INTERSPEECH 2009 BRIGHTON

A Microphone-Independent Visualization Technique for Speech Disorders

Andreas Maier[1,2], *Stefan Wenhardt*[1], *Tino Haderlein*[1,2], *Maria Schuster*[2], *Elmar Nöth*[1]

[1] Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Germany
[2] Abteilung für Phoniatrie und Pädaudiologie, Universitätsklinikum Erlangen, Germany
Andreas.Maier@cs.fau.de

Abstract

In this paper we introduce a novel method for the visualization of speech disorders. We demonstrate the method with disordered speech and a control group. However, both groups were recorded using two different microphones. The projection of the patient data using a single microphone yields significant correlations between the coordinates on the map and certain criteria of the disorder which were perceptually rated. However, projection of data from multiple microphones reduces this correlation. Usually, the acoustical mismatch between the microphones is greater than the mismatch between the speakers, i.e., not the disorders but the microphones form clusters in the visualization. Based on an extension of the Sammon mapping, we are able to create a map which projects the same speakers onto the same position even if multiple microphones are used. Furthermore, our method also restores the correlation between the map coordinates and the perceptual assessment.
Index Terms: visualization, robustness, speech processing.

1. Introduction

The visualization of speakers can reveal the relations between patients with voice disorders in different graduations [1]. Projection of new speakers allows to compare them to the other speakers. This gives a better understanding of the different disorders. Figure 1 shows a map of speakers with different degrees of hoarseness. On the top left, speakers with a tracheoesophageal substitute voice are located [2]. In these patients, the larynx was removed due to cancer. The artificial voice of the laryngectomized speakers can be interpreted as an extreme form of hoarseness. The average age of the laryngectomees was about 60 years. At the top right is an age-matched control group of normal speakers. At the bottom of the map are speakers with chronic hoarseness. On the bottom right, young reference speakers are located. Hence, the axes of the map can be approximately 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.

A problem for the visualization of speech data is the fact that the recording conditions have a great impact. The main factors are the used microphone, the distance between the microphone and the speaker, and the acoustical properties of the recording location. If a speaker was recorded simultaneously by multiple microphones of different quality at different distances, the points in the map which represent the same speaker are spread apart. Figure 2 gives an extreme example using the Sammon mapping: The speakers form two clusters although the speakers were recorded simultaneously with two different microphones [3]. This is caused by the acoustic difference between the microphones which were chosen for the recording. The corresponding representations of the same speaker are far

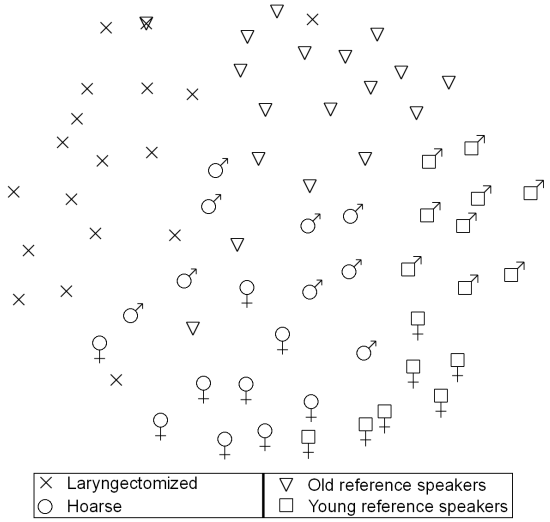


Figure 1: *Visualization of voice disorders: The properties of the speakers' voices are visible in a Sammon map. While the y-axis contains the age of the speaker, the x-axis can be interpreted as the degree of hoarseness.*

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 acoustic properties.

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 mismatches. We propose to gather a representative amount of calibration data which covers most of the acoustical variations of the respective voice disorder and a matching control group. This kind of calibration data can then be used to initialize a new location for the visualization procedure. The 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. New speakers are then projected into the calibrated visualization using just the recording of the new location as described in [5].

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 the parameters of a speaker-adapted speech recognizer. Furthermore, the map should present the recording of one speaker made in different environments at the same or at least a very close position, i.e., minimize the recording influences and therewith restore the

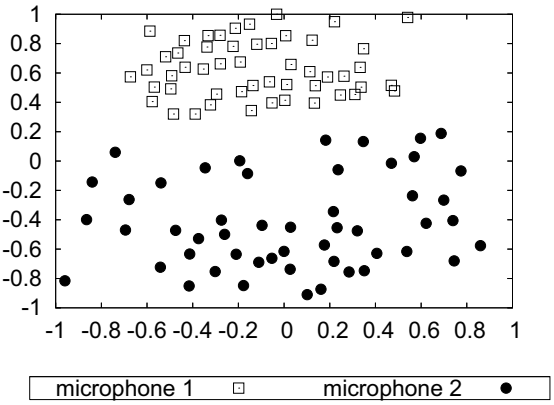


Figure 2: *51 children recorded simultaneously with two different microphones on a Sammon map: remote (rm), and close talk (ct). ct forms one cluster while the rm forms another cluster. Both clusters contain the same speakers [3].*

Table 1: *Serious articulation errors after [7]*

serious articulation errors	
nasalized consonants	the consonants are nasalized, i.e. air is emitted during the articulation of the consonants
glottal articulation (laryngeal replacement)	the closure of the plosives is not done in a labial but in a glottal manner.
backing to uvular	the tongue is shifted backwards towards the uvula
absent pressure consonants	plosives are not formed or weakened during the articulation

meaning of the coordinates.

2. Patients and Methods

2.1. Patients

Data of 31 children with cleft lip and palate (CLP) were recorded at the University Hospital Erlangen using a dnt Call 4U Comfort head set. Furthermore, a control group with 87 children was recorded with a Plantronics Audio USB 510 headset. In order to create matched recordings with both microphones, the data was replayed using a reference loudspeaker and recorded a second time with the respective other microphone.

An experienced speech therapist annotated all words the children with CLP spoke according to several criteria. An overview of the criteria is given in Table 1. The annotation was performed on word level. In order to get a speaker level result, the relative number of occurrences of the criteria was counted to represent the severity of the articulation disorder in the child. Furthermore, the intelligibility of the children was assessed in order to obtain a global outcome parameter for each child. This data set was also investigated in [6] of the automatic detection of speech disorders.

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 [8]. Then the dimensionality is reduced with the Sammon mapping [9]. For the Sammon mapping, an appropriate distance measure has to be chosen. Shozakai et al. chose the Mahalanobis distance [10] between the Gaussian densities of the speech recognizer. The resulting method is called COmprehensive Space Map of Objective Signal (COSMOS) [11].

2.2. Features for Visualization

We use the parameters of the Gaussian mixture densities of a speech recognizer as feature vectors for the visualization. Those densities are adapted with Maximum Likelihood Linear Regression (MLLR) adaption [8] for use with a semi-continuous Hidden Markov Models (SCHMM) speech recognizer that shares all of its $\kappa = 500$ Gaussian densities for all states [12]. The mixture density $f_\kappa(\boldsymbol{x})$ is computed as follows:

$$f_\kappa(\boldsymbol{x}) = \sum_{\kappa=1}^{K} \alpha_{i\kappa} \mathcal{N}_{i\kappa}(\boldsymbol{x}) \quad \text{with} \tag{1}$$

$$\mathcal{N}_{i\kappa}(\boldsymbol{x}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}_{i\kappa}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_{i\kappa}})^\top \boldsymbol{\Sigma}_{i\kappa}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_{i\kappa}})}$$

$\alpha_{i\kappa}$ is the weight for the Gaussian $\mathcal{N}_{i\kappa}(\boldsymbol{x})$, $\boldsymbol{\mu}_{i\kappa}$ is the mean vector, $\boldsymbol{\Sigma}_{i\kappa}$ is the covariance matrix of state i. M denotes the dimension of the feature vectors (here: $M = 24$). The sum over all α_κ equals 1.

2.3. A Distance Metric for SCHMMs

For the computation of the distance between two SCHMMs p and q, there are several applicable metrics. We use the Mahalanobis distance [10] as in [13]. 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

$$d_i(p,q) = \sum_{\kappa=1}^{K} \sqrt{(\hat{\boldsymbol{\mu}}_{i\kappa}(p) - \hat{\boldsymbol{\mu}}_{i\kappa}(q))^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}}_{i\kappa}(p) - \hat{\boldsymbol{\mu}}_{i\kappa}(q))}$$

$$\hat{\boldsymbol{\mu}}_{i\kappa}(p) = \alpha_{i\kappa}(p)\boldsymbol{\mu}_{i\kappa}(p) \tag{2}$$

for mixtures which consist of K Gaussians, with weighted mean vectors $\hat{\boldsymbol{\mu}}_{i\kappa}(p)$ and $\hat{\boldsymbol{\mu}}_{i\kappa}(q)$. $\boldsymbol{\Sigma}$ is the mean covariance matrix of all Gaussians of both mixtures [10], and i is the state number.

Next, the distance between the two SCHMMs has to be computed as the sum of the distance of all states [1]. That leads to the overall distance δ_{pq} between the SCHMMs p and q:

$$\delta_{pq} = \frac{\sum_{i=1}^{N_s} d_i(p,q)}{N_s} \tag{3}$$

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

2.4. Visualization with a Single Microphone

Next, the data is scaled to 2-D or 3-D using the Sammon mapping [9]. It finds the low-dimensional representation which matches the low-dimensional distances θ_{pq} best to the high-dimensional distances δ_{pq}:

$$e_S = s \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \frac{(\delta_{pq} - \theta_{pq})^2}{\delta_{pq}} \tag{4}$$

As low-dimensional distance measure, the Euclidean distance was chosen. s is a problem-dependent scaling factor, and N is the number of recordings.

Since our distance measure of the Sammon mapping is merely dependent on acoustical information, the use of multiple microphones causes distortions to the mapping as shown in Figure 2.

2.5. An Extension to the Sammon Mapping

Now we assume that we have a set of calibration data recorded with each microphone. This kind of data can be generated retrospectively with any kind of microphone at any new location. In order to supply further information to the mapping procedure, an additional term which punishes distances between matching speakers is included into the objective function e_{S} Therefore, the distance between points belonging to the same speaker is minimized.

$$\boldsymbol{G} = \begin{pmatrix} g_{11} & \cdots & g_{1N} \\ \vdots & \ddots & \vdots \\ g_{N1} & \cdots & g_{NN} \end{pmatrix} \tag{5}$$

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

The original error function e_{S} of the Sammon mapping is altered in such a manner that it considers the distance between points 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 a factor which weights the standard Sammon error and the additional error term. A gradient descent is applied to minimize the objective function.

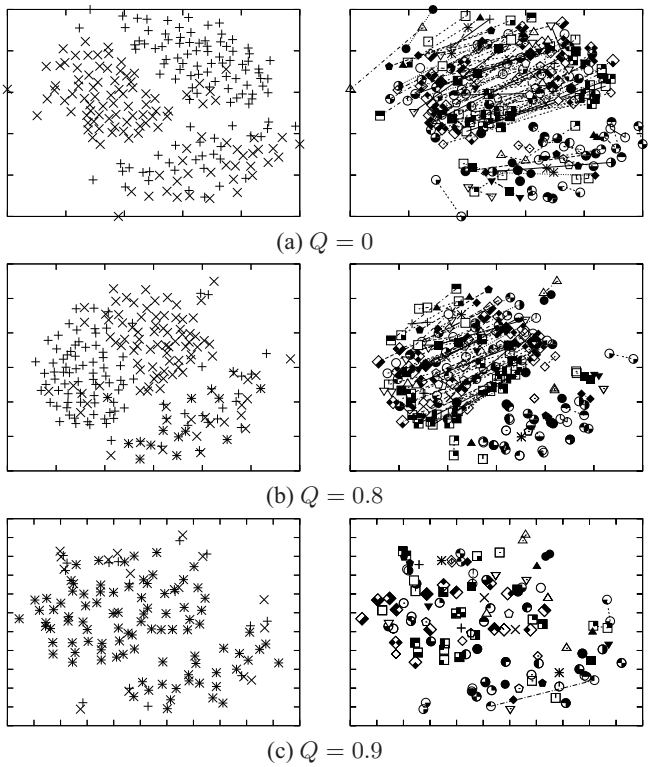
2.6. Quality Metrics for Maps

A major criterion for the visualization of the speakers is that the created map has to be meaningful, i.e., the quality has to be measured. We decided to use three measurements for the evaluation:

- Sammon Error e_{S}: The remaining error computed by the Sammon error function according to Eq. 4. 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, 13].
- Grouping Error e_{Grp}: The average distance between points belonging to the same group (on a map with normalized coordinates in an interval between 0 and 1), i.e., the average distance between a speaker to his own representation recorded with a different microphone.

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

  Note that the normalization is just performed with $\frac{1}{N}$ due to the sparsity of $\boldsymbol{G}$. A grouping error of 0.25 corresponds to an average distance of 25 % of the dimensions of the map between the representations of the same speaker.
- Regression: The regression between the coordinates of a map and a given criterion also provides information on the quality of the map. The regression is computed as the correlation between the least square optimal projection of the coordinates of the map to the given criterion, e.g. the intelligibility.



(a) $Q = 0$

(b) $Q = 0.8$

(c) $Q = 0.9$

Figure 3: *Extended Sammon mapping on data played back with two different microphones: On the left the Plantronics Audio USB 510 microphone is marked with "X" and "+" marks the dnt Call 4U Comfort microphone. Each microphone forms a cluster, although exactly the same speech data is represented. One the right each speaker is represented with a unique symbol. The points which represent the same speaker are connected with a line, i.e., the fewer lines, the fewer the grouping error. With growing Q the grouping error is reduced. With $Q = 0.9$ almost no lines are visible, i.e., the grouping error is close to zero. Note that the two clusters in (c) are the patient (bottom right) and the control group (top left).*

3. Results

With the weight Q, a trade-off between grouping and normal Sammon mapping is created. As Figure 3 shows, the points representing the same speaker move together with growing Q.

Figure 4 shows the development of the grouping and the Sammon error in dependency of Q. The higher Q, the lower is the group error. The Sammon error increases with growing Q. At $Q = 0.9$ a configuration is found where the sum of Sammon and grouping error is minimal as displayed in Figure 3 (c). $Q = 0.9$ seems to put most of the weight on the grouping error.

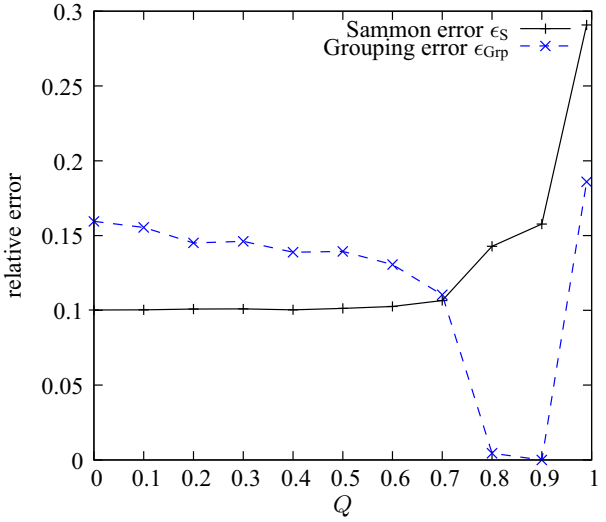


Figure 4: *Development of the Sammon and the grouping error in dependency of Q: While the Sammon error increases steadily with growing Q, the grouping error decreases. With a too high weight of the grouping error, the coordinates become mere random numbers due to the random initialization.*

However, if we recall the definition of e_Q from Eq. 6, and the definition of $\boldsymbol{G}$ from Eq. 5 one can easily see that most of the error sum is caused by the Sammon error and not by the grouping error since $\boldsymbol{G}$ contains only N times an entry with $g_{ij} = 1$ and $(N^2 - N)$ times $g_{ij} = 0$. So if the average error would be equal, i.e.,

$$\sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \frac{(\delta_{pq} - \theta_{pq})^2}{\delta_{pq}} \approx \sum_{p=1}^{N-1} \sum_{q=p+1}^{N} \theta_{pq} \qquad (8)$$

the break-even point between both errors with $N \approx 200$ would be at about $Q = 0.99$. Hence, with $Q = 0.9$ the influence of the Sammon information is still very high.

As shown in Table 2, the visualization of the patient data shows significant correlations to the criteria "laryngeal backing", "weakened plosives" and "intelligibility". However, after addition of the control group to the visualization the correlations drop. The use of the extended Sammon mapping with a Q of 0.9 is able to restore most correlations, esp. the main outcome parameter — intelligibility.

4. Summary

We presented a new method for the robust visualization of speech data. It is not only able to project corresponding speakers with respect to different microphones onto the same position of the map, but also restores the meaning of the coordinates of the map. An online demo is presented at http://peaks.informatik.uni-erlangen.de/visualization.

5. References

[1] T. Haderlein, D. Zorn, S. Steidl, E. Nöth, M. Shozakai, and M. Schuster, "Visualization of Voice Disorders Using the Sammon Transform," in *9th International Conf. on Text, Speech and Dialogue (TSD)*, ser. Lecture Notes in Artificial Intelligence, P. Sojka, I. Kopeček, and K. Pala, Eds., vol. 4188. Berlin, Heidelberg, New York: Springer, 2006, pp. 589–596.

[2] T. Haderlein, *Automatic Evaluation of Tracheoesophageal Substitute Voices*, ser. Studien zur Mustererkennung. Berlin, Germany: Logos Verlag, 2007, vol. 25.

Table 2: *Correlations on a map with only the patient group ("single") and both groups with the Sammon mapping ($Q = 0$) and the extended Sammon mapping ($Q = 0.9$) recorded with two microphones. All measures are computed on the patient subset of the visualization. (*) marks significant correlations at $p < 0.05$ while (**) marks highly significant correlations at $p < 0.01$*

criterion	single	$Q = 0$	$Q = 0.9$
nasalized consonants	0.31	0.10	0.42 (*)
laryngeal replacement	0.54 (*)	0.41 (*)	0.43 (*)
pharyngeal backing	0.61 (**)	0.39 (*)	0.43 (*)
weakened plosives	0.21	0.39 (*)	0.44 (*)
intelligibility	0.52 (*)	0.12	0.55 (**)
marked words	0.28	0.25	0.60 (**)
age	0.32	0.44 (*)	0.39 (*)

[3] A. Maier, J. Exner, S. Steidl, A. Batliner, T. Haderlein, and E. Nöth, "An Extension to the Sammon Mapping for the Robust Visualization of Speaker Dependencies," in *11th International Conf. on Text, Speech and Dialogue (TSD)*, ser. Lecture Notes in Artificial Intelligence, P. Sojka, I. Kopeček, and K. Pala, Eds. Berlin, Heidelberg, New York: Springer, 2008, pp. 381–388.

[4] 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*, vol. 51, no. 5, pp. 425–437, 2009.

[5] A. Maier, M. Schuster, U. Eysholdt, T. Haderlein, T. Cincarek, S. Steidl, A. Batliner, S. Wenhardt, and E. Nöth, "QMOS - A Robust Visualization Method for Speaker Dependencies with Different Microphones," *Journal of Pattern Recognition Research*, vol. 4, no. 1, pp. 32–51, 2009.

[6] A. Maier, F. Hönig, C. Hacker, M. Schuster, and E. Nöth, "Automatic evaluation of characteristic speech disorders in children with cleft lip and palate," in *Interspeech 2008 – Proc. Int. Conf. on Spoken Language Processing, 11th International Conference on Spoken Language Processing, September 25-28, 2008, Brisbane, Australia, Proceedings*, 2008, pp. 1757–1760.

[7] D. Sell, P. Grunwell, S. Mildinhall, T. Murphy, T. Cornish, D. Bearn, W. Shaw, J. Murray, A. Williams, and J. Sandy, "Cleft Lip and Palate Care in the United Kingdom—The Clinical Standards Advisory Group (CSAG) Study. Part 3: Speech Outcomes," *Cleft Palate-Craniofacial Journal*, vol. 32, no. 1, pp. 30–37, 2001.

[8] M. Gales, D. Pye, and P. Woodland, "Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation," in *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, vol. 3. Philadelphia, USA: ISCA, 1996, pp. 1832–1835.

[9] J. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Trans. Computers*, vol. C-18, pp. 401–409, 1969.

[10] P. Mahalanobis, "On the generalised distance in statistics," in *Proceedings of the National Institute of Science of India 12*, 1936, pp. 49–55.

[11] M. Nagino, G. Shozakai, "Building an effective corpus by using acoustic space visualization (cosmos) method," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Philadelphia, USA: IEEE Computer Society Press, 2005, pp. 449–452.

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

[13] M. Shozakai and G. Nagino, "Analysis of Speaking Styles by Two-Dimensional Visualization of Aggregate of Acoustic Models," in *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, vol. 1. Jeju Island, Korea: ISCA, 2004, pp. 717–720.