© Springer-Verlag http://www.springer.de/comp/lncs/index.html
published in: Text, Speech and Dialogue (TSD 2006), Proceedings
LNAI 4188, Springer, Berlin, Heidelberg, 2006, pp. 589–596

Visualization of Voice Disorders Using the Sammon Transform

Tino Haderlein[1], Dominik Zorn[2], Stefan Steidl[2], Elmar Nöth[2], Makoto Shozakai[3], and Maria Schuster[1]

[1] University of Erlangen-Nuremberg, Department of Phoniatrics and Pedaudiology
Bohlenplatz 21, 91054 Erlangen, Germany
Tino.Haderlein@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de
[2] University of Erlangen-Nuremberg, Chair for Pattern Recognition (Informatik 5)
Martensstraße 3, 91058 Erlangen, Germany
[3] Asahi Kasei Corporation, Speech Recognition Department
Atsugi AXT Main Tower 22F, 3050 Okata, Atsugi-shi, Kanagawa 243-0021, Japan

Abstract. The Sammon Transform performs data projections in a topology-preserving manner on the basis of an arbitrary distance measure. We use the weights of the observation probabilities of semi-continuous HMMs that were adapted to the current speaker as input. Experiments on laryngectomized speakers with tracheoesophageal substitute voice, hoarse, and normal speakers show encouraging results. Different speaker groups are separated in 2-D space, and the projection of a new speaker into the Sammon map allows prediction of his or her kind of voice pathology. The method can thus be used as an objective, automated support for the evaluation of voice disorders, and it visualizes them in a way that is convenient for speech therapists.

1 Introduction

Today, automatic speech processing can do much more than simply recognizing speech input. Based on speech, it is possible to find out a user's identity, his or her emotional state, or speech quality. This wide field of possible applications has its basis in the high information load of natural speech that extends far beyond the bare meaning of the spoken word. Still, a field that has been less considered, is the possible benefit to medical or clinical purposes with respect to diagnosis support. There are several scenarios concerning disorders or diseases where methods from speech recognition could be applied successfully for objective analysis. The origins of voice disorders are various, ranging from injuries, inflammation, palsy or neoplasms of the larynx to misuse of the voice or side effects from other diseases. In the USA between 5% and 10% of the population suffer from such disorders [1]. These numbers give an impression of the extent of the problem and the costs connected with it and show that it might be very helpful for speech therapists to get some automated and objective support for the evaluation and classification of pathologic voices or speech. If the results of such an automatic evaluation are merely a sequence of numbers based upon cepstral features, for example, this will be of no help for the technically uneducated medical personnel. Therefore, the goal of our work is to provide a graphical visualization of a small number of features which are extracted from a high number of “technical” features by some adequate dimension reduction. Another important aspect is the ability of comparing a new speaker’s disorder to an existing database of previous speakers.

We created an analysis framework for different kinds of voice disorders as a front-end for traditional speech recognition techniques. The basis of the distance measure between different speakers are the Hidden Markov model parameters of a speech recognition system that are changed when the recognizer is adapted to the current test speaker. Our interest does not focus on recognition or accuracy purposes in the first place (still these can be addressed), but to gain insight into severity and mutual relations of voice disorders. The results of the recognizer adaptation are presented graphically. A mapping technique, the so-called *Sammon* mapping [2], allows the graphical representation of abstract data, unveiling underlying structures and configurations. This method of mapping data is actually not new, but it has never been applied to this concrete problem. Recordings of hoarse speakers and laryngectomized persons were available for testing.

In Sect. 2 the underlying speech recognition system will be described, Sect. 3 defines the distance measure for HMMs needed by the Sammon mapping that will be introduced in Sect. 4. The test data is described in Sect. 5, the results can be viewed in Sect. 6. Section 7 gives a conclusion and a short outlook.

2 Interpolated Semi-Continuous HMMs

The features computed to express the differences between speakers are obtained from the adaptation of a speech recognizer to the current test speaker. Our recognizers are based on semi-continuous Hidden Markov Models (SCHMMs). Unlike discrete HMMs, continuous HMMs represent the output probabilities of their states by continuous probabilistic functions. This improves the recognition results but also heavily increases the number of model parameters. SCHMMs address this problem by sharing a common set of output densities in all states. Each HMM state incorporates these densities by a specific vector of weights. We use the interpolation method from [3] to adapt the output weights of an existing speaker-independent recognizer to individual speaker characteristics with a small amount of adaption data. Unlike in usual acoustic voice evaluation, we do not only use a single, sustained vowel, but a standard text uttered by the respective speaker (see Sect. 5). In this way, we achieve a set of speaker-adapted recognizers. Then we use the output weights of each recognizer for the mapping procedure. The original recognizer was trained on 27 hours of normal laryngeal speech.

This method of feature extraction seems to be rather expensive, but previous unpublished experiments at our institute showed that the features usually used in speech recognition, like cepstral coefficients, are not suitable for this task.

3 A Distance Metric for Semi-Continuous HMMs

The Sammon mapping (Sect. 4) is a non-linear transformation preserving data topology. This topology is represented within the matrix of respective utterance distances. The quality and information quantity of a Sammon map is fully determined by this metric and not by the mapping itself. Thus, it is extremely important to have a suitable distance metric. On the other hand, the distance metric can be chosen without any mathematical restrictions like linearity etc. This is the great advantage of the Sammon Transform against other dimension reduction operations, like PCA.

In our case, we need a good distance calculus for speaker-adapted SCHMMs in order to get the distance between a pathologic voice and the normal voices represented by the baseline recognizer, or between two pathologic voices. We propose a distance measure computed from the distances of the respective elementary SCHMMs of different speaker-dependent speech recognizers. The arithmetic mean of these model distances serves as the final result. So the problem reduces to calculating the distance of the states of two SCHMMs. Distance calculation has to use the interpolation weights but still take into consideration the densities from the recognizer codebook containing the Gaussian output densities. This is due to the varying information load which can be considered higher for densities with low and lower for those with a high variance. If a simple Euclidean distance of the weight vectors were used, this information would get lost and the quality of the distance metric would diminish. The codebook itself is static and common to all speakers.

The basic distance metric is an HMM state distance which is computed in two steps, one for the mean vector of each codebook density and a second one for its covariance matrix.

3.1 Distance of Mean Vectors

Concerning the mean value of the output densities for each HMM state, the approach is straightforward. For each state i of a model p the mean vector $\boldsymbol{m}_{ik}(p)$ of each codebook density k is scaled with the corresponding output weight $c_{ik}(p)$ as introduced in [3]:

$$\hat{\boldsymbol{m}}_{ik}(p) = c_{ik}(p) \cdot \boldsymbol{m}_{ik}(p) \tag{1}$$

Given two HMMs named p and q, the standard Euclidean distance can now be computed between $\hat{\boldsymbol{m}}_{ik}(p)$ and $\hat{\boldsymbol{m}}_{ik}(q)$ which are both of dimension R:

$$_{\text{MEAN}}d_{ik}(p,q) = \sqrt{\sum_{r=1}^{R} (\hat{\boldsymbol{m}}_{ik,r}(p) - \hat{\boldsymbol{m}}_{ik,r}(q))^2} \tag{2}$$

It represents the distance in the mean vectors of the scaled density k of state i between the two HMMs.

3.2 Distance of Covariance Matrices

There are various distance metrics for matrices. For the distance of covariance matrices of two codebook densities we use the Euclidean distance of corresponding weighted column vectors and interprete their arithmetic mean as covariance distance. Analogous to (1), each vector $\boldsymbol{v}_{ik,\rho}(p)$ of column ρ of the covariance matrix is scaled with the corresponding interpolation weight c_{ik}:

$$\hat{\boldsymbol{v}}_{ik,\rho}(p) = c_{ik}(p) \cdot \boldsymbol{v}_{ik,\rho}(p) \tag{3}$$

As the dimension of a codebook density is R, the covariance matrix is of the size $R \times R$, i.e. there are R pairs of corresponding column vectors $\hat{\boldsymbol{v}}_{ik,\rho}(p), \hat{\boldsymbol{v}}_{ik,\rho}(q)$ to be processed. For each pair the Euclidean distance is:

$$_{\text{COVA}}d_{ik,\rho}(p,q) = \sqrt{\sum_{r=1}^{R} (\hat{\boldsymbol{v}}_{ik,\rho r}(p) - \hat{\boldsymbol{v}}_{ik,\rho r}(q))^2} \tag{4}$$

In order to produce a single distance value for two corresponding densities out of the R results $_{\text{COVA}}d_{ik,\rho}(p,q)$ from the column vectors, their arithmetic mean serves as final covariance distance $_{\text{COVA}}d_{ik}(p,q)$ for one codebook density k:

$$_{\text{COVA}}d_{ik}(p,q) = \frac{\sum_{r=1}^{R} {_{\text{COVA}}d_{ik,\rho}(p,q)}}{R} \tag{5}$$

Finally $_{\text{MEAN}}d_{ik}(p,q)$ and $_{\text{COVA}}d_{ik}(p,q)$ are combined to one density distance:

$$d_{ik}(p,q) = \frac{_{\text{MEAN}}d_{ik}(p,q) + {_{\text{COVA}}d_{ik}(p,q)}}{2} \tag{6}$$

In general, $_{\text{MEAN}}d_{ik}(p,q)$ is much smaller than $_{\text{COVA}}d_{ik}(p,q)$. Therefore, the introduction of weights for both values is subject of future work.

3.3 Single State and HMM Distance

The calculations in (1) to (6) are performed for all of the K Gaussian output densities of a state i. In the end, the resulting set of K density distances $d_{ik}(p,q)$ obtained from (6) is averaged and provides a single state distance $d_i(p,q)$:

$$d_i(p,q) = \frac{\sum_{k=1}^{K} d_{ik}(p,q)}{K} \tag{7}$$

In the same way, the HMM distance δ_{pq} between models p and q can be obtained by normalizing the sum of all N state distances:

$$\delta_{pq} = \frac{\sum_{i=1}^{N} d_i(p,q)}{N} \tag{8}$$

The HMM distance in (8) is computed for each pair of elementary HMMs, thus filling up a matrix $\boldsymbol{D}$ holding the speaker distances. This matrix is symmetric, so for n utterances $\frac{n^2-n}{2}$ distances have to be calculated. Limiting the amount of HMM state densities K taken into consideration to some K' when calculating the state distance can reduce computation time. For $K - K' \ll K$ the effect on the resulting HMM distance is negligible.

4 Sammon Mapping

The Sammon mapping performs a topology-preserving reduction of data dimension. It minimizes a stress function between the topology of the low-dimensional Sammon map and the high-dimensional original data. The latter topology is defined by the distances between utterances or speakers, as defined in Sect. 3. The low-dimensional Sammon map is usually visualized as 2-D or 3-D image. With respect to [4], we call a Sammon map a *cosmos*, while a mapped utterance inside a cosmos is called a *star*.

The heart of Sammon's method is its special error function E, yielding a stress factor between the actual configuration of stars in m-dimensional target domain and the original data in d-dimensional space ($m < d$):

$$E = \frac{1}{\sum_{p=1}^{n-1} \sum_{q=p+1}^{n} \delta_{pq}} \sum_{p=1}^{n-1} \sum_{q=p+1}^{n} \frac{(\delta_{pq} - \nu_{pq})^2}{\delta_{pq}} \tag{9}$$

δ_{pq} denotes the distance between HMMs with number p and q, as in (8), ν_{pq} is the distance between stars $s(p)$ and $s(q)$ in the cosmos map. E is within $[0, 1]$, where $E = 0$ means a lossless projection from d- to m-dimensional space. Due to (9), utterances forming clusters in original space will tend to cluster also in destination space. The same holds for utterances being far apart from each other. In order to achieve the final map we apply standard steepest descent to (9).

5 Speech Data

Each test person read the German version of the "North Wind and Sun" passage which is a phonetically rich text with 108 words (71 disjunctive). It is often used in speech therapy in German speaking countries. Dependent on the degree of voice pathology, a speech sample of approx. 35 sec to 3 min duration was thus recorded and then used for the speaker-dependent adaptation of the SCHMMs by recomputing the codebook mixture weights as described in Sect. 2.

We applied the mapping to four different corpora. Firstly, a group of 18 male laryngectomees was investigated. Their larynx had been removed because of laryngeal or hypopharyngeal cancer. These speakers use tracheoesophageal substitute voice, i.e. a shunt valve between the trachea and the pharyngoesophageal segment allows to divert the expiratory air stream into the esophagus and causes voicing by tissue vibrations there. Additionally various voice and speech properties, such as hoarseness, intelligibility, pitch, speech effort etc., were evaluated by five experienced raters on a five-point scale. For a more detailed description of this data set and its recording environment see [5].

The second speaker group were 9 female and 9 male chronically hoarse speakers. Finally, two sets of normal non-pathologic speakers were used as control groups. The first subgroup contains 18 elderly, male persons, they were chosen in order to form an age-matching control group for the laryngectomees. They were recorded in the same environment as the pathologic voices [5]. The second subgroup consisted of 9 young males and 7 females forming an age-matching group

Table 1. The speaker groups

group	#speakers	avg. age	duration
Laryngectomees	18 (18m, 0f)	64.2	21.5 min
Hoarse speakers	18 (9m, 9f)	47.6	18.4 min
Normal sp. (old)	18 (18m, 0f)	65.4	15.5 min
Normal sp. (young)	16 (9m, 7f)	n/a (≈25)	12.5 min

with respect to the training speakers of the baseline recognizer.
Table 1 shows more details concerning the speaker groups.

6 Mappings of Voice Disorders

Figure 1 shows a mapping of all available speakers into a single 2-D cosmos. It is clearly visible that the different speaker groups were almost completely separated into different areas. In addition, the genders of the hoarse and young reference speakers were separated. The degree of voice pathology is growing from right to left, with the hoarse speakers located between the laryngectomees and the normal speakers. The speakers' pitch is growing from the top to the bottom of the cosmos. However, which voice properties are arranged in which direction by the Sammon Transform, is dependent on the data and not known in advance. This phenomenon was already reported in [4] where a cosmos map was suggested to have an unlimited number of axes. Most of them represent complex properties of the data and are thus difficult to describe.

Figure 2 shows an example for the visualization of human and automatic evaluation results which were mapped to a 2-D cosmos of the laryngectomized speakers. The intensity of the stars in the left mapping represents the speech effort rated by the human experts. On the right side the intensity reflects the word accuracy achieved on an SCHMM recognizer (cp. [5]) for each speaker. A strong correlation can be seen between both graphics. Speakers with high effort values are likely to have a low recognition rate and vice versa.

Another experiment was on projecting an unknown speaker into an already existing cosmos of well-known and previously evaluated cases of pathologic voices. If this is possible, then a pre-computed cosmos of various voice disorders can serve as a reference, and the new speaker's degree or even the type of pathology can be determined by the position where the recording is projected into the map. We have slightly modified Sammon's mapping method, so that the existing map stays unchanged and only the new star's coordinates are evaluated based on (9). As an example, Figure 3 shows two maps of the 18 laryngectomized speakers. The map to the left was computed all at once whereas inside the map to the right the marked star has been projected into the cosmos of the 17 remaining speakers. There is no visible difference between both maps. However, the rising mapping error which cannot be reduced if the rest of the map is kept static, will result in more incorrect projections, if a higher number of stars is projected.

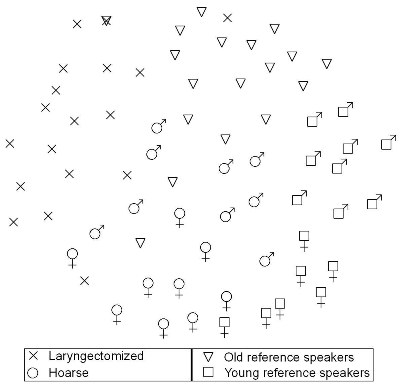

Fig. 1. Cosmos of all speaker groups and their arrangement by the Sammon Transform

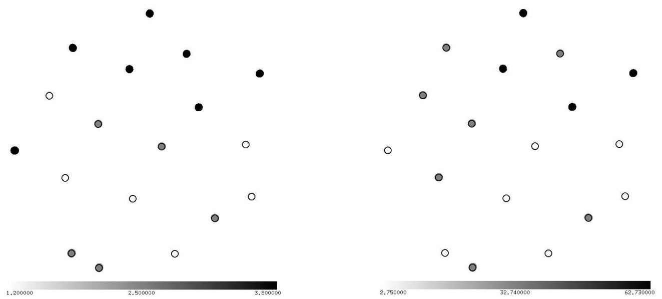

Fig. 2. Cosmos of 18 laryngectomized speakers; *Left:* Distribution of average speech effort rated by 5 human experts. Dark stars mark speakers with high effort. *Right:* Distribution of word accuracy from an automatic speech recognizer. Dark stars mark speakers with low recognition rate. A comparison of both maps shows that recognition performance can be a good indicator for speech effort.

7 Conclusions and Outlook

We believe that voice characteristics are present in the observation probabilities of semi-continuous HMMs. The weights of these HMMs after adaptation to a single speaker serve as input data for the Sammon Transform. It performs a topology-preserving dimension reduction and allows the visualization of high-dimensional feature spaces. Different voice disorders and their extend were clearly separated in 2-D space, and their relations to normal speakers became visible. The projection method allows to insert an unknown speaker into an existing cosmos map with a negligible error. This can serve as an objective and automatic diagnostic support for medical personnel, including adequate visualization of the results.

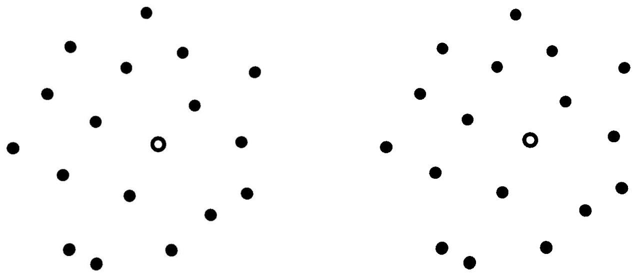

Fig. 3. *Left:* A cosmos of laryngectomized speakers computed all at once. The marked star will be removed and re-projected. *Right:* The star has been re-projected into the map. Its position can be considered identical.

The method can also help to improve the automatic recognition of people with voice disorders, e.g. in dialogue systems. The basis for this idea is a pool of robust prototype recognizers trained on speech with different disorders. When confronted with a new speaker, the system would project the speaker into a cosmos of the prototype recognizers, determine the disorder and select the "closest" recognizer or combine a set of several close recognizers for further processing.

Acknowledgments

This work was partially funded by the German Cancer Aid (Deutsche Krebshilfe) under grant 106266. The responsibility for the contents of this study lies with the authors.

References

1. Ruben, R.: Redefining the survival of the fittest: communication disorders in the 21st century. Laryngoscope **110** (2000) 241–245
2. Sammon, J.: A nonlinear mapping for data structure analysis. IEEE Trans. Computers **C-18** (1969) 401–409
3. Steidl, S., Stemmer, G., Hacker, C., Nöth, E.: Adaption in the Pronunciation Space for Non-Native Speech Recognition. In: Proc. ICSLP, Jeju Island, Korea (2004) 318–321
4. Shozakai, M., Nagino, G.: Analysis of Speaking Styles by Two-Dimensional Visualization of Aggregate of Acoustic Models. In: Proc. ICSLP, Jeju Island, Korea (2004) 717–720
5. Schuster, M., Nöth, E., Haderlein, T., Steidl, S., Batliner, A., Rosanowski, F.: Can You Understand Him? Let's Look at His Word Accuracy – Automatic Evaluation of Tracheoesophageal Speech. In: Proc. ICASSP. Volume I., Philadelphia, PA (2005) 61–64