THE GESTURER IS THE SPEAKER

Binyam Gebrekidan Gebre[⋆⋆] *Peter Wittenburg*[⋆] *Tom Heskes* [†]

[⋆] Max Planck Institute for Psycholinguistics
[†] Radboud University

ABSTRACT

We present and solve the speaker diarization problem in a novel way. We hypothesize that the gesturer is the speaker and that identifying the gesturer can be taken as identifying the active speaker. We provide evidence in support of the hypothesis from gesture literature and audio-visual synchrony studies. We also present a vision-only diarization algorithm that relies on gestures (i.e. upper body movements). Experiments carried out on 8.9 hours of a publicly available dataset (the AMI meeting data) show that diarization error rates as low as 15% can be achieved.

Index Terms— Speaker diarisation, Speaker segmentation and Gesturer diarisation

1. INTRODUCTION

Speaker diarization – the task of determining *who spoke when?* in an audio/video recording– has a number of applications in document structuring and speech processing of broadcast news, debates, movies, meetings and interviews. Some of these applications come in the form of speech and speaker indexing (used for video navigation and retrieval), speaker model adaptation (used for enhancing speaker recognition) and speaker attributed speech-to-text transcription (used for speech translation and message summarization).

The domain focus of speaker diarization application has been changing over the years. Speaker diarization started with telephone conversations and was later followed by broadcast news. Today, conference meetings are receiving the most attention [1, 2]. With these domain changes, the signals used for diarization also began to change. Much of the research in speaker diarization focused on using audio signals [3]. In recent years, however, attention is shifting to using audiovisual signals [2], where the role of video signals is considered more and more. In this paper, we continue this trend and take the role of the video signals to the maximum and the role of the audio to the minimum.

We hypothesize that the gesturer is the speaker and that identifying the gesturer can be taken as identifying the active speaker. This hypothesis arose from the observation that while a speaker may not be gesturing for the whole duration of the speech, a gesturer is usually gesturing at least for some part of or within the vicinity of the duration of the corresponding speech. Section 2 gives evidence for this observation and grounds the hypothesis in gesture–speech synchrony studies.

With the gesture–speech synchrony established, we claim *who gestured when* can be used to answer *who spoke when*. We test for gesture occurrence in the region(s) of the video where there is optical flow. Significant optical flow can be associated with particular regions and these regions are generally the same regions occupied by the speaker(s) and not by the listener(s). This is the core idea of the method described in section 3, where we give details about our assumptions and the methods applied to detect gestures. In section 4, we give an outline of the implementation of the method.

We test the performance of our method using part of the AMI public dataset. A brief description of the dataset is given in section 5. In sections 6 through 7, we present, evaluate and discuss achieved results.

2. GESTURE–SPEECH RELATIONSHIP

When people speak, they gesture and they do so despite different cultural and linguistic backgrounds [4]. With gesture, speakers indicate length, size, shape, direction, distance thereby highlighting essential concepts expressed in words.

Despite differences in the exact interpretation of the relationship between gesture and speech, the gesture literature supports that there is a striking timing relationship between speech and gesture (i.e. gesture and speech execution occur within milliseconds of one another). One leading hypothesis proposes that gesture and speech together form an integrated communication system for the single purpose of linguistic expression. Gesture is linked to the structure, meaning, and timing of spoken language [5].

The arguments for the tight linkage between gesture and speech are the following: *1*) Gestures occur mainly during speech *2*) Delayed auditory feedback (DAF) does not interrupt speech-gesture synchrony *3*) The inborn blind do gesture *4*) Fluency affects gesturing

In the following subsections, we give brief explanations of these arguments and give reference for detailed analysis.

[⋆⋆]The research leading to these results has received funding from the European Commissions 7th Framework Program under grant agreement no 238405 (CLARA).

2.1. Gestures occur mainly during speech

Studies of people involved in conversations show that speakers gesture and listeners rarely gesture [5, 6]. In approximately 100 hours of recording, there were thousands of gestures for the speaker but only one for the listener [5]. In a sample of narrations, about 90% of all gestures occurred during active speech [5]. In a meeting of eight speakers, the occurrence of upper body movement with speech accounted for more than 80% of the total speaking time [6].

2.2. Delayed auditory feedback (DAF)

Gesture and speech remain in synchrony during delayed auditory feedback (DAF). Delayed auditory feedback is the process of hearing one's own speech played over earphones after a short delay (typically, 0.25 seconds). DAF disturbs the flow of speech; speech slows down, becomes hesitant and is subject to drawling and metatheses but despite these interruptions, gesture remains in synchrony with speech [7].

2.3. The inborn blind do gesture

Inborn blind people, who have never seen gesture, do gesture and gesture as frequently as sighted people do [8, 9]. In [8], four children who are blind from birth were tested in 3 discourse situations (narrative, reasoning, and spatial directions) and compared with groups of sighted and blindfolded sighted children. Blind children produced gestures and the gestures they produced resembled those of sighted children in both form and content.

2.4. Fluency affects gesturing

The relationship between gesture production and speech fluency is direct. The number of gestures increases as speech fluency increases and it decreases as speech fluency decreases. Stuttering – a speech disorder, characterized by syllable and sound repetitions and prolongations – is rarely co-produced with gesture. Gesturing is observed to fall to rest (or to stop moving) during the moment of stuttering and then to rise again and resume within milliseconds of resumption of speech fluency [10].

The aforementioned studies provide evidence that speech and gesture are tightly linked in execution. The presence of gesture is evidence of the presence of speech. In the next section, we describe a method to determine gesture occurrence.

3. METHODOLOGY

In order to perform speaker diarization based on the hypothesis: *the gesturer is the speaker*, we need to design modules to determine: *a*) the number of speakers *b*) their location and *c*) whether or not they gestured. The modules can be simple or complex depending on the content of the video and recording conditions. For example, if the video content has people appearing and disappearing unpredictably, then a complex model is needed to track speaker numbers and identities. However, because model complexity is neutral to *the-gesturer-is-the-speaker* concept, we will concentrate our efforts on a simple method that detects and tracks gestures of people in conference meeting videos – where participants usually stay in fixed locations.

In our method, we assume the number of speakers is determined from the first few frames of the video either by human detection algorithms [11] or by a user creating bounding boxes for each speaker. We also assume that the speakers maintain their location or are tracked. Given the (tracked) locations of the speakers, the rest is to define what a gesture is and to determine its occurrence from frame to frame for each speaker/location.

Comparison of any frame with its previous immediate frame shows that each bounded box (i.e. a speaker) will have some movements (arising either from noise or the speaker's gestures). We define gestures to be any movements that last longer than a fixed number of frames (i.e. we exclude brief head or hand movements from consideration). The motivation for the exclusion of isolated and brief movements is to remove noise and to avoid confusion between real gesture and the movements that people make when they relax or when they scratch their head.

4. IMPLEMENTATION

We use low-level features to approximate the signature and movement of head/hands. Specifically, we use corners (i.e. pixels that are significantly different from their surrounding pixels) to detect and track head/hands movements. For tracking the corners, we apply the pyramidal implementation of the Lucas-Kanade algorithm [12, 13]. Because our interest is in gesture (the movement of head/hands), we remove corners that do not move further than one to three pixels.

The following is a pseudo–pseudo–code for determining the active gesturer(s) (and hence speaker(s)).

- Bound regions where there are speakers.
- Detect corners [14] in the bounded regions.
- Track corners using Lucas-Kanade algorithm [12, 13].
- Keep only those that move greater than X pixels.
- Find histogram of motion orientations and keep N-best (N is three corresponding to head, left and right hands).
- Join consecutive motion segments that come from the same region. Uninterrupted gesture sequences from the same speaker constitute a speaking turn.

- Remove motion segments with duration less than Y frames. If a motion segment is not part of a gesture sequence, it is likely to be noise.
- Join consecutive motion segments that come from the same regions (after motion segments less than Y frames are removed).
- Classify motion segments based on region. Motion segments, which have beginning and end times, correspond to speaking times for the speaker in the bounded region.

5. EXPERIMENTS

We applied our algorithm and performed different experiments on the video recordings of the Augmented Multi-Party Interaction (AMI) meetings [15]. The AMI corpus consists of annotated audio-visual data of four participants engaged in a meeting. Each recording of the AMI meeting has a separate video for a center, left and right view of the participants and a separate high resolution video for each participant's face.

For our experiments, we used a subset of the IDIAP meetings (IN10XX and IS1009x) totaling 8.9 video hours. From the different recordings of the same meeting, we selected the left and right camera recordings, each of which has two speakers with visible hands. Figure 1 shows a snapshot of a video that contains the concatenated frames of the left and right camera recording of IN1016 AMI meeting data.

Table 1 gives details of the interaction of the participants in the selected videos. The details concern the length of videos (in minutes), speech-time percentage (speech-time over video length), speech overlap percentage (overlapped speech time over video length), and speaker turn switches (average number of speaker turn switches per minute).

Fig. 1. A snapshot of AMI-IN1016 video data. This represents the expected input to the proposed algorithm.

6. RESULTS AND DISCUSSION

The-gesturer-is-the-speaker diarization system outputs frame numbers and predicted signature(s) of the gesturer(s). This output is evaluated for correctness against manually annotated data in terms of Diarization Error Rate (DER), the metric used in the NIST RT evaluations. In DER-based speaker diarization evaluations, the reference segments are only those with speech. For a realistic comparison, we consider as reference those frames with at least one person gesturing.

Recall that our diarization algorithm discards brief body movements. Figure 2 shows the impact of this discarding on performance for four videos (having the least DERs). The figure shows that as the duration of brief movements (measured in frames) are discarded, the diarization error rates decrease. To give a single DER estimate for each video, we picked the DER at duration of 2.5 seconds (an assumption that is critical for good performance for ICSI-based systems [16]). Table 2 shows the DERs and other informative measures for all tested videos (ranked by DER). Other parameters being equal, the higher the rate of speaker turn switches and the more overlap between the speakers, the worse the DER.

Table 1. Features of experiment videos

Name	Video length (min)	Speech time (%)	Speech overlap (%)	Turn switches (per min)
IN1005	46	94.90	9.53	7.35
IN1016	59	96.95	18.27	12.30
IS1009b	34	87.88	8.97	6.48
IN1012	51	96.89	28.44	12.82
IN1002*	41	93.15	14.31	10.03
IN1007*	40	96.46	22.57	9.43
IS1009c	30	84.16	4.23	4.85
IN1013	51	96.04	26.64	12.88
IN1009	20	89.67	12.61	4.57
IN1014*	61	90.49	12.21	10.00
IN1008*	56	90.81	9.27	12.40
IS1009d*	32	80.83	8.58	8.45
IS1009a*	13	75.15	10.27	3.25

Table 2. Diarization Error Rates (DER)

Name	Gesture time (%)	Gesture overlap (%)	Turn switches (per min)	DER (%)
IN1005	62.54	0.03	1.07	14.52
IN1016	72.45	0.00	1.58	21.62
IS1009b	72.23	0.00	0.78	26.80
IN1012	64.00	0.00	1.67	35.30
IN1002*	63.65	0.00	0.95	37.03
IN1007*	67.06	0.04	1.37	40.41
IS1009c	66.40	0.00	0.70	45.22
IN1013	69.47	0.01	1.42	53.73
IN1009	59.50	0.00	0.67	54.92
IN1014*	71.60	0.00	1.15	58.16
IN1008*	57.80	0.00	1.88	62.47
IS1009d*	68.82	0.00	0.58	63.05
IS1009a*	60.84	0.00	0.28	63.98

The official NIST Rich Transcription 2009 evaluation results for various conditions are described in [16]. For batch audio, the DER ranges between 17.24% and 31.30%. For online audio, the DER is 39.27% and 44.61%. For audiovisual, it is 32.56%. Direct comparison of our results with previous results is hard given the differences in the experimental conditions, set of videos and the sensitivity of the DER [17].

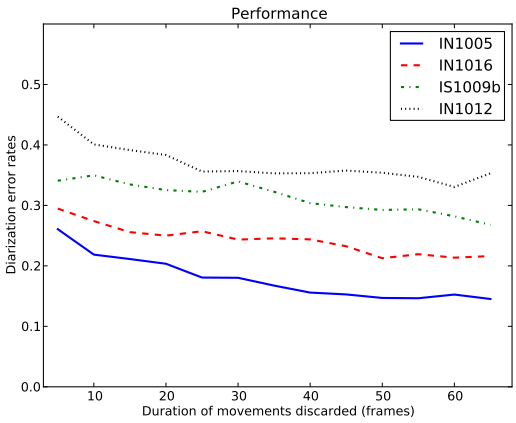

Fig. 2. Shows how diarization error rates decrease as short movements (measured in frames) are discarded.

The diarization method we presented has the advantage of being simple and using only video features. Previous speaker diarization systems are based on the ICSI speaker diarisation system [18] and involve a number of subcomponents [16, 19] for tasks such as filtering (Wiener), modeling (GMMs and HMMs), parameter estimation (Expectation-Maximization), decoding (HMM-Viterbi), clustering (agglomerative hierarchical clustering (AHC) with Bayesian information criterion (BIC)) and feature extraction (such as MFCC, prosody, video features).

Our diarization method does not use any of these subcomponents but uses algorithms for corner detection [14] and tracking [12] under the assumption that upper bodies of stationary or tracked speakers are visible in the video. It is this assumption which limits the application of our diarization method. Where an active speaker becomes invisible in the videos (which is the case for video names marked with *), the diarization error is higher. Furthermore, in videos where the gestures of a person are picked up by the two cameras, which is the case for most videos (because of the camera arrangements), the diarization error becomes higher.

There are two criticisms of using gesture for speaker diarization. One is of the form: *speakers do not all the time gesture*. This is true but gesture is frequent enough that, in some cases, methods can be designed to overcome its absence (e.g. smoothing). In our videos, the diarization algorithm has found that roughly 75% of speech is accompanied by gesture. The other criticism is of the form: *what is a gesture?* This is hard to answer without reference to semantics. In our case, we assumed any movement to be part of a gesture and it seems that this is a reasonable assumption for people in conference meetings. For more complex scenarios, there is a need to differentiate gestural activity from other activities.

7. CONCLUSIONS AND FUTURE WORK

We have presented the speaker diarization problem in a novel way. We hypothesized that the gesturer is the speaker and that gestural activity can be used to determine the active speaker. We provided evidence in support of the hypothesis from two sources *a*) gesture and audio-visual synchrony studies *b*) our experiments on a part of the AMI public dataset. The experiment performance measures confirm the hypothesis.

We have also outlined a method for gestural activity detection based on the location and tracking of corners. The method does not interpret gestures and assumes the background of the speakers is static. Further improvements of the algorithm for understanding gestures under more general recording conditions are left for another study. Future study should examine a probabilistic implementation of the diarization method and include other cues including audio, lip movements and visual focus of attention of speakers (i.e. listeners tend to look at the active speaker).

8. RELATION TO PRIOR WORK

The work presented here has focused on justifying and using gesture for speaker diarization. To the best of our knowledge, this has not been done before and we consider it as our main contribution. Our work is similar to but more general than the work by [20], which focused on using gesturing as a means to perform Voice Activity Detection (VAD). Their main rationale is different from ours. They see audio as the most natural and reliable channel for VAD. They use gesture when audio is unavailable (e.g. in surveillance conditions). We emphasize that gesture is synchronous with speech – and wherever applicable, gesturer diarization can reliably solve the problem of speaker diarization.

The work presented here also included the presentation of a new vision-based speaker diarization method that is different from the ICSI speaker diarization system [18]. The idea of using visual features for speaker diarization is not new. Extensive literature shows that visual features contribute to improved performance [21, 22, 23, 24, 25]. In many cases, however, the visual features are used in combination with audio features and rarely alone.

In summary, our work builds on and extends the literature on two fronts: *a*) emphasis on the use of gesture for speaker diarization *b*) a new vision-only diarization method that performs reasonably well with the advantage of being simpler. Both fronts offer opportunities for research in new directions.

9. REFERENCES

[1] J. Fiscus, J. Ajot, and J. Garofolo, "The rich transcription 2007 meeting recognition evaluation," *Multimodal Technologies for Perception of Humans*, pp. 373–389, 2008.

[2] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 356 –370, feb. 2012.

[3] S.E. Tranter and D.A. Reynolds, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557–1565, 2006.

[4] P. Feyereisen and J.D. de Lannoy, *Gestures and speech: Psychological investigations*, Cambridge University Press, 1991.

[5] D. McNeill, "So you think gestures are nonverbal?," *Psychological review*, vol. 92, no. 3, pp. 350, 1985.

[6] N. Campbell and N. Suzuki, "Working with very sparse data to detect speaker and listener participation in a meetings corpus," in *Workshop Programme*, 2006, vol. 10, p. 1.

[7] D. McNeill, *Gesture and thought*, University of Chicago Press, 2005.

[8] J.M. Iverson and S. Goldin-Meadow, "What's communication got to do with it? gesture in children blind from birth.," *Developmental Psychology*, vol. 33, no. 3, pp. 453, 1997.

[9] J.M. Iverson, H.L. Tencer, J. Lany, and S. Goldin-Meadow, "The relation between gesture and speech in congenitally blind and sighted language-learners," *Journal of nonverbal behavior*, vol. 24, no. 2, pp. 105–130, 2000.

[10] R.I. Mayherry and J. Jaques, "Gesture production during stuttered speech: insights into the nature of gesture-speech integration," *Language and gesture*, vol. 2, pp. 199, 2000.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR Proceedings*. IEEE, 2005, vol. 1, pp. 886–893.

[12] J.Y. Bouguet, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm," *Intel Corporation*, 2001.

[13] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[14] Carlo Tomasi and Jianbo Shi, "Good features to track," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 593–600, 1994.

[15] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, et al., "The ami meeting corpus: A pre-announcement," *Machine Learning for Multimodal Interaction*, pp. 28–39, 2006.

[16] G. Friedland, A. Janin, D. Imseng, X. Anguera Miro, L. Gottlieb, M. Huijbregts, M.T. Knox, and O. Vinyals, "The ICSI RT-09 speaker diarization system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 371–381, 2012.

[17] Nikki Mirghafori and Chuck Wooters, "Nuts and flakes: A study of data characteristics in speaker diarization," in *ICASSP Proceedings*. IEEE, 2006, vol. 1, pp. I–I.

[18] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," *Multimodal Technologies for Perception of Humans*, pp. 509–519, 2008.

[19] M. Huijbregts, D.A. van Leeuwen, and C. Wooters, "Speaker diarization error analysis using oracle components," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 393–403, 2012.

[20] M. Cristani, A. Pesarin, A. Vinciarelli, M. Crocco, and V. Murino, "Look at who's talking: Voice activity detection by automated gesture analysis," in *Workshop on Interactive Human Behavior Analysis in Open or Public Spaces*, 2011.

[21] A. Noulas, G. Englebienne, and B.J.A. Krose, "Multimodal speaker diarization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 79–93, 2012.

[22] G. Friedland, H. Hung, and Chuohao Yeo, "Multimodal speaker diarization of real-world meetings using compressed-domain video features," in *ICASSP Proceedings*, april 2009, pp. 4069 –4072.

[23] H. Hung and S.O. Ba, "Speech/non-speech detection in meetings from automatically extracted low resolution visual features," in *ICASSP Proceedings*, 2010, pp. 830–833.

[24] G. Garau and H. Bourlard, "Using audio and visual cues for speaker diarisation initialisation," in *ICASSP Proceedings*. IEEE, 2010, pp. 4942–4945.

[25] Himanshu Vajaria, Sudeep Sarkar, and Rangachar Kasturi, "Exploring co-occurence between speech and body movement for audio-guided video localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1608–1617, 2008.