Automatic Signer Diarization – the Mover is the Signer Approach

Binyam Gebrekidan Gebre[1], Peter Wittenburg[1], Tom Heskes[2]
[1]Max Planck Institute for Psycholinguistics, Nijmegen
[2]Radboud University, Nijmegen
{bingeb,peter.wittenburg}@mpi.nl,t.heskes@science.ru.nl

Abstract

We present a vision-based method for signer diarization – the task of automatically determining "who signed when?" in a video. This task has similar motivations and applications as speaker diarization but has received little attention in the literature. In this paper, we motivate the problem and propose a method for solving it. The method is based on the hypothesis that signers make more movements than their interlocutors. Experiments on four videos (a total of 1.4 hours and each consisting of two signers) show the applicability of the method. The best diarization error rate (DER) obtained is 0.16.

1. Introduction

Speaker diarization is the task of determining *who spoke when?* in an audio and/or video recording. This task is a dedicated domain of research in the multimedia signal processing community [19, 2]. The applications of speaker diarization include: speech and speaker indexing, document content structuring, speaker recognition, speaker attributed speech-to-text transcription and speech translation.

Most applications and technologies of diarization are driven by spoken language. But spoken language is one of the modalities of human communication. Text and sign languages are the other common modalities. In this paper, we focus on the visual-gestural modality and provide a baseline algorithm for determining *who signed when?* in a video recording of sign language.

In section 2, we present the motivations and applications of signer diarization. Signer diarization applications are similar to those of speaker diarization. In section 4, we propose and implement a method for solving the problem. The method uses no more knowledge than signers' movements (i.e. no sign recognition is involved). The implementation uses corner detection and tracking algorithms. In sections 5 and 6, we report and discuss experiment results.

2. Motivation

Sign language processing lies at the intersection of computer vision and computational linguistics. On the one hand, sign language is visual, lending itself to all issues of computer vision problems. On the other hand, sign language is a proper natural language with its own rules. The intersection offers both challenges and opportunities for developing and testing theories/applications. Signer diarization is one example where the visual aspect of the task makes it simpler than speaker diarization (modeling a speaker's sound is generally hard).

Apart from the insight that signer diarization offers into vision and language, automatically determining *who signed when?* in a video recording of unknown content and unknown signers has immediate applications in sign language technologies. These include information retrieval, machine translation and signer tracking. In general, the applications come in the following forms.

Pre-processing module for single signer-based systems: Signer diarization output can be used as input for signer tracking, signer identification and signer verification algorithms. It can also be used to adapt automatic sign language recognition towards a given signer. Currently, signer-dependent sign language recognition systems outperform signer-independent systems [3, 24, 25, 7, 1]. In this context, automatic signer diarization systems can be used as input to signer adaptation methods.

Signer indexing and rich transcription: Indexing video by signers and the linguistic transcripts makes information search and processing more efficient for both humans and machines. Typical uses of such output are for message summarization, machine translation and linguistic and behavioral analysis (for example, turn-taking studies [16, 6]).

The need for some of the aforementioned applications may not be urgent given sign language recognition is at research stage [7]. But in turn-taking studies [16], humans already perform manual signer diarization. Unfortunately, manual diarization has limitations - it is slow, costly and does not scale with the increasing amount of data. Therefore, there is a need for automatic signer diarization tools.

3. Signer diarization complexity

Given a video of signers recorded using a single camera, automatically determining *who signed when?* is challenging. The challenge arises from signers themselves and the environment (recording conditions).

3.1. Signers

To begin with, the number of signers is unknown and this number may change in time as a participant leaves or joins the conversation. The locations and orientations of signers may change and these changes could take place while signing. Signers may take short signing turns and often enough sign at the same time (overlap in time). The signing spaces of signers may also be shared (overlap in space).

3.2. Environment

The environment includes background and camera noises. The background objects of the video may include dynamic objects – increasing the ambiguity of signing activity. The properties and configurations of the camera induce variations of scale, translation, rotation, view, occlusion, etc. These variations coupled with lighting conditions may introduce noises. These challenges are common in many other computer vision problems too.

4. Method

Sign language is a gestural-visual language. A signer produces a sequence of signs and an interlocutor sees and interprets the sequence. Like a spoken language, a sign language can be described at different levels of analysis such as phonology, morphology, syntax and semantics [21].

The phonemes, which are the basic units of sign languages, are made from a set of hand shapes, locations and movements [17]. These subunits make up the manual signs of a given sign language. The whole message of an utterance is contained not only in manual signs but also in non-manual signs (facial expressions, head/shoulder motion and body posture) [13].

In theory, an automatic signer diarization system can exploit some or all of the basic units from both manual and non-manual signs to perform signer diarization. In practice, however, some sub-units are easier to extract and exploit by the machine. This paper proposes a diarization algorithm based on body movements. The hypothesis is that the active signer makes more movements than the interlocutors.

4.1. Algorithm

The automatic signer diarization algorithm consists of modules that determine: *a*) the number of signers *b*) their identities (or signatures) and *c*) whether or not they signed.

Each module can be simple or complex depending on the content of the video and recording conditions. The most general signer diarization system assumes nothing of the number of signers, their signatures and the video recording conditions. Such a method, besides being more complex to develop, will be inefficient and may even be less performing than a system developed and tailored for a specific instance of video recording conditions [22].

In our diarization system, we make a number of simplifying assumptions about the video recording conditions and provide a mechanism for user involvement using annotation tools like ELAN [15]. The user of the system makes some simple decisions to initialize the system. The user determines the number of signers from the first frame of the video by creating bounding boxes for each signer. These bounding boxes limit the boundaries of the signing spaces for each signer. The diarization system assumes the signers maintain their location (this is a reasonable assumption for videos of interviews and conference meetings) or are tracked [9]. Given the locations of signers, the rest is to define what constitutes signing activity and to determine its occurrence from frame to frame for each signer/location.

What constitutes signing activity? At any frame, each bounded box (i.e. a signer) will have some movement of the head/hands (arising either from signing activity or noise). Movements that last longer than a fixed number of frames are considered to constitute a signing activity. In other words, brief head or hand movements are excluded. The motivation for the exclusion of isolated and brief movements is to remove noise and to avoid confusion between real signs and relaxing movements.

4.2. Implementation

What is a hand/face and what is a movement from implementation perspective? We use corners to detect and track body movements. Corners are shown to be good features for tracking [18]. They have the property that they are different from their surrounding points. A given point in a homogenous image cannot be identified whether or not it has moved in the subsequent frame. Similarly, a given point along an edge cannot be identified whether or not it has moved along that edge. However, the motion of a corner can conveniently be computed and identified [18].

For a given application, not all corners in a video are equally important. For sign activity detection, the interesting corners are the ones resulting from body movements, mainly from head and hand movements. In order to filter out corners irrelevant to body movements, we ignore corners that do not move more than a given threshold. For tracking the movement of corners, we apply the pyramidal implementation of the Lucas-Kanade algorithm [4, 5].

The Lucas Kanade algorithm works based on three assumptions: *1*) *brightness constancy*: a point in a given image does not change in appearance as it moves from frame to frame *2*) *temporal persistence*: the motion of a surface

patch changes slowly in time *3*) *spatial coherence*: neighboring points in a frame belong to the same surface and have similar motion (for example, points of the right hand roughly move together).

The following is a pseudocode for determining the active signer(s).

- Bound a region for each signer.
- Detect corners [18] in the bounded regions.
- Track corners using Lucas-Kanade algorithm [4, 5].
- Keep only those that move greater than X pixels.
- Find histogram of motion orientations and keep N-best (N is, typically, three corresponding to head, left and right hands).
- Join consecutive motion segments that come from the same region. Uninterrupted motion sequences from the same region constitute a signing turn.
- Remove motion segments with duration less than Y frames. If a motion segment is not part of a signing activity, 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 signing times for the signer in the bounded region.

5. Experiments

5.1. Datasets

We ran the signer diarization method on four videos taken from the Language Archive at the Max Planck Institute for Psycholinguistics[1]. Each video has two signers of Kata Kolok [10]. Table 1 shows the details of the interaction of the signers in the videos. The details are extracted from manually annotated data.

5.2. Evaluation metrics

In speaker diarization performance evaluation, the most widely used evaluation metric is the speaker diarization error rate (DER). NIST uses the DER metric to compare different diarization systems[2]. Despite its noisiness and sensitivity [14], we borrow this concept and apply it without any changes except for the name signer diarization error rate (which was speaker diarization error rate).

[1] http://corpus1.mpi.nl/ds/imdi_browser/

[2] http://www.itl.nist.gov/iad/mig//tests/rt/2006-spring/index.html

Table 1. Experiment dataset details

Video	Length	STP	STM	DSS	SO
KN5	≈17	82.89	16.30	62.56	9.61
PiKe	≈18	70.04	15.40	57.62	11.52
ReKe	≈ 24	81.82	19.10	58.13	9.22
SuJu	≈ 24	78.13	15.24	66.39	9.68

Length = Video length in minutes
STP = Signing Time Percentage
STM = # of Signing Turns per Minute
DSS = Dominant Signer Share of sign time
SO = % of Signers Overlap (over sign time)

Diarization error rate is computed as the fraction of signer time that is not attributed correctly to a signer. Equation 1 shows the formula for the DER.

$$DER = \frac{\sum_{s \in S} dur(s)\Big(max\big(N_r(s), N_h(s)\big) - N_c(s)\Big)}{\sum_{s \in S} dur(s) N_r(s)}, \quad (1)$$

where
$dur(s)$ = the duration of segment s,
$N_r(s)$ = the # of reference signers signing in segment s,
$N_h(s)$ = the # of system signers signing in segment s,
$N_c(s)$ = the # of reference signers signing in segment s for whom their matching (mapped) system signers are also signing in segment s. A segment s is the time range where no reference signer or system signer starts signing or stops signing.

In qualitative terms, diarization error rate consists of three types of errors: false alarm signer time fraction (i.e. system predicted signing time that is not in the reference), missed signer time fraction (system failed to predict signing time that is in the reference) and signer error time fraction (signer time that is attributed to the wrong signer).

6. Results and Discussion

The diarization system outputs motion segments (start and end times) for each signer. This output is evaluated for correctness against manually annotated data and the results are reported in terms of Diarization Error Rate (DER). The reference frames are those frames that have manual annotations (70-80% of the video length as shown in table 1)).

Table 2 presents the diarization error rate for each video. The best DER scores are obtained for SuJu, KN5 and ReKe videos. The more the dominance of one signer and the less overlap, the lower the diarization error. The worst DER is obtained for PiKe video. The explanation for the latter result has to do with false alarm errors (movements that are detected by the algorithm but that are not annotated as signs in the manually annotated data).

Examining the video shows the sources of the false alarms. One source is the movement of a child that comes

Table 2. Performance

Video	Y	MS	FA	SE	DER
KN5	13	0.12	0.07	0.05	0.24
PiKe	8	0.11	0.14	0.04	0.29
ReKe	18	0.14	0.05	0.05	0.25
SuJu	10	0.08	0.05	0.03	0.16

Y = minimum signing duration (frames)
MS = fraction of Missed Sign
FA = fraction of False Alarm
SE = fraction of Wrong Signer Prediction
DER = MS + FA + SE

to her mother for part of the video. The other source is the appearance of signing activity of one signer in the signing space of the other signer. The latter source increases false alarms because the signs occur in the other signer's space.

An important parameter of the signer diarization algorithm is the number of frames to remove – parameter Y described in the pseudocode in subsection 4.2. This parameter controls the minimum duration of body movements to consider as signing activity. It is measured in frames and any motion less than Y is considered noise and discarded.

Figure 1 shows the impact of varying the Y parameter on error rates for the four videos. The larger the Y value, the higher the missed signs and the lower false alarms (and vice versa). In other words, the Y value controls the trade-off between false alarms and missed signs. The best Y values that result in the lowest diarization errors are indicated in table 2.

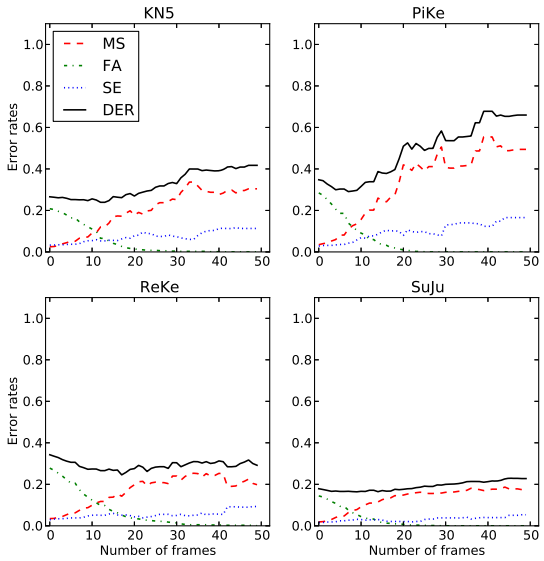

Figure 1. Performance variation as body movements of short duration are discarded.

Apart from the duration of the movements, the presented diarization algorithm does not interpret the movements. This makes it applicable independent of sign languages/signers but it also makes it vulnerable to false alarms. But, as our results indicate, movement is one of the most informative indicators of signing activity or uttering activity, in general. Movements that hearing speakers make, called gestures, are also used to identify speakers [11, 20, 8]. The algorithm outlined in this paper has also been applied in speaker diarization based on the hypothesis that *the gesturer is the speaker* [11].

In standard speaker diarization systems, which are based on iterative segmenting and clustering [23, 12], each speaker is modeled by a GMM model and the segmentation is done using HMM-Viterbi decoding. More specifically, the system starts with K clusters[3] after front-end acoustic processing and removing non-speech segments. At every iteration, two clusters are merged with the best BIC score and after every merging, the whole process of re-training the GMM models (using EM) and segmenting the data (using HMM-viterbi) is repeated.

By contrast, the diarization system presented in this paper uses intuitive heuristics. Each signer is modeled by a location. The motions of each signer (called motion segments) are equivalent to the clusters in speaker diarization. Merging of motion segments is done based on their location and time gaps of their occurrence. The closer the motion segments are both in time and in space, the more likely they belong to a signing activity of one signer. If signing spaces are shared, which is not unlikely, our system fails to disambiguate the sources of signing activity.

7. Conclusions and Future Work

We introduced and motivated the signer diarization problem by drawing similarities with the speaker diarization problem. We also proposed a signer diarization method based on the hypothesis that signers make more body movements than their interlocutors. We implemented the method using corner detection and tracking algorithms. Experimental results show the applicability of the method.

From our results, we can make two conclusions. First, body motion is an inexpensive source of information and as such can serve as a baseline for signer diarization. Second, not all body motions are signing activity. Signing activity detector may need to be applied as preprocessing before the diarization task. Such a detector can be trained on annotated data using features extracted from body posture, head orientations (interlocutors look at the active signer) and audio (signers do sometimes make sound).

Future study should examine an efficient probabilistic implementation of the algorithm outlined in this paper.

[3] K is typically 16 or 40 (more than expected number of speakers)

Acknowledgment

The research leading to these results has received funding from the European Commissions 7th Framework Program under grant agreement no 238405 (CLARA). We thank Stephen C. Levinson and Connie de Vos for contributing ideas and data to this paper. Any errors, however, remain our sole responsibility.

References

[1] S. Akram, J. Beskow, and H. Kjellstrom. Visual recognition of isolated swedish sign language signs. *arXiv preprint arXiv:1211.3901*, 2012. 1

[2] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, 2012. 1

[3] B. Bauer, H. Hienz, and K.-F. Kraiss. Video-based continuous sign language recognition using statistical methods. In *Proceedings of International Conference on Pattern Recognition*, volume 2, pages 463–466. IEEE, 2000. 1

[4] J. Bouguet. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation*, 2001. 2, 3

[5] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 2, 3

[6] J. Coates and R. Sutton-Spence. Turn-taking patterns in deaf conversation. *Journal of Sociolinguistics*, 5(4):507–529, 2001. 1

[7] H. Cooper, E.-J. Ong, N. Pugeault, and R. Bowden. Sign language recognition using sub-units. *Journal of Machine Learning Research*, 13:2205–2231, 2012. 1

[8] 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. 4

[9] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. *International Journal of Computer Vision*, 37(2):175–185, 2000. 2

[10] C. de Vos. *Sign-Spatiality in Kata Kolok: how a village sign language of Bali inscribes its signing space*. PhD thesis, Max Planck Institute for Psycholinguistics, 2012. 3

[11] B. G. Gebre, P. Wittenburg, and T. Heskes. The gesturer is the speaker. In *Proceedings of ICASSP*, 2013. 4

[12] M. Huijbregts, D. van Leeuwen, and C. Wooters. Speaker diarization error analysis using oracle components. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):393–403, 2012. 4

[13] S. K. Liddell. Nonmanual signals and relative clauses in american sign language. *Understanding language through sign language research*, pages 59–90, 1978. 2

[14] N. Mirghafori and C. Wooters. Nuts and flakes: A study of data characteristics in speaker diarization. In *ICASSP Proceedings*, volume 1, 2006. 3

[15] H. Sloetjes and P. Wittenburg. Annotation by category: ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008. 2

[16] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. De Ruiter, K.-E. Yoon, et al. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, 2009. 1

[17] W. Stokoe. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37, 2005. 2

[18] C. Tomasi and J. Shi. Good features to track. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994. 2, 3

[19] S. Tranter and D. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, 2006. 1

[20] H. Vajaria, S. Sarkar, and R. Kasturi. Exploring co-occurence between speech and body movement for audio-guided video localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1608–1617, 2008. 4

[21] C. Valli and C. Lucas. *Linguistics of American Sign Language Text: An Introduction*. Gallaudet University Press, 2001. 2

[22] P. Wittenburg, P. Lenkiewicz, E. Auer, A. Lenkiewicz, B. G. Gebre, and S. Drude. Av processing in ehumanities–a paradigm shift. In *Digital Humanities 2012 Conference*, 2012. 2

[23] C. Wooters and M. Huijbregts. The ICSI RT07s speaker diarization system. *Multimodal Technologies for Perception of Humans*, pages 509–519, 2008. 4

[24] L.-G. Zhang, Y. Chen, G. Fang, X. Chen, and W. Gao. A vision-based sign language recognition system using tied-mixture density HMM. In *International Conference on Multimodal Interfaces: Proceedings of the 6th international conference on Multimodal interfaces*, pages 198–204, 2004. 1

[25] J. Zieren and K.-F. Kraiss. Robust person-independent visual sign language recognition. *Pattern Recognition and Image Analysis*, pages 333–355, 2005. 1