UvA-DARE (Digital Academic Repository)

Similarity visualization for the grouping of forensic speech recordings

Weiland, K.A.; Bouten, J.S.; Veenman, C.J.

DOI

[10.1007/978-3-540-85303-9_16](https://doi.org/10.1007/978-3-540-85303-9_16)

Publication date

2008

Document Version

Submitted manuscript

Published in

Computational Forensics

[Link to publication](#)

Citation for published version (APA):

Weiland, K. A., Bouten, J. S., & Veenman, C. J. (2008). Similarity visualization for the grouping of forensic speech recordings. In S. N. Srihari, & K. Franke (Eds.), *Computational Forensics: Second International Workshop, IWCF 2008, Washington, DC, USA, August 7-8, 2008 : proceedings* (pp. 169-180). (Lecture Notes in Computer Science; Vol. 5158). Springer. https://doi.org/10.1007/978-3-540-85303-9_16

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Similarity Visualization for the Grouping of Forensic Speech Recordings

Klara A. Weiland¹, Jos S. Bouten², and Cor J. Veenman^{1,2}

¹ Intelligent Systems Lab,

University of Amsterdam, Amsterdam, The Netherlands

² Digital Technology & Biometrics Department,

Netherlands Forensic Institute, The Hague, The Netherlands

Abstract. In a forensic phone wiretapping investigation, a major problem is to get the full picture of the speakers involved. Typically, the wiretapped speech recordings are grouped using a clustering tool. The main disadvantage of such an approach is that in a bootstrapped scenario grouping errors accumulate. In this paper, we propose a visual approach to find similar speech recordings that probably stem from the same speaker. We first model the speech recordings and define suitable similarity measures between recordings. Then, through an approximate 2-D visualization of the inter-speech similarities the investigator can identify clear groups of recordings and recordings that are harder to differentiate. We did extensive experiments on phone data of 50 speakers with 2 recordings per speaker. We tested quality of the 2-D visualization in relation to original high dimensional similarities. It turned out that for the original high dimensional similarity measure the nearest recording is almost always the one from the same speaker. In the 2-D visualization, we achieved that on average for all speech recordings a recording of the same speaker is among the 10 nearest recordings.

1 Introduction

Forensic speaker recognition is required in several situations, such as during an investigation to get the social network around a target, or when a questioned recording is compared with a suspect in a case. Since voice is in part a behavioral characteristic, it is not very constant and easy to alter. Also, emotional distress, health and simply aging can alter the characteristics of a speaker's voice significantly. Further, low-quality recordings which introduce noise specific to the recording and transmission device add to this problem.

While automatic speaker recognition systems have been reported to outperform naïve human speaker recognition when the telephone handset is identical for all recordings and the data comes from the same recording session, automatic systems generally are far more sensitive to variation and noise than humans when conditions are mismatched[1][2][3]. This is attributed to the fact that humans use high-level cues while automatic speaker recognition systems operate at a lower, spectral level, where both kinds of variability have a bigger impact[3].

Though in forensics recordings made under good, constant conditions are rarely found, automatic forensic speaker recognition is certainly useful. First, neither of the

cited studies mentions using normalization techniques to remedy the session and channel effects, let alone state of the art techniques. While intra-speaker variability still constitutes a problem and is an important focus of research, automatic speaker recognition performance under adverse conditions has steadily improved in recent years[4][5] and can be expected to continue to do so. Secondly, automatic speaker recognition has many advantages over speaker recognition performed by a human expert. Due to human's sensitivity to language and reliance on high-level features, it is of importance for speaker recognition to be performed by native speakers. If the language is uncommon and its speakers in an area or country thus are more likely to be acquainted or related, it may not be possible to find an unbiased, willing person with enough expertise in speaker recognition to reliably perform the task. Automatic speaker recognition systems are mostly insensitive to problems stemming from language choice. Additionally, automatic speaker recognition systems can process the large amounts of data that often come with forensic casework and where time might be tight.

Whether the speaker recognition is done by an automatic system or by human expert, recognition errors have to be accounted for. That is, in practical forensic situations perfect recognition is an illusion. For evidence evaluation, i.e. when a suspects voice is compared to a questioned recording, the use of a matching score – the likelihood ratio (e.g. [6]) – is therefore becoming more and more accepted. Also in investigation cases, however, the reality of imperfect recognition has to be recognized. Traditional speaker recognition systems work independently of human interaction, taking a number of speech files as an input and giving the judgment on the identities – often augmented with confidence scores – as an output. The current systems do not aim at providing insight into the decision process and do not make the overall similarity relations accessible to human interpretation in an intuitive way. However, as described, human and automatic speaker recognition are complementary in their characteristics. It is therefore desirable to combine the human and automatic speaker recognition process and their respective strengths into a system where human-computer cooperation can lead to mutual improvement. Some approaches to integrating human interaction in automatic speaker recognition exist, but the human contribution here consists merely in feature preselection and the systems require the user to be expert phoneticians [7].

In this paper, we deal with the explicit modeling of uncertainties in the recognition of matching recordings. We propose a visual approach to find similar speech recordings that probably stem from the same speaker. We first model the speech recordings and define a suitable similarity measure between recordings. Through an approximate 2-D visualization of the inter-speech similarities, the investigator can identify clear groups of recordings and recordings that are harder to assign. In the next section, we first describe the problem statement we deal with in this research. In the following section, we describe the components that our system is built up from. In the Section Experiments, we report the elaborate experiments we did to test the plain speaker recognition performance and the effectiveness of the visualization.

2 Problem Statement

The problem we deal with is the presentation of similarities between speech recordings on a 2-D display. The application scenario is a wiretapping investigation over the phone

with up to 50 speakers. The goal is to present the similarities between speech recordings such that: if there is more than one recording of a speaker, then for all these recordings the nearest recording on the 2-D display is from the same speaker. If there is only one recording of a speaker, then no constraints apply.

3 System Components

In this section, we describe the different components of our system. The system starts from a number of audio files, each of which contains speech from only one speaker, and displays the similarity relationships between the speakers in the individual audio files in two dimensions.

At first, the speech signal is segmented into overlapping slices known as frames. We use a common frame length of 30ms with 25% overlap between frames. Then, the frames are classified as being voiced or unvoiced. For the voiced sections we compute appropriate features and learn speech models. In order to compare speech models we need a mechanism to score the similarity between learned models. Finally, all similarity scores are put in a similarity matrix, which will be transformed to several 2-D visualizations. In the next subsection we describe these steps into some detail.

3.1 Extraction of Voiced Sections

Several studies that researched which parts of the speech signal contribute to good performance in speaker recognition systems found that using only the voiced sections yielded better results than using the complete signal or only removing the silent parts [8], [9] and thus the extraction of voiced sections is common in many speaker recognition systems. In our system, we use the combination of short time energy, autocorrelation, and zero crossing rate to establish that a frame is voiced. If all of these measures indicates that a frame is voiced it is flagged as voiced.

3.2 Speech Modeling

For the representation of the speech signals in the segmented frames, we use the Mel-Frequency Cepstral Coefficients (MFCC) [10][11], that are widely used in current speaker and speech recognition systems [12]. To compensate for channel and session effects, Cepstral Mean Subtraction, (CMS)[13][14][15], is applied on the feature matrix by subtracting the mean of each feature dimension from all feature values in that dimension.

For the modeling of the speech recordings, we abstract the set of 18-D MFCC vectors by estimating their 18-D distribution. Commonly, Gaussian Mixture Modeling (GMM) is used for this purpose. In this study we compare two different methods to estimate the Gaussian mixtures. First, we use Expectation Maximization (EM) [16] to optimise the weights, means and covariances of the Gaussians in the mixture. Second, we estimate the mixture by Maximum A Posteriori (MAP) adaptation [17] starting from an EM-trained background GMM, also called an Universal Background Model (UBM) [18].

Since in our case, the MAP adapted speaker models only differ in their means but have identical covariance matrices and weights, the mean vectors are sufficient to fully

describe the speaker and feature-based classification can be applied. A concatenation of all mean vectors in one GMM yields a so-called supervector [19]. For example, if a 128-mixture GMM is adapted using 18-dimensional MFCC data, a supervector with 2304 dimensions results. In principle, any feature based classifier can be applied to learn a speaker classifier. The Support Vector Machine (SVM) has been applied successfully in this domain [19], where the following linear kernel is commonly used:

$$K(x_1, x_2) = \sum_{i=1}^N (\sqrt{w_i} \Sigma^{-\frac{1}{2}} x_i^1)' \cdot (\sqrt{w_i} \Sigma^{-\frac{1}{2}} x_i^2) \quad (1)$$

where w and Σ are the respective weights and covariances of the GMM mixture component. In this case, the diagonal scaling of the means can also be computed on the supervectors beforehand, which means that a regular linear kernel can be used.

An SVM classification model for a certain target speaker is then learned by representing the target class (speaker) through one supervector and a cohort speaker supervectors, which resembles the background model.

3.3 Speech Similarity Measure

Once we have the models of the individual speakers, we need a measure to determine how similar or dissimilar the models and their original data vectors (MFCC) are from each other.

Similarities for SVMs The distance between a trained SVM classifier and a supervector is determined by simply using the classifier on the supervector. The resulting value, the distance from the boundary, indicates the relative fit for both classes on a continuous scale, thus yielding a number that quantifies how close to the speaker on which the SVM was trained the new data lies relative to the distance of the cohort.

Similarities for GMMs The GMM distributions could be compared directly, but measures for the distance between distributions like the Bhattacharyya distance or the chi square metric often impose constraints on the characteristics of the distributions or are expensive to compute. Alternatively, the likelihood expresses the relation between the GMM created from one recording and the data vectors created from another speech recording. Since the GMM components are multiple ordinary normal distributions, the likelihood can be computed for each mixture component on the basis of the general Gaussian likelihood function using the respective means and covariance matrices. These values are summed and, to represent the contribution of each Gaussian to the GMM correctly, weighted, leading to the following definition for the likelihood function for GMMs[16]:

$$p(\vec{x}|\lambda) = \sum_{i=1}^M w_i p(\vec{x}|\lambda_i). \quad (2)$$

The likelihood of a GMM is thus simply a weighted linear combination of the densities of its component distributions.

Note that above equation will yield one value per MFCC vector. Each MFCC vector represents only a fraction of a second of speech and consequently there will be a big number of likelihood values that has to be combined into a single score. Since the

lengths of the MFCC vectors vary depending on the length of the voiced sections in the original utterance, the mean of the log likelihood values is used:

$$\log(p(X|\lambda)) = \frac{1}{T} \sum_t \log(p(\vec{x}|\lambda)) \quad (3)$$

An alternative used in speaker recognition is the Bayesian Information Criterion (BIC)[20]. It is based on the log likelihood score, but additionally penalises the model complexity. Since in our case the complexity of the model is fixed and the number of MFCC vectors per recording is similar, the added value of the BIC score is limited. We do not consider it further in this study.

The likelihood score as described lacks one important characteristic compared to the SVM scores: They are not normalized with respect to the cohort or background model that represents the average speaker. This means that the score only represents the relative distance between a model and some data, but this score does not relate to how likely an average, nondescript speaker is to produce an utterance that leads to the same feature vectors and therefore how significant the original score is.

This problem is easily remedied by redefining the score as a ratio of likelihoods scores[17]:

$$ll_{ratio} = \frac{\log(p(X|\lambda))}{\log(p(X|\lambda_{background}))} \quad (4)$$

The likelihood ratio thus expresses the magnitude of the score relative to the score of the background speaker, making scores comparable across the evaluation of the same data for different models.

Symmetric Similarities Ideally, both described similarity measures are symmetric. That is, given two sets of data (MFCC) vectors X_1 and X_2 , the similarity $s(X_1, X_2) = s(X_2, X_1)$, where $s(X_i, X_j)$ is computed through SVM or GMM modeling. Because the data per speaker is limited and the modeling is imperfect, in practice the similarity is not symmetric. Since this is undesirable for the visualization, we make the similarity symmetric as follows:

$$s'(X_1, X_2) = s(X_1, X_2) + s(X_2, X_1) \quad (5)$$

3.4 Visualization

For the visualization of the similarities between speaker models or between models and data vectors from recordings, we need to transform the original similarities to similarities or distances in 2-D. In the following, two dimension reduction algorithms used in the experiments, Multidimensional Scaling[21] and Isomap[22] will be described. These non-linear dimension reduction techniques were chosen, since on the one hand it is an established technique (MDS) and on the other hand a more recent related technique (Isomap). The similarities between the two dimension reduction algorithms will be explained in the following.

Multidimensional Scaling Multidimensional Scaling is an established set of techniques that find wide application in various scientific fields like Psychology and Economy. It takes a distance matrix and tries to find a configuration of points in lower dimensional space whose Euclidean distances match the distances in the input matrix.

There are multiple variants of MDS, here a nonlinear variant using a Sammon mapping which is performed on a symmetric distance matrix with zero diagonal is described and evaluated.

The MDS procedure consists in the minimization of an error function, the stress. We consider the original metric MDS with the following optimization or stress function:

$$Stress = \sum_{i=1}^N [d(x_1, x_2) - d'(x_1, x_2)]^2 \quad (6)$$

where $d(x_1, x_2)$ is the distance between x_1 and x_2 in the original space and $d'(x_1, x_2)$ is their distance in the mapped space.

Isomap The Isomap algorithm [22] extends MDS by first calculating the geodesic distances in the high-dimensional space and then applying metric MDS on the resulting distance matrix. The idea behind the nonlinear dimension reduction technique Isomap is to find coordinates in lower dimensional space that preserve the intrinsic geometry of the data by finding the underlying manifold, a topological space that is locally but not globally Euclidean. Due to this property of the manifold, the distance between adjacent points is calculated using the Euclidean distance, while the distance between points that are far away from each other, the geodesic distance is calculated. This is done by first constructing a graph in which nearby points are connected via edges, and then calculating the distance between the two points along the shortest path between them.

Fig. 1. Geodesic and Euclidean distance illustrated

As an illustration, consider the dataset in figure 1 where the circles indicate the data points. The direct distance between the two red points, indicated by the red line, ignores the underlying spiral form of the data and the Euclidean distance calculated between the points. The geodesic distance on the other hand can be approximated by building a graph structure where each point is connected to its nearest neighbor and then summing the distances between the edges of the shortest path, here indicated by the magenta line. The geodesic distance here is the sum of the Euclidean distances between the magenta circles and the first and last red circle.

4 Experiments

To test the speech modeling and the visualization that is based on it, we did extensive experiments. Below, we first describe the dataset used and in the following section the results.

4.1 Dataset

In absence of true forensic wiretapped recordings, all tests were run on data from the SwitchBoard-2 phase 1 corpus³ which is frequently used in the NIST speaker recognition task. Each file had speech from only one speaker, meaning that speaker segmentation was not needed, and contained a concatenation of consecutive turns recorded during a telephone conversation.

The phone conversations were collected by the Linguistic Data Consortium, the speakers are all US Americans, mostly from the Mid-Atlantic area of the US and college students. Conversations lasted about 5 minutes in total, a proposed topic for each conversation was given but not enforced. As is typical for telephone data, the audio was encoded at a sampling rate of 8KHz and 8 bits per sample. All files were initially about 60 seconds in duration, of which roughly a third was discarded when removing the unvoiced sections, silence and noise.

There were three segments per speaker, two of them from the same session. The third segment was recorded during a separate sessions but the same handsets were used. We do not report the mixed session results, because of space limitations. Also, for the visualizations there were no systematic differences between the methods with the same session results that we report here.

172 files, 86 female and 86 male recordings, were separated from the rest of the files and used as data for the computation of the cohorts and background models which were thus balanced in terms of gender[17].

Performance measure The wiretapping scenario, for which our system is designed, typically has up to 50 speakers per case. Therefore, we randomly drew the same session training and test recordings of groups of 50 speakers from the 162. We repeated this 50 times and averaged the results to become less sensitive to a specific sample of 50 speakers. Moreover, random effects in the GMM estimation and MDS optimization are averaged out too.

To quantify how well the recordings of the same speaker are found as nearest recording by the system (as required in the Problem Statement), we propose the nearest neighbor score. This is a score that counts for how many recordings the recording of the same speaker is among the k -nearest recordings. In other words, the nearest neighbor score (NN-score) is a function of k . In our setting, there are in total models of 100 recordings. Then, at $k = 100$ the score becomes always 1. Further, we define the Nearest Neighbor Hit-score (NNH-score) as the k value for which the NN-score reaches 1.

³ <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC98S75>

4.2 Results

For each type of model, one distance matrix for the best performing number of mixtures and for each session condition was selected and the dimension reduction algorithms were applied. For the EM-GMM, 32 mixture components were chosen, while for the SVM and MAP-GMM, distance matrices created from 256 mixture component models were selected.

Table 1. Table with Nearest Neighbor Hit-scores in original high-dimensional space and after dimension reduction

Method	High dimensional	MDS	Isomap			
			3	7	10	15
EM-GMM	1.00	23.50	12.36	18.06	11.34	18.28
MAP-GMM	1.00	46.26	8.80	12.46	34.74	13.52
SVM	1.00	23.14	6.60	49.32	17.74	18.32

Besides the target number of dimensions, the MDS algorithm has no parameters. Isomap on the other hand additionally has as parameter the number of neighbors for the underlying nearest neighbor graph. In the experiments we set this parameter to 3, 7, 10 and 15 neighbors, respectively. Moreover, as the MDS and Isomap models are optimized nondeterministically, every experiment was run 50 times and the result of the run with the lowest overall remaining stress was chosen.

Table 1 gives the mean NNH-scores (averaged over 50 samples) for the models in the original high-dimensional space and after dimension reduction through MDS and Isomap. The table clearly shows that for same session conditions, the system has close to perfect average performance. Also, MDS is outperformed by Isomap for all modelings. The neighborhood size for Isomap has an effect on performance, but there is no clear relation. The type of model appears to have no systematic influence on the performance of the different dimension reduction techniques. Isomap(3) has the best performance for MAP-GMM and SVM modeling. Also for EM-GMM this dimension reduction scheme is close to the best.

Visualizations For each model type, 2-D plots representing the speaker distances after dimension reduction with each of the techniques were created. In the case of Isomap, the neighborhood size yielding the lowest NNH-score was used.

Two types of plots were created, one displaying each speech file as a symbol connected to the speaker's other speech file in the set by a line. The other plot type shows the name of the speech files with the prefix, 'M' or 'F', indicating a male and female speaker respectively, and where identical numbers mean that two files came from the same speaker. The plots in figure 2(a) show the MDS visualization where the speaker models are 32-mixture EM-estimated GMMs.

Remarkably, gender classification emerges although all steps of processing are gender-neutral or balanced in gender. That is, the recordings of male and female speakers

(a) MDS

(b) Isomap (10 neighbors)

Fig. 2. Visualization of the speaker similarities using EM-GMM modeling are clearly separated in the 2-D plot, while the gender of the speakers is not disclosed at any stage in the system.

Figure 2(b) shows the Isomap visualizations. These visualizations show many cases where matching models are very close together, but also a number of cases where files from the same speaker have a relatively large distance between them. All in all, the distances in this plot are more varied than those in the MDS plot.

Figure 3(a) and 3(b) show the speaker spaces resulting from the application of the two dimension reduction techniques to distance matrices created from 256-mixture MAP-adapted GMMs. In the MDS plot, again, apart from a small number of outliers, distances between matching models are small and relatively similar. As the label plot shows, speakers here are not automatically arranged according to gender, but rather in several groups with the same gender.

As in the other two models, the MDS visualization (figure 4(a)) based on distances determined by SVM classification shows relatively small, constant distances between matching files with few big outliers in terms of distance. The models again are separated by gender.

(a) MDS

(b) Isomap (3 neighbors)

Fig. 3. Visualization of the speaker similarities using MAP-GMM modeling

5 Conclusion

In this paper, we proposed a system for visualizing the similarities between forensic speech recordings. The system explicitly models and shows the uncertainty in the speaker recognition process, so that the user can decide on the basis of the visualisation which recordings stem from the same speaker. To this end, the user may apply additional knowledge.

In order to be able to visualize speech similarities we built several speech models that are based on Gaussian Mixtures and defined similarity measures between models. Finally, we applied two dimension reduction techniques for visualization; Multi Dimensional Scaling (MDS) and Isomap, which is based on a nearest neighbor graph.

We conducted extensive experiments using the SwitchBoard-2 phase 1 corpus, which showed that the speaker recognition performance in same session conditions was close to perfect. Though this is not realistic, such a base-line shows what performance we can obtain under good conditions. This then also holds for the dimension reduction and visualization. Although none of the dimension reduction techniques was able to represent the distances found in high-dimensional space, it was found that the right choice of dimension reduction technique leads to two-dimensional visualizations that display the speaker similarities with high accuracy. The MDS technique in general shows a few outliers for which the matching speech recording is far away. With the Isomap tech-

Fig. 4. Visualization of the speaker similarities using SVM modeling

nique this hardly happened, so from this study we conclude that Isomap is the most suitable. This also follows from the Nearest Neighbor Hit-scores that we introduced in this paper.

Interestingly, representations were arranged by gender in most visualizations, while the gender of the speakers was not disclosed to the system. This further indicates that the visualization display meaningful speaker similarities.

Finally, further tests are needed with more realistic data and involving forensic investigators in the interpretation of the visualizations.

References

1. Schmidt-Nielsen, A., Crystal, T.H.: Human vs. machine speaker identification with telephone speech. In: Proceedings of the 5th International Conference on Spoken Language Processing, Sydney, Australia (1998)
2. Ezzaidi, H., Rouat, J.: Speaker identification by computer and human evaluated on the SPIDRE corpus. *Canadian Acoustics* **28**(3) (2000)
3. Alexander, A., Dessimoz, D., Botti, F., Drygajlo, A.: Aural and automatic forensic speaker recognition in mismatched conditions. *Forensic Linguistics: The International Journal of Speech, Language and the Law* **12**(2) (2005)

4. Przybocki, M.A., Martin, A.F., Le, A.N.: NIST speaker recognition evaluation chronicles – part 2. In: Proceedings of the Odyssey 2006. Speaker and Language Recognition Workshop, San Juan, Puerto Rico (2006) 1–6
5. Martin, A.F., Przybocki, M.A.: The NIST speaker recognition evaluations: 1996-2001. In: Proceedings of the Odyssey 2001. Speaker and Language Recognition Workshop, Crete, Greece (2001)
6. Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., Ortega-Garcia, J.: Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language* **20** (2006) 331–355
7. Bimbot, F.B., Fredouille, J.F., Gravier, C., Magrin-Chagnolleau, G., Mcignier, I., Merlin, S., Ortega-Garcia, T., Petrovska-Delacretaz, J., Reynolds, D.A.: A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing* (2004) 430–451
8. Kim, J.K., Shin, D.S., Bac, M.J.: A study on the improvement of speaker recognition system by voiced detection. In: Proceedings of the 2002 45th Midwest Symposium on Circuits and Systems. Volume 3., Tulsa, USA (2002) 324 – 327
9. Krishnamachari, K., Yantorno, R.: Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions. In: Proceedings of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems, Honolulu, USA (2000)
10. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics Speech and Signal Processing* **28**(4) (1980)
11. Campbell, J.: 8: Speaker Recognition. In: Biometrics - Personal Identification in Networked Society. Springer US (2002)
12. Reynolds, D.A.: An overview of automatic speaker recognition technology. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, USA (2002) 4072 – 4075
13. Atal, B.: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America* **55**(6) (1974) 1304–1312
14. Furui, S.: Speaker Recognition. In: Survey of the State of the Art in Human Language Technology. Cambridge University Press (1998) 36 – 41
15. Johnsen, M.H., Svendsen, T., Harborg, E.: Experiments on cepstral mean subtraction (CMS) and rasta-filtering applied to SAMPA phoneme recognition. In: COST 249, Nancy, France (1995)
16. Reynolds, D.A., Rose, R.: Robust text independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing* **3**(1) (1995) 72 – 83
17. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian Mixture Models. *Digital Signal Processing* **10** (2000) 19 – 41
18. Barras, C., Gauvain, J.L.: Feature and score normalization for speaker verification of cellular data. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume 2., Hong Kong, China (2003) 49 – 52
19. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support Vector Machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters* **13**(5) (2006) 308 – 311
20. Schwartz, G.: Estimation of the dimension of a model. *Annals of Statistics* **6** (1978) 461 – 464
21. Torgerson, W.S.: Theory and methods of scaling. New York: John Wiley (1958)
22. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. *Science* **290**(5500) (2000) 2319 – 2323