An Extension to the Sammon Mapping for the Robust Visualization of Speaker Dependencies

Andreas Maier, Julian Exner, Stefan Steidl, Anton Batliner, Tino Haderlein, and Elmar Nöth

Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5)
Martensstraße 3, 91058 Erlangen, Germany
andreas.maier@informatik.uni-erlangen.de

Abstract. We present a novel method for the visualization of speakers which is microphone independent. To solve the problem of lacking microphone independency we present two methods to reduce the influence of the recording conditions on the visualization. The first one is a registration of maps created from identical speakers recorded under different conditions, i.e., different microphones and distances in two steps: Dimension reduction followed by the linear registration of the maps. The second method is an extension of the Sammon mapping method, which performs a non-linear registration during the dimension reduction procedure. The proposed method surpasses the two step registration approach with a mapping error ranging from 17 % to 24 % and a grouping error which is close to zero.

1 Introduction

The facets of voices and speech are very complex. They comprise various stationary and dynamic properties like 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. Therefore, the visualization in two or three dimensions was shown to be very effective in many fields of application:

- The visualization often allows to gain further insight in the structure of the data. For example the speaking style like loudness and rate-of-speech of different speakers can be analyzed [1].
- The visualization can also be used to select a subset of representative training speakers which cover all of its areas to reduce the number of training speakers. In a first step few data of many speakers are collected of which only the representative ones are included for a second recording session in order to collect more data. In this manner the recognition performance stays in the same range as if the second session would have been done with all speakers [2].
- The visualization can reveal the relations between patients with voice disorders in different graduations [3]. This gives the medical personal a better understanding of the different disorders. Projection of a new speaker allows to compare him to the other speakers.

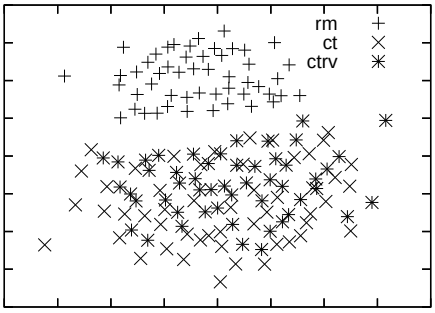


Fig. 1. 51 speakers recorded simultaneously with three different microphones: remote (rm), artificial reverberation (ctrv), and close talk (ct). ct and ctrv form one cluster while the rm forms another cluster. All three clusters contain the same speakers.

A problem for the visualization of speech data is the fact that the recording conditions have a great impact. The recording conditions consist mainly of the used microphone, the distance between the microphone and the speaker, and the acoustical properties of the recording location. If a speaker uttering a sentence was recorded simultaneously by multiple microphones of different quality at different distances, the points representing the same speaker are spread across the result of the visualization. Fig. 1 gives an extreme example using the Sammon mapping: The speakers form two clusters although the speakers were recorded simultaneously. This is caused by the acoustic difference between the two microphones which were chosen for the recording. The two corresponding representations of the same speaker are far away from each other in this visualization. The dominating factor is the microphone. In general, all visualizations of data collected in different acoustic conditions show similar effects in different graduations depending on the discrepancy between the acoustics.

If applied in a medical environment, for example with our fully automatic internet speech evaluation software [4], recordings are often performed at multiple locations simultaneously, e.g. in multi-site studies. Therefore, a method is desirable which removes or reduces these recording differences. The mismatch of the recording conditions can be reduced if a set of known calibration data is replayed with a standardized loudspeaker at a new location. In this manner, the effect of the new microphone and the recording conditions can be "learned" and removed from the visualization. In this paper we chose for simultaneously recorded data as we wanted to exclude the disturbances which might be created by the playback with a loudspeaker.

In order to create a visualization of the data the dimension has to be reduced to a two- or three-dimensional space. As a representation of a speaker we chose for the parameters of a speaker-adapted speech recognizer. Furthermore, the map should present same speakers at the same or at least a very close position, i.e., minimize the recording influences. To minimize the interferences of the recording conditions, two approaches are presented: The first one employs the standard Sammon mapping for the dimension reduction and a linear trans-

formation of the data points in the low dimensional domain in order to project corresponding ones as close to each other as possible, i.e., a registration of the maps. 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 the distance of points belonging to the same group already during the dimension reduction.

All methods were evaluated using the Aibo database. It consists of children speech recorded with a head-set microphone and the microphone of a video camera. A third recording condition was simulated using artificial reverberation. The Aibo data show very strong differences between the recordings conditions and are therefore ideal for the demonstration of our method.

2 Material

The database used in this work contains emotional speech of children. In a Wizard-of-Oz experiment children in the age of 12 to 14 years were faced with the task to control a Sony AIBO$^{\mathrm{TM}}$ robot by voice [5]. In total 51 pupils (21 male and 30 female) of two different schools were recorded in the German language. The whole scenery was recorded by a video camera in order to document the experiment and a head-mounted microphone (UT 14/20 SHURE UHF). The close talking version is referred to as *ct*. From the sound track of the video tape a second version of the AIBO corpus was extracted: a distant-talking version (*rm*) was obtained. In this manner no second manual transliteration was necessary because the transcription of the distant-talking and the close-talking version is the same. The distance between the speaker's position and the video camera was approximately 2.5 m. 8.5 hours of spontaneous speech data were recorded.

Artificial reverberation is used to create disturbances which resemble those caused by reverberation in a real acoustic environment. It is applied to the signal directly before the feature extraction. The idea is to convolve the speech signal with impulse responses characteristic for reverberation in typical application scenarios e.g. a living room. Thus a reverberated signal can be computed. These impulse responses can be measured in the proposed target environment or generated artificially. In current research the artificial reverberation was found to improve the robustness of speech recognizers to acoustic mismatches [6]. We applied the same twelve impulse responses as in the previously mentioned work.

In this manner the recordings from the close talk microphone were artificially reverberated to simulate another recording condition (*ctrv*). This way, three speech recognizers are adapted for each of the children from a speaker-independent one and are used for the creation of the visualizations.

3 Methods

3.1 Reduction of Dimensionality

The Sammon transformation (ST) is a nonlinear method for mapping high dimensional data to a plane or a 3-D space [7]. As already mentioned, the ST uses

the distances 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. Doing so, the ST 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):

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

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

δ_{pq} is the high dimensional distance between the high dimensional features p and q stored in a distance matrix $\boldsymbol{D}$, θ_{pq} is the Euclidian distance between the corresponding stars p and q in the map. For the computation of the high dimensional distance between two speech recognizers we use the Mahalanobis distance [8] as in [1]. s is a scaling factor derived from the high dimensional distances:

$$s = \frac{1}{\sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \delta_{pq}} \tag{3}$$

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 [9]. In [2] this method is referred to as "COSMOS" (COprehensive Space Map of Objective Signal).

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

The first approach to reduce the influence of the recording conditions, is the use of a linear registration. The idea is to use utterances from several speakers and record 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 linear registration is applied, aiming to reduce the distance between the stars belonging to one speaker. The objective function for the registration is

$$e_{\mathrm{REG}}(h_i, h_j) = \frac{1}{N_m} \sum_{i=1}^{N_m} \theta_{p^{h_i} p^{h_j}} \tag{4}$$

for the two maps recorded with microphone h_i and h_j, each consisting of $N_m = \frac{N}{H}$ stars. $\theta_{a_i b_i}$ is the Euclidian distance between the star p^{h_i} of the map from h_i and star p^{h_j} of microphone h_j.

$$n_i' = An_i + t \tag{5}$$

with the transformation matrix A and the translation vector t.

The error is minimized using gradient descent. For the projection of a new star into a map the dimensionality has to be reduced first according to the Sammon mapping. Then, the registration can be performed according to Eq. 5.

A non-linear registration 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 affiliation is used, i.e., stars representing the same speaker form a group. Therefore, a grouping error is introduced to extend the objective function. The recording is of the same structure as for the linear approach. A group weight g_{ij} indicates whether the stars, respectively the high dimensional features, belong to the same group. Thus, $g_{ij} = 1$, 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.

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

$$e_Q = s \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \left[Q g_{pq} \theta_{pq} + (1-Q)(1-g_{pq}) \frac{(\delta_{pq} - \theta_{pq})^2}{\delta_{pq}} \right] \tag{6}$$

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 to optimize the error criterion.

In allusion to the name of the method for speaker visualization as presented in [1] we refer to our method as QMOS.

3.3 Quality Metrics for the Visualization

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 computed by the Sammon error function according to Eq. 1. This error is used to describe the loss of the mapping from the high dimensional space to the low dimensional space. In the literature this term was shown to be a crucial factor to describe the quality of a representation [1–3].
- 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).

$$e_{\mathrm{Grp}} = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \theta_{ij} g_{ij} \tag{7}$$

Both errors are relative to the maximal error and can therefore also be interpreted as percentages, i.e. an error of 0.14 corresponds to 14 % of the maximal error.

Table 1. Metrics for maps created from all data with the different visualization methods: Both versions of COSMOS have a very high grouping error. QMOS surpasses both methods in grouping and Sammon error while having a lower grouping error.

method	e_S	e_{Grp}
COSMOS	**0.09**	0.40
COSMOS + reg.	0.21	0.21

method		Error	
	Q	e_S	e_{Grp}
QMOS	0.00	**0.09**	0.40
QMOS	0.60	0.09	0.36
QMOS	0.75	0.16	0.07
QMOS	0.87	**0.18**	**0.01**
QMOS	0.96	0.24	**0.00**

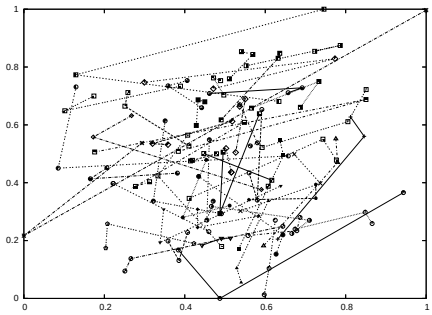


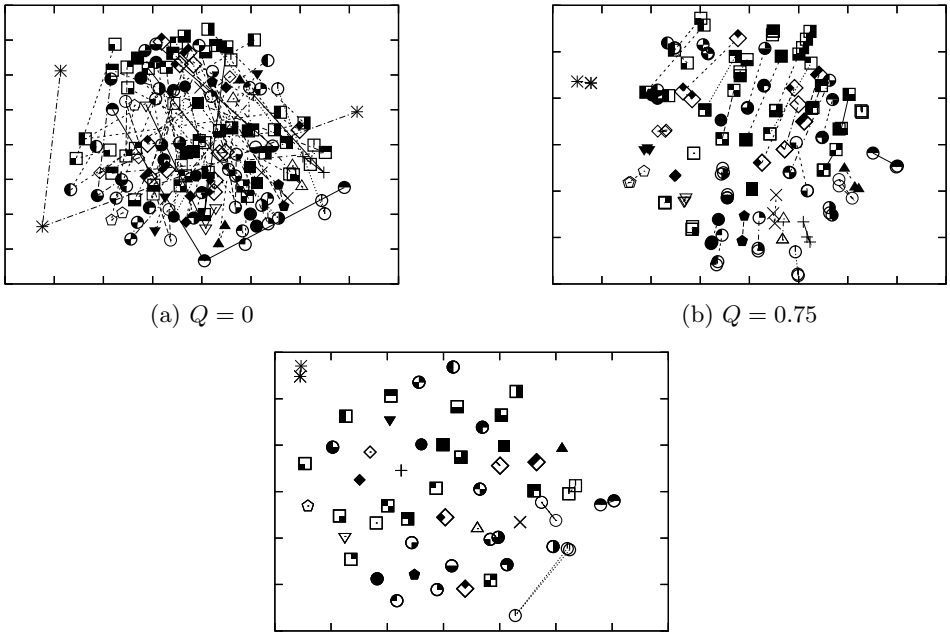
Fig. 2. Visualization computed with COSMOS and a linear registration: Points representing the same speaker are connected with lines. The visualization method does not yield a good visualization. The speakers are almost randomly distributed in each map.

4 Results

For all experiments the speech data, together with a transliteration of the spoken text, is used to adapt a speech recognizer for each speaker, using MLLR adaption of the Gaussian mixture density for the output probabilities. The mixture densities are used to compute distances or directly as features.

Evaluation was performed with and without linear registration. Table 1 shows the results. The method with the best Sammon error is of course the Sammon mapping. The visualizations of the registered maps can be seen in Fig. 2. The linear method fails to project the same speakers close to each other. The visualization cannot be interpreted properly.

Since the QMOS method is dependent on the weighting factor Q it has to be determined experimentally. Table 1 shows the dependency between the group and the Sammon error. The trade-off between grouping accuracy and reduction of the Sammon error has to be determined. The effect of the weight on the visualization is shown in Fig. 3. The optimal value of the group 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 very good. Note that there are several configurations of Q which yield a very low sum of grouping and Sammon error i.e. one can choose from several optimal values Q depending on the problem one wishes to visualize.



(a) $Q = 0$ (b) $Q = 0.75$

(c) $Q = 0.87$

Fig. 3. Extended ST on Aibo data with three different weight factors: The points belonging to one speaker are connected with lines. (a) shows the same map as Fig. 1.

5 Discussion

We evaluated the visualization methods for speakers in different acoustic conditions. The example we chose is difficult since the differences of the acoustic conditions in the AIBO database are large. Unfortunately, a signal-to-noise ratio cannot be computed between all versions of the AIBO corpus since the data is not always frame-matched. As reported in [10] the baseline recognition rates for matched conditions differ a lot (77.2 % WA on ct, 63.1 % WA on ctrv, and 46.9 % WA on rm). They are even worse if training and test set are not from the same version (12.0 % WA with the ct recognizer on the rm test set). Using these data we created a suitable mapping task since we wanted to create a method for visualization which can also handle extreme cases.

The linear method for the registration could only yield maps in which the corresponding speakers are in a corresponding region, i.e., the speakers are not projected into the opposite site of the map. Investigation of other linear methods such as PCA, LDA, and ICA showed no better results. A configuration with a lower group or Sammon error than the proposed method could not be obtained.

6 Conclusion

We successfully created a new method for the robust visualization of speaker dependencies: Using our novel method it is possible to create a single map al-

though the data was collected with different microphones. The method can even handle very strong differences in the acoustic conditions.

It was shown that the QMOS method is a good method to reduce the influence of recording conditions on a visualization (grouping error was almost zero) while keeping the mapping error low (Sammon error $e_S < 0.25$). It performs better than linear registration in minimizing the grouping error and has a Sammon error that is about half of the error in the linear methods. The key to create an appealing map with well balanced Sammon and grouping error is to choose the right weight factor which is of course problem dependent. If the factor is too low, the grouping error will be large, if it is too high, the Sammon error will be large and the map will not be a good visualization anymore.

Our method is ideal for the integration into a clinical environment since it is the only method which could handle the acoustical mismatches.

References

1. 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)*, Jeju Island (Rep. of Korea), 2004, vol. 1, pp. 717–720.
2. G. Nagino and M. Shozakai, "Building an effective corpus by using acoustic space visualization (cosmos) method," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05)*, 2005, pp. 449–452.
3. T. Haderlein, D. Zorn, S. Steidl, E. Nöth, M. Shozakai, and M. Schuster, "Visualization of Voice Disorders Using the Sammon Transform," in *Proc. Text, Speech and Dialogue; 9th International Conference*, P. Sojka, I. Kopecek, and K. Pala, Eds. 2006, number 4188 in Lecture Notes in Artificial Intelligence, pp. 589–596, Springer, Berlin, Germany.
4. A. Maier, E. Nöth, A. Batliner, E. Nkenke, and M. Schuster, "Fully Automatic Assessment of Speech of Children with Cleft Lip and Palate," *Informatica*, vol. 30, no. 4, pp. 477–482, 2006.
5. 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*, ELRA, Ed., 2004, pp. 171–174.
6. A. Maier, T. Haderlein, and E. Nöth, "Environmental Adaptation with a Small Data Set of the Target Domain," in *Proc. Text, Speech and Dialogue; 9th International Conference*, P. Sojka, I. Kopecek, and K. Pala, Eds. 2006, number 4188 in Lecture Notes in Artificial Intelligence, pp. 431–437, Springer, Berlin, Germany.
7. J. Sammon, "A nonlinear mapping for data structure analysis," in *IEEE Transactions on Computers C-18*, 1969, pp. 401–409.
8. P.C. Mahalanobis, "On the generalised distance in statistics," in *Proceedings of the National Institute of Science of India 12*, 1936, pp. 49–55.
9. W. Naylor and B. Chapman, "WNLIB Homepage," 2008, http://www.willnaylor.com/wnlib.html, last visited 17/01/2008.
10. A. Maier, C. Hacker, S. Steidl, E. Nöth, and H. Niemann, "Robust parallel speech recognition in multiple energy bands," in *Pattern Recognition, 27th DAGM Symposium, August 30 - September 2005, Vienna, Austria, Proceedings*, G. Kropatsch, R. Sablatnig, and A. Hanbury, Eds. 2005, vol. 3663 of *Lecture Notes in Computer Science*, pp. 133–140, Springer, Berlin, Germany.