QMOS: a robust visualization method for speaker dependencies with different microphones

Andreas Maier, Maria Schuster, Ulrich Eysholdt, Tino Haderlein, Tobias Cincarek, Stefan Steidl, Anton Batliner, Stefan Wenhardt, Elmar Nöth

Journal of Pattern Recognition Research 1 (2009) 32-51

Received Oct 17, 2008. Revised Apr 2, 2009. Accepted Apr 12, 2009.

QMOS - A Robust Visualization Method for Speaker Dependencies with Different Microphones

Andreas Maier, Maria Schuster, Ulrich Eysholdt, Tino Haderlein
Abteilung für Phoniatrie und Pädaudiologie, Universität Erlangen-Nürnberg Bohlenplatz 21, 91054 Erlangen, Germany *andreas.maier@cs.fau.de*

Tobias Cincarek
Speech and Acoustics Processing Laboratory, Nara Institute of Science and Technology 8916-5, Takayama-cho, Ikoma-shi, Nara, Japan

Stefan Steidl, Anton Batliner, Stefan Wenhardt, Elmar Nöth
Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg Martensstraße 3, 91058 Erlangen, Germany

Abstract

There are several methods to create visualizations of speech data. All of them, however, lack the ability to remove microphone-dependent distortions. We examined the use of Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and the COmprehensive Space Map of Objective Signal (COSMOS) method in this work. To solve the problem of lacking microphone independency of PCA, LDA, and COSMOS, we present two methods to reduce the influence of the recording conditions on the visualization. The first one is a rigid registration of maps created from identical speakers recorded under different conditions, i.e. different microphones and distances. The second method is an extension of the COSMOS method, which performs a non-rigid registration during the mapping procedure. As a measure for the quality of the visualization, we computed the mapping error which occurs during the dimension reduction and the grouping error as the average distance between the representations of the same speaker recorded by different microphones. The best linear method in leave-one-speaker-out evaluation is PCA plus rigid registration with a mapping error of 47 % and a grouping error of 18 %. The proposed method, however, surpasses this even further with a mapping error of 24 % and a grouping error which is close to zero.

Keywords: Speech intelligibility, speech and voice disorders, speech evaluation, dimensionality reduction, Sammon mapping, QMOS, COSMOS, COmprehensive Space Map of Objective Signal

1. Introduction

The facets of voices and speech are very complex. They comprise various stationary and dynamic properties such as frequency, energy, and even more complex structures such as prosody. In order to comprehend these characteristics, the high dimensionality of the speech properties has to be reduced. The visualization in two or three dimensions was shown to be very effective in many fields of application.

The visualization can reveal the relations between patients with voice disorders in different graduations [5]. Each of the patients reads a standard text. This speech material is then used for the visualization. Projection of new patients, i.e. speakers, allows to compare them to the other speakers. This gives a better understanding of the different disorders. Fig. 1 shows a map of speakers with different degrees of hoarseness. On the top left, speakers with a substitute voice are found. In these patients the larynx was removed due to cancer.

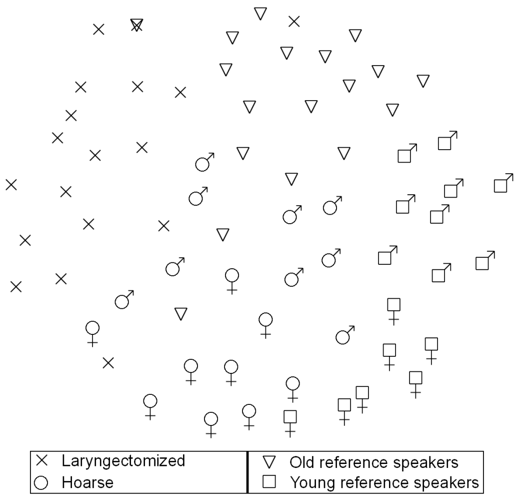


Fig. 1: Visualization of voice disorders: The properties of the speaker's voices are visible in the map. While the y-axis contains the age of the speaker, the x-axis can be interpreted as the degree of hoarseness of the speakers.

The artificial voice of the laryngectomized speakers can be interpreted as an extreme form of hoarseness. Average age of the laryngectomees is about 60 years. At the top right, an age-matched control group of normal speakers is located. At the bottom of the map are speakers with chronic hoarseness. On the bottom right, young reference speakers are found. Hence, the axes of the map can be roughly interpreted as the age on the y-axis and the degree of hoarseness on the x-axis. All data were gathered with the same microphone with the same recording setup. The map was computed with the COSMOS method as described in Section 2.3.3

For routine clinical use [13], however, this poses a serious problem. Modern Internet technologies allow for the recording of speech data at various locations simultaneously in multi-site studies [8]. This also means that all data are recorded in different conditions with different hardware and therewith varying quality of A/D conversion and microphones. Recording conditions also have a great influence. Major factors are the distance between the microphone and the speaker and the acoustical properties of the recording environment. Given a speaker who was recorded simultaneously by multiple microphones of different characteristics at different distances, the points representing the same speaker in the map are spread across the result of the visualization. Fig. 2 gives an extreme example: The data were recorded simultaneously with a close-talking microphone and a distant-talking microphone (distance of about 2.5 m, cf. Section 2). The room had strong reverberations which are only audible on the distant-talking microphone. The speakers form two clusters which is caused by the acoustic difference between the two microphones. The two corresponding representations of the same speaker are far away from each other in this visualization. The dominating factor is the microphone in this example.

In order to alleviate this problem we investigated several techniques (cf. Fig. 3). Data of the same speakers were recorded at two or more different locations or with different microphones. We want to reduce the dimensionality of the data in order to create a visualization. In a first step we take the audio data of each speaker and a speaker-independent speech recognizer. The speaker-dependent data is used to create a speaker-dependent speech rec-

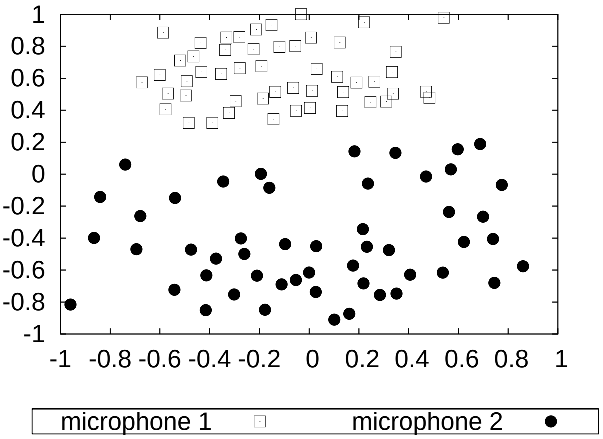


Fig. 2: 51 speakers recorded simultaneously with two different microphones (remote and close-talk recordings): The two microphones form two clusters although both clusters contain the same speakers.

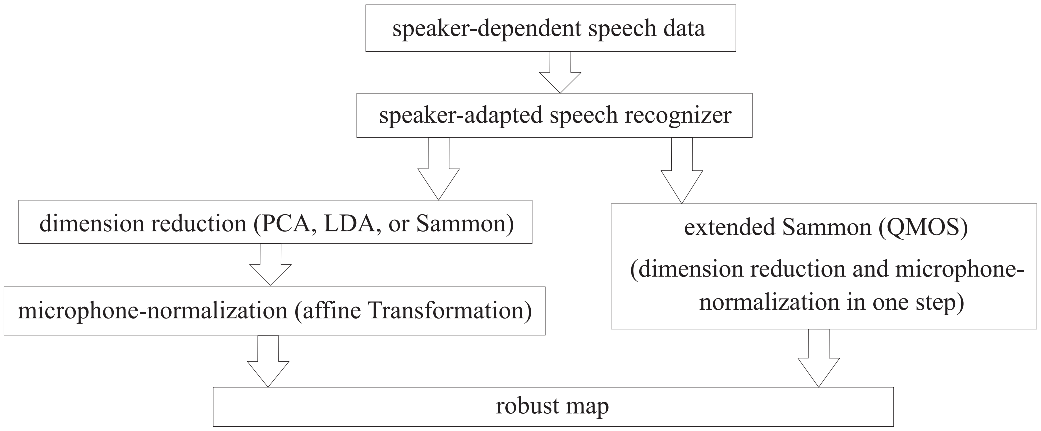

speaker-dependent speech data
speaker-adapted speech recognizer
dimension reduction (PCA, LDA, or Sammon)
microphone-normalization (affine Transformation)
extended Sammon (QMOS)
(dimension reduction and microphone-normalization in one step)
robust map

Fig. 3: Steps to create a visualization: first the audio data of each speaker has to be processed to create speaker specific features (here: the parameters of a speaker adapted speech recognizer). Then, the dimension has to be reduced. In the last step the microphone influence has to be reduced. QMOS performs dimension reduction and the elimination of the microphone-dependency in a single step.

ognizer. In this manner we get a time-invariant speaker-dependent model. The parameters of the probability density functions of the this model can be interpreted as an abstract representation of the speaker. However, this representation is still very high dimensional as we will see in the following. Hence, the dimensionality still has to be reduced. Therefore, we apply common dimension reduction techniques like the PCA, the LDA, or the Sammon mapping (cf. [15]).

Additionally, we want to reduce the influence of the recording conditions. In an ideal map a speaker should be projected onto the same position even if he was recorded with a different microphone or in a different room. In this manner one could collect one validated database of voice and speech disorders. Later on, if a new recording site also wants to compare its data with the validated database, a method is required to make the previously recorded data comparable to the new site. In the following we will investigate two approaches to realize this: The first one employs a linear transformation of the data points in order to

Table 1: The recording conditions used in this work

set acronym	environment
ct	close-talking
ctrv	close-talking artificially reverberated
rm	distant-talking

project corresponding ones as close to each other as possible. The second one extends the Sammon mapping by a grouping term which causes the same speakers to be projected as close to each other as possible i.e. it uses the prior knowledge about the group membership and punishes points belonging to the same speaker if they are apart from each other.

A last requirement of the map is that we want to project new speakers also into same coordinates, even if the new speaker was collected with only one of the incorporated microphones. Furthermore, this projection also has to be robust, i.e., if the speaker is collected again with one of the other microphones it should also be projected into the same location as projected before. This feature of the map will be tested in leave-one-speaker-out evaluation.

All methods were evaluated using a children database described in the next chapter. It is an appropriate database for the reduction of noise conditions since it consists of children speech recorded with a head-set microphone and a microphone of a video camera. A third recording condition was simulated using artificial reverberation as presented in [4].

2. Speech Data and Methods

The database used in this work contains speech of children. The children were at the age from 12 to 14 years [1]. In total 51 pupils (21 male and 30 female) of two different schools were recorded in the German language. The speech data were recorded by a video camera in order to document the experiment and a head-mounted microphone (UT 14/20 SHURE UHF). The close-talk version is referred to as *ct*. From the sound track of the video tape, a second version of the corpus was extracted: A distant-talk version (*rm*) was obtained. The distance between the speaker's position and the video camera was approximately 2.5 m. In total, 8.5 hours of spontaneous speech were recorded. The data were segmented automatically at long pauses into utterances of several words. The resulting utterances contain 3.5 words on average. This corpus with 12,858 utterances in total was split into a training, a validation, and a test set with 8,374, 1,310, and 3,174 utterances, respectively. The size of the vocabulary is 850 words plus 350 word fragments. A category-based 4-gram language model was trained on the transliteration of the training set and has a perplexity of 50 on the test set.

Artificial reverberation is used to create distortions which resemble those caused by reverberation in a real acoustic environment. It is applied to the close-talking signal directly before feature extraction. The idea is to convolve the speech signal with the impulse responses of an environment typical for the application, e.g. a living room. In this way, a reverberated signal can be computed. For this application, impulse responses were measured in a specific environment. In current research artificial reverberation was found to improve the robustness of speech recognizers against acoustic mismatches of training and test data [9, 4]. In both articles the training data was reverberated using the same twelve impulse responses from assumed speaker positions shown in Fig. 4. The responses differ in the distance, the angle to the microphone, and the reverberation time T_{60}. The reverber-

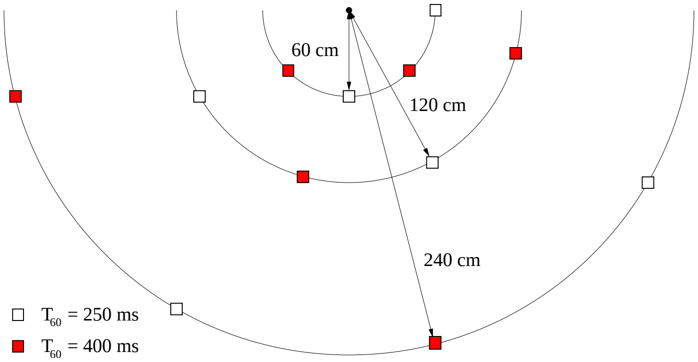


Fig. 4: Speaker positions for the impulse responses used in this work (black dot: microphone; squares: assumed speaker positions)

ation time is defined as the time that passes until the signal decays to 10^{-6} of its initial sound energy after the signal source has been switched off. This corresponds to a reduction of 60 dB.

Each response is applied to $\frac{1}{12}$ of the training data. So training data are created which cover a broad variety of possible reverberation. Additionally, the recordings from the close talk microphone were artificially reverberated to simulate another recording condition (*ctrv*). In this way, three speech recognizers are trained for each child. These recognizers are used for the creation of the visualizations. A signal-to-noise ratio cannot be computed between all versions of the corpus since the data is not always frame-matched. Table 1 lists the recording conditions available on the data.

To create a visualization the first step is to compute characteristic features from individual speech samples. Next, the dimensionality of those features has to be reduced to two or three (for 2-D or 3-D visualization). These features are obtained from Gaussian mixture densities of a speech recognizer which is adapted for each speaker [3, 9]. In this work we examined common methods for the dimensionality reduction: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and the Sammon mapping [15] (cf. Fig. 3). For both linear methods (PCA and LDA), the parameters of the Gaussian densities are directly transformed to the low-dimensional space. For the Sammon mapping, an appropriate distance measure has to be chosen. Shozakai et al. chose the Mahalanobis distance [6] between the Gaussian densities of the speech recognizer. The resulting method is called COmprehensive Space Map of Objective Signal (COSMOS) [10].

2.1 Features for Visualization

First, the 16 bit, 16 kHz speech signals are processed and 24 dimensional Mel-frequency cepstral coefficient (MFCC) feature vectors are extracted from them every 10 msec. The windows length is 256 samples, i.e., 16 msec. As features 12 static and 12 delta coefficients are computed. The first MFCC and its delta is replaced by the energy of the signal. Furthermore, cepstral mean subtraction (CMS) is applied.

In order to get a time-invariant representation of each speaker, we use the parameters of Gaussian mixture densities of a speech recognizer as feature vectors for the visualization. Those densities are adapted with Maximum Likelihood Linear Regression (MLLR) adaption [3] for use with a semi-continuous Hidden Markov Models (SCHMM) speech recognizer that shares all of its 500 Gaussian densities for all states of the acoustical polyphone[1] models [17]. Note, that this procedure is similar to the use of an universal background model [2]. The mixture density $f_\kappa(\boldsymbol{x})$ has the following form:

$$\begin{aligned} f_\kappa(\boldsymbol{x}) &= \sum_{\kappa=1}^{K} \alpha_{i\kappa}\mathcal{N}_{i\kappa}(\boldsymbol{x}) \quad \text{with} \\ \mathcal{N}_{i\kappa}(\boldsymbol{x}) &= \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}_{i\kappa}|^{1/2}} e^{-\dfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_{i\kappa}})^\top \boldsymbol{\Sigma}_{i\kappa}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_{i\kappa}})} \end{aligned} \tag{1}$$

where $\alpha_{i\kappa}$ is the weight for the Gaussian $\mathcal{N}_{i\kappa}(\boldsymbol{x})$, $\boldsymbol{\mu}_{i\kappa}$ the mean vector, $\boldsymbol{\Sigma}_{i\kappa}$ the covariance matrix of state i. M denotes the dimension. The sum over all α_κ equals 1.

For the linear dimensionality reduction methods (PCA and LDA), Gaussian mixture parameters can be directly interpreted as feature vectors. However, since each Gaussian has 600 parameters (24 in the mean vector and 576 in the covariance matrix) and there are 500 state-tied Gaussians, the total number of parameters is 300,000. Hence, the linear transformations are first applied to all Gaussians to reduce the number of parameters per Gaussian from 600 to two. In a second transformation step, the dimension is further reduced to two dimensions using the same type of linear transformation but trained for the reduction from 1000 parameters to two. This procedure is necessary since the storage requirements for a full 300,000 × 300,000 matrix as computed in a full transformation would be too high. Furthermore, the sheer number of parameters which would have to be estimated from just a few observations would also cause great numerical difficulties.

For the case of the Sammon method, the processing can be performed in a single step. However, a distance measure between each of the SCHMM recognizers has to be defined.

2.2 A Distance Metric for SCHMMs

For the computation of the distance between two SCHMMs m and n, there are several applicable metrics. In our case we use the Mahalanobis distance [6] as in [18]. Since all codebooks are adapted from the one original codebook by a linear transformation (MLLR) the correspondences between the distributions are known. To calculate the distance between two Gaussian mixtures, we use

$$d_i(m,n) = \sum_{\kappa=1}^{K} \sqrt{(\hat{\boldsymbol{\mu}}_{i\kappa}(m) - \hat{\boldsymbol{\mu}}_{i\kappa}(n))^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}}_{i\kappa}(m) - \hat{\boldsymbol{\mu}}_{i\kappa}(n))} \tag{2}$$

$$\hat{\boldsymbol{\mu}}_{i\kappa}(m) = \alpha_{i\kappa}(m)\boldsymbol{\mu}_{i\kappa}(m) \tag{3}$$

for mixtures which consist of K Gaussians, with weighted mean vectors $\hat{\boldsymbol{\mu}}_{i\kappa}(m)$ and $\hat{\boldsymbol{\mu}}_{i\kappa}(n)$. $\boldsymbol{\Sigma}$ is the mean covariance matrix of all Gaussians of both mixtures [6] and i is the state number.

[1] A polyphone is an acoustical model whose phonetic context may vary according to the amount of training data which was available at the time of the training. We constructed a polyphone for each phoneme sequence which could be observed at least 50 times in the training data [17]. Training data and the recognizer is described in detail in [19].

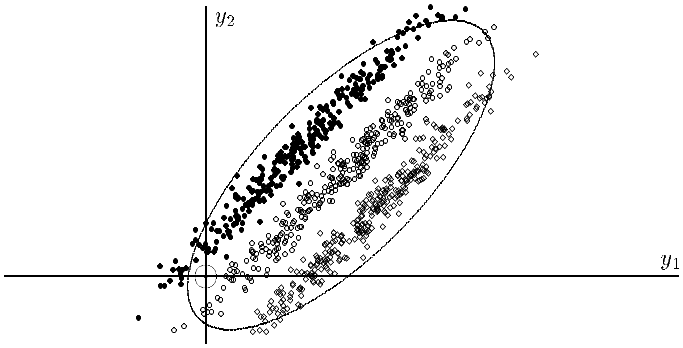


Fig. 5: The adidas problem is a set of points belonging to multiple classes which have the highest scatter in a direction orthogonal to the axis with the best discrimination. Hence, the dimension reduction with PCA would not yield an axis with good discrimination. In this example the class information would be lost [17].

Next, the distance between the two SCHMMs has to be computed as the sum of the distance of all states [5]. That leads to the overall distance δ_{mn} between the SCHMMs m and n:

$$\delta_{mn} = \frac{\sum_{i=1}^{N_s} d_i(m,n)}{N_s} \tag{4}$$

where N_s is the number of states. In this manner a symmetric distance matrix D is computed which holds all mutual distances between the SCHMMs.

2.3 Reduction of Dimensionality

2.3.1 PCA

PCA is a linear transformation that transforms the data to a new coordinate system such that the largest variance comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. The first principal component is assumed to be the most important one for separation of data of different classes. But this is not true in general, as can be seen in the so called adidas problem (cf. Fig. 5).

We have a set of ν feature vectors $\mathbf{c}_i (i = 1 \ldots \nu)$ which has to be transformed by a matrix $\mathbf{T}$ so that the resulting feature vectors $\mathbf{c}'_i$ show maximum spread in their first component. This results in an eigenvalue problem

$$\mathbf{\Sigma} \mathbf{e}_j = \lambda_j \mathbf{e}_j \tag{5}$$

with the covariance $\mathbf{\Sigma}$ of the features $\mathbf{c}_i$ and eigenvectors $\mathbf{e}_j$ and eigenvalues λ_j. The transformation matrix $\mathbf{T}$ for reducing the feature vector dimension to ξ consists of the ξ eigenvectors that correspond to the ξ largest eigenvalues.

2.3.2 LDA

While PCA does not use any information about the class of each feature (given by the speaker information in our case), LDA uses it to maximize the ratio of between-class scatter $\mathbf{B}$ and within-class scatter $\mathbf{W}$ [17]:

$$\det(B)/\det(W) \to \max \tag{6}$$

The within-class and between-class scatter are defined as

$$\mathbf{W} = \sum_{i=1}^{K} p_i \mathbf{\Sigma}_i, \qquad \mathbf{B} = \sum_{i=1}^{K} p_i$$

where p_i is the a priori probability that a sample belongs to class i $(i = 1 \dots K)$, μ_i and $\mathbf{\Sigma}_i$ are the mean and the covariance, respectively, of the samples of class i and μ is the mean of all samples.

Solving (6) results in a generalized eigenvalue problem:

$$\mathbf{Be} = \lambda \mathbf{We} \tag{7}$$

Again, the ξ eigenvalues that correspond to the ξ largest eigenvalues define the transformation $\mathbf{T}$.

2.3.3 Sammon Mapping

Unlike the PCA, the Sammon mapping is a nonlinear method for mapping high-dimensional data to a plane or a 3-D space [15]. In the late 1970's, a fast-converging algorithm for a generalized Sammon mapping was presented by Niemann and Weiß [12]. The Sammon mapping uses the distances (cf. Section 2.2) between the high-dimensional data to find a lower dimensional representation — called "map" in the following — that preserves the topology of the original data, i.e. keeps the distance ratios between the low-dimensional representation — called "star" in the following — as close as possible to the original distances. Hence, the Sammon mapping is cluster preserving. To ensure this, the function e_{S} is used as a measurement of the error of the resulting map (2-D case):

$$e_{\mathrm{S}} = s \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \frac{(\delta_{pq} - \theta_{pq})^2}{\delta_{pq}} \quad \text{with} \tag{8}$$

$$\theta_{pq} = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \tag{9}$$

δ_{pq} is the high-dimensional distance between the high dimensional features p and q stored in a distance matrix $\mathbf{D}$, θ_{pq} is the Euclidian distance between the corresponding stars p and q in the map. s is a scaling factor derived from the distances in the high-dimensional space:

$$s = \frac{1}{\sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \delta_{pq}} \tag{10}$$

The transformation is started with randomly initialized positions for the stars. Then the position of each star is optimized, using a conjugate gradient descent library [11].

2.4 Reduction of the Influence of the Recording Conditions in the Visualization

2.4.1 Rigid Registration

The first approach to reduce the influence of the recording conditions is the use of a rigid registration. The idea is to use utterances from several speakers and record each of them under different conditions. Then the features generated from the recordings are transformed into a 2-D (or 3-D) map. The map is split according to the recording conditions $(h_1 \dots h_H)$, and afterwards a rigid registration is applied, aiming to reduce the distance between the stars belonging to one speaker. The objective function for the registration is

$$e_{\mathrm{REG}}(h_i, h_j) = \frac{1}{N_m} \sum_p \theta_{p^{h_i} p^{h_j}} \tag{11}$$

for the two maps recorded in condition h_i and h_j, each consisting of $N_m = \frac{N}{H}$ stars. $\theta_{p^{h_i} p^{h_j}}$ is the Euclidian distance between the star p^{h_i} of the map from h_i and star p^{h_j} of acoustic condition h_j. The registration is considered to be a Euclidian transformation from h_i to h_j [16].

$$p^{h_j} = A_{h_i}^{h_j} p^{h_i} + t_{h_i}^{h_j} \tag{12}$$

with the transformation matrix $A_{h_i}^{h_j}$ and the translation vector $t_{h_i}^{h_j}$, i. e. it performs rotation and translation only.

The error is minimized using gradient descent. Using a non-rigid registration is not reasonable because this will lead to a great increase of the mapping error, i.e. loss of the structure of the map.

For the projection of a new star into a map, the dimensionality has first to be reduced according to the chosen method. Then, the registration can be performed according to Eq. 12.

2.4.2 Non-rigid Registration

The non-rigid approach can be included into the optimization process of the Sammon mapping: To minimize the distance between stars belonging to the same speaker, additional information about the group membership is used. Therefore, a grouping error is introduced to extend the objective function.

$$\boldsymbol{G} = \begin{pmatrix} g_{11} & \cdots & g_{1N} \\ \vdots & \ddots & \vdots \\ g_{N1} & \cdots & g_{NN} \end{pmatrix} \tag{13}$$

g_{ij} indicates whether the stars, or respectively the high-dimensional features, belong to the same group, i.e. the same speaker. Hence, $g_{ij} = 1$ if the feature vector j corresponds to speaker i, else $g_{ij} = 0$. Remember that one speaker is recorded in our application by multiple microphones, so there are more recordings for one speaker. Therefore, $\boldsymbol{G}$ looks like a sparse matrix.

The original error function of the Sammon mapping is altered such that it considers the distance between stars that belong to the same group. So a new error function e_Q is formed:

$$e_Q = s \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \left[Q g_{pq} \theta_{pq} + (1-Q)(1-g_{pq}) \frac{(\delta_{pq} - \theta_{pq})^2}{\delta_{pq}} \right] \tag{14}$$

where g_{pq} is the group indicator and Q is the weight factor which balances the standard Sammon error to the additional error term.

Again, gradient descent is applied. Taking partial derivatives leads to the following gradient:

$$\frac{\partial e_{\mathrm{SE}}}{\partial q_x} = s\sum_{p=1}^{n-1}\sum_{q=p+1}^{n}\left[Qg_{pq}\frac{-(p_x-q_x)}{\sqrt{(p_x-q_x)^2+(p_y-q_y)^2}}+\right.$$
$$\left.(1-Q)(1-g_{pq})\frac{2(p_x-q_x)(\delta_{pq}-\theta_{pq})}{\sqrt{(p_x-q_x)^2+(p_y-q_y)^2}}\right] \quad (15)$$

The derivatives for the other coordinates are formed analogously. Let $\mathbf{D}$ be the matrix of all the distances between the speaker features δ_{pq} in all recording conditions h_i $(i = 1 \dots H)$ (cf. Fig. 6):

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_{h_1}^{h_1} & \dots & \mathbf{D}_{h_1}^{h_H} \\ \vdots & \ddots & \vdots \\ \mathbf{D}_{h_H}^{h_1} & \dots & \mathbf{D}_{h_H}^{h_H} \end{pmatrix} \quad (16)$$

where $\mathbf{D}_{h_j}^{h_i}$ denotes the distances of all speakers recorded in condition h_i to the speakers recorded in condition h_j.

For the projection of a new speaker p' who was only recorded in condition h_i a new distance matrix $\hat{\mathbf{D}}$ is required. This matrix is like the distance matrix $\mathbf{D}$ except each partial matrix $\mathbf{D}_{h_j}^{h_k}$ $(j, k = 1 \dots H)$ has one more row and column for the new speaker. However, only matrix $\hat{\mathbf{D}}_{h_i}^{h_i}$ can be computed directly since all distances in h_i were observed. All distances in $\hat{\mathbf{D}}_{h_j}^{h_i}$ and $\hat{\mathbf{D}}_{h_i}^{h_j}$ $(j \neq i)$ except the one concerning the speaker himself in the other recording conditions are known. For the projection, all distances of the speaker to all other speakers in all other recording environments are required. This, however, would mean that the speaker has to be recorded in the other conditions as well since their differences might differ a lot. In order to compute distances to and between unknown recording conditions h_j and h_k, some kind of interpolation has to be chosen. For this interpolation we postulate the following:

1. The topology within one recording condition resembles the topology of the other recording conditions at least to some extent i.e. a linear scaling can be performed:

$$\hat{\mathbf{D}}_{h_j}^{h_j} \approx m(\mathbf{D}_{h_j}^{h_j}, \mathbf{D}_{h_i}^{h_i})\mathbf{D}_{h_i}^{h_i} \quad \text{with} \quad (17)$$

$$m(\mathbf{D}_{h_j}^{h_k}, \mathbf{D}_{h_i}^{h_s}) = \frac{\sum\limits_{p,q=1}^{N/H}(\delta_{pq}(\mathbf{D}_{h_j}^{h_k}) - \mu(\mathbf{D}_{h_j}^{h_k}))(\delta_{pq}(\mathbf{D}_{h_i}^{h_s}) - \mu(\mathbf{D}_{h_i}^{h_s}))}{\sum\limits_{p,q=1}^{N/H}(\delta_{pq}(\mathbf{D}_{h_i}^{h_s}) - \mu(\mathbf{D}_{h_i}^{h_s}))^2} \quad (18)$$

$$\mu(\mathbf{D}_{h_j}^{h_k}) = \frac{H}{N}\sum_{p,q=1}^{N/H}\delta_{pq}(\mathbf{D}_{h_j}^{h_k}) \quad (19)$$

   $\delta_{pq}(\mathbf{D}_{h_i}^{h_s})$ refers to the distance between stars p and q in block $\mathbf{D}_{h_i}^{h_s}$.

2. The distance between the speaker and himself projected to another acoustic environment can be approximated by the average distance from recording environ-

ment h_j to h_i:

$$\begin{aligned} \delta_{p'p'}(\hat{\mathbf{D}}^{h_j}_{h_i}) &\approx t(\mathbf{D}^{h_j}_{h_i}, \mathbf{D}^{h_i}_{h_i}) && (20) \\ t(\mathbf{D}^{h_k}_{h_j}, \mathbf{D}^{h_s}_{h_i}) &= m(\mathbf{D}^{h_k}_{h_j}, \mathbf{D}^{h_s}_{h_i})\mu(\mathbf{D}^{h_k}_{h_j}) \\ &\quad -\mu(\mathbf{D}^{h_s}_{h_i}) && (21) \end{aligned}$$

3. The distances between h_k and h_j $(k, j \neq i)$ can be interpolated from the nearest neighbor p^* of the desired point p' in known conditions $\mathbf{D}^{h_s}_{h_i}$ and a scaling factor $\upsilon(\mathbf{D}^{h_k}_{h_j}, \mathbf{D}^{h_s}_{h_i})$.

$$\begin{aligned} \delta_{p'q}(\hat{\mathbf{D}}^{h_j}_{h_k}) - \delta_{p^*q}(\hat{\mathbf{D}}^{h_j}_{h_k}) &\approx \upsilon(\mathbf{D}^{h_k}_{h_j}, \mathbf{D}^{h_s}_{h_i})(\delta_{p'q}(\mathbf{D}^{h_s}_{h_i}) \\ &\quad -\delta_{p^*q}(\mathbf{D}^{h_s}_{h_i})) && (22) \end{aligned}$$

Therefore, a set of two recording conditions has to be found where the distances resemble the distances of the unknown combination. In order to determine the similarity between two blocks different measures can be applied. We decided for the Pearson correlation r [14] between the corresponding entries of the blocks i.e.

$$\mathbf{D}^{h_s}_{h_i} = \begin{cases} \underset{\mathbf{D}^{h_{\hat{s}}}_{h_i}}{\arg\max}\, r(\mathbf{D}^{h_{\hat{s}}}_{h_i}, \mathbf{D}^{h_j}_{h_k}), & \text{if } k, j \neq i \\ \mathbf{D}^{h_i}_{h_i}, & \text{else} \end{cases} \tag{23}$$

with $\hat{s} = 1 \ldots H$. Note, that we must choose block $\mathbf{D}^{h_i}_{h_i}$ as the most similar block if one of the acoustic conditions is already h_i, since $\mathbf{D}^{h_i}_{h_i}$ is the only block which contains all distances, because all of them could be observed. Using the similar distance block $\mathbf{D}^{h_s}_{h_i}$ or $\mathbf{D}^{h_i}_{h_s}$ which is analogously defined, the closest point in the known domain p^* is found:

$$p^* = \underset{q \neq p'}{\arg\min}\, \delta_{p'q}(\mathbf{D}^{h_s}_{h_i}); \tag{24}$$

In order to transform the distances from one block to another, the average scaling $\upsilon(\mathbf{D}^{h_k}_{h_j}, \mathbf{D}^{h_s}_{h_i})$ between the distances of blocks $\mathbf{D}^{h_s}_{h_i}$ and $\mathbf{D}^{h_j}_{h_k}$ has to be determined:

$$\upsilon(\mathbf{D}^{h_k}_{h_j}, \mathbf{D}^{h_s}_{h_i}) = \frac{1}{(N-1)^2} \sum_{p,q \neq p'} \frac{\delta_{pq}(\mathbf{D}^{h_s}_{h_i})}{\delta_{pq}(\mathbf{D}^{h_j}_{h_k})} \tag{25}$$

The interpolation of the actual distance $\delta_{p'q}(\hat{\mathbf{D}}^{h_j}_{h_k})$ is now computed as postulated in Eq. 22:

$$\begin{aligned} \delta_{p'q}(\hat{\mathbf{D}}^{h_j}_{h_k}) &= \delta_{p^*q}(\hat{\mathbf{D}}^{h_j}_{h_k}) + \upsilon(\mathbf{D}^{h_k}_{h_j}, \mathbf{D}^{h_s}_{h_i})(\delta_{p'q}(\mathbf{D}^{h_s}_{h_i}) \\ &\quad -\delta_{p^*q}(\mathbf{D}^{h_s}_{h_i})) && (26) \end{aligned}$$

Since $\hat{\mathbf{D}}$ is a symmetric matrix, $\hat{\mathbf{D}}^{h_j}_{h_k} = \hat{\mathbf{D}}^{h_k \top}_{h_j}$ is valid. This equality can be employed to save computation time.

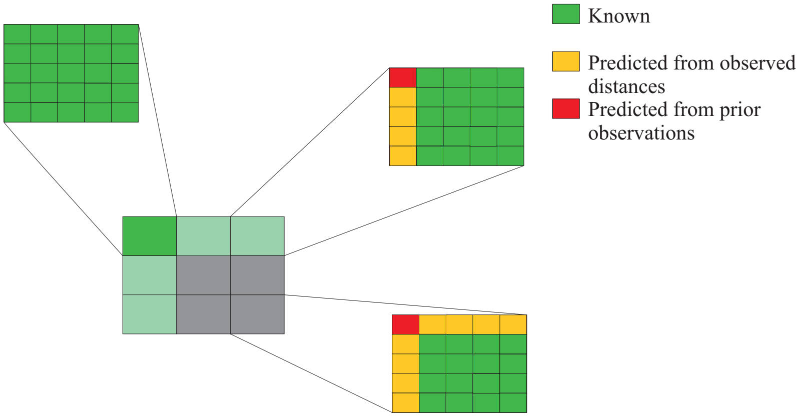

🟩 Known
🟨 Predicted from observed distances
🟥 Predicted from prior observations

Fig. 6: Structure of $\hat{\mathbf{D}}$ in the interpolation for unseen data with $H = 3$ and $N = 15$: The red values have to be approximated from prior knowledge while the yellow values are estimated from observed data. All green distances are observable and can hence be directly computed.

Based on these postulates, the projection of a new star which was just observed by a single microphone becomes possible since all entries of $\hat{\mathbf{D}}$ can now be computed. Choosing the right weight factor Q (cf. Eq. 14) is crucial to achieve a good reduction of the influence of the recording conditions while keeping the mapping error low. Also the factor must not be chosen too high, for otherwise it would corrupt the results when projecting new points into an existing map. With a good choice for Q the acoustic differences are removed from the map. In analogy to COSMOS the method is referred to as QMOS (Quality-independent Map Of Speakers).

Note that the use of the inter-acoustical differences from Eq. 22 is a crucial part in the computation since the estimates from Eqs. 17 and 20 are very coarse (cf. Fig. 6). Using Eq. 22 a lot of additional information is incorporated into the estimation process which would be lost if we assume for example $\hat{\mathbf{D}}^{h_i}_{h_j} = \mathbf{0}$ $(i \neq j)$ where $\mathbf{0}$ is the null matrix would be used. So only the diagonal blocks of the distance matrix would have to be computed. However, since $\hat{\mathbf{D}}^{h_i}_{h_j} = \mathbf{0}$ would end in a division by zero in the gradient function as well as in the cost function, a new distance matrix $\mathbf{D}^* = \sum_i \hat{\mathbf{D}}^{h_i}_{h_i}$ has to be created and only postulate (1) is required. The dimension of the new distance matrix is $\frac{N}{H} \times \frac{N}{H}$. Subsequently, the normal Sammon mapping as defined in Eq. 8 can be performed to compute the lower dimensional representation. This method is referred to as COSMOS$_\Sigma$. As shown in Section 3 QMOS outperforms COSMOS$_\Sigma$.

2.4.3 Quality Metrics for Maps

The measurement of the quality of a visualization is a very difficult task. In our case we decided to use two measurements for the evaluation:

- Sammon Error e_{S}: The remaining error is computed by the Sammon error function according to Eq. 8. This error is used to describe the error of the topology in the low-dimensional representation compared to the high-dimensional space. In the literature this term was shown to be a crucial factor to describe the quality of a representation [18, 10, 5]. Since the scaling of the maps influences the Sammon error, all maps were scaled in order to match their average Euclidean distances with the average distances of the high dimensional data, i.e. the best linear scaling in terms of the Sammon error was created. In this manner a comparable Sammon error can be computed for the

Table 2: Metrics for maps created from all data with the different visualization methods and rigid registration: Grouping works best with the PCA and rigid registration while COSMOS has the smallest Sammon error.

	e_{S}	e_{Grp}
PCA	0.28	0.45
PCA + reg.	0.47	**0.18**
LDA	0.25	0.28
LDA + reg.	0.44	**0.18**
COSMOS	**0.09**	0.40
COSMOS + reg.	0.21	0.21

linearly transformed maps which might be scaled sub-optimally regarding the Sammon error. All Sammon errors presented in the result section can also be interpreted as percentages, i.e. a Sammon error of 0.25 corresponds to 25 % of the total sum of the high dimensional distances.

- Grouping error e_{Grp}: The average distance between stars belonging to the same group (on a map with normalized coordinates in an interval between 0 and 1).

$$e_{\text{Grp}} = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \theta_{ij} g_{ij} \tag{27}$$

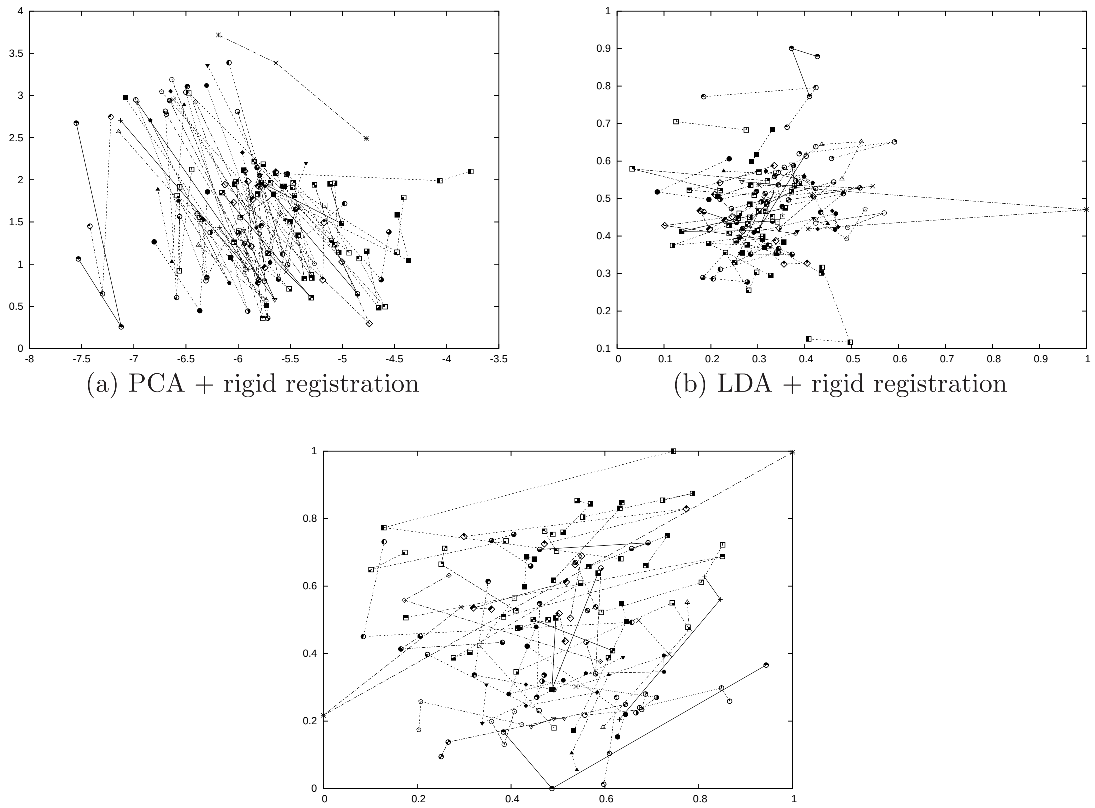
Note that the normalization is just performed with $\frac{1}{N}$ due to the sparsity of $\mathbf{G}$. A grouping error 0.25 corresponds to an average distance of 25 % of the maximum distance in the map between the representations of the same speaker.

In order to test whether the mapping method is also applicable for unseen data, the experiments were also conducted in Leave-One-Speaker-Out (LOO) evaluation. Therefore, the mapping is performed as in matched conditions with all but one speaker. This speaker is then projected into the map with the previously estimated transformation matrices in case of the linear transformations. For the non-linear mappings the previously described interpolation methods are applied. This is done for all recording conditions. e_{S} and e_{Grp} are then computed using the LOO predicted representation of the speaker in all conditions. Hence, the grouping error contains now the differences between the different conditions if only one microphone was seen for the projection of the speaker. Iteration of this process for all speakers yields the mean and the standard deviation of the LOO evaluated maps.

3. Results

The recording conditions differ a lot between the versions of the data. This can be seen in the recognition rates of a speech recognition system trained with these data (using two thirds of the data as training and one third as test). As reported in [7], the baseline recognition rate for close-talking data is 77.2 % WA. The recognition drops to 12.0 % WA if the *ct* recognizer is evaluated on the *rm* test set. In matched conditions speech recognition performs better with 63.1 % WA on *ctrv* data and 46.9 % WA on *rm*). However, the recognition rate on the *rm* data is almost of the rate of the *ct* data.

For all experiments the speech data, together with a transliteration of the spoken text, was used to adapt a speech recognizer for each speaker, using MLLR adaption of the Gaussian



(a) PCA + rigid registration

(b) LDA + rigid registration

(c) COSMOS + rigid registration

Fig. 7: Visualizations computed with rigid registration: Points representing the same speaker are connected with lines. None of the visualization methods yields a good visualization. The speakers are almost randomly distributed in each map.

mixture density for the SCHMM output probabilities. The mixture densities were used to compute distances (cf. Section 2.2) or directly used as features (cf. Section 2.1). Then a map was computed using the previously described methods. In the map all information from the three recording conditions (ct, ctrv, and rm) should be represented optimally. Evaluation criteria were the grouping error i.e. the distances between the stars representing the same speaker in each condition and the Sammon error i.e. the preservation of the high-dimensional topology.

The evaluation was performed for all dimensionality reduction techniques with and without registration. The performance of the different methods, transforming a set of known data, differed a lot. Table 2 shows the results. The method with the lowest grouping error is PCA plus rigid registration. The method with the best Sammon error is COSMOS. The visualizations of the rigid registered maps can be seen in Fig. 7. All three methods fail to project the same speakers close to each other. None of the visualizations can be interpreted properly.

Since the QMOS method is dependent on the weighting factor Q, the factor has to be determined experimentally. Table 3 shows the dependency between the grouping error and the Sammon error. The trade-off between grouping accuracy and reduction of the Sammon error has to be determined. The higher Q the higher the Sammon error and the lower the grouping error (cf. Fig. 8). The effect of the weight on the visualization is shown in

Table 3: Comparison between the different mapping methods in matched and LOO condition: All methods except QMOS fail to handle the LOO condition. For PCA, LDA, and COSMOS the rigid registration was applied.

method		known data		LOO Projection	
	Q	e_{S}	e_{Grp}	e_{S}	e_{Grp}
PCA + reg.		0.47	0.18	0.47±0.00	0.18±0.00
LDA + reg.		0.44	0.18	17.76±9.15	0.06±0.05
COSMOS + reg.		0.21	0.21	0.24±0.01	0.14±0.01
COSMOS$_{\Sigma}$		**0.19**	**0.00**	0.20±0.00	0.06±0.02
QMOS	0.00	**0.09**	0.40	0.10±0.00	0.40±0.01
QMOS	0.20	0.09	0.40	0.10±0.00	0.40±0.01
QMOS	0.40	0.09	0.39	0.10±0.00	0.39±0.01
QMOS	0.50	0.09	0.37	0.10±0.00	0.47±0.01
QMOS	0.60	0.09	0.36	0.10±0.00	0.36±0.00
QMOS	0.70	0.09	0.33	0.10±0.01	0.36±0.03
QMOS	0.75	0.16	0.07	0.16±0.03	0.22±0.10
QMOS	0.77	**0.17**	**0.**02	0.21±0.00	0.04±0.01
QMOS	0.80	**0.17**	**0.02**	0.21±0.00	0.04±0.01
QMOS	0.82	**0.**17	**0.02**	**0.21±0.00**	**0.03±0.01**
QMOS	0.85	0.18	0.03	0.22±0.00	0.04±0.00
QMOS	0.87	**0.**18	**0.01**	0.22±0.00	0.04±0.01
QMOS	0.90	0.18	0.03	0.23±0.00	0.03±0.01
QMOS	0.95	0.19	0.01	0.24±0.00	0.02±0.01
QMOS	0.96	0.24	**0.00**	**0.24±0.00**	**0.00±0.00**
QMOS	0.99	0.22	**0.00**	0.28±0.04	0.05±0.05

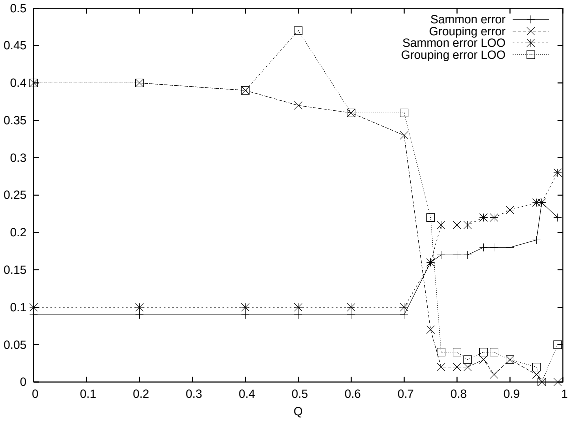
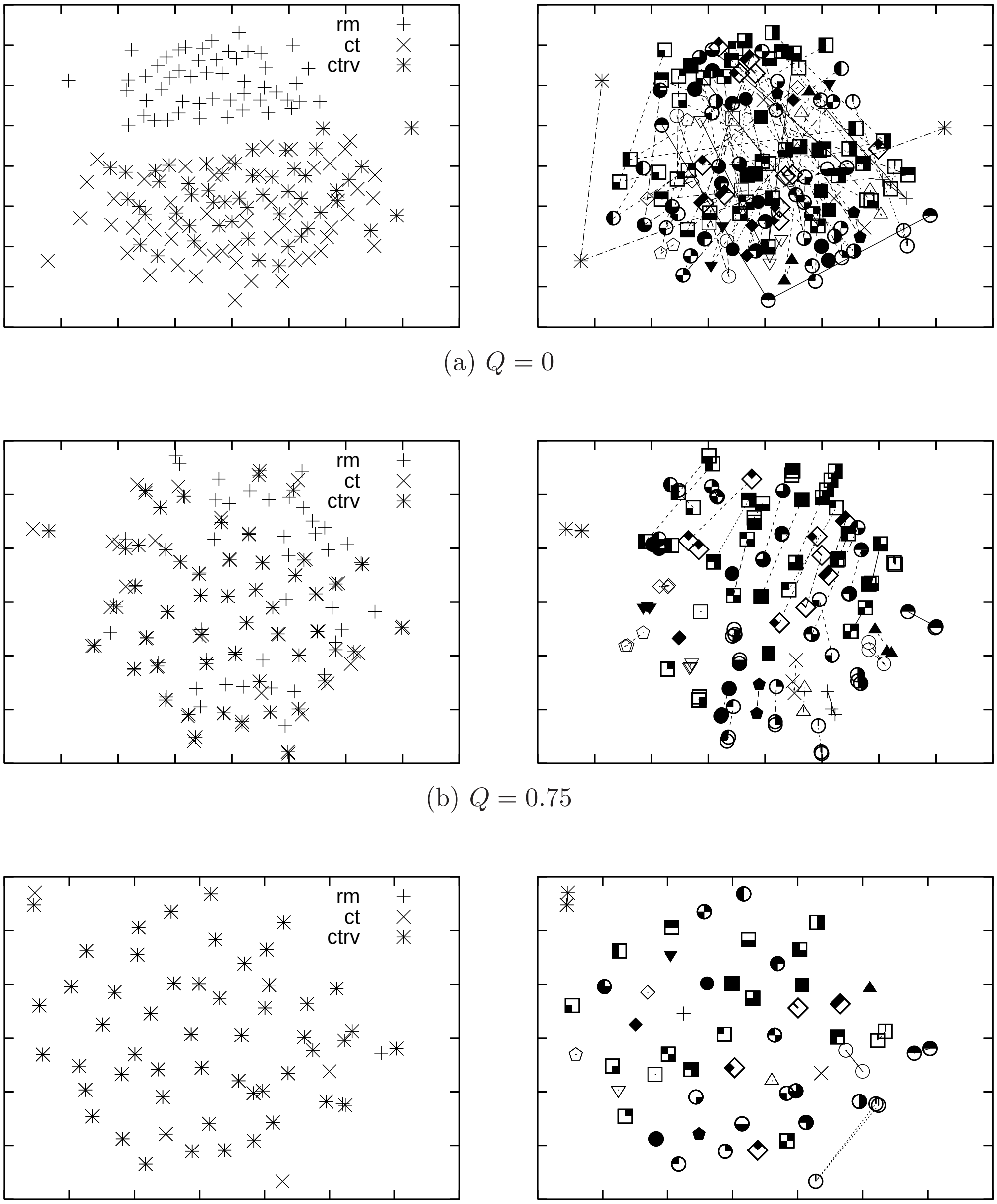


Fig. 8: Sammon and grouping errors in dependency of Q (cf. Table 3)

Fig. 9. The optimal value of the grouping error is at $Q = 0.87$ with a grouping error of only 0.01. At that position the trade-off between grouping and Sammon error is also optimal. Note that there are several configurations of Q which yield optimal sums of the grouping error and the Sammon error, i.e. one can choose from several values of Q depending on the problem one wishes to visualize.

Table 3 also displays the leave-one-speaker-out evaluation in order to test for the stability of the method. As can be seen, the PCA with rigid registration performs as bad as in



(a) $Q = 0$

(b) $Q = 0.75$

(c) $Q = 0.87$

Fig. 9: QMOS with three different weight factors. On the left side, the recording conditions are marked as close-talk (ct), ct + artificial reverberation (ctrv), and remote talk (rm). Note that almost all stars are at the same position with $Q = 0.87$ and can, hence, not be distinguished. On the right side, the stars belonging to one speaker are connected with lines. The fewer lines, the lower is the grouping error. The symbols represent each individual speaker.

Table 4: Correlations between the different blocks of the distance matrices $\mathbf{D}$ and its best predictors. Note that in the line and in the column of the condition which is to be predicted only N_m values have to be predicted. The correlations for the blocks where the most predictions $(2N_m - 1)$ are required are printed in bold face.

(a) Prediction of h_{ct}

$\mathbf{D}$	h_{ct}	h_{ctrv}	h_{rm}
h_{ct}	1 (known)	0.90 ($\mathbf{D}^{h_{\mathrm{ct}}}_{h_{\mathrm{ct}}}$)	0.71 ($\mathbf{D}^{h_{\mathrm{ct}}}_{h_{\mathrm{ct}}}$)
h_{ctrv}	0.90 ($\mathbf{D}^{h_{\mathrm{ct}}}_{h_{\mathrm{ct}}}$)	**0.98** ($\mathbf{D}^{h_{\mathbf{ct}}}_{h_{\mathbf{ct}}}$)	**0.99** ($\mathbf{D}^{h_{\mathbf{ct}}}_{h_{\mathbf{rm}}}$)
h_{rm}	0.71 ($\mathbf{D}^{h_{\mathrm{ct}}}_{h_{\mathrm{ctrv}}}$)	**0.99** ($\mathbf{D}^{h_{\mathbf{rm}}}_{h_{\mathbf{ct}}}$)	**0.87** ($\mathbf{D}^{h_{\mathbf{ct}}}_{h_{\mathbf{ct}}}$)

(b) Prediction of h_{ctrv}

$\mathbf{D}$	h_{ct}	h_{ctrv}	h_{rm}
h_{ct}	**0.98** ($\mathbf{D}^{h_{\mathbf{ctrv}}}_{h_{\mathbf{ctrv}}}$)	0.93 ($\mathbf{D}^{h_{\mathrm{ctrv}}}_{h_{\mathrm{ctrv}}}$)	**0.99** ($\mathbf{D}^{h_{\mathbf{ctrv}}}_{h_{\mathbf{rm}}}$)
h_{ctrv}	0.93 ($\mathbf{D}^{h_{\mathrm{ctrv}}}_{h_{\mathrm{ctrv}}}$)	1 (known)	0.63 ($\mathbf{D}^{h_{\mathrm{ctrv}}}_{h_{\mathrm{ctrv}}}$)
h_{rm}	**0.99** ($\mathbf{D}^{h_{\mathbf{rm}}}_{h_{\mathbf{ctrv}}}$)	0.63 ($\mathbf{D}^{h_{\mathrm{ctrv}}}_{h_{\mathrm{ctrv}}}$)	**0.85** ($\mathbf{D}^{h_{\mathbf{ctrv}}}_{h_{\mathbf{ctrv}}}$)

(c) Prediction of h_{rm}

$\mathbf{D}$	h_{ct}	h_{ctrv}	h_{rm}
h_{ct}	**0.87** ($\mathbf{D}^{h_{\mathbf{rm}}}_{h_{\mathbf{rm}}}$)	**0.88** ($\mathbf{D}^{h_{\mathbf{rm}}}_{h_{\mathbf{rm}}}$)	0.92 ($\mathbf{D}^{h_{\mathrm{rm}}}_{h_{\mathrm{rm}}}$)
h_{ctrv}	**0.88** ($\mathbf{D}^{h_{\mathbf{rm}}}_{h_{\mathbf{rm}}}$)	**0.85** ($\mathbf{D}^{h_{\mathbf{rm}}}_{h_{\mathbf{rm}}}$)	0.89 ($\mathbf{D}^{h_{\mathrm{rm}}}_{h_{\mathrm{rm}}}$)
h_{rm}	0.92 ($\mathbf{D}^{h_{\mathrm{rm}}}_{h_{\mathrm{rm}}}$)	0.89 ($\mathbf{D}^{h_{\mathrm{rm}}}_{h_{\mathrm{rm}}}$)	1 (known)

matched conditions with known data. The projection, however, is very stable with a standard deviation which is close to 0. LDA shows a very low grouping error but a very high Sammon error which is caused by the projection of the new data. Since the data were not seen during the computation of the transformation matrix, the new points are projected far away from the other stars. Hence, in the resulting map the previously projected stars are all very close to each other which causes the grouping error to be very low. The Sammon error increases a lot and is on average 18 times higher than the sum of all high-dimensional distances i.e. the result is mostly independent of the high-dimensional configuration. Rigid registration of the COSMOS maps shows a similar performance in the LOO case as in the matched case. The Sammon error is quite low, but the grouping error is high. This is also the case for the COSMOS$_\Sigma$ mapping although it performs slightly better. The first configurations with low Q show a very low Sammon error, but their grouping error is quite high. With increasing Q the grouping error drops while the Sammon error grows. With $Q = 0.82$ and $Q = 0.96$, two very stable configurations are found. One allows a greater variation in the grouping of the points while the other one allows a greater deviation in the dimensionality reduction error. The projection error is about half of the PCA Sammon error while the grouping error is reduced close to zero even in LOO evaluation. If Q is chosen too high, both errors increase ($Q = 0.99$).

Table 4 shows the prediction quality according to postulates 1,2, and 3 (cf. Section 2.4.2). The prediction of the blocks outside the diagonal of $\mathbf{D}$ is performed according to the nearest neighbor postulate. Good predictors of all blocks can be found. The best prediction block is printed in brackets behind the value. The correlations between the blocks range from 0.99 to 0.63 which are all significant (significance level p<0.01).

4. Discussion

In this study we evaluated visualization methods for speakers in different acoustic conditions. For this purpose we chose a children database which consists of 25.5 hours ($3 \times 8.5\,\mathrm{h}$) in total. Speech data was recorded in three different ways showing high differences due to the recording conditions.

We present common methods for the visualization of speaker data. None of them could provide a reliable result. The linear methods for dimensionality reduction (LDA and PCA) and the rigid registration could only yield maps in which the corresponding speakers are in a corresponding region. The grouping error, however, is far too high to interpret the map. For the PCA and rigid registration, this was also the case in the LOO evaluation. LDA could not handle unseen data. It caused an extreme increase in the projection error. Investigation of other linear methods such as Independent Components Analysis (ICA) showed even worse results since the selection criterion for the components poses a major problem. Selection of components which reduced the grouping error extremely increased the Sammon error and vice versa. Only the proposed method could yield a satisfying representation of the data. A configuration with a lower group or Sammon error than the proposed method could not be obtained. Therefore, results with ICA are not presented in this article.

The non-linear dimensionality reduction with the Sammon mapping surpassed the linear methods in the projection error in all cases. The grouping error, however, was worse even when a rigid registration was performed. So registration had to be included into the process in a non-linear manner. On the one hand, the summation of corresponding distances and therewith declaring these points as identical worked well in matched conditions (COSMOS$_{\Sigma}$) with a grouping error of 0. On the other hand, the method had a rather high grouping error in the LOO evaluation. However, the error is only half as high compared to the rigid registration. In the LOO evaluation, only QMOS was able to create an accurate representation of the speakers even when the recording was just performed with one microphone. This was achieved by linear interpolation in the distance domain instead of the map domain. The average grouping error was much lower than the minimal error of the other transformations (PCA) in LOO evaluation while keeping the projection error in the same range as the other non-linear methods (e.g. COSMOS$_{\Sigma}$).

In the future QMOS will be applied on clinical data. First a database with validated data has to be collected using a single microphone. Next, the data can be used as calibration data. Then new recording stations can use these data to compare new speakers with the validated data set.

5. Conclusion

We successfully created a new method for the robust visualization of speaker dependencies: Using our method it is possible to display speech data although the data were collected with different microphones. Furthermore, the method can even handle very strong differences in the acoustic conditions which occurred in the data.

The QMOS method is the only method to reduce the influence of recording conditions on a visualization (grouping error was almost zero) while keeping the mapping error low (Sammon error $e_{\mathrm{S}} < 0.25$). It performs better than rigid registration in minimizing the grouping error with a Sammon error that is about half of the error in the linear methods

(PCA and LDA). The key to create an appealing map with well-balanced Sammon and grouping error is how to choose an appropriate weight factor Q.

Finally, the data collected with different microphones can be projected into one single map. This allows for unified maps of speaker dependencies which enable analyses of speech data even if they were collected in different acoustic conditions.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (German Research Foundation; grant SCHU2320/1-1).

References

[1] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russell, and M. Wong. You stupid tin box – children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proceedings of the 4th International Conference of Language Resources and Evaluation LREC 2004*, pages 171–174, 2004.

[2] T. Bocklet, A. Maier, J. Bauer, F. Burkhardt, and E. Nöth. Age and Gender Recognition for Telephone Applications based on GMM Supervectors and Support Vector Machines. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 1605–1608, Las Vegas, USA, 2008. IEEE Computer Society Press.

[3] M. Gales, D. Pye, and P. Woodland. Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, volume 3, pages 1832–1835, Philadelphia, USA, 1996.

[4] T. Haderlein, E. Nöth, W. Herbordt, W. Kellermann, and H. Niemann. Using Artificially Reverberated Training Data in Distant-Talking ASR. In V. Matoušek, P. Mautner, and T. Pavelka, editors, *Text, Speech and Dialogue; 8th International Conference TSD 2005*, volume 3658 of *Lecture Notes in Artificial Intelligence*, pages 226–233. Springer, Berlin, Germany, 2005.

[5] T. Haderlein, D. Zorn, S. Steidl, E. Nöth, M. Shozakai, and M. Schuster. Visualization of Voice Disorders Using the Sammon Transform. In P. Sojka, I. Kopecek, and K. Pala, editors, *Proc. Text, Speech and Dialogue; 9th International Conference*, volume 4188 of *Lecture Notes in Artificial Intelligence*, pages 589–596. Springer, Berlin, Germany, 2006.

[6] P.C. Mahalanobis. On the generalised distance in statistics. In *Proceedings of the National Institute of Science of India 12*, pages 49–55, 1936.

[7] A. Maier, C. Hacker, S. Steidl, E. Nöth, and H. Niemann. Robust parallel speech recognition in multiple energy bands. In G. Kropatsch, R. Sablatnig, and A. Hanbury, editors, *Pattern Recognition, 27th DAGM Symposium, Vienna, Austria, Proceedings*, volume 3663 of *Lecture Notes in Computer Science*, pages 133–140. Springer, Berlin, Germany, 2005.

[8] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth. PEAKS – A System for the Automatic Evaluation of Voice and Speech Disorders. *Speech Communication*, 51(5):425–437, 2009.

[9] A. Maier, T. Haderlein, and E. Nöth. Environmental Adaptation with a Small Data Set of the Target Domain. In P. Sojka, I. Kopecek, and K. Pala, editors, *Proc. Text, Speech and Dialogue; 9th International Conference*, volume 4188 of *Lecture Notes in Artificial Intelligence*, pages 431–437. Springer, Berlin, Germany, 2006.

[10] G. Nagino and M. Shozakai. Building an Effective Corpus By Using Acoustic Space Visualization (COSMOS) method. In *Proceedings of ICASSP 2005 - International Conference on Acoustics, Speech, and Signal Processing*, pages 449–452, 2005.

[11] W. Naylor and B. Chapman. WNLIB Homepage, 2005.
http://www.willnaylor.com/wnlib.html, last visited 07/14/2007.

[12] H. Niemann and J. Weiss. A fast-converging algorithm for nonlinear mapping of highdimensional data to a plane. *IEEE Trans. Computers*, C-28:142–147, 1979.

[13] E. Nöth, A. Maier, T. Haderlein, K. Riedhammer, F. Rosanowski, and M. Schuster. Automatic Evaluation of Pathologic Speech—from Research to Routine Clinical Use. In *Text, Speech*

and Dialogue, Lecture Notes in Artificial Intelligence, pages 294–301, Berlin Heidelberg, 2007. Springer.

[14] K. Pearson. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society of London*, 187:253–318, 1896.

[15] J. Sammon. A nonlinear mapping for data structure analysis. In *IEEE Transactions on Computers C-18*, pages 401–409, 1969.

[16] J. Schmidt. *3-D Reconstruction and Stereo Self-Calibration for Augmented Reality.* Logos Verlag, Berlin, 2006.

[17] E.G. Schukat-Talamazzini. *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen.* Vieweg, Braunschweig, Germany, 1995.

[18] M. Shozakai and G. Nagino. Analysis of Speaking Styles by Two-Dimensional Visualization of Aggregate of Acoustic Models. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, volume 1, pages 717–720, Jeju Island (Rep. of Korea), 2004.

[19] G. Stemmer. *Modeling Variability in Speech Recognition.* Logos Verlag, Berlin, Germany, 2005.