GMM-UBM based open-set online speaker diarization

Conference Paper · January 2010

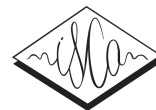Source: DBLP

3 authors, including:

Frank Wallhoff
Jade University of Applied Sciences, Oldenburg, Germany
106 PUBLICATIONS   **944** CITATIONS

SEE PROFILE

GMM-UBM based open-set online speaker diarization

Jürgen Geiger, Frank Wallhoff and Gerhard Rigoll

Institute for Human-Machine Communication
Technische Universität München
80290 Munich, Germany
{geiger, wallhoff, rigoll}@mmk.ei.tum.de

Abstract

In this paper, we present an open-set online speaker diarization system. The system is based on Gaussian mixture models (GMMs), which are used as speaker models. The system starts with just 3 such models (one each for both genders and one for non-speech) and creates models for individual speakers not till the speakers occur. As more and more speakers appear, more models are created. Our system implicitly performs audio segmentation, speech/non-speech classification, gender recognition and speaker identification. The system is tested with the HUB4-1996 radio broadcast news database.

Index Terms: Speaker diarization, Gaussian mixture models, open-set speaker recognition

1. Introduction

In audio indexing, it is not only needed to transcribe speech, but also to extract some more data, called meta-data. This includes, for example, speaker turns. Finding speaker turns and identifying the speakers is known as speaker diarization. In simple words, speaker diarization answers the question of "who spoke when". Speaker diarization has been the subject of NIST Rich Transcription evaluations [1].

In speaker diarization, the task is to segment an audio stream into homogenous segments where only one speaker is active and to classify this speaker. Speaker diarization can be divided into offline and online speaker diarization. In the offline variant, the whole audio stream is processed at the same time. This means, for example, that all segments must be present before the resulting speakers are compared and clustered. In online speaker diarization, the segments must be processed as soon as they are created, meaning that as soon as a new segment of the audio stream arrives, it must be assigned to a speaker.

Most of the offline speaker diarization systems work the following way: a complete audio stream is segmented in smaller homogenous parts, each of them containing only one speaker. After the complete audio stream has been segmented, the segments are compared and clustered. In this way, one cluster is created for every speaker. This is done for example using the Bayesian Information Criterion (BIC) [2]. Online speaker diarization is different: the system cannot wait for all segments to arrive before the clustering process begins. Therefore, no hierarchical clustering algorithm (as in offline speaker diarization) can be applied. More sophisticated clustering algorithms have to be used. In [3] a leader-follower clustering for k-means clustering and a dispersion-based speaker clustering are proposed for online speaker diarization.

In this work, we propose a system that can perform online speaker diarization. In addition, our system performs open-set recognition. We use a model-based approach, applying Gaussian mixture models (GMMs). As a starting point for our system, just three models exist: one for *male*, one for *female* and one for *garbage* (for example music). The models for the occurring speakers are created during runtime as the corresponding speakers occur. This means that we have an open-set system.

Gaussian mixture models (GMMs) in addition with universal background models (UBMs) and maximum aposteriori (MAP) adaptation (also known as Bayesian adaptation [4]) have been proposed for speaker verification in [5]. From a large training database, a UBM is created and to enroll a new speaker in the verification system, very little data is needed, because the speakers' GMM is created from the UBM with MAP adaptation. We propose to use these techniques for a real-time open-set speaker diarization system.

A system which is similar to our system has been presented in [6]. However, there are some substantial differences. In [6], a different model adaptation technique is used and the main difference is that we use another database. In [6], a database of recordings of European Parliament plenary speeches was used, whilst we use a radio broadcast news database. In our database, not only speech occurs, but also music, which is responsible for a lot of errors. Another similar system has been proposed in [3]. However, in [3], only speaker clustering has been performed. The system has used the reference segmentation instead of performing segmentation itself. In [7], GMMs with MAP adaptation have been used for offline speaker diarization.

One important property of our system is that in contrast to many other speaker diarization systems like proposed in [8], our system performs online speaker diarization. Furthermore, it does not only perform online speaker clustering or audio segmentation, but carries out both steps. Also, as our system works with GMMs, the result of one pass of the system are not only speaker clusters, but also complete trained GMMs that can be used for speaker recognition. Due to the database we use, the system has to work not only with clean speech, but also with speech overlapped by music. One possible application for our system is a robot which automatically learns to distinguish between different speakers.

In Section 2, an overview of the system is given as well as a detailed description of all its modules. Experiments and results are presented in Section 3, before a conclusion is drawn in Section 4.

2. System Description

2.1. Overview

The system introduced in this paper can be divided into an offline part and an online part. In the first phase, the offline part of the system is used to train GMMs for each gender (*male* and *female*) and *garbage*. These three models are used as an initial-

26 – 30 September 2010, Makuhari, Chiba, Japan

ization for the online phase, where the *male* and *female* models play the role of universal background models (UBMs).

The online part of the system does speaker clustering. This is done as explained in the following. Initially, the continuous audio stream is segmented. Each time a new segment arrives, a model-based classification is performed and a speaker label is assigned to the segment based on the classification result. To be able to perform model-based classification, a model has to be built for each speaker. The two gender models as well as the *garbage* model are constructed in the offline phase of the system. Models for individual speakers are generated sequentially in the online phase. This is shown in Fig. 1. When the first segment of the audio stream arrives, recognition with the existing three models is performed. If the segment is classified as *male* or *female*, a new speaker model is created by copying the corresponding gender model and then adapting the model with the audio data of the segment, using MAP adaptation. Here, the first audio segment is classified as *male*, thus a new speaker model is created by copying the *male* model and adapting it with the audio data of segment 1. We now have the gender models, the *garbage* model and the model for the first speaker, named *s001*. If an audio segment is classified as *garbage*, as is segment 2 in this example, no model adaptation is performed. Segment 3 is classified as *female* here, leading to a new speaker model *s002*. The next segment is recognised as *s001*. In this case, the model is once again adapted with the new audio data. The system continues to work in this way: Whenever a new segment is recognized as *male* or *female*, a new speaker model is created by copying and adapting the gender model. If an already seen speaker is recognised, its model is adapted as well.

Our system implicitly performs several tasks: audio segmentation, speech/non-speech classification, gender recognition, speaker novelty detection and speaker identification. Audio segmentation is performed energy-based, as described below. Speech/non-speech classification is done model-based: The decision between the *garbage* model and the speaker models constitutes the decision between speech and non-speech. The decision between the two gender models and any of the speaker models (which have been derived by one of the gender models) corresponds to the gender recognition. For example, the process of creating new models could be turned off to get a gender recognition system. Speaker novelty detection is achieved by the maximum likelihood decision in the classification step. Consider the case where several speaker models have already been created. Then, if one of the gender models has the highest likelihood in the classification step for a new audio segment, the segment is classified as being from a new speaker.

The system works online: There is no clustering process that uses the whole audio recording. The recording is processed online and labeled with speaker names. After a complete pass with one recording, the performance can be evaluated. The advantages of the system are the following: we don't need any trained speaker models (beside the gender and garbage models), the models are created on the fly. Additionally, the number of possible speakers need not be known to the system. This is known as open-set speaker recognition. And a result of the online pass of the system are trained speaker models (GMMs) that can be used for speaker recognition. The model adaptation process of the system was implemented with HTK [9].

2.2. Segmentation and Feature Extraction

To segment the audio stream, we applied energy-based segmentation (a modified version of the segmentation as implemented in [9]): framewise, each audio frame is either declared
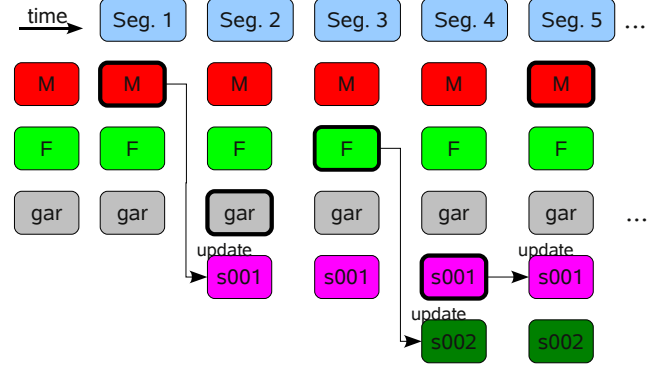

Figure 1: System operation. At every time step, the model the segment is classified as is highlighted by bold border lines.

as speech/garbage or silence. When the energy of the frame exceeds a certain threshold, it is declared as speech/garbage, otherwise as silence. A couple of rules are then used to determine starting and end points of speech/garbage segments. For example, we applied a maximum segment length, guaranteeing for a low system latency. Different maximum segment lengths are evaluated in section 3.

As acoustic features, we use 12 standard MFCCs (+ Energy) with delta and acceleration coefficients, which sums to a total of 39 features. We use a frame rate of $10\ ms$ and window size of $25\ ms$.

2.3. Speech/Non-Speech Classification

As noted above, the system presented in this work does speech/non-speech classification. This is achieved in the classification step by the decision for one of the GMMs. The gender models and the models for invidual speakers stand for *speech*, whereas the *garbage* model stands for *non-speech*.

2.4. Gender Identification

In the training phase of our system, gender GMMs are trained (male,female) and one GMM for music and silence etc., named *garbage*. The models act as universal background models (UBM). In the online part of the system, gender recognition is implicitly performed: Each segment is classified as either *male*, *female*, *garbage* or one of the newly created speakers. Each of the new speaker models is created from either *male* or *female*, determining its gender.

2.5. Adaptation Process

The model adaptation process is the main part of the system. In the recognition phase of the system, the speaker models are constantly adapted with the new audio data. In order to create a new model for a new speaker, the corresponding gender model is copied and adapted with the new data of this speaker. We use MAP adaptation the same way as it is used in GMM systems with universal background models (UBM) [5] to adapt the means, mixture weights and variances of the speaker GMMs. The mean of mixture component m of the GMM is adapted using Eq. (1):

$$\hat{\boldsymbol{\mu}}_m = \frac{N_m}{N_m + \tau}\bar{\boldsymbol{\mu}}_m + \frac{\tau}{N_m + \tau}\boldsymbol{\mu}_m, \qquad (1)$$

where $\hat{\boldsymbol{\mu}}_m$ is the adapted mean, $\bar{\boldsymbol{\mu}}_m$ is the mean of the observed adaptation data, $\boldsymbol{\mu}_m$ ist the old mean of the GMM, τ is a weighting factor and N_m is the occupation likelihood of the adaptation

data for mixture component m.

The adaptation of mixture weight w_m for mixture m follows Eq. (2):

$$\hat{w}_m = \left(\frac{N_m}{N_m + \tau}\frac{N_m}{T} + \frac{\tau}{N_m + \tau}w_m\right)\gamma, \qquad (2)$$

where $\hat{w}_m$ is the adapted mixture weight, w_m is the original mixture weight, T is the length of the adaptation data, and γ is a normalizing factor, which is needed to ensure that all mixture weights sum to 1.

Eq. (3) is used to adapt the variances of the GMMs:

$$\hat{\boldsymbol{\sigma}}_m^2 = \frac{N_m}{N_m + \tau}\mathrm{E}_m(\mathbf{x}^2) + \frac{\tau}{N_m + \tau}(\boldsymbol{\sigma}_m^2 + \boldsymbol{\mu}_m^2) - \hat{\boldsymbol{\mu}}_m^2. \quad (3)$$

Here, $\hat{\boldsymbol{\sigma}}_m^2$ is the adapted variance, $\boldsymbol{\sigma}_m^2$ is the old variance and $\mathrm{E}_m(\mathbf{x}^2)$ is the expected value of the squared observation vector $\mathbf{x}^2$.

The weighting factor τ was optimised on the development test data. Smaller values of τ lead to higher adaptation, which means that the new mean is nearer at the mean of the adaptation data than at the old mean. Using extreme small values of τ causes the system to produce a new speaker for almost every audio segment, because in this case, the new model is too much adapted to the small amount of adaptation data and not general enough to be able to recognize other segments from this speaker. If τ is set to zero, the new speaker model corresponds to a model trained only with the adaptation data. Conversely, higher values of τ lead to less adaptation. In the case of $\tau = \infty$, the old model is just copied in order to get the new model, while neglecting the adaptation data.

3. Experiments

3.1. Database

For testing purposes we use the HUB4-1996 radio broadcast news database [10]. This database consists of radio broadcast recordings. We use a total of 11 recordings, 6 for training the gender and garbage models, 3 for development test and 2 for evaluation test. Each recording is about 28 minutes long.

Note that in this system, there is not only a simple gender recognition, but a speech/non-speech classification as well, because of the *garbage* class. The composition of the three files of the development test set is: 74 % *male*, 19 % *female* and 7 % *garbage*. The *garbage* parts consist mostly of music segments and pauses between speaker turns. Additional difficulty is added because sometimes speech is overlapped with music.

3.2. Speech/non-speech classification and Gender Recognition Results

In order to get results for the speech/non-speech classification and gender recognition performance of our system, we start the online phase of our system without using the adaptation technique. In this way, no individual speaker models are created and recognition is performed with just the three models *male*, *female* and *garbage*. To get the speech/non-speech classification error rate, the amount of time that *male* or *female* are classified as *garbage* and vice versa is summed up and divided by the total length of the audio test material.

The gender recognition performance can be obtained by summing up the amount of time that *male* and *female* are mixed up. The speech/non-speech classification and gender recognition error rates with different number of mixtures for the GMMs are shown in Table 1.

mixtures	speech/non-s.	gender
1	9.1	23.0
2	14.2	20.2
4	5.5	22.7
8	5.1	19.9
16	4.1	12.1
32	3.9	11.7
64	3.7	7.1
128	4.3	5.9
256	4.6	4.8
512	4.5	4.1
1024	4.7	3.6

Table 1: Speech/non-speech classification and gender recognition error rates on the development test set for different number of mixtures (in %)

As can be seen in the table, more mixtures generally promise better results. However, the speech/non-speech classification performance stagnates at a medium number of mixtures and even gets worse a bit for higher numbers. Therefore, we decide to work with 128 mixtures. The classification performance is good enough and computational effort is small enough to let the system work in real-time. In sum, the error rate with 128 mixtures is 10.2 %.

One part of the errors occur due to segmentation errors: If the segmentation step erroneously undersegments the audio stream, there are segments which contain more than one speaker, for example a female speaker followed by a male speaker. In the classification step, it is inevitable to make a small error for this segment, as only one label can be assigned to each segment. Another large part of the errors occur because of music overlapping with speech.

3.3. Online Recognition Results

The online phase of the system including the adaptation process is the main part to be evaluated. The system makes similar segmentation, speech/non-speech classification and gender identification errors as shown above. Thus, when comparing the clustering results with similar works, e.g. [6] or [3], it must be taken into account that the system presented in this work performs the whole processing loop, while for example in [3], the reference segmentation is used. We choose the trained gender and garbage models with 128 mixes for the online test, as this number represents the best trade-off between error rate and computation time. The best results can be achieved with adaptation parameter τ from Eq. (1) chosen to $\tau = 135$.

In the recognition process, the speakers are labeled *s001*, *s002* etc., depending on the order of their appearance. These labels do not necessarily match with the speaker labels in the reference transcription. In order to evaluate the recognition results, we need an assignment of recognised speakers to speakers in the reference transcription. In the online phase of the system, a confusion matrix C is filled: for every speaker in the transcription, the amount of time for every speaker it is labeled with is collected. Ideally, if no recognition errors are made, every speaker in the reference transcription gets only one label, which is not used for segments of other speakers. For the confusion matrix this would mean that in every row and every column, there is only one non-zero value.

As the labels in the recognition process are chosen arbitrarily, they are renamed, such that the misclassification rate is minimised. The columns of the confusion matrix are switched

such that *trace(C)* is maximised, which is done with the following algorithm:
1. look for maximum value c_{ij} in C
2. copy column C_j into new matrix C^* to position C_i and delete it in C
3. repeat steps 1 and 2 until all columns from C are copied to C^*

With this procedure, an assignment between speakers in the transcription and recognized speakers is made such that the system performance can be evaluated.

The misclassification rate e is calculated by summing up the number of frames for each speaker that are classified as another speaker divided by the total number of frames:

$$e = \frac{\sum_{i=1}^{N}\sum_{\substack{j=1\\i\neq j}}^{M} f_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{M} f_{ij}}, \qquad (4)$$

with N being the true number of speakers while M is the number of speakers hypothesized by the system and f_{ij} equals the length of all audio frames that belongs to speaker i and were classified as speaker j.

Another measure for the performance of the clustering process is the cluster purity p, which is calculated as:

$$p = \frac{\sum_{j=1}^{M} \max_{1<i<N} f_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{M} f_{ij}} \qquad (5)$$

The cluster purity shows how well the clustering process generates clusters which contain only one single speaker.

Table 2 shows the results of the proposed speaker diarization system for the three files of the development test set and the two files of the evaluation test set. Shown is the true number of speakers N as well as the hypothesized number of speakers M, the misclassification rate e for three different maximum segment lengths in the segmentation step and the cluster purity p.

audio file	N	M	miscl. rate e			cl. purity p
			1s	2s	3s	
dev. test file 1	15	19	42.3	21.9	29.9	82.3
dev. test file 2	27	20	52.7	38.4	45.7	70.9
dev. test file 3	17	21	51.4	29.8	32.1	83.1
dev. test Ø			48.8	30.0	35.9	78.8
eval. test file 1	18	22	42.1	35.0	40.1	72.7
eval. test file 2	23	20	42.2	55.4	38.1	70.3

Table 2: True and hypothesized number of speakers, misclassification rates for different maximum segment lengths (in %) and cluster purity (in %)

As can be seen, the system is capable of creating a reasonable number of speaker clusters. Misclassification rate e and cluster purity p both contain errors from each of segmentation, gender recognition and speaker recognition. The best results are achieved with a maximum segment length of 2s.

4. Conclusion and Future Work

We have constructed an open-set online speaker diarization system by training gender models (GMMs) and a *garbage* model on the HUB4-1996 radio broadcast news database and using these models as an initialization for an online system which creates models for individual speakers during runtime as the speakers occur. The speaker models are created by copying the gender model and adapting it using MAP adaptation. As a result, we have obtained a misclassification rate of 30.0 % for the online system on our development test set.

The system has problems with speech overlapped by music: If a speaker occurs several times in the recording, sometimes with overlapping music and sometimes without, the system tends to create two different speaker models here, one for the speaker with overlapping music and one without. In order to reduce the influence of overlapping music, we are planning to integrate a music removal module in our system.

In our ongoing research, we want to try to use the proposed method for an acoustic event detection and classification system to classify acoustic events instead of speakers. Furhermore, we are planning to implement the system on a robot to use it in a real environment.

5. Acknowledgements

6. References

[1] NIST, "Benchmark Tests: Rich Transcription (RT)," http://www.nist.gov/speech/tests/rt/.

[2] A. Tritschler and R.A. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in *Sixth European Conference on Speech Communication and Technology, Eurospeech*, 1999.

[3] D. Liu and F. Kubala, "Online speaker clustering," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. (ICASSP'03)*, 2003, vol. 1, pp. 572–575.

[4] R.O. Duda and P.E. Hart, *Pattern classification and scene analysis*, Wiley, New York, 1973.

[5] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[6] K. Markov and S. Nakamura, "Never-ending learning system for on-line speaker diarization," in *IEEE Workshop on Automatic Speech Recognition & Understanding, 2007. ASRU*, 2007, pp. 699–704.

[7] R. Sinha, S.E. Tranter, M.J.F. Gales, and P.C. Woodland, "The cambridge university march 2005 speaker diarisation system," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[8] S.E. Tranter and D.A. Reynolds, "Speaker Diarisation for Broadcast News," in *Odyssey Speaker and Language Recognition Workshop*, 2004, pp. 337–344.

[9] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.2)," *Cambridge University Engineering Department*, 2002.

[10] D. Graff, "An overview of Broadcast News corpora," *Speech Communication*, vol. 37, no. 1-2, pp. 15–26, 2002.