# FAST AND ROBUST SPEAKER CLUSTERING USING THE EARTH MOVER'S DISTANCE AND MIXMAX MODELS

*Thilo Stadelmann*[1,2]

[1]SFB/FK615
University of Siegen
D-57068 Siegen, Germany
stadelmann@informatik.uni-marburg.de

*Bernd Freisleben*[1,2]

[2]Dept. of Mathematics & Computer Science
University of Marburg
D-35032 Marburg, Germany
freisleb@informatik.uni-marburg.de

## ABSTRACT

Speaker clustering is the task of assigning a unique label to all speech segments in a video uttered by the same speaker. There are two key challenges: processing speed and robustness in the presence of noise. In this paper, we present an approach to significantly improve the processing speed of a hierarchical speaker clustering algorithm by using the earth mover's distance (EMD) as the distance measure. By extending the well-known MIXMAX speaker model such that the EMD can be applied, noise robustness is achieved. Experimental results show that the runtime of the proposed EMD approach decreases by more than a factor of 120 compared to a likelihood ratio based distance measure while the clustering performance remains nearly the same.

## 1. INTRODUCTION

Speaker clustering algorithms are essential for (a) gathering data for speaker adaptation to improve automatic speech recognition (ASR) performance and (b) to identify persons in audio or video material for surveillance or retrieval purposes. Our goal is to use a speaker clustering algorithm as the basis of building an audio-based person indexing system for videos. The two main challenges of this task are speed and robustness. Since speaker clustering is only one step in a chain of operations to analyze a video, its runtime has to be as small as possible. Furthermore, to obtain an approach that works under varying conditions, a speaker clustering algorithm must be robust against different types of noise.

In this paper, we present an approach to improve the speed of a hierarchical speaker clustering algorithm. The basic idea is to compare speaker models directly rather than relying on the underlying feature vectors and use the earth mover's distance (EMD) [1] known from the image retrieval domain to measure the distance between speaker models. To achieve robustness in the presence of noise, a method is proposed to use the EMD with a special model based noise cancellation scheme, the MIXMAX model [2]. Experimental results for a 47 minute test video show that the runtime of the proposed EMD approach outperforms a likelihood ratio based distance measure by more than a factor of 120 while the clustering performance remains nearly the same.

The paper is organized as follows. Section 2 discusses related work in the domain of speaker clustering. Section 3 briefly reviews the used speaker modelling techniques. Section 4 presents the new clustering method using the EMD, while section 5 reports experimental results. Section 6 concludes the paper and outlines areas for future research.

## 2. RELATED WORK

In [3], Jin et al. present a hierarchical speaker clustering system for ASR improvement consisting of Gaussian mixture models (GMM) for speaker representation and the generalized likelihood ratio test (GLR) as the distance measure. The authors report improvements in the word error rate as high as with hand labeled data using their unsupervised system. The same techniques where used by Solomonoff et al. [4].

Ajmera and Wooters report on their unsupervised speaker-segmentation and -clustering system in [5]. They use hidden Markov models (HMM) to represent the data, where each state represents a single speech segment and is modeled by a GMM. To merge states, the Bayesian information criterion (BIC) is used to determine the pair of nearest clusters (states). This introduces an automatic stopping criterion, so that the algorithm can be regarded as being robust against wrong parameter settings.

In [6], Liu and Kubala introduce their online speaker clustering algorithm. It clusters a new segment immediately after it has been processed rather than first collecting all segments. In contrast to the computational complexity of a hierarchical approach, which increases exponentially with the number of speech segments, their method's complexity increases only linearly. It also shows better results in terms of cluster purity and misclassification rate while still using GMMs and GLR.

## 3. ROBUST SPEAKER MODELLING

In this section, the speaker modelling techniques used in this paper are briefly reviewed.

### 3.1. Gaussian Mixture Model

GMMs are widely used for speaker modelling due to their ability to model arbitrarily shaped probability density functions (pdf). They consist of a mixture of $M$ Gaussians each with $D$-dimensional mean $\vec{\mu}$ and typically diagonal covariance $\vec{\sigma}$, weighted by a factor $w$ so that the overall mass is 1. The likelihood of a set of $D$-dimensional feature vectors $X = \{\vec{x_1}..\vec{x_T}\}$ is given by $p(X|\lambda)$, which can be computed per dimension using the one-dimensional Gaussian pdf $g(..)$ due to the diagonal covariances.

$$\lambda_{GMM} = \{w_i, \vec{\mu_i}, \vec{\sigma_i}\}, \quad i = 1..M \tag{1}$$

$$p(X|\lambda) = \prod_{t=1}^{T} \sum_{i=1}^{M} w_i \cdot \prod_{d=1}^{D} g(x_{t_d}, \mu_{i_d}, \sigma_{i_d}) \tag{2}$$

### 3.2. MIXMAX Model

The MIXMAX model as proposed by Rose et al. [2] consists of a standard GMM $\lambda^s$ as the speaker model and a second GMM $\lambda^b$ to model the accompanying log-additive background noise. The advantage is that no a priori clean speech models are needed: during the speaker model estimation phase, the noisy speech mixtures get "masked" by the background mixtures rather than cleaned. In the likelihood computation, the feature vectors are scored against the combined speaker-background model. The more a speaker mixture is masked by noise, the less it contributes to the final likelihood score. Here, $G()$ is the $1D$-Gaussian error function.

$$\lambda_{MIXMAX} = \{\lambda^s_{GMM}, \lambda^b_{GMM}\} \quad (3)$$

$$p(x_{t_d}|i,j,\lambda) = g(x_{t_d}, \mu^b_{j_d}, \sigma^b_{j_d}) \cdot G\left(\frac{x_{t_d} - \mu^s_{i_d}}{\sigma^s_{i_d}}\right) + g(x_{t_d}, \mu^s_{i_d}, \sigma^s_{i_d}) \cdot G\left(\frac{x_{t_d} - \mu^b_{j_d}}{\sigma^b_{j_d}}\right) \quad (4)$$

$$p(X|\lambda) = \prod_{t=1}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} w^s_i \cdot w^b_j \cdot \prod_{d=1}^{D} p(x_{t_d}|i,j,\lambda) \quad (5)$$

## 4. A NEW APPROACH TO SPEAKER CLUSTERING

The online speaker clustering algorithm presented in [6] has the drawback of not having all relevant data available when making its decision about which clusters to merge. A hierarchical method that first collects all speaker models can make the globally best choice rather than working only locally. It therefore is more powerful at the expense of having exponential runtime. However, each step in the hierarchical method consists of only two single activities, distance computation and merging. Merging clusters is rather simple because it mainly consists of copying the data, thus the distance computation yields most room for improvement. If one succeeds in significantly reducing the runtime of the distance computation, even hierarchical clustering can be feasible for applications where speed is required. The rest of this section discusses popular distance measure candidates as well as our approach to reduce the runtime of the distance computation.

### 4.1. Generalized Likelihood Ratio

The standard dissimilarity measure for speaker clustering is the GLR. Assuming that $X$ and $Y$ are two sets of speech feature vectors used to build up two speaker GMMs $\lambda_x$ and $\lambda_y$, GLR can be expressed as

$$d_{GLR}(\lambda^x, \lambda^y) = \log\left(\frac{L(X|\lambda^x) \cdot L(Y|\lambda^y)}{L(X \cup Y|\lambda^{x \cup y})}\right) \quad (6)$$

where $L(.)$ is the likelihood-function and $X \cup Y$ indicates the concatenation of both segments. Keeping in mind equations (2) and (5), it is obvious that this computation can take quite some time, since it involves nested loops over all feature vectors and mixtures as well as the training of the new model $\lambda_{x \cup y}$.

### 4.2. Cross Likelihood Ratio

The CLR is commonly used [4] if the GLR is regarded as computationally too expensive. Since it does not require a new model to be trained, it is faster but also less accurate than the GLR.

$$d_{CLR}(\lambda^x, \lambda^y) = \log\left(\frac{L(X|\lambda^x)}{L(X|\lambda^y)}\right) + \log\left(\frac{L(Y|\lambda^y)}{L(Y|\lambda^x)}\right) \quad (7)$$

### 4.3. Beigi/Maes/Sorensen Distance

Considering the runtime problems when using likelihood-based distance measures and the effort that has been made to build a speaker model, it is appealing if one could compare two models directly on the basis of their parameters. Since a GMM forms a pdf, the Kullback-Leibler (KL) distance between distributions comes to mind, but it cannot be computed directly on GMMs because they lack a closed form solution [7]. Beigi et al. addressed this problem in [8] by introducing a method that extends the KL distance (or any other measure) between single mixture components of GMMs to a distance between the entire models.

$$d_{KL}(\lambda^x_i, \lambda^y_j) = \frac{1}{2} \cdot (\vec{\mu}^y_j - \vec{\mu}^x_i)^T((\Sigma^y)^{-1} + (\Sigma^x)^{-1})(\vec{\mu}^y_j - \vec{\mu}^x_i) + \frac{1}{2} \cdot tr((\Sigma^x)^{-1}\Sigma^y + (\Sigma^y)^{-1}\Sigma^x - 2 \cdot I) \quad (8)$$

$$W^x_i = w^x_i \cdot \min_{j=1..N}\left(d_{KL}(\lambda^x_i, \lambda^y_j)\right) \quad (9)$$

$$W^y_j = w^y_j \cdot \min_{j=1..M}\left(d_{KL}(\lambda^y_j, \lambda^x_i)\right) \quad (10)$$

$$d_{BMS}(\lambda^x, \lambda^y) = \frac{\sum_{i=1}^{M} W^x_i + \sum_{j=1}^{N} W^y_j}{\sum_{i=1}^{M} w^x_i + \sum_{j=1}^{N} w^y_j} \quad (11)$$

Here, $\Sigma$ is the (full) covariance matrix, $I$ is the identity matrix and $tr()$ is the trace function; $w^x_i$ is the weight of the $i$th mixture $\lambda^x_i$ in a GMM $\lambda^x$. The Beigi/Maes/Sorensen (BMS) distance is fast and accurate and allows the comparison of GMMs with different sizes. Its major drawback is that it is not freely available for commercial applications due to patent protection rights.

### 4.4. Earth Mover's Distance

The EMD has been introduced by Rubner et al. [1] as a metric for image retrieval. It is defined between collections of distributions called "signatures": $S = \{w_i, c_i\}, i = 1..M$. Here, $w_i$ is the weight of the centroid $c_i$, which can be any vector or set representing a cluster centroid. Loosely spoken, the EMD measures the amount of work needed to transport one element of mass from one distribution (regarded as a "hill") to the other (regarded as a "hole"). This explanation and the perceptually meaningful results in its original domain inspired many authors to adopt the EMD for their problem. Among others, the EMD has been applied successfully to the tasks of music similarity computation [9] and phoneme matching [10]. To compute the EMD, the optimal flow $F = (f_{ij})$ of mass from signatures $S^x$ to $S^y$ has to be found according to the following rules:

$$f_{ij} \geq 0, \qquad i = 1..M, j = 1..N \quad (12)$$

$$\sum_{j=1}^{N} f_{ij} \leq w^x_i, \qquad i = 1..M \quad (13)$$

$$\sum_{i=1}^{M} f_{ij} \leq w^y_j, \qquad j = 1..N \quad (14)$$

$$\sum_{i=1}^{M}\sum_{j=1}^{N} f_{ij} = \min\left(\sum_{i=1}^{M} w^x_i, \sum_{j=1}^{N} w^y_j\right) \quad (15)$$

$$F = \arg\min_{F}\left(\sum_{i=1}^{M}\sum_{j=1}^{N} d_{ij} \cdot f_{ij}\right) \quad (16)$$

The optimal flow is the one minimizing the amount of work (represented by the argument in equation (16)) to be done according

to the ground distance matrix $D = (d_{ij})$, which has to be computed before. Once the flow is found using the transportation-simplex method, the EMD between to signatures $S^x$ and $S^y$ is given by

$$d_{EMD}(S^x, S^y) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} d_{ij} \cdot f_{ij}}{\sum_{i=1}^{M}\sum_{j=1}^{N} f_{ij}} \quad (17)$$

Like the BMS distance, the EMD is able to compare signatures of differing size. If the overall mass of both signatures is identical, the EMD is a true metric. Furthermore, every metric between two Gaussians can be used as the ground distance. We use the KL distance because it showed superior results compared to the Euclidean or Mahalanobis distance in preliminary experiments.

### 4.5. Using the EMD for Speaker Clustering

By regarding each mixture component of a GMM as a cluster centroid and the mixture's weight as this cluster's mass, it is straightforward to put it in signature form and compute an EMD between two GMMs. However, problems arise when applying this simple rule to the MIXMAX model: its advantage of masking noisy mixtures is not fully represented in the model's parameters alone, but mainly arises from the method of likelihood computation via equations (4) and (5). We propose the following method to mimic this noise masking process during the EMD computation:

Equation (4) is the probability that the $d$th component of the current observation, $x_{t_d}$, is modelled by speaker model mixture $i$ and background model mixture $j$. Equation (18) now gives the probability that this current observation in the current state $\{i, j\}$ is equal to the unobservable, uncorrupted clean speech sample component $s_{t_d}$, i.e. that it is noise-free:

$$p(x_{t_d} = s_{t_d}|i, j, \lambda) = \frac{g(x_{t_d}, \mu^s_{i_d}, \sigma^s_{i_d}) \cdot G\left(\frac{x_{t_d} - \mu^b_{j_d}}{\sigma^b_{j_d}}\right)}{p(x_{t_d}|i, j, \lambda)} \quad (18)$$

We now extend the parameters of the speaker GMM by a vector $\vec{m} = (m_1..m_M)$ that we call the mask level:

$$m_i = \frac{\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{d=1}^{D} 1 - p(x_{t_d} = s_{t_d}|i, j, \lambda)}{T \cdot N \cdot D} \quad (19)$$

The mask level is computed during model estimation while the feature vectors are still available. A level of 0 for a mixture $i$ means that this mixture is noise-free while a level of 1 means that it is fully corrupted by noise. Before EMD (or BMS distance) computation, we multiply each speaker model mixture's weight with the factor $1 - m_i$. In this way, the more a mixture is masked by noise, the less it contributes to the final distance.

## 5. EXPERIMENTAL RESULTS

### 5.1. Test Corpus

We use a subset of the MPEG-7 video content set [11] for our performance evaluation, namely the Portugese night journal video 'jornaldanoite1'. The 'jornaldanoite1' video includes some difficulties for a standard speaker clustering system, particularly many interviews (ca. 50% of the overall time) under non-ideal outdoor conditions, leading to a relatively low SNR.

To study the effect of additive noise on our algorithms, we also conducted experiments with a short German news video called 'news-2' and its derivatives, which have been mixed with differing types of coloured noise in some scenes. Detailed information about all videos can be found in Table 1.

| Video | Length [s] | øSNR [dB] | min SNR [dB] | max SNR [dB] |
|---|---|---|---|---|
| news2 (0) | 244 | 13.6 | 6.3 | 19.7 |
| news2 (1) | 244 | 12.9 | 6.3 | 19.7 |
| news2 (2) | 244 | 11.15 | 6.3 | 16.4 |
| news2 (3) | 244 | 12.4 | 6.3 | 19.7 |
| news2 (4) | 244 | 10.27 | 6.3 | 15.08 |
| news2 (5) | 244 | 12.4 | 6.3 | 19.7 |
| news2 (6) | 244 | 9.97 | 6.3 | 14.2 |
| news2 (7) | 244 | 12.68 | 6.3 | 19.7 |
| news2 (8) | 244 | 12.09 | 6.3 | 19.7 |
| news2 (9) | 244 | 10.81 | 3.7 | 18.06 |
| jornaldanoite1 | 2855 | 7.99 | 0.67 | 26.61 |

**Table 1**. Overview of the used corpus. The video 'news2' has been used in different versions: (0) is the original version, (1)-(9) have been partly augmented with different amounts of of coloured noise: Low, medium and high brown noise, pink noise and white/brown noise, respectively. The SNR values are per segment.

### 5.2. Speaker Clustering Framework

Our speaker clustering system operates on the 16kHz/16bit audio track of each video. The audio track is high- and lowpassed to fit into the frequency range of 50-7000Hz, then preemphasized with a factor of 0.97. We segment it into 32ms long frames with 16ms overlap and use a 512 point FFT to convert each frame into one of the following feature vector types: 20 MFCCs (mel frequency cepstral coefficients) for GMM modelling or 24 log filterbank energies for MIXMAX modelling. The frequency scale for the filterbank is ExpoLog [12] in both cases. We discovered that the typical termination criteria for hierarchical clustering proposed in the literature, BIC and WCD [6], constantly overestimated the number of speakers in our case by far. We therefore rely on groundtruth data to terminate clustering at the right point as well as to make the speech/non-speech decision for each frame and to detect speaker changes. Silence and unvoiced speech are removed using an enhanced version of the adaptive silence detector proposed in [13].

### 5.3. Evaluation

We evaluated the following three performance criteria:

- Time: The elapsed time for the entire process from feature extraction to speaker clustering.
- Recall: $100 \cdot \frac{\#correctsegments}{\#availablesegments}$
- Precision: $100 \cdot \frac{\#fittingsegments}{\#clusteredsegments}$

Here, a segment of speech is an area of continous speech interrupted by less than 75ms of non-speech. The number of available segments is the count of segments long enough to be analyzed (min. 1 second of length). Segments are regarded as *fitting* if they belong to any cluster in which segments of their speaker are in the majority. Segments are regarded as *correct*, if they are fitting and belong to the cluster containing the most segments of this speaker. Clustered segments are those which are included in any cluster.

The experiments on the 'news2' derivatives were conducted to investigate to which extent the degradation of the SNR influences the clustering performance. With a SNR level of min. 12dB per scene, our system was able to reach 100% recall and precision using the MIXMAX model and any distance measure. With a SNR

lower than 10.5dB, the performance dropped heavily because segments are clustered according to background noise rather than according to voice.

The experiments with the longer 'jornaldanoite1' video focussed on speed. The results can be found in Table 2, where the time column indicates the measured wallclock time on an Intel 1.8 GHz Pentium 4 PC with 512 MB memory running Windows/XP and an implementation written in C++. In the case of the GMM, the EMD shows the best overall recall/precision pair while being only negligibly slower than our implementation of the BMS distance according to [8]. This can be due to the fact that we use Rubner et al.'s [1] reference implementation of the EMD and need to copy all data into the authors' format prior to distance computation. The runtime of the EMD is about a factor of 61 faster than the runtime of the GLR. When MIXMAX speaker modelling is used, CLR and EMD are at nearly the same performance level considering both recall and precision, only outperformed by the GLR with a 0.87% and 6.25% better recall and precision, respectively, compared to the EMD. The BMS distance performs worst, being 10.26% and 9.97% below the EMD in terms of recall and precision, respectively.

This indicates that the clustering performance of the EMD is only slightly worse than that of the GLR, but the speed differences are significant: the EMD is 61 times faster than the GLR in case of the GMM, and even 124 times faster in case of the MIXMAX model. This difference arises from the fact that computing a GLR using a MIXMAX model is much more expensive than for a simple GMM, but computing an EMD is very much the same for both (the runtime difference between the GMM/EMD and MIXMAX/EMD approach arises only from a longer time for model training in the MIXMAX case). Of course, we still use a hierarchical clustering approach with a runtime rising exponentially with the number of processed segments (for the 'news2' derivatives, the factor is only 3.7). But in practice, waiting 10 minutes for the results of our MIXMAX/EMD approach compared to more than 20 hours in case of MIXMAX/GLR is a quite significant improvement.

| Method | Time [s] | Recall [%] | Precision [%] |
|---|---|---|---|
| GMM/GLR | 5686 | 68.0 | 91.07 |
| GMM/CLR | 2989 | 67.65 | 90.54 |
| GMM/BMS | 90 | 72.52 | 84.82 |
| GMM/EMD | 93 | 74.09 | 85.89 |
| MIXMAX/GLR | 74218 | 74.78 | 92.68 |
| MIXMAX/CLR | 16169 | 69.91 | 90.54 |
| MIXMAX/BMS | 560 | 63.65 | 76.76 |
| MIXMAX/EMD | 598 | 73.91 | 86.43 |

**Table 2**. Experimental results on the 'jornaldanoite1' video with SNR-values as stated in Table 1.

## 6. CONCLUSIONS

In this paper, we proposed an approach to accelerate hierarchical speaker clustering by using different distance measures. We compared commonly known likelihood based measures (GLR, CLR) with methods which directly operate on the speaker model's parameters. For this reason, the earth mover's distance (EMD) has been applied to speaker distance computation for the first time. We also developed a method to profit from the MIXMAX noise modelling scheme even when using the EMD. Our results then showed an increase in clustering speed by a factor of 120 on a 47 minute test video.

There are several aspects for future work. For example, we observed that the commonly used clustering termination criteria failed on our data. Furthermore, there is still room for greater noise robustness. Finally, the speed problem can be further addressed by combining the online- and hierarchical clustering schemes to take advantage of both their strenghts.

## 7. ACKNOWLEDGEMENT

This work is financially supported by the Deutsche Forschungsgemeinschaft (SFB/FK 615, Teilprojekt MT).

## 8. REFERENCES

[1] Y. Rubner, C. Tomasi, and L.J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *International Journal of Computer Vision*, vol. 40, pp. 99–121, 2000.

[2] R.C. Rose, E.M. Hofstetter, and D.A. Reynolds, "Integrated Models of Signal and Background with Application to Speaker Identification in Noise," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 245–258, 1994.

[3] H. Jin, F. Kubala, and R. Schwartz, "Automatic Speaker Clustering," in *Proc. of the DARPA Speech Recognition Workshop*, 1997, pp. 108–111.

[4] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering Speakers by Their Voices," in *IEEE Proc. of ICASSP*, 1998, pp. 757–760.

[5] J. Ajmera and C. Wooters, "A Robust Speaker Clustering Algorithm," in *IEEE ASRU Workshop*, 2003, pp. 411–416.

[6] D. Liu and F. Kubala, "Online speaker clustering," in *IEEE Proc. of ICASSP*, 2004, vol. 1, pp. 333–336.

[7] J. Goldberger and H. Aronowitz, "A Distance Measure Between GMMs Based on the Unscented Transform and its Application to Speaker Recognition," in *Proc. of Interspeech*, 2005, pp. 1985–1989.

[8] H.S.M. Beigi, S.H. Maes, and J.S. Sorensen, "A Distance Measure Between Collections of Distributions and its Application to Speaker Recognition," in *IEEE Proc. of ICASSP*, 1998, vol. 2, pp. 753–756.

[9] S. Baumann, *Artifcial Listening Systems: Modellierung und Approximation der individuellen Perzeption von Musikähnlichkeit*, Ph.D. thesis, Technical University of Kaiserslautern, Germany, 2005.

[10] N. Srinivasamurthy and S. Narayanan, "Language-Adaptive Persian Speech Recognition," in *Proc. of EUROSPEECH*, 2003, pp. 3137–3140.

[11] MPEG-7 Requirement Group, "Description of MPEG-7 Content Set," *ISO/IEC JTC1/SC29/WG11/N2467*, 1998.

[12] S.E. Bou-Ghazale and J.H.L. Hansen, "A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress," *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 429–442, 2000.

[13] Y. Li, S.S. Narayanan, and C.C.J. Kuo, "Content-Based Movie Analysis and Indexing Based on Audiovisual Cues," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, pp. 1073–1085, 2004.