the ideal models to select the most likely speaker for each solo speech segment. Then, at the end, we revisit overlapping segments and attribute more speakers to them, based on the top few most likely models. Thus, overlap speech is ignored during model training, but is still labelled with speaker ids.

4. RESULTS

Oracle Number of Speakers: As Table 1 shows, knowing the number of speakers has little effect on performance and in sometimes degrades it. Slightly too many clusters can actually be better, if each speaker is well represented – i.e., speaker-attributed clusters have high purity and the *extra* clusters are small. Continuing until the oracle number of speakers is reached may result in incorrect cluster merges.

Oracle SAD: One of the more substantial contributing factors to overall DER was found to be the intial SAD. The automatic method was subject to Missed Speech error in particular. Adding an oracle segmentation, of course, completely eliminates all *Miss_speech* and *FA_speech* error. However, as observed in Fig.3, it is worth noting that this does not propagate on to a substantial reduction in *SpkErr*. This suggests that, while still important, the performance of the diarization algorithm itself is not highly dependent on SAD. Importantly, this also indicates that it is safe to use oracle SAD when investigating other components of the system.

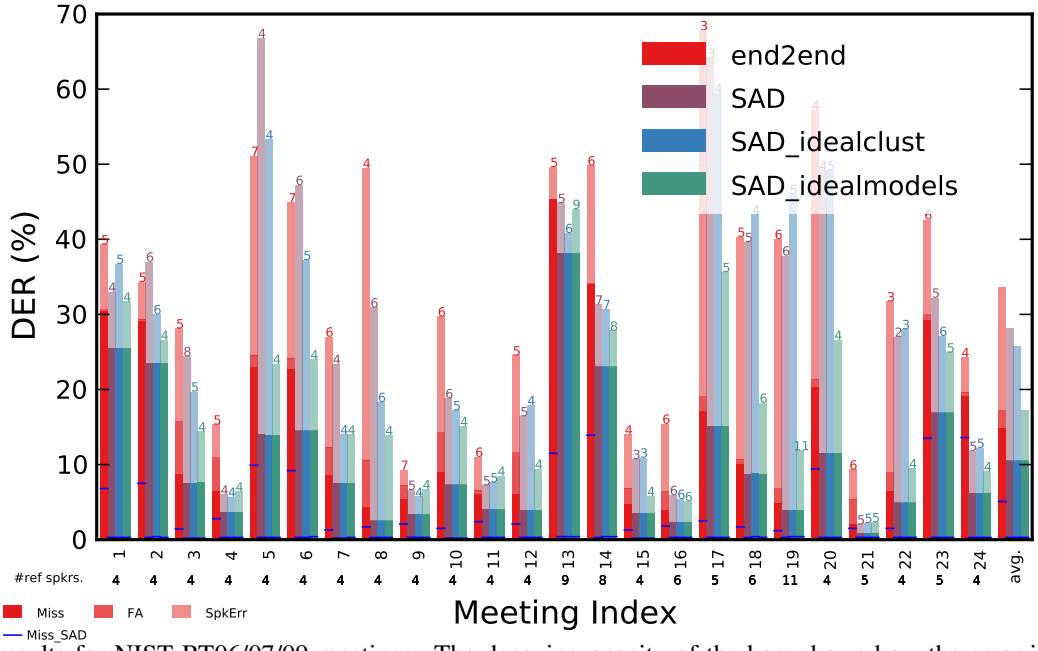

Fig. 3. DER results for NIST RT06/07/09 meetings. The decaying opacity of the bars shows how the error is composed of Missed and False Alarm speaker time (inclusive of overlap) as well as speaker error (due to speaker id misclassification). The blue horizontal bar indicates the amount of missed speech error contributed by SAD. Above each bar the number of hypothesised speakers is shown and the reference is provided parallel to the x-axis.

Ideal Cluster Initilization: While the average *SpkErr* shows a reasonable reduction Fig.3, the inclusion of ideal cluster initialisations has greatest effect for meetings where the End-to-End system gave a high *SpkErr*. This suggests that poor cluster initiliastion, whereby initial clusters all have a low purity, may be non-recoverable.

Ideal Models: Providing the system with ideal models trained on each speaker's data substantially reduces *SpkErr*, confirming the models do work. Fig.2 shows the effect of varying the amount of data used to train the models. As little as 10% improves over baseline. Worryingly, more iterations *degrade* performance. This suggests cluster purity is critical to the clustering process: impurities introduced at each iteration cannot be accommodated, and the models do not recover.

Overlap Segmentation: As we see in Table 1, ignoring overlap (0OL) is costly (19.11% in this case), especially when using *Miss_Spkr* to calculate DER. By attempting to get at least 1 speaker right per overlap region, we halve that error. An average minimum 10.50% DER is always incurred if only 1 speaker at a time is possible, but by getting at least the 2nd speaker correct, we halve the error again.

5. CONCLUSIONS

5.1. Relation to Huijbregts and Wooters

As Huijbregts and Wooters [3] also found, results are highly dependent on the evaluation data (i.e., high variation in DER between meetings) and some system components can be sensitive to the performance of preceeding ones. Like them, we found that SAD can be a major contributor to DER by directly contributing *Miss_speech*, but we would add that subsequent components actually have little dependence on its performance.

5.2. New Findings and Future Work

One of the key findings from our experiments is the importance of estimating the speaker models on pure data: speech from just one speaker. If this could be achieved, dramatic reductions in DER would result (Fig.2). Even if only a fraction of the data for each speaker could be reliably identified, free from the polluting effects of data from other speakers, than large improvements would still be expected. Methods for estimating some form of *confidence* in speaker homogeneity when seeding clusters with data should therefore work well, even if that entails rejecting a large proportion of the data.

Our ideal models are strong enough to allocate multiple speakers to overlap regions. So another focus of future research should be in overlap-speech detection. Systems which do not consider overlap will always concede substantial error.

The take-home message, given that further iterations degrade models that were initially pure (Fig.2), is that the final set of speaker models should not necessarily be trained on all data to be diarized, but only on reliably-identified pure data.

6. REFERENCES

[1] NIST, *Spring 2006 (RT-06S) Rich Transcription Meeting Recognition Evaluation Plan*, 2006.

[2] N. Mirghafori and C. Wooters, "Nuts and Flakes: a Study of Data Characteristics in Speaker Diarization," in *Proc. IEEE Int Acoustics, Speech and Signal Processing Conf. ICASSP 2006*, 2006, vol. 1.

[3] Marijn Huijbregts and Chuck Wooters, "The blame game: performance analysis of speaker diarization system components.," in *INTERSPEECH*. 2007, pp. 1857–1860, ISCA.

[4] G. Friedland, A. Janin, D. Imseng, X. Anguera, L. Gottlieb, M. Huijbregts, M. Knox, and O. Vinyals, "The ICSI RT-09 Speaker Diarization System," *IEEE Transactions on Audio, Speech, and Language Processing*, , no. 99, 2011, Early Access.

[5] Deepu Vijayasenan, Fabio Valente, and Petr Motlícek, "Multistream speaker diarization through Information Bottleneck system outputs combination.," in *ICASSP*. 2011, pp. 4420–4423, IEEE.

[6] 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. IEEE Int Acoustics Speech and Signal Processing (ICASSP) Conf*, 2010, pp. 4958–4961.

[7] Hanwu Sun, Bin Ma, Swe Zin Kalayar Khine, and Haizhou Li, "Speaker diarization system for RT07 and RT09 meeting room audio.," in *ICASSP*. 2010, pp. 4982–4985, IEEE.

[8] Xavier Anguera, Chuck Wooters, and Jose M. Pardo, "Robust Speaker Diarization for Meetings: ICSI RT06s evaluation system," *Interspeech 2006*, 2006.

[9] X. Anguera, C. Woofers, and J. Hernando, "Speaker diarization for multi-party meetings using acoustic fusion," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding*, 2005, pp. 426–431.

[10] André Gustavo Adami, Lukás Burget, Stéphane Dupont, Harinath Garudadri, Frantisek Grézl, Hynek Hermansky, Pratibha Jain, Sachin S. Kajarekar, Nelson Morgan, and Sunil Sivadas, "Qualcomm-ICSI-OGI features for ASR.," in *INTERSPEECH*, John H. L. Hansen and Bryan L. Pellom, Eds. 2002, ISCA.

[11] Erich Zwyssig, Steve Renals, and Mike Lincoln, "Determining the number of speakers in a meeting using microphone array features.," in *ICASSP*. 2012, pp. 4765–4768, IEEE.

[12] Steve J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*, Cambridge University Press, 2006.

[13] Xavier Anguera Miró, Simon Bozonnet, Nicholas W. D. Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker Diarization: A Review of Recent Research.," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.