Estimating Dominance in Multi-Party Meetings Using Speaker Diarization

Hayley Hung*, *Member, IEEE,* Yan Huang, *Member, IEEE,*
Gerald Friedland, *Member, IEEE,* Daniel Gatica-Perez, *Member, IEEE,*

***Abstract*—With the increase in cheap commercially available sensors, recording meetings is becoming an increasingly practical option. With this trend comes the need to summarize the recorded data in semantically meaningful ways. Here, we investigate the task of automatically measuring dominance in small group meetings when only a single audio source is available. Past research has found that speaking length as a single feature, provides a very good estimate of dominance. For these tasks we use speaker segmentations generated by our automated faster than real-time speaker diarization algorithm, where the number of speakers is not known beforehand. From user-annotated data, we analyze how the inherent variability of the annotations affects the performance of our dominance estimation method. We primarily focus on examining of how the performance of the speaker diarization and our dominance tasks vary under different experimental conditions and computationally efficient strategies, and how this would impact on a practical implementation of such a system. Despite the use of a state-of-the-art speaker diarization algorithm, speaker segments can be noisy. On conducting experiments on almost 5 hours of audio-visual meeting data, our results show that the dominance estimation is robust to increasing diarization noise.**

I. Introduction

FROM an initial encounter between unacquainted individuals, a dialog begins, which can start from a contest of who can maintain eye contact for the longest, to who speaks first [1]. These two examples in particular can be viewed as typical behaviors for establishing hierarchy, which is not necessarily inherent to the group, and must be established through verbal or non-verbal interactions [2]. Specifically, the innate behavior in humans to establish their status within a group can be viewed as dominance. Studying this particular type of behavior in groups is useful for assessing the effectiveness of teams or as a cue for searching or browsing many recorded meetings. For example, the most dominant person could be causing their team to perform less effectively or the least dominant could be encouraged to take a more active role in future meetings. In other cases, if someone wants to find a recording of a particular meeting, sometimes cues which are related to memories of the interactions other than dates and locations, might help to find the information more quickly and easily. Dominance can also be used to indicate the hierarchical position of a person for previously unseen groups.

In speech processing, there has been much work on using just non-verbal cues to classify aspects of human behavior such as involvement [3] or frustration and anger [4]. Here, we draw on evidence both in social psychology and ubiquitous computing that non-verbal cues, specifically speaking length, is a very good non-verbal indicator of dominance [5], [6]. In practical situations having a microphone for each person may not be feasible or indeed practically desirable. There may only be a single microphone, requiring the audio signal to be temporally segmented and associated with the correct speaker. 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 limitations of high computational complexity if improved estimates of the speaker segmentations are required.

The work presented in this paper studies closely, how estimating the dominance of participants in a group meeting using just a single source can be affected by: (i) different strategies for increasing the efficiency of the diarization algorithm using an algorithm developed by Huang et al. [7]; and (ii) the experimental conditions. In this paper, we enrich this work by providing a more detailed study of the relation between the diarization error rate (DER) and dominance estimation performance under the same experimental conditions. We present a fully automatic system that is practical to use and does not necessarily require user intervention.

We study the variations in performance for different and apposing tasks to understand better the differing nature of the two behavior types. In social psychology, it has been noted that dominant people tend to be more verbally and physically active while submissive people are less so [2]. However, it was also observed by social psychologists that inferring the behavior from less dominant people can be difficult since they interact less actively, leading to a lower confidence in judgments [6]. This paper studies how such variations in behavior from extreme cases of dominant behavior are linked to not only speaking length, but also how estimates of the speaking length for an unknown number of meeting participants can affect the two opposing classes of behavior (most and least dominant).

Also, inferring dominant behavior between interacting individuals, is known to be a subjective task, which can vary across individuals and also between those observing and participating in the interaction [2]. This has a significant impact for automated systems where human judgments are required for evaluation purposes.

The novelty of this work is listed below:

- A fully automatic, computationally efficient method of estimating the most dominant person from a single microphone.
- An extensive evaluation of the performance trade-offs using speaker diarization for previous automated dominance estimation tasks.
- An examination of the differing degrees of variation that exist in the annotations of dominant behavior to quantify how annotator variability can affect automated judgments given different conditions and strategies of diarization.
- Experiments on two different dominance tasks, namely estimating the most and least dominant person when only a single microphone source is available.

It is important to note that no language-based cues are required since we rely solely on the nonverbal information of each person as a cue for dominance.

The rest of this paper is structured as follows: Section II discusses related work in social psychology on defining the characteristics of dominant behavior and in particular, why speaking length is a good

H. Hung is with the University of Amsterdam and this work was carried out while she was working at Idiap Research Institute, Switzerland email:H.Hung@uva.nl

D. Gatica-Perez is with Idiap Research Institute, Switzerland and Ecole Polytechnique de Lausanne (EPFL), Switzerland

Yan Huang is with Microsoft Corporation, Redmond, USA and this work was conducted while she was with the International Computer Science Institute (ICSI)

G. Friedland is with the International Computer Science Institute, Berkeley, USA

indicator of dominant behavior; Section III details related work in the area of automated dominance estimation; Section IV provides details about our experimental approach; Section V-B describes the data and annotation procedure for our experiments; Section VI provides details of the speaker diarization approaches that are used and the experimental conditions that we consider; Section VIII and IX provides and discusses our results using the various diarization strategies, experimental conditions, and dominance tasks; we summarize, compare and discuss in Section X; and we conclude in Section XII.

II. Dominance in Social Psychology

Over several decades, social psychologists have tried to characterize dominant behavior in face to face discussions. Often, it is used synonymously with power, influence, status and domineeringness. However, some psychologists such as Dunbar and Burgoon have argued otherwise by suggesting that perceived dominance is a set of "expressive, relationally based communicative acts by which power is exerted and influence achieved" [2] (p208). More specifically, Dunbar and Burgoon suggested that while power and status are properties that exist through a long-term establishment of hierarchy, dominance is viewed as "necessarily manifest. It refers to context and relationship-dependent interactional patterns in which one actors assertion of control is met by acquiescence from another" (p.208) [2]. This idea of assertion and acquiescence was suggested previously by Rogers-Millar and Millar [8] who defined domineeringness and dominance as two separate control variables; domineeringness is the proportion of 'one-up' maneuvers a person performs during a conversational interaction; dominance is the ratio of 'one-up' to 'one-down' maneuvers.

Dunbar and Burgoon [2] quantified the effect of different non-verbal cues on a person's perceived dominance levels. These cues were categorized as vocalic and kinesic features, referring to speech (e.g. speaking time, loudness or energy, speaking rate, pitch vocal control or interruptions [9]) and gesture based cues (e.g. body movement, posture and elevation, facial expressions, gestures or eye gaze [10]) respectively.

More specifically, Schmid Mast conducted a meta-analysis of 40 articles containing 45 studies in social psychology performed over 5 decades, concluding that dominance could be inferred through speaking time [5]. This meta-analysis resulted in 45 examples of dominance being expressed through speaking time, which could be quantified by their effect size using the product-moment correlation coefficient. Measures of the effect size are commonly used in meta-analyses to quantify the statistical significance of experimental findings from different data. 2,850 participants were involved in the studies. Schmid Mast found that dominance was expressed through speaking time more in role-based dominance scenarios (e.g. manager/employee or teacher/student) compared to cases where dominant personality traits were observed. The highest effect sizes of dominance expressed through speaking time, extracted from the 'assigned' (role-based) and 'actual' (trait-based) dominance studies were 0.76 and 0.31 respectively.

Interruptions can also be viewed as an individual's attempts to 'grab' the conversational floor or assert themselves. In particular, West and Zimmerman [9] found that those who interrupted more tended to be more dominant. However, Tannen also presented examples to suggest that interruptions could be co-operative as well [11]. For the work described here, we assume that there are no speaker overlaps since handling interruptions complicates the diarization process and would make it more difficult for us to analyze how the diarization algorithm relates to the final dominance estimation output. Therefore, interruptions were not considered in this study. However, Beattie [12] observed that in the context of interruptions in political interviews, people are more likely to remember the way that a politician delivers a speech, than exactly what they said [12]. He also stated that shy people tend to "have longer pauses between turns and speak less frequently and for a shorter percentage of the time" (p. 94).

Studying the prosodic features of the voice more closely, contrasting findings have been made into whether certain characteristics of the voice are more correlated with dominant behavior. For example, both low and high vocal fundamental frequencies (F_0) have been associated with dominant behavior [13], [14], [15], while a high F_0 was an indicator of submissiveness. There has also been research to show that loudness of the vocal signal, greater pitch and a faster speaking rate is correlated with perceptions of dominance for someone reading both a confident and doubtful piece of text [15]. Faster speaking rate is also indicative of competence, which is also Buller and Aune suggested was linked to dominance [16]. However, they also found that perceptions of competence through a faster speaking rate was only perceived by observers who were good at understanding and interpreting nonverbal cues. The studies listed above illustrate some evidence for certain prosodic features to indicate dominance. However, the findings were conducted using subjects who were asked only to listen to tapes of actors, who articulated their voices differently, depending on the experiments. While other signals for dominance such as competence and confidence can be useful, they do not necessarily become apparent through an interaction so it is difficult to know if such findings would be difficult to conclude if such findings can be applied to the understanding of meeting dynamics. The study in this paper, follows the evidence from the study of Schmid Mast [5], by using speaking length as a measure of dominance. The reasons and advantages of concentrating on this feature are that it is fast to compute, performs relatively robustly, and can be used when only a single microphone is available.

Perceived dominance is also an aspect of dominance that has been investigated by social psychologists. Dominant behavior can be perceived either by observers of an interaction or the participants of the interaction itself [10]. For example Dovidio et al. [10] found that people could perceive visual dominance displayed by others. More details on understanding dominance from a social psychology perspective can be found in [2], [17].

III. Computational Approaches for Dominance Modeling

Early work in automatic dominance modeling in conversations was done by Basu et al. [18] who studied debates in 5-participant meetings. They used a combination of manually and automatically extracted audio-visual features such as speaking status, turns, and visual activity patterns from skin-color blob-tracking. They modeled exchanges on a dyadic basis using Markov chains. They showed preliminary results using human interaction data where two out of five participants were pre-selected to debate for one minute before the debate was opened to the rest of the participants. This lead to an artificially constrained conversational setting where there would always be a larger amount of discussion between the first two interacting participants.

Semantically higher level features for determining dominance rankings from meetings were proposed by Rienks et al. [19] but were extracted using manual speech transcriptions of the meetings so no automated audio feature extraction was attempted. They modified the task to labeling each participant with low, medium and high levels of dominance according to human annotations. Using a support vector machine (SVM) approach, they found good performance for estimating dominance levels. After this, Rienks et al. [19] conducted a study to compare the performance of two different methods for estimating influence in meetings.

Recently, Otsuka et al. [20] used non-verbal cues based on automatically extracted gaze patterns, to explain pair-wise influence in group discussions. Using a Dynamic Bayesian Network (DBN), audio and visual sources were combined to estimate the conversational context and therefore the gaze of participants during the conversation. Measures for interpersonal influence were then calculated based on these gaze patterns. They used 10 minutes of conversational data of pre-defined topics collected from two 4-participant groups. The participants were asked to come to a conclusion on each topic after 5 minutes. There was no quantitative evaluation of their method. In summary, in all previous work, no attempts were made to use distant microphones, relying instead on relatively clean, good-quality signals from close-talk microphones.

Hung et al. investigated how different audio and visual cues could be used for finding the most dominant person in a meeting [21]. They showed that the speaking length performed the best as a single feature for indicating dominance. These preliminary investigations were carried out by using thresholded speaking energy values from individual headset microphones to determine speaking status. Following from this, Jayagopi et al. [6] completed a more comprehensive study which investigated audio-visual cues for dominance estimation using both an unsupervised and supervised model. Again, audio features were recorded from headset microphones so speaking activity for each individual could be extracted relatively cleanly. Finally, they found that speaking length still had superior performance.

After the investigation of different cue types, Hung et al. [22] investigated how the dominance estimation would vary if only one microphone was available. While the scenario is more challenging, it is also more practically desirable since little hardware is required and there is no need for specialist equipment. In the work presented here, we enhance these experiments by considering how variations in the annotations of dominance could effect the estimation performance and we also consider the tasks of estimating both the most and least dominant person. We also provide a much more thorough analysis of how the results vary, which extends and enriches the work in [22]. In addition the experiments we present here differ from our preliminary experiments [22] where the diarization algorithm was performed on longer meeting sessions that could range between 15 and 35 minutes. Therefore, the test data is much more challenging when only 5 minutes of data is used for building speaker models.

IV. Our Approach

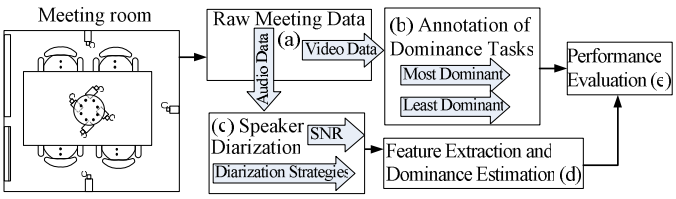

Fig. 1. Flow diagram of our approach. A description of each block of the flow diagram are provided in the main text.

We investigate several different aspects of the problem of estimating the dominant person in conversational settings where only audio data from a distant microphone is available. Practically speaking, a system would be easier to use if it was fast and easy to set up. Both these criteria can be affected by hardware constraints such as where microphones can be placed relative to all the meeting participants as well as power consumption constraints where minimum time should be spent on analyzing the meeting data before storing it for future reference. Our approach is summarized in Figure 1 and below:

(a): Section V-A describes the audio-visual meeting data that was captured and the scenario that was used to record the data.

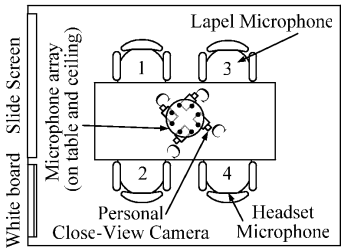

Fig. 2. Plan of the meeting room. Only audio sources were used for automated dominance estimation. The cameras were used for human annotations.

(b): Section V-B describes the annotation procedure for determining the dominance of participants in the meeting data. Through this, two different dominance tasks with two additional sub-tasks are identified.

(c): Section VI describes how, using speaker diarization, a single audio source can be divided into speaker clusters, where each represents a person and when they speak. To assess the performance of our dominance estimation technique, we modify the speaker diarization algorithm in different ways for faster performance. In addition, we adjust the audio source conditions (i.e the signal-to-noise ratio (SNR), which is sensitive to the distance of the source from the speakers) to see how the dominance estimation performance will be affected.

(d): Section VIII-A describes how dominant people are estimated from the speaker clusters generated from the speaker diarization algorithm.

(e): Section VIII describes and discusses the results.

V. Data And Annotation

A. The Data

A subset of the AMI corpus [23] where five different exclusive sets of 4-participant meetings were used. Each group was asked to design a remote control over a number of sessions varying over 15-35 minutes. The sessions were not scripted and the participants were allowed to move freely in the room to encourage natural behavior. All meetings were carried out in the room shown in Figure 2. The room contains a table, a slide screen, and a white board. A circular microphone array containing eight evenly distributed microphones is set in the middle of the table and a linear array with four microphones is set in the ceiling. Participants were also asked to wear both headset and lapel omni-directional microphones, which were attached via long cables to enable freedom of movement around the room. Cameras were mounted on three sides of the room and on the table. The video sources were used only for annotation purposes. Ground truth speaker segmentations for each participant were manually generated.

B. The Dominance Annotation Procedure

The dominance tasks and the annotation procedures used in our experiments were presented by Jayagopi et al. [6]. In the data set, 59 non-overlapping five-minute meeting *segments* were extracted from 11 *sessions*. These were used for human annotations of the dominance task. There were 21 annotators in total, who 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 synchronized audio and multi-view video streams where three synchronized videos from the rear and side cameras were shown, as illustrated in Figure 3. Annotators were not given any initial definition of dominance.

C. Defining Dominance Tasks

Using the annotation analysis from Jayagopi et al. [6], The two dominance tasks we used are summarized and defined in Table I below. Within each dominance task there are two sub-tasks that correspond to meetings where there is (i) **Full** agreement among annotators who labeled the same meeting, and (ii) **Maj**ority where at least 2 out of the 3 annotators agreed. We also provide the number and proportion of meetings that were used for each sub-task.

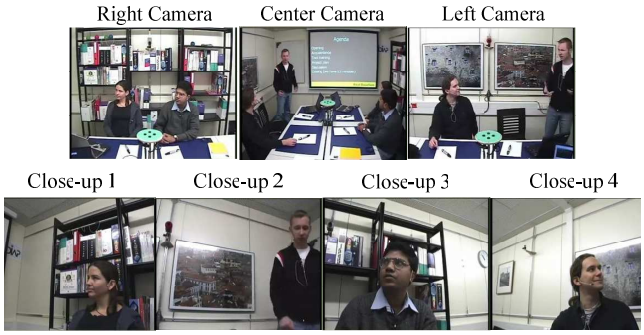

Fig. 3. Example screen-shots of the seven camera views available in the meeting room. Top row: the right, center and left cameras which were used for annotation; Bottom row: the view from each of the close up cameras.

Dominance Estimation Task	**Sub-Tasks**	**Number of Meetings**	**Proportion of Total Meetings (%)**	**Self-reported Mean Annotator Confidence**
Most	**Full**-agreement	34	57.6	1.74
	Majority-agreement	57	96.6	1.85
Least	**Full**-agreement	31	52.5	2.11
	Majority-agreement	54	91.5	2.4

TABLE I
DOMINANCE TASKS AND CORRESPONDING DATA-SETS.

VI. SPEAKER DIARIZATION

From a single audio source with an unknown number of speakers, speaker diarization segments the signal into speaker-homogeneous regions with the goal of answering the question "who spoke when?" [24]. We use the speaker diarization algorithms of Huang et al. [7] to extract speaker clusters from the single audio stream using different computationally efficient strategies.

A. ICSI Speaker Diarization System

The ICSI speaker diarization system uses an agglomerative clustering method with the Bayesian Information Criterion (BIC) [25] to both identify individual speakers and the number of speakers in a given audio stream. In this system, each individual's voice is modeled by a Gaussian Mixture Model (GMMs) of frame-based cepstral features (MFCCs) [24], [26] are used to create estimates of each speaker's voice. The system extracts MFCC features from audio, discriminates between speech and non-speech regions (speech activity detection), and then uses an agglomerative clustering approach to perform the segmentation and grouping in one step. The final output is a set of clusters (one for each speaker), with the estimated speaking patterns (speech and silence) for each. Further details can be found in [27], [26].

The algorithm is first initialized using k clusters with the initial segmentation generated by uniformly partitioning the audio into k segments of the same length. Then it iteratively performs re-segmentation, model re-training, and cluster merging as follows:

Re-Segmentation: The Viterbi algorithm is used to search for the optimal path through different speaker states and obtain an updated speaker segmentation. A minimum duration constraint of 2.5s is enforced in this procedure.

Model Re-Training: Given the new segmentation, the speaker GMMs are re-estimated using the Expectation Maximization (EM) algorithm.

Cluster Merging: We determine which two clusters should be merged and when the merging should stop using the BIC. A Merge Score, which is based on the BIC, is calculated for each merge hypothesis. The pair-wise merge which produces the best improvement in the merge score is identified as the best pair of merge candidates. If no merge improves the merge score, the algorithm terminates.

The output consists of a set of clusters where for each, a speech segment hypothesis is provided in terms of the start and end times, and the label of the speaker cluster. The speaker diarization performance is measured by the Diarization Error Rate (DER) which is defined by the National Institute of Standards and Technology (NIST) (http://nist.gov/speech/tests/rt/rt2004/fall). The DER is decomposed into three components: misses (speaker appears in the reference, but not in the hypothesis), false alarms (speaker appears in the hypothesis, but not in the reference), and speaker-errors (the mapped reference speaker is not the same as the hypothesized speaker). To calculate the DER, a dynamic programming procedure is used to find the optimal one-to-one mapping between the hypothesis and the ground truth segments so that the total overlap between the reference speaker and the corresponding mapped hypothesized speaker cluster is maximized, i.e. the DER is the ratio of the non-overlapped region and the total length of the audio.

B. Rapid Speaker Diarization Using Fast-Match

Although the system described in Section VI-A achieves high performance in terms of accuracy, it does not meet the real-time requirement of downstream applications. To achieve the goal of robust, rapid speaker diarization, a fast-match framework for fast speaker diarization was proposed in [7]. It uses a computationally inexpensive method to reduce the merge hypothesis space of the more expensive and accurate search. Two fast-match strategies have been explored to significantly reduce the computational load of the BIC-based model order selection process, each of which can be used interchangeably. The first strategy uses the pitch-correlogram [28], to capture speaker variances by looking at the statistics of pitch patterns at the frame level. The second strategy uses KL-divergence to compare two probabilistic distributions which characterize the speaker clusters. The best result is achieved using the KL-divergence fast-match strategy, which speeds up the baseline system by 41 % without affecting the speaker Diarization Error Rate (DER).

C. Speaker Diarization on the AMI data

We conducted speaker diarization experiments on each of the 5-minute meeting segments. In one track of these experiments, the system was run in a completely data-driven fashion using BIC to determine when the merging should stop. We refer to this track subsequently, as the 'Automatic' case. In the second track ('Fixed'), since the stopping criterion for the cluster merging is data-driven, it is possible for the number of clusters to be unequal to the number of participants in the meeting. Thus a more controlled procedure is used to constrain the number of clusters at the end of the algorithm, i.e. the merge can only stop when the number of clusters drops down to n ($n \leq 4$). With the integration of prior knowledge about the number of true speakers, we hypothesized that the fixed case would enforce a better estimate of the number of speakers and hence better diarization performance.

D. Experimental Conditions

We tested the diarization algorithm using increasingly noisy signals to see how this affected the performance of both the diarization and dominance tasks. The various experimental conditions can be categorized into a Single Distant Microphone (SDM) setting and a Mixed Individual Close-talk Microphone, as summarized in Table II. For the close-talk microphone case, a single audio stream is obtained by mixing individual head-set microphone data through a basic summation across all 4 streams at each sample, i.e. Mixed Headset (H) or Mixed Lapel (L). For the single distant microphone condition, a single microphone is selected randomly from either the microphone array on the table (T) or that of the ceiling (C). For these 4 experimental conditions, a range of different signal-to-noise ratios (SNR) were represented, as shown in Table II. While the

mixed signals may appear artificial, the results from these conditions can be used to estimate what the performance would be if all the participants in the meeting were situated closer to the single microphone source. Note that the SNR decreases as the distance between each participant and the microphone source increases and for condition C, the participants are likely to be situated between 1.5-2m away from the microphone.

Source Types	SNR(dB)
Mixed Individual Close-talk Microphones	
H: Mixed **H**eadset	31
L: Mixed **L**apel	22
Single Distant Microphone	
T: Single Array Microphone:**T**able	21
C: Single Array Microphone:**C**eiling	18

TABLE II
SUMMARY OF VARIOUS EXPERIMENTAL CONDITIONS

E. Diarization Error Rate across Different Experiments

To begin our investigations, we provide a summary of the DER performances given our different experimental conditions and algorithmic strategies as shown in Table III. The terms 'KL', 'PC', and 'No' refer to the KL-divergence fast match, Pitch Correlogram fast match, and No fast match respectively. Conceptually these computationally efficient strategies can be thought of as fast, medium, and slow methods. The results have been colour coded where lighter colors indicate better performance. The upper part of Table III also shows the different experimental conditions and their corresponding diarization error rates (DER), the signal to noise ratios (SNRs), and speed increases relative to real-time. The rows and columns of the results table have been labeled with letters and numbers for easy reference. This labeling system will be used for subsequent results tables.

(a)

Source	SNR (dB)	**Fixed number of speaker clusters**			**Automatic speaker cluster estimation**		
		KL	PC	No	KL	PC	No
H	31	33.17	32.17	32.52	33.78	32.83	33.16
L	22	34.71	34.19	34.94	36.47	35.91	36.35
T	21	35.34	34.94	34.94	36.14	36.19	36.16
C	18	35.94	36.22	34.85	35.96	36.89	36.55
		1	**2**	**3**	**4**	**5**	**6**

(b)

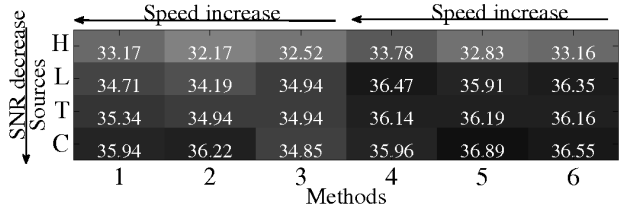

TABLE III
DIARIZATION RESULTS (DER) IN NUMBERS AND COLOUR CODING. LIGHTER COLORS REPRESENT BETTER PERFORMANCE. SEE TABLE II FOR THE CORRESPONDING SOURCE LABELS. 'KL', 'PC', AND 'NO' REPRESENT THE FAST MATCHING SCHEMES THAT USE EITHER KULLBACK-LIEBLER DIVERGENCE, PITCH CORRELOGRAMS OR NEITHER RESPECTIVELY TO DETERMINE LIKELY MERGE HYPOTHESES.

VII. HYPOTHESES

Given the number of different conditions and diarization strategies that could be employed, and also dominance estimation tasks, we provide a set of expected outcomes from varying these factors.

[H1]:Reduction in the signal to noise ratio of the input source leads to worse dominance estimation performance since the DER is affected.

[H2]:Reducing the computational complexity of the speaker diarization algorithm will lead to worse dominance estimation performance.

[H3]:Higher variability in the annotations leads to worse dominance estimation performance since human judgments are not unanimous.

[H4]:Due to lower annotator confidence and also the expectation that the least dominant person will speak less, it will be more difficult to estimate this person in the meetings compared to the most dominant person estimation task.

VIII. ESTIMATING THE MOST DOMINANT PERSON

In this section, we describe the method for estimating the dominant person, how the evaluation is performed, and also show our results. Since determining whether someone is dominant is quite subjective, we compare the results with different annotator agreement to study how the estimations will be affected with greater annotator variability.

A. 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 five-minute meeting segment. We found this simple computational strategy to be robust, effective, and fast [21]. Moreover, we found this to be more accurate in predicting the dominant person than more elaborate strategies such as that described in [19]. Moreover, the use of speaking time as a measure of dominance is supported by findings in social psychology, as discussed in Section II. The use of a static measure of dominance is also useful for minimizing the direct effect of temporal mis-alignments of a person's speaking status.

B. Speaker Cluster/Person Association

Since we have no prior information about the seating order of the participants in the meeting, it was not possible to know which speaker cluster corresponded to which person so there are two problems that need to be addressed. Firstly, for the case where model selection is done automatically, the speaker diarization algorithm can estimate more clusters than the number of speakers due to its reliance on the BIC score. So, a one-one association of clusters to participants is not always possible. Secondly, we needed to perform some cluster to person association to identify the dominant person.

The two problems were solved by only choosing the cluster with the longest speaking length as that of the most dominant person. Then for evaluation, two methods were proposed. For the first method, once the dominant person was chosen, the associated speaker turns pattern was matched against all speaker segmentations from the ground truth. The channel which gave the smallest sum of square distances was labeled the most dominant person. The second method used the speaker cluster-to-person labels that were generated during the calculation of the DER for evaluating the estimates. Note that the final DER was calculated based on a one-to-one mapping of each of the speaker labels to clusters, regardless of the estimated number of clusters. In our case, we assumed the longest cluster is always mapped to a speaker. This latter evaluation method was used to observe if an improvement in performance could be gained from taking advantage of the dynamic programming technique used to maximize the DER across all speakers in the meeting. It is important to emphasize here that while both approaches are not fully automatic for the purposes of evaluation, the method is still automatic, if the goal is only to extract the audio track of the most dominant person.

C. Full Agreement Among Annotators

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. The average classification accuracy for each experiment is shown in Table IV. The best and worst results were 74% and 62% respectively. It was encouraging to see that out of the 5 cases where the best performance was obtained, two corresponded to condition T where an SDM was used. These results were also achieved using the fastest diarization strategy. For over half of the experimental conditions, the performance was equal to or better than the session-based experiments reported in [22]. The segment-based results showed a lack of sensitivity to the SNR, which contradicts

our expectations states in H1. The baseline result, estimated using the individually extracted headset speaker segmentations, yields a performance of 85%, showing a significant decrease in performance when a single audio source is used.

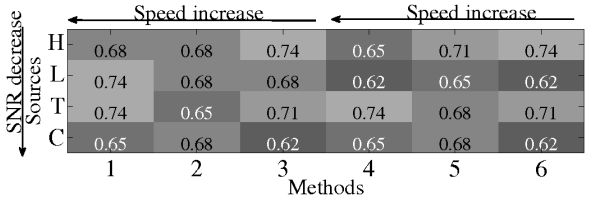

TABLE IV

RESULTS FOR THE MOST DOMINANT PERSON TASK. HIGHER PERFORMANCE IS SHADED LIGHTER. SEE TABLE II FOR THE CORRESPONDING SOURCE LABELS AND TABLE III FOR THE CORRESPONDING DIARIZATION STRATEGY LABELS.

The mappings of the clusters to speakers, which are generated as a by-product of the DER calculation were used to evaluate the estimated dominant cluster. The results are shown in Table V. Here we see consistently better performance compared to just matching the speaking patterns of the most dominant cluster. The highest average performance of 79% was obtained for the fastest diarization strategy and also for condition T, with the second worst SNR.

1) Using Speaker Turns rather than Speaking Length: We also used another speaking activity feature. There are other speech features that could be extracted such as speaker turns. A turn is considered to be an interval of time for which a person's speaking status is true. Taking the person with the total number of speaker turns to be the most dominant, has been found to be effective [6]. Also, using this feature allows us to observe how the dominance estimation would be affected by representing the temporal accuracy of the diarization estimates. Speaker turns represent the ability of each participant to 'grab the floor'. The estimated most dominant person was evaluated by matching the cluster with the greatest number of turns to the ground truth speaker segmentations. Note that only the speaker turns which were greater than 4 seconds were considered since the feature was found to be much more discriminative with this constraint. The shorter turns could be roughly approximated to back-channels, which tend to support rather than disagree with what is being said.

The results are shown in Table VI we see that there is not a significant difference between the results, though the majority of conditions showed a decrease in performance compared to those using speaking length (see Table IV). This could be due to the greater sensitivity of the speaker turns feature to the temporal accuracy of the turn-taking patterns.

The performance of both the speaking length and speaker turns features were studied in more detail by observing correct and incorrect estimates of the the most dominant person, as shown in Figure 4. For both feature types, the accumulated value using the ground truth speaker segmentations are generated and shown in the

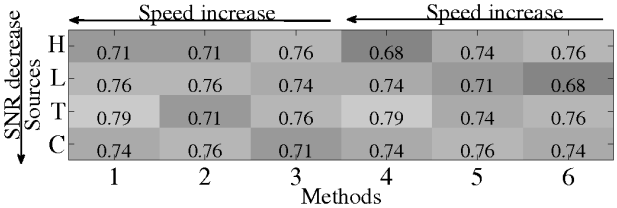

TABLE V

PERFORMANCE WHEN ESTIMATING THE MOST DOMINANT PERSON USING THE SPEAKER-CLUSTER MAPPINGS WHICH WERE A BY-PRODUCT OF THE DER COMPUTATION. SEE TABLE II FOR THE CORRESPONDING SOURCE LABELS AND TABLE III FOR THE CORRESPONDING DIARIZATION STRATEGY LABELS.

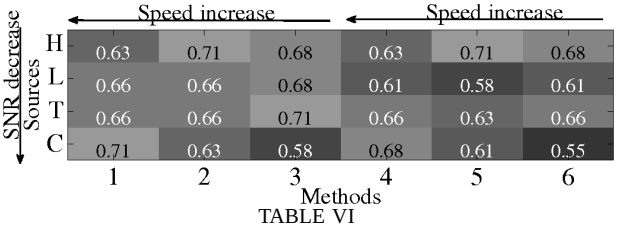

TABLE VI

PERFORMANCE WHEN ESTIMATING THE MOST DOMINANT PERSON USING THE TOTAL SPEAKER TURNS. SEE TABLE II FOR THE CORRESPONDING SOURCE LABELS AND TABLE III FOR THE CORRESPONDING DIARIZATION STRATEGY LABELS.

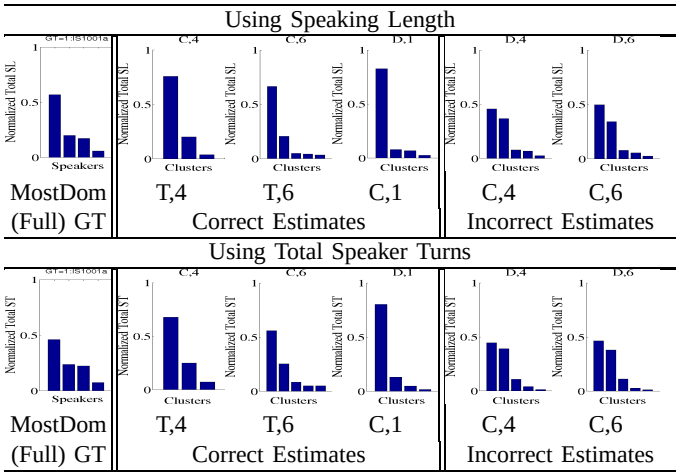

Fig. 4. Selected examples comparing correctly and incorrectly estimations of the most dominant person using the full agreement data set. Two different feature types are shown for the same meeting. For each row, the speaking length (or total speaker turns) generated using ground truth segmentations is shown in the left-most graph. The bar charts following that, show the speaking lengths or speaker turns for all clusters with the corresponding experimental conditions and diarization speed using the same labeling conventions as Table III. All features are ranked in descending order.

left-most column. In all cases the feature values are normalized across the total for all participants and ranked in descending order. The same experimental conditions are shown for comparison across feature types. Note that the cluster ordering is not related to the actual speakers: the highest value must be matched against the ground truth speaker segmentations to identify the best matching speaker. The labels below each graph indicate which experimental conditions were examined, according to the same labeling system as Table III. For the examples showing the speaking length, the cluster with the highest value tends to be higher than that of the dominant speaker that was generated using the ground truth. Also, the dominant person was estimated correctly, regardless of the number of estimated clusters. For the two incorrect estimates, the features do not match as closely with the features generated using the ground truth, compared to the conditions where the most dominant person was estimated correctly. Finally, the total speaker turns provided slightly less discrimination between the most dominant and non-dominant participants. This may explain the slightly worse performance of this feature compared to using speaking length.

Following the experiments in this section, we decided not to continue using the speaker-cluster mappings that were generated from the DER calculations. Though using the speaker-cluster mappings which were generated as a by-product of the DER improved the results, this method has a drawback, since not all the clusters are mapped to speakers and the shortest cluster may not have an associated speaker from the DER calculations. Therefore, this affects the estimation of the least dominant person. It is important to note that while the speaker diarization can only estimate clusters for people who speak, this did not affect our results since each participant spoke

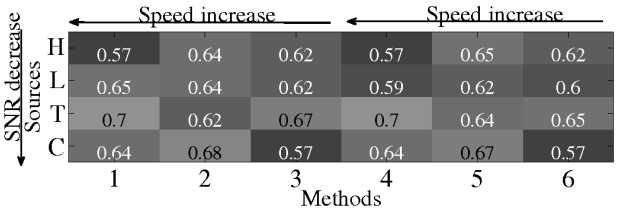

TABLE VII

THE RESULTS FOR THE MOST DOMINANT PERSON TASK WHERE THE MAJORITY OF ANNOTATORS AGREED. SEE TABLE II FOR THE CORRESPONDING SOURCE LABELS AND TABLE III FOR THE CORRESPONDING DIARIZATION STRATEGY LABELS.

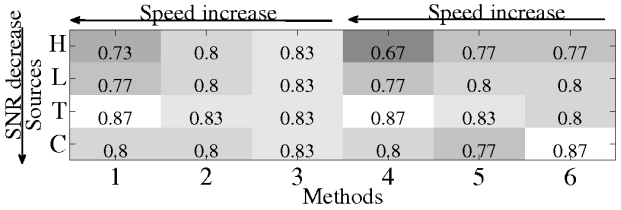

TABLE VIII

RESULTS SHOWING THE PERFORMANCE OF ESTIMATING THE LEAST DOMINANT PERSON WHEN THERE WAS FULL ANNOTATOR AGREEMENT. SEE TABLE II FOR THE CORRESPONDING SOURCE LABELS AND TABLE III FOR THE CORRESPONDING DIARIZATION STRATEGY LABELS.

at least once in each 5-minute segment (which was sufficient for them to be detected). These relatively quiet participants provide evidence for very small clusters which are generally not included in the DER calculation since they are so small compared to other clusters and do not improve the DER enough. Therefore, unlike the most dominant person estimation task, it did not seem appropriate in general, to evaluate our method using the speaker-cluster associations which are also only produced as a by-product of the DER calculation.

While using the total speaker turns led to comparable performance, it did not provide better performance than just using the speaking length. We decided that using speaking length as a single nonverbal cue for the rest of the experiments would be more appropriate since there has been more previous research that speaking length is a good feature for estimating dominance [5], [6].

D. Majority Agreement in Annotations

We studied the performance of the dominance task when at least two out of the three annotators for each meeting agreed on the most dominant person. The results in Table VII show that the best performance was 70% from experiments (T,1) and (T,4) which both used the audio source with the second worst SNR. Comparing with the baseline, there is a drop in performance in absolute terms, of 7%. The worst score was 57%, which occurred for experiments (H,1 and 4), (C,3 and 6), indicating no particular dependency on the speed strategy or SNR of the input source.

As expected, from our hypothesis H3, there was a systematic drop in performance between this dominance task and that using the full agreement data, which is shown by the overall darker shade of the results table and also Fig. 7. This suggests that a higher variability in human judgments 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%. However, for the majority case, the drop in performance was much less than for the full agreement case when speaker diarization was used.

IX. Estimating the Least Dominant Person

In this section, we discuss our results for the least-dominant person classification task. The experiments that were carried out were identical to the most-dominant case so the discussion in this section will be more brief. We conducted experiments on the least dominant person classification task with full-agreement data (31 meetings) and majority-agreement data (54 meetings). For the model, the person that corresponds to the lowest proportion of speaking time among all participants is classified as the least dominant.

A. Full Agreement among Annotators

The results for the least dominant person estimation are shown in Table VIII. Here the two best-performing conditions at 87% was achieved by condition (C,6), (T,1) and (T,4) and all three conditions use a SDM and in two cases, the fastest diarization strategy was used. Compared to the baseline results, which achieved a classification accuracy of 84%, there were 3 cases which performed slightly better. This was encouraging particularly because on closer inspection, both the mean and maximum performance for the SDM case T, matched or out-performed the baseline results. This result was surprising, given our hypothesis H4. However, one might say that since there is fewer observations from less dominant people, the estimates of their speaking status have fewer opportunities to be wrong. In some cases, the least dominant talked so little that despite noisy estimates, their speaker time was still much less than those who were more dominant.

B. Majority Agreement among Annotators

For the case were there was majority agreement among the annotators, the performance was much lower than that of the full agreement case, which also echos the results from the most dominant person estimation task and our hypothesis H3. Here the best score was 63%, which was obtained only when using the audio signal with the best SNR and the number of participants needed to be known a priori. Again, for the majority agreement case, the performance was often similar or better to the baseline case (59%) computed using turn-taking patterns extracted from headset microphones. We could explain this improvement on the baseline performance since smaller utterances tend to be more difficult to detect. Therefore, someone who tends to speak less, may be detected less often as speaking, leading to a higher level of discrimination between their speaking length and that of more dominant people in the meeting.

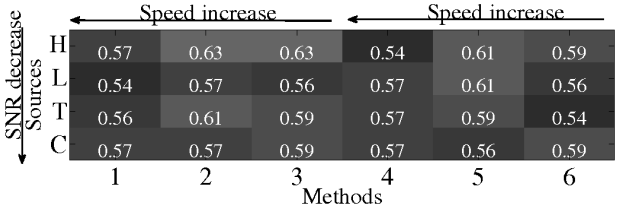

TABLE IX

PERFORMANCE WHEN ESTIMATING THE LEAST DOMINANT PERSON WITH MAJORITY AGREEMENT AMONG ANNOTATORS. SEE TABLE II FOR THE CORRESPONDING SOURCE LABELS AND TABLE III FOR THE CORRESPONDING DIARIZATION STRATEGY LABELS.

X. Studying the Results Further

Following our results for each dominance task, and the various experimental conditions and diarization strategies, there are observations to make across all of these tasks, which follow the hypotheses that we presented in Section VII. These can be divided into 4 categories: the effect of annotator variability on performance; the variations in performance between estimating the most and least dominant person; the effect of increasingly noisy audio sources; and the effect of different fast diarization strategies. These variations and corresponding results are summarized in Figures 5, 6, and 7.

A. Varying the SNR of the Input Source

In a practical situation, the distance of the microphone from each speaker can vary greatly. Many practical constraints can hinder the ease of use of a system. Therefore, knowing to what extent a worse signal affects the estimation results is useful. 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 single source (SDM). This contradicts

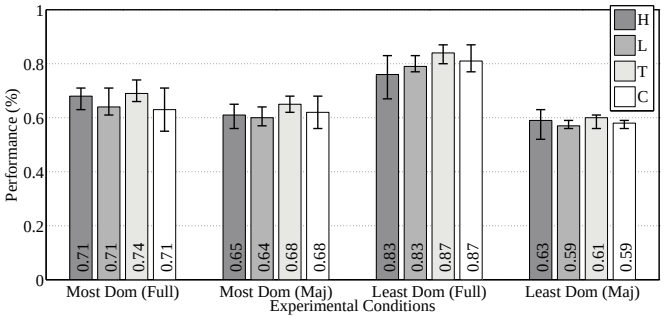

Fig. 5. Comparison of the mean, best and worse performing results for each source type. Each bar shows the best performance for each source condition and dominance task, by the height and labeled number. The markers on each bar indicate the best and worse performance in each category. The source condition for each type represent: Mixed Headset signal (H), Mixed Lapel signal (L), Single Distant Microphone from Table (T) and Single Distant Microphone from Ceiling (C).

H1 of our hypotheses. We can observe this in more detail in Figure 5 where a summary of the performance for all 4 dominance tasks under different experimental conditions are shown. For each cluster of bars, the SNR of each source type decreases from left to right. The level of each bar shows the mean performance while the markers indicate the highest and lowest performance values for each source condition and dominance task. We would expect the performance to decrease as the SNR decreases since this behavior is observed for the DER, as shown in Table III previously. However, the opposite is true in the case of estimating the least dominant person when there is full annotator agreement.

In all dominance tasks, experimental condition T (using single distant microphone on the table), matched or produced better results on average, compared to the two mixed sources (with lower SNR). Also, for the SDM case with the worse SNR (condition C), when the full range of diarization strategies were considered, the performance was able to match those of the other SDM condition (T) in all but the least dominant task where there was majority agreement. In almost all cases, the performance was worse than the baseline results, though there were some cases with the least dominant estimation task with full agreement where the performance was better. In addition, each of these cases correspond to the SDM conditions.

B. Varying Diarization Speed

In practical situations, it is desirable to have algorithms that work quickly. Therefore, if we can study the performance of different computationally efficient diarization strategies, we can understand the trade-offs between speed and accuracy. Figure 6 shows a comparison of the dominance estimation performance when the speaker diarization strategy is modified for different speeds of execution. For the most dominant estimation task, the fastest diarization strategy performed best on average, though the number of speakers was known beforehand. Also in both tasks where the most dominant person is estimated, the performance was higher on average for the fastest strategy (KL) when the number of clusters was estimated automatically. In general, the performance was slightly worse when the number of clusters was determined automatically rather than when fixing them to be ≤ 4. In contrast, for the task of estimating the least dominant person when there was full agreement, the average performance when using the fixed or automatic cases did not differ greatly. The worst and best average performances for this dominance task was observed for the fastest diarization strategy, indicating less stability in the results when the KL method is used. This could be due to the difficulty of modeling speakers when they speak very little so eliminating merge hypotheses too soon could be detrimental to the formation of the shorter clusters.

Overall, varying the diarization strategies appeared to have the least impact on the results. This was particularly encouraging since the fastest strategy with an automated estimate of the number of speaker clusters gave comparable performance to not using a fast strategy (where better a DER would be expected). This result was also surprising given our hypothesis H2.

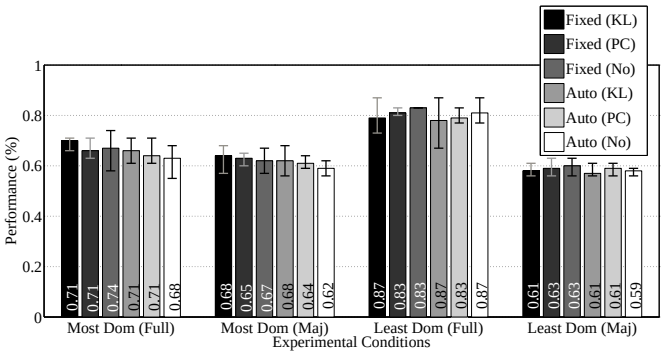

Fig. 6. Summary of the mean, best and worse performing results for each diarization strategies for different dominance tasks. The fixed case corresponds to fixing the number of clusters to be ≤ 4, while the 'auto' case allowed the algorithm to stop naturally. Each bar shows the highest performance for each source condition and dominance task, which is indicated more clearly by the number labeled on each. The horizontal markers on each bar indicate the best and worse performance in each category and the bars are clustered according to the 4 dominance tasks. 'KL', 'PC' and 'No' represent respectively, the different diarization strategies KL-fast matching, Pitch-Correlogram fast matching and No fast matching.

C. Effect of Fixing the Number of Expected Speaker Clusters

The effect of fixing the number of expected speaker clusters on the dominance estimation task is shown in Figure 6. We observed that the performance was sometimes better when the number of final speaker clusters was fixed but in general, the difference was minimal. Also, some of the best performing conditions using automatically estimated speaker numbers had comparable performance to the fixed case.

D. Effect of Annotator Variability on Performance

Studying how the annotator variability effects the dominance estimation is important since perceptions of dominance are subjective so estimates of dominance can only be as accurate as human perceptions. However, despite this, we can still use these differing test sets to analyze whether the estimation method that we employ is reliable, even with less agreement among the annotators. In all cases (including baselines), increasing the annotator variability led to a systematic decrease in performance, which is in line with H3 of our hypotheses. However, the decrease in performance between the full and majority agreement cases was greater for the least dominant task compared to that of the most dominant. Also, when observing the majority agreement cases, the performance of the least dominant person estimation task is comparable to the baseline while for the most dominant case, the automated performance is much lower than its corresponding baseline. Comparing the baseline and automated results for the full-agreement cases, the most dominant estimation performance is not as close to the baseline compared to the least dominant case. However, compared to the majority agreement case, for the least dominant estimation task, source conditions H and L do not appear to perform as closely to the corresponding baseline performance. Studying Figure 7 closer, the variations of performance across experiments changes fairly consistently, despite the increase in annotator variability. Therefore, we observe a stable behavior despite variability in the annotations.

E. Differences in Estimating the Most and Least Dominant Person

Estimating both the most and least dominant person is useful for ensuring that participants in a meeting are able participate. However, while the behavior of both types of people contrast

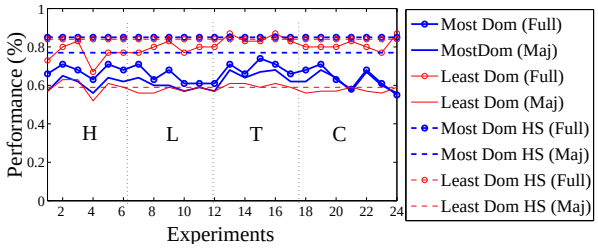

Fig. 7. Summary for comparison of the mean performance for all the dominance tasks and experimental conditions and diarization schemes. The baseline results using segmentations generated from individual headset microphones, indicated by 'HS', are also shown. The experiments are ordered from left to right, and then top to bottom, according to Table III.

considerably, understanding why their performance differs is also useful for understanding both roles better. For the baseline case using the full-agreement data, estimating the most dominant person performed slightly better than the least dominant case. However, for the automated case, the reverse was true. The reason for this could be related again to the low speaking times of the least dominant participants, leading to relatively more robust measurements of these speakers, even when using far-field microphones since there would be fewer occasions for the diarization estimates to the incorrect. Observing the results in both Figures 5 and 6, we see that the least dominant person task, when there is full annotator agreement in the data, leads to much higher performance compared to the all other dominance tasks. In terms of statistical significance, the highest results for the most and least dominant tasks with full agreement where no fast-match strategy was used, was significant at the 6% level. All other results were not significant at the 10% level.

XI. Further Investigations of Diarization and Dominance Estimation Errors

We have seen so far that there appears to be no correlation between reduced SNR, higher DER, and the dominance estimation performance. In this section, we investigate the likely cause of this lack of sensitivity. On inspecting the errors between various speakers, it was found that in general, the absolute amount of errors is proportional to the person's speaking length. In addition, the person who spoke the longest had errors leading to a reduction in speaking time while the rest of the speakers had proportionate increases. Therefore we devise a method of simulating the diarization error in a controlled manner in order to see how increasing the noise would affect the various dominance tasks that we have considered in the previous sections. We present a selection of interesting results that demonstrate why the dominance estimation performance is not particularly sensitive to the diarization error.

We devise here a method of simulating how increased error could affect the distribution of the total speaking length for each person based on their rank in terms of their speaking length according to the ground truth. To simulate noise, we considered the scenario where n frames are randomly chosen from the entire meeting where at least one person was speaking. This speaking frame is then assigned to one of the other speakers, who is chosen randomly, weighted by their total speaking time. This noise model simulates the tendency for more errors to occur for the people who talk more and shows the worst case scenario in terms of the risk of two of the distributions overlapping completely. Note that the errors we simulated here are not emulating diarization error or the errors caused by differing signal to noise ratio in the input audio signal. However, since our aim is to understand why the dominance estimation appears not to be sensitive to errors, investigation of how errors might affect the distribution of the total speaking lengths for each of the participants will help us to explain our results better.

A. The Effect of Errors on Estimating the Most Dominant Person

Figure 8 shows the distributions of the total speaking length as increasing noise is added to the ground truth speaking segmentations where there was full agreement among annotators on the most dominant person. We see that as the noise levels increase, the distributions of the first and second longest speaking time begin to overlap. The means of the distributions for the person who speaks least and second least also drift higher. By the time the speaking status was contaminated by 85% errors, the two distributions were fully overlapping. This represents a much higher level of errors than the 36.89% DER that was found for experiment (C,5) in Table III and demonstrates that the behavior itself holds the potential for a much higher noise tolerance level than those presented in our experiments.

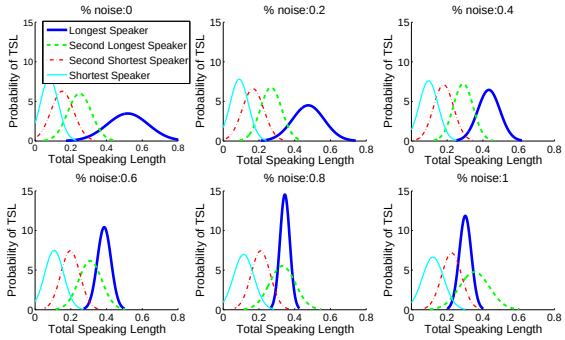

Fig. 8. The effect of increased noise on the total speaking length of each person. Simulated noise is added increasingly from the top left (0%) to bottom right (100 %) graphs. We see that as the noise level is increased, the distribution of the person with the highest speaking length drifts towards the person with the second highest speaking length since the speaking activity of the longest speaker is passed to the other speakers.

B. The Effect of Errors on Estimating the Most Dominant Person with Majority Agreement

For the same experiments, but now using the most dominant majority agreement data, the noise level for which there was full overlap between the longest and second longest speaker distributions was around 5% lower in absolute terms than for the data set with full agreement for the most dominant person. Closer inspection of the distribution of the total speaking time of the data where only 2 annotators agreed on the dominant person showed that the variance of the total speaking time for the person with the longest speaking length was larger and overlapped the distributions of the other speakers more, as shown in Figure 9(b). This suggests that for the cases when only 2 of 3 annotators agreed, other criteria than speaking length may have been used to label the most dominant person. From the annotation process, we collected a free-form description from each annotator about what their definition of dominance was. These included cues such as whether someone had an authoritative tone or if someone seemed to lead the conversation of the group.

C. Distribution of Total Speaking Time using Diarization Estimates

In terms of the general decrease in performance of the estimation of the most dominant person when speaker diarization was used compared to the baseline (85%), we observed that the diarization estimates led to flatter distributions for the total speaking time of the most dominant speakers. An example of this behavior is shown in Figure 9(c) where the diarization experiment (T,4) was used to generate the total speaking time distributions from the most dominant full agreement data set. The conditions that used the table microphone in particular had almost fully overlapping distributions for the second shortest and shortest speaker. This may explain the slightly better dominance estimation performance for the most dominant tasks when the table microphone was used.

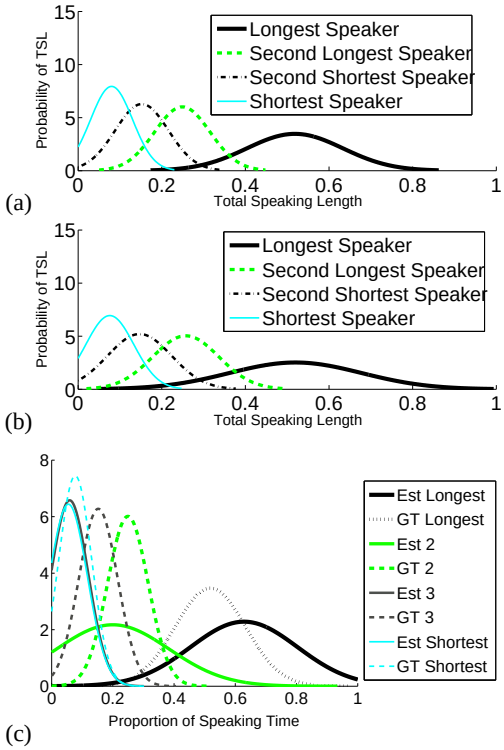

Fig. 9. (a):Distribution of the ground truth total speaking length for the most dominant person when there was full annotator agreement. (b):Distribution of the ground truth total speaking length for most dominant person when 2 of 3 annotators agreed. (c): Distributions of total speaking length based on diarization estimates (Est) and ground truth (GT) of most dominant person with full agreement. The diarization estimate shown here used the KL Fast match approach with the table microphone and automatically estimated the number of speakers.

D. The Effect of Errors on Estimating the Least Dominant Person

We conducted the same noise simulation experiments with the data set used for the least dominant person task when there was full agreement and a similar behavior was observed. The distribution of the person who spoke the least remained fairly well separated from the rest, though a shift in the probability density function towards that corresponding to the person with the highest speaking length was seen. When using the diarization estimates from the table microphones, we found that the distributions of the shortest and second shortest speaking lengths were far more separated. In fact, there were many cases where the distribution for the shortest speaking length was pushed even lower than the reference speaker. This would probably explain the slightly better performance using the estimates (87%) than the baseline method (84%).

XII. Conclusion and Final Remarks

Our study has shown that use of speaker diarization to estimate the dominant and least dominant speaker in group conversations is possible. We tested on systems that were fully automatic, which had comparable performance to methods that required some user intervention. The effectiveness of the system when just a single microphone source is used also emphasizes the ease-of-use of the system. Given that dominance itself is a complex notion which one would consider a semantically high-level behavior, we demonstrate an effective solution which is practical, fast and backed by strong evidence in the social psychology literature. The findings in the study have highlighted some surprising and interesting results, which are summarized in Table X.

In addition, we carried out noise simulation experiments to show the robustness of the conversational dynamics to increase diarization noise. We observed that the distribution of the total speaking lengths, particularly that of the person who spoke longest, is very robust to our simulated diarization errors. When observing the distribution of the total speaking time for the full agreement data compared to the case where 2 of the 3 annotators agreed on the most dominant person, we observed a higher variance for the longest speaking length, which suggests that for the majority agreement data set, the annotators may have used different criteria other than speaking length to judge who the most dominant person was.

Variable	Outcome for Dominance Estimation
SNR	Not Particularly Sensitive. Works best in T condition.
Diarization speed	Using KL-divergence to trim the cluster merge hypothesis space gave best performance.
Fixed vs Automatic clusters	Forcing the number of estimated speaker clusters to less than or equal to the number of speakers did not always lead to better dominance estimation performance.
Full agreement vs Majority agreement	Systematic decrease in performance when annotations had majority agreement.
Most Dominant vs Least Dominant	Least dominant person much easier to estimate.

TABLE X
SUMMARY OF DOMINANCE ESTIMATION PERFORMANCE GIVEN DIFFERENT VARIABLES. SEE TABLE II FOR THE CORRESPONDING SOURCE LABELS.

[H1]: Reduction in the signal to noise ratio of the input source leads to worse dominance estimation performance since the DER is affected: Contrary to our hypothesis that worse performance is generally achieved when an input signal has a lower SNR, our findings have shown that our dominance estimation task is not as sensitive to such conditions. This was further explained by our experiments in Section XI where the diarization errors from using the table microphone caused the distribution of the person who spoke second least amount of time to be shifted so that it completely overlapped the distribution of the person who spoke the least.

[H2]: Reducing the computational complexity of the speaker diarization algorithm will lead to worse dominance estimation performance: Contrary to H2, the performance did not decrease as the result of using faster diarization strategies and in some cases was better, than using the slowest strategy. A similar trend was observed for the DER when the true SDM cases were used. One possible explanation could be that the KL method is less sensitive to noisy data compared to using the BIC score. Further investigation in was provided in Section XI.

[H3]: Higher variability in the annotations leads to worse dominance estimation performance since human judgments are not unanimous: We found that increasing annotator variability to the data led to a decrease in performance when estimating the most or least dominant person. However, for the least dominant person case, the decrease in performance was much greater when comparing the use of the full and majority agreement data. This showed that estimating the least dominant person was more sensitive to annotator judgments, which could be due to having less observable behavior make judgments on. We also found that despite the decreased confidence in the annotations of the least dominant person, performance for the full agreement task was consistently better compared to all other dominance tasks since having less observable behavior meant that there was fewer estimates that could be made incorrectly by the diarization algorithm. For the majority cases however, both the most and least dominant tasks performed similarly, though the most dominant task in this case performed slightly better, which indicates again the sensitivity of annotating the least dominant person with little data. Overall, our findings are encouraging since for the fastest strategy with an automated estimation of the number of speaker clusters and single distant source with low SNR performed well,

sometimes providing the best performance results achievable.

[H4]: Due to lower annotator confidence and also the expectation that the least dominant person will speak less, it will be more difficult to estimate this person in the meetings compared to the most dominant person estimation task: The least dominant person was easier to estimate than the most dominant person due to the low levels of observations that could be affected by erroneous estimates from the diarization algorithm. This was further corroborated by our experiments that simulated the diarization noise in Section XI where Figure 8 shows that as the noise increases, the distribution of the person who speaks the least remains relatively close to zero, despite the decrease in the mean value of the distribution of the longest speaker. Also, small utterances were more difficult to detect for the least dominant person so this led to even fewer observations, as shown in Figure 9(c) where the distributions of the people who speak the least and second least are fully overlapped when diarization estimates are used. This probably made the behavior of the least dominant person significantly different from the most dominant person. Compared to the results using ideal audio conditions (HS), the estimates of the least dominant person were closer than the respective baseline results for the most dominant person.

A. Limitations and Future work

Currently, the method is not able to identify either the location or visual identity of the person speaking. Therefore, a fully automated way of performing speaker cluster/seat association will be investigated in future work, using both video and audio cues. In addition, if a person does not speak, they will not be detected by our system: It is always the person who is detected as speaking the least, who is the least dominant. However, using video sensors, it would be possible to detect silent but visually active participants in a meeting. Preliminary work by Hung et al. [29], [30] shows that it is possible to detect the dominant person audio-visually using only speaking length to estimate the dominant person. It may also be possible to use contextual cues and video cues to improve the diarization performance and possibly enhance our results for the various dominance tasks. Since the method can already run in real-time, it would be desirable to make the system perform on-line and in real-time to address the problem of 'who is speaking now'. An on-line diarization such as that proposed in [31] would be a promising direction to follow.

B. 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). We would also like to thank Oriol Vinyals for his contributions to the rapid speaker diarization work.

REFERENCES

[1] E. Rosa and A. Mazur, "Incipient status in small groups," *Social Forces*, vol. 58, no. 1, pp. 18–37, September 1979.

[2] 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.

[3] B. Wrede and E. Shriberg, "Spotting" Hot Spots" in Meetings: Human Judgments and Prosodic Cues," in *in EUROSPEECH*, 2003, pp. 2805–2808.

[4] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. International Conference on Speech and Signal Processing*, 2002, pp. 2037–2040.

[5] M. Schmid Mast, "Dominance as expressed and inferred through speaking time," *Human Communication Research*, no. 3, pp. 420–450, July 2002.

[6] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 17, no. 3, pp. 501–513, 2009.

[7] 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 *IEEE Automatic Speech Recognition Understanding Workshop*, 2007.

[8] E. Rogers-Millar and F. M. III, "Domineeringness and dominance: A transactional view," *Human Communication Research*, vol. 5, no. 3, pp. 238–246, 1979.

[9] C. West and D. H. Zimmerman, *Language, Gender, and Society*. Newbury House, 1983, ch. Small Insults: A study of interruptions in cross-sex conversations between unaquainted persons, pp. 103–117.

[10] J. F. Dovidio and S. L. Ellyson, "Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening," *Social Psychology Quarterly*, vol. 45, no. 2, pp. 106–113, June 1982.

[11] D. Tannen, *Gender and Discourse*. Oxford Univesrity Press, 1993, ch. Interpreting Interruption in Conversation, pp. 53–83.

[12] G. W. Beattie, "Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted," in *Semiotica*, vol. 39, no. 1-2, 1982, pp. 93–113.

[13] K. J. Tusing and J. P. Dillard, "The sounds of dominance: Vocal precursors of perceived dominance during interpersonal influence," *Human Communication Research*, vol. 26, no. 1, pp. 148 – 171, January 2000.

[14] C. Aronovitch, "The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker," *The Journal of Social Psychology*, vol. 99, pp. 207–220, 1976.

[15] K. R. Scherer, H. London, and J. Wolf, "The voice of confidence: Paralinguistic cues and audience evaluation." *Journal of Research in Personality, 7*, pp. 31–44, 1973.

[16] D. Buller and R. Aune, "The effects of vocalics and nonverbal sensitivity on compliance: A speech accomodation theory explanation," *Human Communication Research*, vol. 14, no. 3, pp. 301–322, Spring 1988.

[17] J. K. Burgoon and N. E. Dunbar, "Nonverbal expressions of dominance and power in human relationships," in *The Sage Handbook of Nonverbal Communication*, V. Manusov and M. Patterson, Eds. Sage, 2006.

[18] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, "Learning human interactions with the influence model," MIT Media Laboratory Technical Note, Tech. Rep., 2001.

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

[20] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase, "Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns," in *Proc. ACM CHI Extended Abstract*, Montreal, Apr. 2006, pp. 1175–1180.

[21] 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, pp. 835–838.

[22] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez, "Estimating the dominant person in multi-party conversations using speaker diarization strategies," in *International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 2197–2200.

[23] J. 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 *Machine Learning for Multimodal Interaction*, 2005, pp. 28–39.

[24] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. of International Conference on Acoustics, and Speech Signal Processing*, 2005.

[25] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA speech recognition workshop*, 1998.

[26] X. Anguera, B. Peskin, and M. Aguilo, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in *NIST Machine Learning for Multimodal Interaction, Meeting Recognition Workshop*, 2005.

[27] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proc. IEEE Automatic Speech Recognition Understanding Workshop*, 2003.

[28] N. Jhanwar and A. Raina, "Pitch correlogram clustering for fast speaker identification," *EURASIP Journal on Applied Signal Processing*, pp. 2640–2649, 2004.

[29] H. Hung, Y. Huang, C. Yeo, and D. Gatica-Perez, "Associating audio-visual activity cues in a dominance estimation framework," in *Computer Vision and Pattern Recognition Workshop on Human Communicative Behaviour*, 2008.
[30] H.Hung and G. Friedland, "Towards audio-visual on-line diarization of participants in group meetings," in *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications in conjunction with ECCV*, Marseille, France, October 2008.
[31] O. Vinyals and G. Friedland, "Towards semantic analysis of conversations: A system for the live identification of speakers in meetings," in *Proceedings of IEEE International Conference on Semantic Computing*, August 2008.

Hayley Hung Hayley Hung is a Marie Curie postdoctoral research fellow at the University of Amsterdam in The Netherlands. Between 2007-2010, she was a postdoctoral researcher at Idiap Research Institute, Switzerland. She graduated with an MEng degree in Electronic and Electrical Engineering from Imperial College, London in 2002 and PhD in computer vision from Queen Mary University of London in 2007, which was funded by the EPSRC, UK and QinetiQ Ltd. In 2009 she won the Institute of Engineering and Technology (IET) written premium competition. She received the West Midlands Runner Up Young Engineer for Britain Award for designing an automatic aligning antenna in 1997. She was awarded the Lucent Technologies Global Science Scholarship in 1998. She is a member of both the IEEE and IET.

Yan Huang Yan Huang (M08) received the M.S.E. degree in electrical engineering from The Johns Hopkins University, Baltimore, MD, in 2001, and the M.S. degree in computer science from the University of California, Berkeley, in 2007. Previously, she has been with Li Creative Technologies, Inc., International Computer Science Institute (ICSI), Berkeley, CA, Panasonic Speech Technologies Laboratory, Santa Barbara, CA, and the Center of Language and Speech Processing, Baltimore, MD. She has been working on various components of large vocabulary speech recognition systems, robust speaker recognition, speaker diarization, and speech synthesis. Currently she is a Speech Scientist with Microsoft, Redmond, WA, where she focuses on unsupervised, semi-supervised acoustic model training and user feedback loop model in voice search. Her major research interest is in machine learning and its applications in speech and language processing.

Gerald Friedland Gerald Friedland is a staff research scientist at the International Computer Science Institute (ICSI), an independent non-profit research lab associated with the University of California at Berkeley where he, among other functions, is currently leading the Speaker Diarization research. He was a site coordinator for the EU-funded AMIDA and Swiss-funded IM2 projects which sponsored the research on multimodal meeting analysis algorithms, including the research presented in this article. Apart from speech, his interests also include image, video and multimedia processing. As a member of the IEEE and IEEE Computer Society, Dr. Friedland is involved in the organization of various ACM and IEEE conferences. He is also co-founder and program director of the IEEE International Summer School for Semantic Computing at UC Berkeley. Dr. Friedland is the recipient of several research and industry recognitions, among them the Multimedia Entrepreneur Award by the German government and the European Academic Software Award. Most recently, he won the first prize in the ACM Multimedia Grand Challenge 2009. He received his diplom and doctorate (summa cum laude) in computer science from Freie Universitaet Berlin in 2002 and 2006, respectively.

Daniel Gatica-Perez Daniel Gatica-Perez (S'01, M'02) received the B.S. degree in Electronic Engineering from the University of Puebla, Mexico in 1993, the M.S. degree in Electrical Engineering from the National University of Mexico in 1996, and the Ph.D. degree in Electrical Engineering from the University of Washington, Seattle, in 2001, receiving the Yang Research Award for his doctoral work. He is now a senior researcher at Idiap Research Institute, Martigny, Switzerland, where he directs the Social Computing group. His recent work has developed statistical methods to analyze small groups at work in multisensor spaces, populations using cell phones in urban environments, and on-line communities in social media. He has published over 100 refereed papers in journals, books, and conferences in his research areas. He currently serves as Associate Editor of the IEEE Transactions on Multimedia, Image and Vision Computing, Machine Vision and Applications, and the Journal of Ambient Intelligence and Smart Environments. He is a member of the IEEE.