ESTIMATING THE DOMINANT PERSON IN MULTI-PARTY CONVERSATIONS USING SPEAKER DIARIZATION STRATEGIES

Hayley Hung, Daniel Gatica-Perez

IDIAP Research Institute
Martigny
Switzerland
(hhung,gatica)@idiap.ch

Yan Huang, Gerald Friedland

International Computer Science Institute (ICSI),
Berkeley
USA
(yan,fractor)@icsi.berkeley.edu

ABSTRACT

In this paper, we apply speaker diarization strategies from a single source to the task of estimating the dominant person in a group meeting. Previous work has shown that speaking length is strongly correlated with perceived dominance. Here we investigate this in more depth by considering two dominance tasks where there is full and majority agreement amongst ground-truth annotators. In addition, we investigate how 24 different speed-up and algorithmic strategies, and source types lead to interesting outcomes when applied to dominance estimation. We obtained the best performance of 77% using our slowest scheme and a single distant microphone (SDM). Within the top 3 out of 24 performing experiments in both dominance tasks, we show that we can use the furthest SDM, with no prior knowledge of the number of speakers and the fastest diarization scheme, which performs 1.3 times faster than real-time.

Index Terms— speaker diarization, dominance modelling,

1. INTRODUCTION

With the increasing audio and video capture of humans in various living and working environments, there is a need to quickly understand these patterns of behaviour for automated categorisation. From a work perspective, we may be keen to understand human behaviour in terms of teams and group dynamics in a task-orientated environment. In particular, classifying the roles that participants play in meetings is of interest. In this paper, we concentrate on the task of automatically finding the most dominant person in a group meeting scenario. More specifically, we investigate the case where a single distant microphone is the only available audio source at the meeting.

Early work in automatic dominance modeling in conversations was suggested by Basu et al. [1]. They showed preliminary results using human-human interaction data where two out of five participants were pre-selected to debate for one minute, leading to a rather artificially constrained conversational setting. Their model detected who the two debating participants were using only manually labeled speaker turns, speaking energy, as well as other audio-visual cues. Another study of group dominance in scripted meeting scenarios [2] used audio and speech transcription-based features. Participant speaking length was shown to perform well as a baseline measure of ranked dominance in 30 five-minute scripted meetings. Semantically higher level features for determining dominance rankings from meetings were proposed in [3] but were extracted using manual speech transcriptions of the meetings so no automated audio feature extraction was attempted. In our previous work [4], we investigated how different audio and visual cues could be used for finding the most dominant person in a meeting. Preliminary investigations were carried out by using thresholded speaking energy values from individual headset microphones to determine speaking status. In addition, we compared this with the performance obtained using automated speaker diarization on a single audio source. However, we only considered a test data set where there was no variability in the annotations and a more detailed investigation of different single audio sources, and algorithmic strategies was not considered.

Given that we only have a single audio source, we must find a method to separate the signal into different speakers as well as their respective turn-taking patterns. In addition, the speech signals from individuals are likely to be significantly attenuated relative to the ambient noise, which leads to potential difficulties in disambiguating speakers, particularly during periods of overlapping speech. Automated speaker diarization is a well known solution to this problem but is affected by the required computational complexity. We have recently investigated methods of improving the diarization accuracy as well as increasing computation time with very promising results [5]. To this end, we investigate the performance trade-offs of estimating the most dominant person in a group meeting where the estimated speaking length is taken from the cluster outputs from different diarization strategies under various input conditions.

Although automated dominance estimation in group discussions is a relatively uncharted area, research on social dominance has been conducted in social psychology for several decades. Perceived dominance has been defined as 'expressive, relationally based communicative acts by which power is exerted and influence achieved' [6].

The novelty of this work is two-fold. Firstly, to our knowledge, there has been no work which conducts an extensive evaluation of the performance trade-offs using a speaker diarization for automated dominance estimation. Secondly, we also examine varying degrees of ambiguity that are possible in the annotation of the most dominant person to quantify how annotator variability can affect our automated judgements. It is important to note that no language-based cues are required since we rely solely on the estimated speaking length of each person as a cue for dominance.

2. DATA AND ANNOTATION

We use a subset of the AMI corpus [7] where five different exclusive sets of participants were used. Each group contained 4 participants who were asked to design a remote control over a number of sessions of varying length. Meeting over 15-35 minute non-scripted sessions encouraged team members to behave naturally. All meetings were carried out in a fully equipped meeting room as shown in Figure 1. The room contains a table, slide screen and white board. A circular microphone array containing eight evenly distributed sources is set in the middle of the table and one with four microphones is set in the ceiling. Participants were also asked to wear both headset and

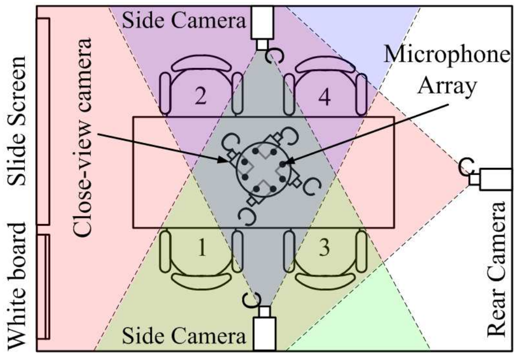

Fig. 1. Plan view of the meeting room set up. Only audio sources were used for automated dominance estimation.

lapel omni-directional microphones, which were attached via long cables to enable freedom of movement around the room. Cameras are mounted on three sides of the room and on the table to capture localised, or group visual behaviour. The video sources were used only for annotation purposes.

A total of 59 non-overlapping five-minute meeting *segments* from 11 *sessions* were provided for multi-observer annotation. 21 annotators were split into groups of three such that each group always annotated the same segments. For each watched segment, annotators were asked to rank the participants, from 1 (most) to 4 (least), according to their level of perceived dominance. They watched each segment using a video player with synchronised audio and multi-view video streams where three synchronised videos from the rear and side cameras were amalgamated. Annotators were not given any initial definition of dominance but were asked to provide a free-form verbal description on completion of the annotations.

In our previous work [4], we only considered the case where *all* annotators agreed on who the most dominant person was. Here we investigate the problem more deeply by probing the variability of labeling dominant behaviour. So in addition to the sub-set of 34 five-minute meetings where all annotators agreed on the most dominant person, we also investigate the cases where at least two out of three annotators agreed. For this majority-case subset, there were 57 out of 59 meetings, which indicates significant concensus amongst annotators. In addition, considering this larger data set allowed us to observe the degradation in performance when the variability of the annotations was increased.

3. AUTOMATED SPEAKER DIARIZATION

Speaker diarization, or the "who spoke when" task, tries to find speaker-homogeneous regions from speech-based audio signals. Our system uses a combination of agglomerative clustering using the Bayesian Information Criterion (BIC) and Gaussian Mixture Models (GMMs) of the frame-based cepstral features (MFCCs) from a single audio source. The agglomerative clustering approach starts with a large number of initial clusters and proceeds by an iterative procedure of cluster merging, model re-training and re-alignment. In the cluster merging step, a BIC score is used to determine whether two clusters should be merged. The algorithm stops when no further merging improves the BIC score. Experiments were also considered when the number of speakers was known. In this case, merging stops only when the number of clusters left and the true number of speakers is equal. A more detailed description can be found in [5].

This approach achieves satisfactory accuracy in terms of Speaker Diarization Error (DER). However, it exhibits inherent complexity due to the iterative cluster merging and sophisticated model selection procedure, which takes 62% of the total run-time and was identified as a bottleneck for the whole system. As a result, the original system was several times slower than real-time [8].

To achieve the goal of real-time performance, two fast-match techniques are used as a search-space tailoring method before applying the more computationally expensive BIC-merge score computation to the reduced set of more probable hypotheses. The two techniques, which are described in detail in [5], are called Pitch-correlogram Fast-Match (PCFM) and KL-divergence Fast-Match (KLFM). Since both filtering techniques have different influences on the DER, they will be applied to the dominance tasks in the remainder of this paper.

4. EXPERIMENTS AND RESULTS

4.1. Unsupervised Dominance Estimation

We associate the label of the most dominant person with that who had the longest total speaking length at the end of each meeting. We found this simple computational strategy to be robust, effective and fast [4]. Moreover, we found this to be more accurate in predicting the dominant person than more elaborate strategies such as that described in [2].

4.2. Experimental Conditions

The various experimental conditions can be categorized into a Single Distant Microphone (SDM) setting and a Mixed Individual Close-talk Microphone (MICM), as summarized in Table 1. For the MICM case, a single audio stream is obtained by mixing individual close-talk microphone data, i.e. Mixed Headset (MH) or Mixed Lapel (ML) using a basic summation. For the SDM condition, a single microphone is selected from a microphone array from either the table (AT) or ceiling (AC) sources.

MICM	SDM
Mixed Headset (MH)	Single Array Microphone:Table (AT)
Mixed Lapel (ML)	Single Array Microphone:Ceiling (AC)

Table 1. Summary of various experimental conditions

4.3. Diarization Error Rate across different experiments

To begin our investigation, we provide a summary of the variation in DER performance given our different experimental conditions and algorithmic strategies. This is shown in Table 2, which shows the different experimental conditions and their corresponding diarization error rates (DERs), signal to noise ratios (SNRs) and speed increases relative to real-time. The terms 'KLFM', 'PCFM', and 'NoFM' refer to the KL-divergence Fast Matching, Pitch Correlogram Fast Matching and No Fast Matching respectively. Note that these have been calculated from all the original 15-35min meeting *sessions* where the diarization was calculated from the full *sessions* rather than five-minute *segments*. In Table 2, we also show a colour-coded representation of the results where lighter colours indicate better performance. It is interesting to note that using a fixed number of speaker clusters yielded worse DER in 7 out of 12 experiments and that a higher SNR is related with a lower DER.

4.4. Speaker cluster/person association

Since we have no prior information about the seating of participants in the meeting, we needed to associate a speaker cluster with the correct person. There were two problems with this task. Firstly, for the case where model selection is done atomically, the speaker diarization algorithm can estimate more clusters than the number of speakers due to its reliance on the BIC score. Secondly, we needed to perform cluster/person association. The first problem was solved by only choosing the cluster with the longest speaking length as

Source	SNR (dB)	**Fixed number of speaker clusters**			**Automatic speaker cluster estimation**			
		KLFM	PCFM	NoFM	KLFM	PCFM	NoFM	
MICM MH	31	27.7	28.1	27.1	27.5	27.0	26.9	**A**
MICM ML	22	29.4	28.0	27.4	28.9	28.4	28.6	**B**
SDM AT	21	34.8	37.8	34.0	36.2	36.8	35.0	**C**
SDM AC	18	34.1	33.0	32.4	33.3	32.6	32.9	**D**
speed-up: (x>RT)		1.2	0.9	0.8	1.3	1	0.8	
		1	**2**	**3**	**4**	**5**	**6**	

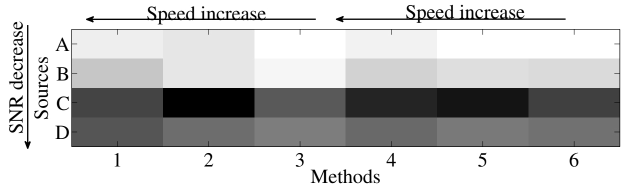

Table 2. Diarization results (DER) in numbers and also with colour coding below. Lighter colours represent better performance.

that of the most dominant person. Once the dominant person has been chosen, the associated speaker turn pattern was matched against all speaker segmentations from the personal headset microphones. These were extracted by thresholding speaker energy values. The channel which gave the smallest sum of square distances was labeled the most dominant person. It is important to emphasise here that while this means that the approach is not fully automatic for the purposes of evaluation, the method can still be automated, for example, if the only result that is required is the audio track of the most dominant person.

4.5. The Dominant Person Task (Full Agreement in Annotations)

We firstly targeted the task of finding the most dominant person from the 34-meeting data set containing all cases where all three annotators who annotated the meeting agreed on who the most dominant person was. Table 3 shows a colour-ranked visualisation of the results where lighter colours indicate higher performance. The rows and columns of the results table have been labeled with letters and numbers for easy reference.

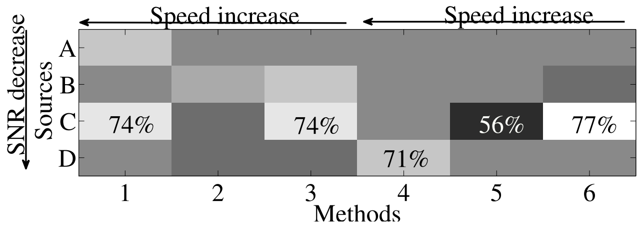

Table 3. Results for the most dominant person task where there was full agreement amongst annotators. A colour-coded representation of the performance in descending order is shown where higher performance is a shaded lighter.

The best performance was 77% which was obtained for experiment (6,C) where it is interesting to note that this case is taken from a single distant microphone from the table array with fully automatic estimation of the number of speaker clusters and no fast matching was applied. This was 8% lower than if audio sources from an ideal situation are used where individual close-talk microphone sources were used. Using the corresponding KLFM scheme with experiment (4,C), the performance dropped to 65% but gained a speed increase of 1.3 times faster than real-time. These results show an encouraging performance since the signal to noise ratio (SNR) is relatively low for the SDM AC source compared to the MICM cases. Encouragingly, the third best (71%) performing conditions used the fastest diarization strategy, fully automated speaker cluster estimation, and had the worse SNR. The second best performance of 74% was obtained from experiments (1,C) and (3,C), which required the number of speakers to be known beforehand. The mid-speed increase strategy, PCFM performed the least well and also gave the worst performance (56%) for experiment (5,C).

A summary of the overall performance by source type is shown in Table 4 where we see that the performance does not appear to decrease consistently with decreased SNR or increased speed. Indeed, on closer inspection of all the experiments in the context of their DERs and dominance performance, we found that while a decrease in SNR lead to a systematic drop in DER, this was not true for the dominance estimation. In our previous work [4], we computed

Source	**Mean**	**Standard Dev**	**Max**	**Min**
MICM MH	0.66	0.02	0.71	0.65
MICM ML	0.66	0.03	0.71	0.62
SDM AT	0.68	0.07	0.77	0.56
SDM AC	0.65	0.03	0.71	0.62

Table 4. Table showing the mean, standard deviation, and maximum and minimum values for each input source type for most the dominant person task where there was full agreement amongst annotators.

the speaker diarization output using a delay sum of the headset signals. Using this method, we achieved a performance of 74% while in these experiments, the best score is 77%. It is encouraging to see that without any beam-forming techniques, we obtain very similar performance. Note that the 3% increase corresponds to only one meeting difference in performance between the two cases so it is difficult to make any conclusive conclusions about whether the the performance is necessarily better.

4.6. The Dominant Person Task (Majority Agreement in Annotations)

In the following task, we studied the performance of the dominance task when at least two out of the three annotators who annotated the same meeting agreed on the most dominant person. From studying the results in Table 5 in more detail, we can see that the best performance was 63% from experiments (5,B) and (6,C) where more noisy input sources were used but one case also used a mixed headset signal. Comparing this with the ideal case where individual close-talk microphones were used, this showed a 14% drop in performance. Again, the worst performance was experiment (5,C) with 51% but the KL fast match method (4,C) performed better despite having the greatest speed increase over the original diarization algorithm. Experiment (4,D) performed well with the second best score of 61%. In addition, some experiments using a fixed number of speaker clusters performed equally well, but required prior knowledge of the number of participants in the meeting. It is interesting to note that there is

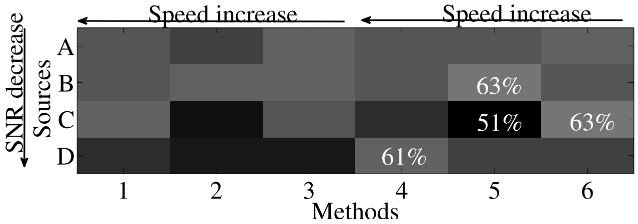

Table 5. The results for the Most Dominant person task where the majority of annotators agreed. A colour-coded representation of the performance in descending order of rank where higher performance is shown with a lighter colour.

a systematic drop in performance between this dominance task and that in the previous section, which is shown by the overall darker shade of the results table. This suggests that a higher variability in the human judgement leads to a more challenging data set; the drop in performance can also be seen from the individual headset results where the performance dropped to 77% from 85%.

Studying the performance for this dominance task across different source types in Table 6, we found that in general, a decrease in SNR lead to a drop in performance. In both dominance tasks, the SDM AT source showed greatest variability in performance, yielding both the best and worse performance from all experiments in their task category.

Source	Mean	Standard Dev	Max	Min
MCIM MH	0.60	0.01	0.61	0.58
MICM ML	0.61	0.01	0.63	0.60
SDM AT	0.57	0.04	0.63	0.51
ADM AC	0.57	0.02	0.61	0.54

Table 6. Mean, standard deviation, and maximum and minimum value for each input source types for the task where the majority of annotators agreed on the most dominant person.

Figure 2 shows a summary of the decrease in performance across the different dominance tasks where all experiments for each task were ranked in descending order. In addition, the performance using speaker segmentations from individual headset microphones are also provided for comparison. While in both dominance tasks, using automated speaker diarization leads to reduced performance, we see that there is a consistent drop for the more ambiguous sub-set. In addition, the increased variability of the test set seems to lead to less sensitivity to the experimental conditions.

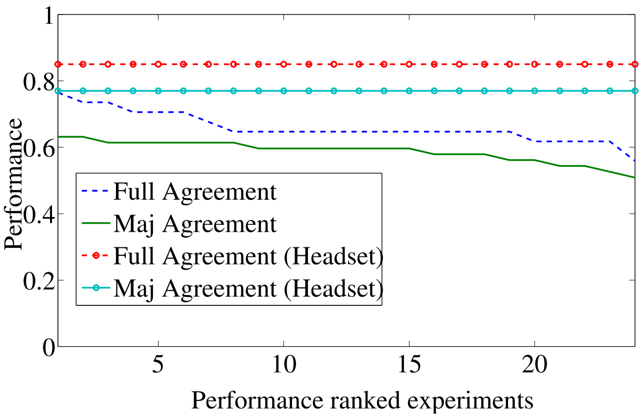

Fig. 2. Graph showing the degradation in performance over all the experimental conditions. For comparison, the performance using speaking segmentations generated from individual headset microphones has also been included for the different dominance tasks.

5. CONCLUSION AND FUTURE WORK

As expected, there was a decrease in performance between using the ideal case of speaker segmentations generated using the individual headset microphones compared to the single source cases. For the dominance task where full agreement was observed among annotators, the performance decrease was at least 8% while for the majority-agreement case, the drop was even greater, at 14%. This highlights the challenge of trying to estimate speaker turns using a single audio source rather than from individual ones, particularly during overlapping speech.

Overall, the results show some surprising findings. In particular, while the DER appears to be more strongly dependent on the SNR, this does not seem to be the case for the dominance tasks where the best performing experimental conditions used a SDM. As expected, these cases corresponded to no fast matching but surprisingly, estimated the number of speaker clusters automatically. Surprisingly, while the DER suffered with decreased SNR, it appeared that the performance on the dominance tasks was not as directly affected. We also observed cases in the various speed-up strategies where the performance increased even though the speed also increased. However, the contrary was also observed. The best overall compromise between performance and speed in both cases was shown for the SDM AC case, which had the worst SNR.

There was also a consistent decrease in performance with increased variability in the test data, indicating the stability of the annotations and shows that there may be other cues which play an important role in determining dominant behaviour in more ambiguous meeting scenarios.

At the moment, the DER has not been calculated for the subsets of meetings which were used for the dominance tasks. We plan to investigate more concretely, possible correlations between the DER and dominance performance and speed-increase strategies for the corresponding subset of meeting *segments*. We would also like to find a fully automated way of performing speaker cluster/seat association will be investigated, using both video and audio cues.

Acknowledgments

This research was partly funded by the US VACE program, the EU project AMIDA , the Swiss NCCR IM2, and the German Academic Exchange Service (DAAD).

6. REFERENCES

[1] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, "Learning human interactions with the influence model," in *NIPS*, 2001.

[2] D. Zhang, D. Gatica-Perez, S. Bengio, and D. Roy, "Learning influence among interacting Markov chains," in *NIPS*, 2005.

[3] Rutger Rienks, Dong Zhang, Daniel Gatica-Perez, and Wilfried Post, "Detection and application of influence rankings in small group meetings," in *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*. 2006, pp. 257–264, ACM Press.

[4] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. Ba, J-M Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez, "Using audio and video features to classify the most dominant person in a group meeting," in *ACM Multimedia*, 2007.

[5] Y. Huang, O. Vinyals, G. Friedland, C. Müller, N. Mirghafori, and C. Wooters, "A fast-match approach for robust, faster than real-time speaker diarization," in *ASRU*, 2007.

[6] N. E. Dunbar and J. K. Burgoon, "Perceptions of power and interactional dominance in interpersonal relationships," *Journal of Social and Personal Relationships*, vol. 22, no. 2, pp. 207–233, 2005.

[7] J.C. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraiij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, M. McCowan, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus: A pre-announcement," in *Proc. MLMI*, 2005.

[8] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Proceedings of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop*, 2007.