# MOTION HISTORY IMAGES FOR ONLINE SPEAKER/SIGNER DIARIZATION

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

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

## ABSTRACT

We present a solution to the problem of online speaker/signer diarization - the task of determining *who spoke/signed when?*. Our solution is based on the idea that gestural activity (hands and body movement) is highly correlated with uttering activity. This correlation is necessarily true for sign languages and mostly true for spoken languages. The novel part of our solution is the use of motion history images (MHI) as a likelihood measure for probabilistically detecting uttering activities. MHI is an efficient representation of where and how motion occurred for a fixed period of time. We conducted experiments on 4.9 hours of a publicly available dataset (the AMI meeting data) and 1.4 hours of sign language dataset (Kata Kolok data). The best performance obtained is 15.70% for sign language and 32.46% for spoken language (measurements are in DER). These results show that our solution is applicable in real-world applications like video conferences.

***Index Terms*—** Speaker diarization, signer diarization, motion history images, motion energy images

## 1. INTRODUCTION

Human communication (or conversations) among $N$ speakers can take place in text, speech and sign language. In any of these modalities, determining *"who said when?"* is a challenging problem. In written works (e.g. fiction books), tracking the number of characters and their conversations is hard because of many reasons including anaphora resolution [1]. In spoken languages, determining *"who said when?"* has also proven hard despite the attention and research dedicated to it [2, 3]. In visual languages, even though there is little research into it, recent work shows that it is also a hard problem because of non-communicative body movements [4].

In this paper, we propose a novel solution to the problem of online speaker and signer diarization. We are interested in this problem because it has applications in human-to-human or human-to-computer interactions. For example, in video conferences, we would like to focus automatically on the active speaker/signer. In human-robot interactions, we would like the robot to look at the person speaking/signing. And, in information retrieval, we would like to index and search by speakers/signers.

The aforementioned applications and others have motivated extensive research into speaker diarization and have resulted into many solutions and tools [2, 3, 5, 6, 7, 8]. The novel part of our solution is the application of motion history images [9] in solving both speaker and signer diarization problems. Motion History Image (MHI) is an efficient way of representing arbitrary movements (coming from many frames) in a single static image. This type of represenation has been used for various action recognition tasks [9, 10, 11]. The strength of MHI is its descriptiveness and real-time representation. It is descriptive because it can tell us where and how motions occurred. It is real-time because its computational cost is minimal. The rest of the paper gives more details about MHI and its application in speaker/signer diarization.

## 2. GESTURING/SIGNING REPRESENTATION

When hearing people speak, they mostly gesture. When the deaf sign, they inherently make movements (signing entails movement). In either case, our goal in a diarization system is to determine where motion occurs and to decide if it indicates an uttering activity. How do we separate body motion from others? The paper assumes that the motion of the body is separated from background motion in a preprocessing step or that the background is static.

For conference settings or meeting data, it is safe to assume that motions come mainly from humans engaged in conversations. For such scenarios, background substraction [12] or frame differencing is enough. In our experiments, we applied frame differencing and obtained qualitatively similar results compared with those coming from a background subtraction algorithm based on Gaussian Mixture Model.

After finding the foregroud (moving) objects, how do we efficiently and conveniently represent motion that indicates *a*) where it occurred (space)? *b*) when it occurred (time)? We use Motion History Image (MHI) [9]. MHI is a single stacked image that encodes motion that occurred between every frame pair for the last $\tau$ number of frames. The type of information encoded in the MHI can be binary and, in such a case, it is called Motion Energy Image (MEI). MEI indicates where the motion has occurred in any of the $\tau$ frames. We use this MEI to tell us which person is speaking or signing. MEI does not tell us how the motion occurred. For this informa-

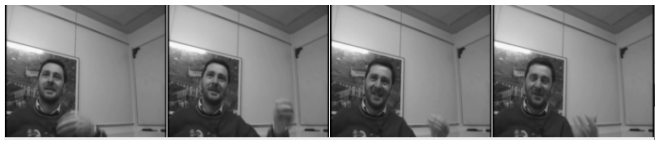
tion, we need to use Motion History Image (MHI). MHI is an image whose intensities are a function of recency of motion. The more recent a motion is, the higher its intensity. More formal definitions of MEI and MHI are given in the following subsections.

### 2.1. Motion Energe Image

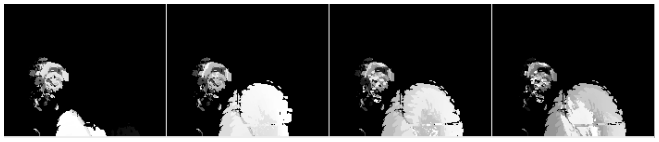
To represent where motion occurred, we form a Motion Energy Image and it is constructed as follows. Let $I(x, y, t)$ be an image sequence, and let $D(x, y, t)$ be a binary image sequence indicating regions of motion (for example, generated by frame differencing). Then the binary MEI $E(x, y, t)$ is defined as follows:

$$E_\tau(x, y, t) = \bigcup_{i=0}^{\tau-1} D(x, y, t - i) \tag{1}$$

where $\tau$ is the temporal extent of motion (for example, a fixed number of frames). Figure 1 (c) shows an image example of a MEI for a speaker who is also gesturing.



(a) Frames



(b) MHI

(c) MEI

**Fig. 1**. Examples of visualizations of MHI and MEI images. (a) shows selected frames of a video taken from AMI meeting data. (b) shows the MHI of 25 frames - recent motions are brighter. (c) shows the MEI of 25 frames - white regions correspond to motion that occcurred in any of the last 25 frames.

### 2.2. Motion History Image

To represent how motion occurred, we form a Motion History Image (MHI) as follows:

$$H_\tau(x, y, t) = \begin{cases} \tau & \text{if } D(x, y, t) = 1 \\ 0 & \text{else if } H_\tau(x, y, t) < (\tau - \delta) \end{cases} \tag{2}$$

where $\tau$ is the current timestamp and $\delta$ is the maximum time duration constant ($\tau$ and $\delta$ are converted to frame numbers based on frame rate). Figure 1 (b) shows an example of a MHI for a speaker who is also gesturing. Note that a MEI image can be generated by thresholding a MHI above zero.

## 3. THE ONLINE DIARIZATION SYSTEM

In an online diarization system, we want to determine who at any time is speaking/signing given we have video observations from 0 to $t$. Let each person's state be represented by $x_t^i$ (binary values of speaking or not speaking) and let $z_{0:t}^i$ be measurements (of the video frames) for each person $i$, the objective is then to calculate the probablitity of $x_t^i$ at time $t$ given the observations $z_{0:t}^i$ up to time $t$:

$$p(x_t^i|z_{0:t}^i) = \frac{p(z_t^i|x_t^i)p(x_t^i|z_{0:t-1}^i)}{p(z_t^i|z_{0:t-1}^i)} \tag{3}$$

where $p(z_t^i|z_{0:t-1}^i)$ is a normalization constant. In equation 3, there are two important probability distribtution: one is $p(x_t^i|z_{0:t-1}^i)$, we refer to it as converstation dynamics and the other is $p(z_t^i|x_t^i)$ and we refer to it as the gesture model.

### 3.1. Conversation dynamics

Conversation among $N$ speakers imposes its own dynamics on speakers. A given speaker is more likely to continue to speak in the next frame than stop or be interrupted by others. We encode this type of dynamics as follows:

$$p(x_t^i|z_{0:t-1}^i) = \sum_{x_{t-1}} p(x_t^i|x_{t-1}^i)p(x_{t-1}^i|z_{0:t-1}^i) \tag{4}$$

where $p(x_{t-1}^i|z_{0:t-1}^i)$ is the posterior from the previous time and $p(x_t^i|x_{t-1}^i)$ is the conversation dynamics. We assume that a speaker is $90\%$ more likely to continue speaking than not. Similarly, a silent person is more likely to continue to be silent (listening). We encode this assumption in a transition matrix as follows:

$$p(x_t^i|x_{t-1}^i) = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \tag{5}$$

### 3.2. Gesture model: gamma distribution

For both speaker and signer diarization systems, we assume that MEI is a strong indicator of an utterance. The higher the energy (the sum of MEI individual values), the higher the probability of an utterance. We model this type of relationship using gamma distribution with shape parameter $\boldsymbol{k}$ and scale parameter $\boldsymbol{\theta}$.

$$p(z_t^i|x_t^i;\boldsymbol{k},\boldsymbol{\theta}) = \frac{(z_t^i)^{\boldsymbol{k}_x-1}\exp(-\frac{z_t^i}{\boldsymbol{\theta}_x})}{\boldsymbol{\theta}_x^{\boldsymbol{k}_x}\Gamma(\boldsymbol{k}_x)} \quad \text{for } z_t^i, \boldsymbol{k}, \boldsymbol{\theta} > 0 \tag{6}$$

where $x = x_t^i$, $z_t^i$ is the number of 'on' pixels in a MEI for speaker or signer $i$ and $x_t^i$ is a binary random variable whose values represent speaking and non-speaking status of each person. Each state of $x_t^i$ has its own gamma distribution whose parameter values are learned from speaking and non-speaking manually annotated data.

## 4. EXPERIMENTS

### 4.1. Datasets

#### *4.1.1. Spoken language data*

We ran our algorithm on seven video recordings ($\approx$ 4.9 hours). These videos are taken from a publicly available corpus called the AMI corpus [13]. The AMI corpus consists of annotated audio-visual data of a number of participants engaged in a meeting. We selected seven meetings which had four participants (IN10XX and IS1009). The upper body of each participant is recorded using a separate camera and we put them together before diarization.

#### *4.1.2. Sign language data*

We also ran our diarization algorithm on four video recordings ($\approx$ 1.4 hours) of Kata Kolok, a sign language used in northern Bali [14]. Each video has two participants conversing in sign language and is recorded from a single fixed camera. In these videos, there is no boundary between signers. Signing space is sometimes shared - making the task of diarization even more difficult.

We solved this difficulty by clustering MEI 'on' pixels into a prefixed $K$ centers, set equal to the number of signers. We implemented a sequential k-means that updates the centers of clusters (signing space) in an online fashion as follows:

$$\boldsymbol{C}_t^i = \boldsymbol{C}_t^i + \frac{1}{n_{0:t}^i}(\boldsymbol{P}_t^j - \boldsymbol{C}_t^i) \tag{7}$$

$\forall j$ with $\boldsymbol{C}_t^i$ closest to $P_t^j$. $\boldsymbol{C}_t^i$ is the $x$-$y$ center point for signer $i$ at time $t$ and $n_{0:t}^i$ is the total count of $x$-$y$ points for signer $i$ for times $0:t$. $\boldsymbol{P}_t$ refers to a location with non-zero value of MEI at time $t$ and $P_t^j$ stands $C_t^i$.

### 4.2. Evaluation metrics

Diarization error rate (DER) is the metric that is widely used to evaluate speaker diarization systems. Despite its noisiness and sensitivity [15], it has been used by NIST to compare different diarization systems. DER consists of three types of errors: false alarm, missed speaker time and speaker error.

## 5. RESULTS AND DISCUSSION

### 5.1. Speaker diarization

The output of our speaker diarization system is given by probability values - one for each person per frame. We say that a person is speaking when the probability value for that person is the largest. The assumption is that at any time frame, only one person is speaking (unless more than one person has the same largest probability). Figure 2 shows a snapshot example of the output of the diarization system after running it on IN1016-AMI meeting data. In this figure, we can clearly see that the person that is gesturing is the speaker and the MHI clearly reflects this observation. But is that always the case? Table 1 shows that a person could be moving without speaking or that they could be speaking without gesturing.

**Table 1**. Speech and motion overlap for the seven videos

| Speech? | Motion? | Overlap |
|---|---|---|
| Yes | Yes | 0.98 |
| No | Yes | 0.77 |

DER = 196.75
Motion for each speaker is defined as sum(MEI) $> 0$

Table 2 gives performance scores of the diarization system after running it on seven videos. Performance scores range from 31.90% to 59.90% DER. Previous state-of-the-art scores for online diarization using audio range between 39.27% DER (for multiple microphones) and 44.61% DER (for a single microphone) [16]. Our scores, which use only gestures, are roughly close to previous scores. Notice that in table 2, the scores for FA are close to 0. This resulted a) from forcing our system to assume that only one person is speaking at any time and b) from evaluating the performance on speech-only segments. The non-zero FA scores in the table resulted from speakers sharing the same largest probability.

**Table 2**. Online speaker diarization results

| Video | Miss | FA | Spkr | DER | DER\{FA} |
|---|---|---|---|---|---|
| IN1005 | 2.90 | 0.00 | 38.40 | 41.24 | 41.30 |
| IN1009 | 5.50 | 0.00 | 54.40 | 59.90 | 59.90 |
| IN1012 | 11.00 | 0.00 | 40.30 | 51.34 | 51.30 |
| IN1013 | 12.80 | 0.00 | 36.40 | 49.23 | 49.20 |
| IN1016 | 6.70 | 0.50 | 33.50 | 40.66 | 40.20 |
| IS1009b | 2.60 | 0.50 | 29.30 | 32.46 | 31.90 |
| IS1009c | 1.80 | 0.00 | 45.30 | 47.14 | 47.10 |
| ALL | 6.80 | 0.20 | 38.80 | 45.72 | 45.60 |

MS = Missed Sign, FA = False Alarm
Spkr = Speaker error
DER = MS + FA + Spkr
DER\{FA} = DER without FA


(a) Frames

(b) MHI

**Fig. 2**. Output of the online diarizer on IN1016 meeting video. (a) shows original frames with the active speaker identified. The vertical bar shows the relative confidence in the prediction of *who is speaking?* (b) shows the MHI of the active speaker.

### 5.2. Signer diarization

Like the speaker diarization output, the output of the signer diarization system is also given by probability values. We say that a person is signing when the probability value for that signer is the largest. The performance scores for signer diarization are given in table 3. These error scores are lower than those reported in our previous work, where we used corner detection and tracking [4].

**Table 3**. Online signer diarization results

| Video | Miss | FA | Sgnr | DER | DER\{FA} |
|---|---|---|---|---|---|
| KN5jan7 | 5.80 | 0.00 | 9.90 | 15.67 | 15.70 |
| PiKe4jan7 | 7.80 | 0.00 | 14.80 | 22.63 | 22.60 |
| ReKe10jan7 | 6.90 | 0.00 | 13.00 | 19.93 | 19.90 |
| SuJu16jan7 | 7.10 | 0.00 | 15.00 | 22.18 | 22.10 |
| ALL | 6.90 | 0.00 | 13.30 | 20.17 | 20.20 |

One main difference between signer diarization and speaker diarization is that whenever there is signing, there is definitely motion. This fact is confirmed by table 4, which also shows that there can be significant motion in the absence of signing. Non-signing motion makes signer diarization a non-trivial problem.

**Table 4**. Sign and motion overlap for the four videos

| Sign? | Motion? | Overlap |
|---|---|---|
| Yes | Yes | 1.00 |
| No | Yes | 0.94 |

DER = 121.66

Motion for each signer is defined as sum(MEI) $> 0$

## 6. CONCLUSIONS

This study proposed and showed the use of motion history images (MHI) as a representation of gestural activity in an online speaker or signer diarization system. MHIs can efficiently represent where, how and how long motion occurred. The study claimed that these properties make MHIs applicable in online speaker and signer diarization systems, where motion is an integral part of uttering activity. Experiments on speaker and signer diarization problems using real data indicate that our solution is applicable in real-world applications (for example, video conferences).

Future work on diarization can extend our work in two ways. One way is by adding in extra information (for example, speech in the case of speaker diarization, or gaze in the case of signer diarization, where interlocutor(s) must be looking at the signer to be part of the conversation). The second way to extend our work is to modify our model of conversation dynamics. In our conversation model, each person has an independent model of *speaking/signing*. But one can enrich the model by adding in parameters to model the relationship of listening and speaking. Such a model can, for example, encode the idea that a speaker is less likely to continue speaking if another just started speaking.

## 7. RELATION TO PRIOR WORK

The work presented here has focused on using MHI for both speaker and signer diarization. To the best of our knowledge, this is our contribution. This work is similar to our previous work [17], where we first justified and used gestures for speaker diarization. Our previous work performs speaker diarization by tracking corners, filtering out motionless corners and classifying them based on the location of the speakers. The core of our previous system depends on corner detection and Lucas-Kanade tracking. These operations are computationally expensive [18, 19]. By contrast, our current diarization system is much less computationally intensive because of use of Motion History Image (MHI) [9, 10, 11].

In terms of the modeling framework, our work is similar to [5], who used a probabilistic framework that utilizes multimodal information to perform online speaker diarization. The difference is that they use SIFT descriptors [20] to model the visual aspect of the multimodal information, while we use MHI. Other video features like compressed MPEG-4 features have also been used in the multimodal speaker diarization literature [21, 22, 2, 23]. We contribute to this literature by drawing attention to the advantages of using motion history images [9, 10, 11] in speaker and signer diarization.

In summary, our work builds on and extends the literature in two ways: *a*) emphasis on the use of MHI for speaker and signer diarization *b*) an online diarization system that works on visual data (speaker and signer diarization). The *C++* source code will be made available.

## 8. REFERENCES


[1] Ruslan Mitkov, *Anaphora resolution*, vol. 134, Longman London, 2002. 1

[2] Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012. 1, 4

[3] Sue E Tranter and Douglas A Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006. 1

[4] B.G. Gebre, P. Wittenburg, and T. Heskes, "Automatic signer diarization - the mover is the signer approach," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, 2013, pp. 283–287. 1, 4

[5] Athanasios Noulas and Ben JA Krose, "On-line multimodal speaker diarization," in *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 2007, pp. 350–357. 1, 4

[6] Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, and Sylvain Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," 2013. 1

[7] Sylvain Meignier and Teva Merlin, "Lium spkdiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, 2010, vol. 2010. 1

[8] Deepu Vijayasenan and Fabio Valente, "Diartk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings.," in *INTERSPEECH*, 2012. 1

[9] James W Davis and Aaron F Bobick, "The representation and recognition of human movement using temporal templates," in *1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1997, pp. 928–934. 1, 4

[10] Gary R Bradski and James W Davis, "Motion segmentation and pose recognition with motion history gradients," *Machine Vision and Applications*, vol. 13, no. 3, pp. 174–184, 2002. 1, 4

[11] Md Atiqur Rahman Ahad, *Motion History Images for Action Recognition and Understanding*, Springer, 2013. 1, 4

[12] Pakorn KaewTraKulPong and Richard Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Video-Based Surveillance Systems*, pp. 135–144. Springer, 2002. 1

[13] 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. 3

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

[15] 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. 3

[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. 3

[17] Binyam Gebrekidan Gebre, Peter Wittenburg, and Tom Heskes, "The gesturer is the speaker," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 3751–3755. 4

[18] Carlo Tomasi and Jianbo Shi, "Good features to track," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 593–600, 1994. 4

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

[20] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004. 4

[21] Félicien Vallet, Slim Essid, and Jean Carrive, "A multimodal approach to speaker diarization on tv talk-shows," 2013. 4

[22] N Seichepine, S Essid, C Fevotte, and O Cappe, "Soft nonnegative matrix co-factorizationwith application to multimodal speaker diarization," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3537–3541. 4

[23] G. Friedland, H. Hung, and Chuohao Yeo, "Multimodal speaker diarization of real-world meetings using compressed-domain video features," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 4069–4072. 4