Speaker Diarization Using Gesture and Speech

Binyam Gebrekidan Gebre[1]*, Peter Wittenburg*[1]*, Sebastian Drude*[1]*, Marijn Huijbregts*[2]*, Tom Heskes*[2]

[1]Max Planck Institute for Psycholinguistics, The Netherlands
[2]Radboud University, The Netherlands
{firstname.lastname}@mpi.nl,{marijn.huijbregts@let,t.heskes@science}.ru.nl

Abstract

We demonstrate how the problem of speaker diarization can be solved using both gesture and speaker parametric models. The novelty of our solution is that we approach the speaker diarization problem as a speaker recognition problem after learning speaker models from speech samples corresponding to gestures (the occurrence of gestures indicates the presence of speech and the location of gestures indicates the identity of the speaker). This new approach offers many advantages: comparable state-of-the-art performance, faster computation and more adaptability. In our implementation, parametric models are used to model speakers' voice and their gestures: more specifically, Gaussian mixture models are used to model the voice characteristics of each person and all persons, and gamma distributions are used to model gestural activity based on features extracted from Motion History Images. Tests on 4.24 hours of the AMI meeting data show that our solution makes DER score improvements of 19% on speech-only segments and 4% on all segments including silence (the comparison is with the AMI system).
Index Terms: speaker diarization, gestures, speaker recognition, gaussian mixture models, motion history images

1. Introduction

Speaker diarization is the task of determining *who spoke when* from an audio/video recording. It is used in many systems such as information retrieval and speech recognition. In information retrieval, it is used to facilitate indexing and searching of audio-visual recordings. In speech recognition, it is used to enhance the readability of speech transcription by structuring the transcription in speaker turns.

The standard problem formulation of speaker diarization is as follows: given an audio or audio-video recording, the task is to determine the number of speakers and the segments of speech corresponding to each speaker. In this formulation, the state-of-the-art technique used to solve the problem is based on the ICSI system [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. The ICSI system performs three main tasks: speech/non-speech detection, speaker segmentation, and speaker clustering. The latter two tasks are performed iteratively using an agglomerative clustering technique using HMMs, GMMs and BIC.

The assumption in the ICSI-based systems is that the number of speakers and speaker models remain unknown (uncertain) all along the length of signals. However, this assumption may not hold for particular scenarios where such information is known a priori, which is the case in our experiments, or can be reliably estimated at initial stages. In meeting videos, the number of speakers can be determined from a few video frames using standard human/face detection algorithms [11]. Furthermore, speaker models, as this paper will show, can also be estimated for each person based on the occurrence of gestures.

In our previous work [12, 13], we performed speaker diarization on meeting videos based on the hypothesis that the person who is gesturing is also the speaker. In theory, this should work because there is a tight relationship between speech and gesture [14], but, in practice, the hypothesis has limitations: speakers can speak without gesturing and gesture recognition, by itself, is a challenging problem (e.g. people may appear to be gesturing when they move for other other reasons).

The goal of this paper is to solve these limitations by using the best of both worlds. Predictions based on gestures are used to develop speaker models with first pass on the data. On subsequent passes of the data, the learned speaker models are iteratively used to classify the frames of the speech and adapt speaker models. The iteration is not more than 3.

Summary of contributions: a) speaker diarization is formulated as continuous speaker identification b) speaker models are learned on the first pass of the data, based on predictions of *who is speaking* using gestures c) on the second or more passes of the data, speaker models are used to identify the speaker. The rest of the paper gives more details.

2. Speech-gesture representation

Given that the signals from speech and gesture are different (e.g. audio is 1-dimensional and video is 2-dimensional), how can we represent them such that they can be used for efficient computation and integration? For audio, we use MFCCs and for gestures, we use Motion History Images (MHI) [15, 13].

2.1. Speech representation: MFCC

Speech is a time-varying signal and as such is not suitable for speaker recognition. We, therefore, convert the speech signal to MFCCs (Mel Frequency Cepstral Coefficents) [16]. MFCCs are widely used features in speaker and speech recognition. We extract MFCC features as follows (the numbers correspond to the parameter values we selected). Our speech signal, which is sampled at 16 kHz, is divided into a number of overlapping frames, each 20 ms long (320 samples) with an overlap of 10 ms (160 samples). After multiplying each frame with a Hamming window, each frame is FFT-transformed (Fast Fourier Transform). The resulting power spectrum is then warped according to Mel-scale using 26 overlapping triangular filters producing filterbank outputs. The amplitudes of the DCT (Discrete Cosine Transform) of the logarithms of the filterbank outputs make the MFCC features. In our experiments, we take the first 20 MFCC coefficients (including the energy coefficient C_0) plus their first and second order derivatives for a total of 60-dimensional MFCC feature vector per speech frame. The HTK toolkit is used to compute the coefficients [17, 18].

2.2. Gesture representation: MHI

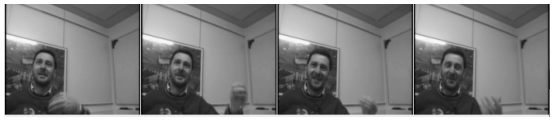
To represent gestures, we use Motion History Image (MHI) [15, 13]. MHI is a single stacked image that encodes motion that occurred between every frame pair for the last τ number of frames (where τ is the number we can fix ourselves). The type of information encoded in the MHI can be binary and, in such a case, it is called Motion Energy Image (MEI) or it can be scalar. In the latter case, it is called Motion History Image.

2.2.1. Motion Energy 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 (we perform frame differencing). Then the binary MEI $E(x, y, t)$ is defined as follows:

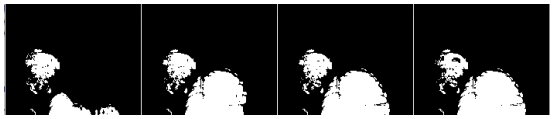
$$E_\tau(x, y, t) = \bigcup_{i=0}^{\tau-1} D(x, y, t - i), \tag{1}$$

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



(a) Frames

(b) MHI



(c) MEI

Figure 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 occurred in any of the last 25 frames.

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

3. Our diarization system

At a high-level, our diarization system performs the following steps:

1. Train a UBM on all audio data of the given recording.
2. Based on the location of gestures in the video, determine which speech sample belongs to which person (i.e. perform speaker diarization based on gestures).
3. Adapt UBM to create speaker models based on current predictions.
4. Use the current speaker models to identify to which speaker the next speech sample belongs (i.e. perform speaker diarization based on speaker models).
5. Repeat steps 3 and 4 N times each time using the latest diarization predictions and speaker models. In our experiments, $N = 3$.

3.1. Diarization using gestures

Given video and the number of speakers, we wish to infer, based on gestures, which person is speaking at time t . The inference is made using probabilistic models as follows. Let each person's state (speaking or non-speaking) be represented by z_t^i and let $v_{0:t}^i$ be video measurements (i.e. gestures) for person i, the objective is then to calculate the probability of z_t^i given $v_{0:t}^i$:

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

where $p(v_t^i|v_{0:t-1}^i)$ is a normalization constant, $p(z_t^i|v_{0:t-1}^i)$ is referred to as a conversation dynamics model and $p(v_t^i|z_t^i)$ is referred to as the gesture model. The person with the highest probability, $p(z_t^i|v_{0:t}^i)$, is the gesturer and hence, the speaker. The gesture and conversation dynamics models are described below.

3.1.1. Gesture model

We use gamma distributions to model gestural and non-gestural activities. The assumption is that MEI is a strong indicator of gestural activity. The higher the energy (the sum of MEI values), the higher the probability of gestural activity. A gamma distribution has a shape parameter $\boldsymbol{k}$ and scale parameter $\boldsymbol{\theta}$:

$$p(v_t^i|z_t^i; \boldsymbol{k}, \boldsymbol{\theta}) = \frac{(v_t^i)^{k_z-1} \exp(-\frac{v_t^i}{\theta_z})}{\theta_z^{k_z}\Gamma(k_z)} \quad \text{for } v_t^i, k_z, \theta_z > 0, \tag{4}$$

where $z = z_t^i$, v_t^i is the count of 'on' pixels in a MEI of speaker i and $x_t^i \in \{0, 1\}$ represents the probability of gestures for speaking and non-speaking person. The gamma distributions for speaking and non-speaking are the same for all speakers and their parameter are learned from annotated development data.

3.1.2. Conversation dynamics

In a conversation, the act of speaking has its own dynamics. The current speaker is more likely to have been speaking for a longer time than just the current frame. We encode this type of dynamics as follows:

$$p(z_t^i|v_{0:t-1}^i) = \sum_{z_{t-1}} p(z_t^i|z_{t-1}^i)p(z_{t-1}^i|v_{0:t-1}^i), \tag{5}$$

where $p(z^i_{t-1}|v^i_{0:t-1})$ is the posterior from the previous time and $p(z^i_t|z^i_{t-1})$ is the conversation dynamics. For simplicity, we set the conversation dynamics to a fixed matrix based on a heuristics: a speaker is 90% more likely to remain in the same state (speaking or non-speaking) as shown below:

$$p(z^i_t|z^i_{t-1}) = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}. \quad (6)$$

3.2. Diarization using speaker models

The diarization based on gestures gives output at the rate of video frame rate (40 ms). The MFCC features we get from audio come at the rate of 10ms. To make the two streams compatible, we take four MFCC feature vectors and replace them with their average vector. Given the average MFCC feature vectors, we determine which person is speaking at time t using maximum likelihood:

$$\hat{i}(t) = \arg\max_i \sum_{t'=t-\Delta}^{t+\Delta} \log p(\boldsymbol{a}_{t'}|\boldsymbol{\lambda}^i), \quad (7)$$

where delta, Δ, is a window of frames included for making predictions at time t and $\boldsymbol{\lambda}^i = \{\boldsymbol{w}^i, \boldsymbol{\mu}^i, \boldsymbol{\Sigma}\}$ is a speaker model for speaker i. In our experiments, Δ is set to 50 (2 seconds). The speaker models are derived from a UBM as described below.

3.2.1. Universal Background Model

A Universal Background Model (UBM) is a Gaussian Mixture Model (GMM) model. A GMM model is a weighted sum of M component densities:

$$p(\boldsymbol{a}_t|w_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \sum_{j=1}^{M} w_j \mathcal{N}(\boldsymbol{a}_t, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \quad (8)$$

where w_j are the mixture weights satisfying $\sum_{j=1}^{M} w_j = 1$ and $\mathcal{N}(\boldsymbol{a}_t, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ are the individual component densities. Each density component j a D-variate Gaussian of the form:

$$\mathcal{N}(\boldsymbol{a}_t, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \frac{\exp\{-0.5(\boldsymbol{a}_t - \boldsymbol{\mu}_j)(\boldsymbol{\Sigma}_j)^{-1}(\boldsymbol{a}_t - \boldsymbol{\mu}_j)\}}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_j|^{1/2}}, \quad (9)$$

where $\boldsymbol{\mu}_j$ is the mean vector and $\boldsymbol{\Sigma}_j$ is the covariance matrix.

In our system, the UBM is trained on audio features (MFCC features) from all speakers of a recording (including the silences). The UBM serves two purposes: first, it is used to derive speaker-dependent GMM models. Second, it is used to serve as a background or negative speaker model, against which each particular speaker model is compared to determine if they are speaking. Our UBM model consists of 64 60-variate Gaussian components. The covariance type is diagonal. The minimum variance value of the covariance matrix is limited to 0.01 to avoid spurious singularities [19]. Parameters of the UBM are estimated using EM algorithm [20, 21].

3.2.2. Adaptation of Speaker Models

The UBM, represented by $\boldsymbol{\lambda} = \{\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}_{\text{ubm}}$, is trained on all audio samples of a given recording. To make it model a particular speaker i, we need speech samples from speaker i and an adaptation technique. Initially, speech samples are collected for each speaker based on the occurrence of their gestures but later speech samples are collected based on speaker models. In either case, the adaptation technique is the same; we use a form of Bayesian parameter adaptation [22, 23]. Given $\boldsymbol{\lambda}$ and training speech samples for speaker i, $A^i = \{\boldsymbol{a}^i_1, \boldsymbol{a}^i_2, \ldots, \boldsymbol{a}^i_T\}$, we compute the responsibilities of each mixture component m^i in the UBM as follows:

$$p(m^i|\boldsymbol{a}_t, \boldsymbol{\lambda}) = \frac{w_m \mathcal{N}(\boldsymbol{a}^i_t, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{j=1}^{M} w_j \mathcal{N}(\boldsymbol{a}^i_t, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (10)$$

$p(m^i|\boldsymbol{a}_t, \boldsymbol{\lambda})$ and $\boldsymbol{a}_t$ are then used to compute sufficient statistics for the weight and mean of speaker i as follows[1]:

$$n^i_m = \sum_{t=1}^{T} p(m^i|\boldsymbol{a}_t, \boldsymbol{\lambda}). \quad (11)$$

$$E^i_m(\boldsymbol{a}) = \frac{1}{n^i_m} \sum_{t=1}^{T} p(m^i|\boldsymbol{a}_t, \boldsymbol{\lambda}) \boldsymbol{a}^i_t. \quad (12)$$

Using $E^i_m(\boldsymbol{a})$ and n^i_m, we can now adapt the UBM sufficient statistics for mixture m for speaker i as follows:

$$\hat{w}^i_m = [\alpha^i_m n^i_m / T + (1 - \alpha^i_m) w_m] \gamma^i. \quad (13)$$

$$\hat{\boldsymbol{\mu}}^i_m = \alpha^i_m E^i_m(\boldsymbol{a}) + (1 - \alpha^i_m) \boldsymbol{\mu}_m. \quad (14)$$

γ^i is a normalisation factor to ensure that the adapted mixture weights, $\hat{w}^i_m$, sum to unity:

$$\gamma^i = \frac{1}{\sum_{j=1}^{M} \hat{w}^i_j}. \quad (15)$$

α^i_m is an adaptation coefficient used to control the balance between old and new estimates for the weights and means. For each mixture m^i, a data-dependent adaptation coefficient is fixed as:

$$\alpha^i_m = \frac{n^i_m}{n^i_m + r}, \quad (16)$$

where r is a relevance parameter and is set to 16. For more details on these parameters, see [23].

4. Experiments

4.1. Datasets

We validate our proposed solution on test data of seven video recordings (≈ 4.24 hours), taken from a publicly available corpus called the AMI corpus [24]. The AMI corpus consists of annotated audio-visual data of a number of participants engaged in a meeting. The selected videos (IB4XXX) have four participants. The upper body of each participant is recorded using a separate camera and we put them together before diarization. For audio, we use the mixed-headset single wave file per video. Our development data consists of 4.9 hours of videos coming from IN10XX and IS1009x. The development data are used to learn parameter values when necessary.

4.2. Evaluation metrics

We report our scores using Diarization error rate (DER). DER consists of false alarm, missed speech and speaker errors [25]. DER is known to be noisy and sensitive [26] but is still widely used in many evaluations [7, 3]. A perfect diarization system scores 0% DER, but a very bad system (e.g. a system that predicts every speaker is speaking all the time) can go over 100%.

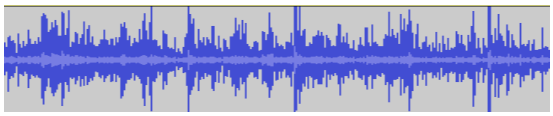
[1] covariance parameter is kept the same for all speakers; adapting it with new data decreased performance

5. Results and Discussion

Figure 2 illustrates how training speech samples are collected for adapting speaker models based on predictions using gestures. The figure clearly shows that the person that is gesturing is the speaker and the MHI visualization clearly reflects it. As table 1 shows, this is not always true (i.e. a person could be moving without speaking or that they could be speaking without gesturing). Hence, the need to pass through the data iteratively (adapting speaker models and making predictions).

(a) Frames

(b) MHI



(c) Speech

Figure 2: A snapshot of IN1016-AMI meeting data: (a) Original frames with the person gesturing identified. (b) The MHI of the gesturing person. (c) The speech waveform belongs to the person gesturing and is used to adapt a speaker model for that person. Each adapted speaker model is then used to identify the speaker for the given audio frames using maximum likelihood.

Table 1: Speech and motion overlap on all test videos

Speech?	**Motion**?	**Overlap**
Yes	Yes	0.96
No	Yes	0.82

DER = 72.09
Motion for each speaker is defined as sum(MEI) > 0

After the first diarization using gestures, we have adapted speaker models. Based on equation 7, we then use the adapted speaker models to score each audio feature vector – a person is said to be speaking at frame t when the likelihood for that person is the largest in a window spanning $\pm$ 50 frames[2] (4 seconds). The scoring is repeated 3 times: new diarization results are used to adapt speaker models and new adapted speaker models are used to make new diarization. Based on this procedure, DER scores are given in tables 2 and 3. The best scores of our system come after 3 iterations and are better than the baseline scores (18.79% vs 23.28% and 29.87% vs 31.18%). The baseline system is the AMI system [27, 28], which is based on an agglomerative clustering and segmentation technique.

[2]Note that the assumption is that only one person is speaking at any frame. The alternative to this assumption is to set a threshold for likelihood, which may be necessary to handle overlapped speech.

Table 2: Speaker diarization scores evaluated on speech-only segments. Each column in the speaker models section is a diarization score of speaker models that are adapted using diarization results from the previous column.

Diarization Error Rates (%)					
			Speaker models		
Name	**Baseline**	**Gesture**	**1st**	**2nd**	**3rd**
IB4001	19.76	53.81	33.51	27.06	23.76
IB4002	54.40	58.42	52.03	48.12	40.86
IB4003	12.20	44.53	16.13	10.48	10.35
IB4004	39.05	49.68	32.33	27.14	24.79
IB4005	13.56	37.69	17.89	18.70	19.63
IB4010	18.15	50.52	19.34	13.29	12.92
IB4011	14.59	45.76	11.53	10.64	10.37
ALL	23.28	48.04	24.14	20.20	18.79

Table 3: Speaker diarization scores evaluated on all segments including silences. Evaluating our system on silence segments increases DER as a result of increase in False Alarms.

Diarization Error Rates (%)					
			Speaker models		
Name	**Baseline**	**Gesture**	**1st**	**2nd**	**3rd**
IB4001	38.26	82.50	61.27	54.78	51.48
IB4002	100.20	104.76	97.62	93.71	86.39
IB4003	13.20	48.89	18.13	12.47	12.34
IB4004	41.15	59.44	37.16	31.94	29.61
IB4005	16.16	47.66	23.80	24.61	25.55
IB4010	20.75	56.18	25.42	19.37	19.00
IB4011	17.59	52.57	18.27	17.38	17.09
ALL	31.18	60.99	35.23	31.28	29.87

6. Conclusions

This study proposed a solution to the speaker diarization problem based on the exploitation of the best of two worlds: gestures and speech. The use of gestures enables the formulation of the diarization problem in a non-standard way. A UBM is first trained on all audio feature vectors of a given recording. The UBM is then adapted to different speakers based on the speech samples corresponding to their gestures. Finally, the adapted speaker models are used to perform diarization (then adaptation, then diarization and then adaptation ...). This new approach has comparable state-of-the-art performance and is faster (avoids agglomerative clustering).

Future work can extend our work in many ways. One way is by enriching the gesture model. Our current gesture model is quite efficient [13] but may fail to distinguish true gestures from random body movements. The other way is to make an online version of our system. Our current system makes multiple passes through the data but this may not be necessary: speaker models do not need much more than 90 seconds of training samples [19] and the UBM, which, in our current system, is trained on the whole audio recording, could be trained on general population. The speaker models could then be adapted online as more gestures and speech samples arrive.

7. References

[1] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-multiple speaker clustering using hmm," in *INTERSPEECH*. Citeseer, 2002.

[2] G. Friedland, H. Hung, and C. Yeo, "Multi-modal 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.

[3] 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*, vol. 20, no. 2, pp. 356–370, 2012.

[4] S. E. Tranter and D. 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.

[5] S. Meignier and T. Merlin, "Lium spkdiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, vol. 2010, 2010.

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

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

[8] G. Friedland, A. Janin, D. Imseng, X. Anguera Miro, L. Gottlieb, M. Huijbregts, M. 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.

[9] M. Huijbregts, D. 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.

[10] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," 2013.

[11] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[12] B. G. Gebre, P. Wittenburg, and T. Heskes, "The gesturer is the speaker," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 3751–3755.

[13] B. G. Gebre, P. Wittenburg, T. Heskes, and S. Drude, "Motion history images for online speaker/signer diarization," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2014.

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

[15] J. W. Davis and A. 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.

[16] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[17] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book (for htk version 3.4)," *Cambridge university engineering department*, vol. 2, no. 2, pp. 2–3, 2006.

[18] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*. Entropic Cambridge Research Laboratory Cambridge, 1997, vol. 2.

[19] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, 1995.

[20] A. P. Dempster, N. M. Laird, D. B. Rubin *et al.*, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 291–298, Apr 1994.

[23] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[24] 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.

[25] X. A. Miro, *Robust speaker diarization for meetings*. Universitat Politècnica de Catalunya, 2007.

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

[27] D. A. Van Leeuwen and M. Huijbregts, "The AMI speaker diarization system for NIST RT06s meeting data," in *Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 371–384.

[28] M. Huijbregts, "Segmentation, diarization and speech transcription: Surprise data unraveled," PhD Thesis, University of Twente, Nov. 2008, publisher: Centre for Telematics and Information Technology University of Twente, publisherlocation: Enschede, ISSN: 1381-3617, ISBN: 978-90-365-2712-5, Numberofpages: 172.