EVALUATION AND ASSESSMENT OF SPEECH INTELLIGIBILITY ON PATHOLOGIC VOICES BASED UPON ACOUSTIC SPEAKER MODELS

Tobias Bocklet[1], *Tino Haderlein*[1,2], *Florian Hönig*[1], *Frank Rosanowski*[2], *Elmar Nöth*[1]

[1] Chair of Pattern Recognition (Computer Science 5), University of Erlangen-Nuremberg
Martensstraße 3, 91058 Erlangen, Germany
[2] Department of Phoniatrics and Pedaudiology, University of Erlangen-Nuremberg
Bohlenplatz 21, 91054 Erlangen, Germany

tobias.bocklet@informatik.uni-erlangen.de

ABSTRACT

We describe a GMM-UBM-based evaluation system for pathologic voices that uses standard cepstral features. Per speaker one GMM is created and its components are used to create a so-called GMM supervector. The supervector of each speaker is labeled with the intelligibility values obtained by human evaluation and is used to train an SVR. We studied different GMM supervectors containing different GMM components. On a database of 85 pathologic speakers, we achieved a correlation between the automatic system and the expert listeners of $r = 0.83$ when using a 13312-dimensional supervector containing the values of the diagonal covariance matrices of 26-dimensional Gaussians.

Index Terms— Gaussian Mixture Models, Support Vector Machines, Acoustic Analysis, Pathologic Voices

1. INTRODUCTION

For speech therapy, it is very important to evaluate a patient's voice by a speech therapist. Objective, automatic measures are very helpful for this task. In previous works we showed that the word accuracy (WA) of an automatic speech recognition system can be employed as objective intelligibility measure for tracheoesophageal substitute voices [1]. These are voices where the larynx has been removed completely and the source of voice is located in the pharyngo-esophageal segment. A correlation of $|r| = 0.8$ to human evaluations could be achieved. If the cancer is detected in early and intermediate stages, a partial laryngectomy, i.e., a partial removal of the larynx, is a sufficient means to stop the propagation of the cancer. The advantage is that in most cases at least one of the vocal folds or the vestibular folds can be preserved (see Figure 1). Compared to the total laryngectomy with tracheoesophageal substitute voice, the voice after partial laryngectomy sounds almost normal but commonly hoarse. This work focuses on a database of 85 pathologic speakers, before and after partial laryngectomy.

In a previous study, the correlation between the WA and the subjective raters was low ($|r| = 0.6$) [2]. It is assumed that due to the higher and more uniform voice quality of partially laryngectomized persons the WA alone is not a sufficient feature to achieve high correlations to the intelligibility scores of the expert listeners on a small evaluation text of 107 words.

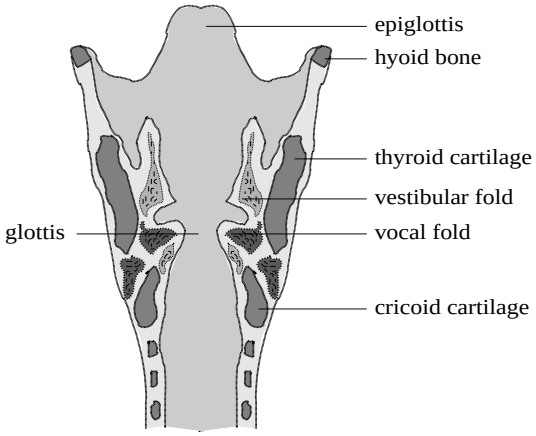

Fig. 1. Anatomy of an intact larynx

In this work we present an approach which is only based on the acoustic properties of affected speakers. To our knowledge this approach has not been used before for speech assessment. We assume that there is a similarity between these automatically computed acoustic properties and the intelligibility scores from expert listeners. The acoustic features of a speaker are modeled by *Gaussian Mixture Models* (GMMs). The system uses a *Universal Background Model* (UBM) that is adapted by *Maximum A Posteriori* (MAP) adaptation to speaker-specific spectral features. Based on the speaker-specific GMM, supervectors are extracted. Supervectors contain certain components of the GMMs. These supervectors are labeled with the experts' intelligibility score and are used as input vectors for a *Support Vector Regression* (SVR). In order to evaluate the acoustics of a speaker equally, it is important that every patient reads/speaks the same text or a text, that contains all phonemes of a certain language. Assuring this, all speaker GMMs are created with the same information and should be similar for speakers with similar pathology. Patient's with different states of pathology are supposed to be represented by different GMMs.

The remainder of the paper is organized as follows. Section 2 introduces the dataset used for training and evaluation. Section 3 gives some details on the subjective evaluation performed by expert

rater	MS	RO	SA	SU	VD
r	0.82	0.76	0.80	0.86	0.83

Table 1. Correlation between intelligibility scores of one rater and the mean value of the others

listeners and shows their inter-rater agreements. Section 4 describes the evaluation system and the different employed GMM-based supervectors. In Section 5 we present leave-one-speaker-out results achieved with these supervectors. The paper concludes with a summary and proposed future research.

2. DATA

The dataset contains recordings of 85 patients. 75 of them are males and 10 are females. They suffer from cancer in different regions of the larynx. 65 of them had already undergone partial laryngectomy, 20 speakers were still awaiting surgery. The former group was recorded on the average 2.4 months after surgery. The average age of all speakers was 60.7 years with a standard deviation of 9.7 years. The youngest and the oldest person were 34 and 83 years old, respectively.

Each person read the text "Der Nordwind und die Sonne", a phonetically balanced text with 108 words (71 disjunctive) which is used in German speaking countries in speech therapy. It contains all phonemes of the German language. The English version is known as "The North Wind and the Sun" [3]. The speech data were sampled with 16 kHz and an amplitude resolution of 16 bit. They were recorded in a quiet room in the university clinics in Erlangen and digitally stored on a server by a client/server-based system [4].

3. SUBJECTIVE EVALUATION

Five experienced phoniatricians and speech scientists evaluated each speaker's intelligibility according to a 5-point Likert scale [5] with the labels "very high", "high", "moderate", "low", and "none". For each patient five different intelligibility labels were collected. Each rater's decision was converted to an integer number between 1 and 5. As reference for the automatic evaluation, the mean value of the five ratings was calculated for each patient. The speech data were labeled with this reference value and used in the automatic system.

Table 1 shows the correlation between the intelligibility score of one rater and the mean values of the four other experts. The correlations vary from 0.76 to 0.86. This is similar to the agreement of highly pathologic speakers, i.e., speakers with tracheoesophageal substitute voices [6]. So the agreement within the five raters can be regarded as reliable.

4. EVALUATION SYSTEM

In this paper we employ an evaluation system which is based on *Support Vector Regression* (SVR) and *Gaussian Mixture Models* (GMM). We first published the system in [7] where it was used to determine the age of children. For that kind of evaluation, we assumed that the acoustics are changing during the growth of the vocal tract. We based our evaluation only on a shift of the mean vectors from the UBM to the mean vectors of a speaker GMM. In this work, our main intention was that the acoustics of a pathologic speaker differs to the acoustics of a normal speaker. Normal speakers have the possibility to pronounce a vowel or word always in the same way. Pathologic speakers have difficulties in doing that. So not only the mean vectors but also the variances seem to be important to perform an evaluation of pathologic speech. In this section we describe the system in detail.

First, short-time features in form of *Mel-Frequency Cepstrum Coefficients* (MFCCs) and *Perceptive Linear Prediction* (PLP) are extracted from the speech signal. GMMs are employed to model these features for each speaker, i.e., a GMM for each speaker is adapted from a UBM. Such a system has been used for voice disorder assessments in [8, 9]. In our system the GMM is then used as meta feature vector for an SVR, i.e., GMM-based supervector. We evaluated different kinds of supervectors. They are described in Section 4.2.

4.1. Feature Extraction

The first type of features we studied were MFCCs. A Hamming window with a size of 16 ms and a time shift of 10 ms is applied to the speech signal. The Mel spectrum with 25 triangle filters is calculated afterwards. At the end, a 24-dimensional feature vector, which contains log energy, 11 Mel-frequency cepstral coefficients and the first-order derivatives of these 12 static features is created ($D = 24$).

The second set of features is revised PLP (RPLP) [10], a simplified and improved variant of PLP [11] employing the Mel filter-bank instead of the Bark filter-bank. We took the first 13 cepstral coefficients of the PLP model spectrum and their first-order derivatives which results in a feature dimension of $D = 26$.

Since we use the same filter-bank and a simplified version of PLP, the sole difference between RPLP and MFCC is that RPLP performs an additional spectral smoothing step by applying linear prediction (LP) on the Mel filter-bank spectrum and obtaining the cepstral coefficients from the resulting PLP model spectrum [10].

4.2. GMM-based Supervectors

The basic idea behind the GMM supervector approach is to model the acoustic features of a speaker by a GMM. This is the standard approach in speaker identification/verification [12]. A GMM λ is composed of M unimodal Gaussian densities:

$$p(\boldsymbol{c}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \omega_i p_i(\boldsymbol{c}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \qquad (1)$$

$$= \sum_{i=1}^{M} \omega_i \cdot \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma_i}|^{(1/2)}} e^{-(1/2)(\boldsymbol{c}-\boldsymbol{\mu_i})^T \boldsymbol{\Sigma_i}^{-1}(\boldsymbol{c}-\boldsymbol{\mu_i})}, \qquad (2)$$

where ω_i denotes the weight, $\boldsymbol{\Sigma_i}$ the covariance matrix and $\boldsymbol{\mu_i}$ the mean vector of the i-th Gaussian density.

4.2.1. GMM Training

After extraction of the acoustic features, i.e., MFCCs or PLPs, a UBM is created with all the available training data, using the EM algorithm [13]. The UBM is then employed as an initial model for a MAP adaptation [14]. The MAP adaptation adjusts the UBM to the speaker-dependent training data in a single iteration step and combines these new densities with the UBM parameters based on a relevance parameter. The relevance parameter considers the number of speaker-specific feature vectors. Finally, a GMM is created for each

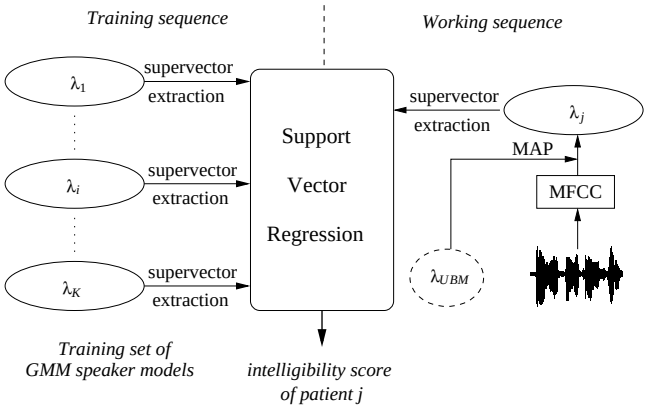

Fig. 2. Principle of the Support Vector Regression (SVR) system

speaker. We did not observe major differences between full and diagonal covariance matrices. Because of that we only present the results achieved with diagonal covariance matrices, and the notation $\mathbf{\Sigma}$ refers to the values of a diagonal covariance matrix.

4.2.2. Different Realizations of Supervectors

A GMM-based supervector is constructed by a selection of the GMM components, i.e., weights ω_i, mean vectors $\boldsymbol{\mu_i}$ and diagonal covariance matrices $\mathbf{\Sigma_i}$ of each Gaussian density i. In terms of age recognition, as described in [7], the supervector contained only the mean vectors. For the evaluation of pathologic voices, we expected the other GMM components, e.g. covariance matrices and weights, to be very important for an automatic intelligibility assessment. Due to the fact that persons with speech disorders have difficulties in pronouncing certain phones, we expected the variance, which represents the acoustic variety of a certain phoneme, to be more important for our task. In order to examine the importance of the different aspects, we defined different types of GMM-based supervectors with different GMM component combinations:

- ω (dimension: M)
- $\boldsymbol{\mu}$ (dimension: $D \cdot M$)
- $\mathbf{\Sigma}$ (dimension: $D \cdot M$)
- $\boldsymbol{\mu} + \mathbf{\Sigma}$ (dimension: $D \cdot M \cdot 2$)
- $\omega + \boldsymbol{\mu}$ (dimension: $M + D \cdot M$)
- $\omega + \mathbf{\Sigma}$ (dimension: $M + D \cdot M$)
- $\omega + \boldsymbol{\mu} + \mathbf{\Sigma}$ (dimension: $M + D \cdot M \cdot 2$)

4.3. Support Vector Regression System

Figure 2 shows the principle of the training and the working sequence of the SVR system. For each speaker of the training set, a GMM λ is created. A GMM-based supervector is created for each GMM. These supervectors are used as input vectors for the SVR training. They can be regarded as a mapping from the acoustics of a speaker (MFCCs or PLPs) to a higher-dimensional feature vector which represents the speaker himself or his/her characteristics. In the case of this paper, the property of interest is the intelligibility. If the intelligibility of a speaker j is to be determined, a speaker model λ_j is derived from the background model λ_{UBM}. Note that the background model is the same that is used to create the GMMs of the training speakers. Again, the GMM supervector is created for speaker j, and an intelligibility prediction is performed by SVR [15]. For our experiments we employed a linear kernel.

5. RESULTS

We evaluated our system on 85 patients with pathologic voices as described in Section 2. We employed the MFCC and PLP approach for acoustic feature extraction. The first part of Table 4.3 shows the results achieved with MFCCs, and the second part contains the results with PLP features. We based the experiments on the different supervectors described in Section 4.2. With MFCCs a maximum correlation to the mean value of the 5 human expert raters of $r = 0.77$ was achieved when using a supervector composed of weights, mean vectors and diagonal covariance matrices of the speaker GMM. The number of Gaussian densities in this experiment was 128, so the dimension of the supervector was $128 \cdot (24 \cdot 2 + 1) = 6272$. The best results with PLP features were achieved when using supervectors that contain the diagonal covariance matrices of GMMs with 256 Gaussians. The correlation to the human experts was $r = 0.83$ in this case.

Hence, a system based on RPLP features seems to have a clear advantage compared to an MFCC-based system (correlation of 0.83 vs. 0.77). Due to the low number of patients (85 speakers) this argument can be weakened; the improvements are not significant ($p \geq 0.1$; only five patterns are different).

However, a reason for the better performance of RPLP could be the "peak-hugging property" of LP in combination with the Mel-cepstrum features: The LP model spectrum is more influenced by the peaks than by the dips [16]. This leads to a more accurate and robust modeling of the vocal tract transfer function.

Another fact that should be mentioned is that the covariances of the Gaussians are very important to assess pathologic voices. The PLP-based systems achieved the best correlation of $r = 0.83$ when the supervector contained the 256 diagonal entries of the covariance matrices. Comparing this value with the results achieved by a supervector containing the 256 mean vectors ($r = 0.75$) the improvements are significant at the $p \leq 0.1$ significance level. Patients with pathologic voices often have difficulties in articulating one and the same phone equally at different times. Phones are modeled by weighted mixtures of different Gaussians. A different pronunciation of a specific phone in different realizations leads to Gaussians with a higher variance. So the shape and size of the variance is very important for the assessment of pathologic voices.

When comparing the results of the two systems with a supervector that contains the covariance matrix entries, one can see, that the PLP-based system outperforms the MFCC-based system. In case of GMMs with 256 Gaussians, the PLP-based system achieves a correlation coefficient of 0.83 and the MFCC-based system 0.73 respectively. This again is significant at the $p \leq 0.1$ significance level. Patients with a stronger pathology have difficulties in articulating one and the same phone similar in different realizations. This leads to a higher variance of the voice. As mentioned above, PLP features are representing the phones more precisely, i.e., different realizations are noticeable. So the higher variance of the voice is modeled directly within the GMMs, which affects the covariance matrices of the Gaussian densities. A combined supervector of means and co-

# of Gaussians	supervectors						
	ω	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	$\boldsymbol{\mu}+\boldsymbol{\Sigma}$	$\omega+\boldsymbol{\mu}$	$\omega+\boldsymbol{\Sigma}$	$\omega+\boldsymbol{\mu}+\boldsymbol{\Sigma}$
results with MFCC features							
128	0.49	0.73	0.73	0.76	0.74	0.74	**0.77**
256	0.63	0.71	0.73	0.74	0.72	0.74	0.75
512	0.70	0.72	0.73	0.74	0.73	0.76	0.75
1024	0.66	0.71	0.69	0.73	0.72	0.72	0.73
results with PLP features							
128	0.47	0.75	0.80	0.79	0.76	0.80	0.79
256	0.63	0.75	**0.83**	0.80	0.76	0.81	0.80
512	0.57	0.76	0.82	0.79	0.76	0.81	0.79
1024	0.57	0.71	0.72	0.77	0.74	0.72	0.77

Table 2. Pearson Correlations dependent on the supervector and the number of Gaussian densities with MFCC and PLP features

variance values does not improve the PLP-based system, so it seems that the covariance matrix entries of the PLP-based system are an adequate means to assess the voices of pathologic speakers.

In case of a supervector that contains the values of the diagonal covariance matrices of the Gaussian densities, the PLP-based system outperforms the MFCC-based system. But with a supervector that contains only the weights of Gaussian densities, the MFCC-based system shows slightly better correlation coefficients. Regarding GMMs with 512 densities, the differences between the MFCC-based system and the PLP-based system are significant at the $p \leq 0.05$ level. We do not have an explanation for this fact yet. Future work will focus on that.

6. CONCLUSION

In this paper we described a system for the recognition and analysis of pathologic voices based only on the acoustic properties of a patient's voice. We evaluated the system on a dataset of 85 patients with pathologic voices. The system models the acoustic properties of a person's speech by a GMM and uses this GMM as a meta feature in SVR. With this system we achieved a correlation between the GMM parameters and the intelligibility score of human experts of up to 0.83. Due to the fact that the system is only based on the acoustics of a voice, it is most likely that the system is language independent when a test covering most/all phonemes of a certain language is used. Future work will focus on this task.

7. REFERENCES

[1] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nöth, "Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 70, pp. 1741–1747, 2006.

[2] T. Haderlein, E. Nöth, A. Maier, M. Schuster, and F. Rosanowski, "Influence of Reading Errors on the Text-Based Automatic Evaluation of Pathologic Voices," in *Proceedings Text, Speech and Dialogue; 11th International Conference*, Petr Sojka, Ales Horak, Ivan Kopecek, and Karel Pala, Eds., Berlin, 2008, Lecture Notes of Artificial Intelligence, pp. 325–332.

[3] International Phonetic Association (IPA), "Handbook of the International Phonetic Association," Cambridge University Press, 1999.

[4] A. Maier, T. Haderlein, M. Schuster, and E. Nöth, "PEAKS – a Platform for Evaluation and Analysis of all Kinds of Speech Disorders," in *Proc. 41st Annual Meeting of the Society for Biomedical Technologies of the Association for Electrical, Electronic & Information Technologies (BMT 2007)*, Aachen, Germany, 2007.

[5] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 140, 1932, Columbia University, New York, NY (USA).

[6] Tino Haderlein, *Automatic Evaluation of Tracheoesophageal Substitute Voices*, Logos Verlag, 2007.

[7] T. Bocklet, A. Maier, and E. Nöth, "Age Determination of Children in Preschool and Primary School Age with GMM-Based Supervectors and Support Vector Machines/Regression," in *Proceedings Text, Speech and Dialogue; 11th International Conference*, Petr Sojka, Ales Horak, Ivan Kopecek, and Karel Pala, Eds., Heidelberg, 2008, vol. 1 of *Lecture Notes in Artificial Intelligence*, pp. 253–260.

[8] C. Fredouille, G. Pouchoulin, J.F. Bonastre, M. Azzarello, A. Giovanni, and A. Ghio, "Application of Automatic Speaker Recognition techniques to pathological voice assessment (dysphonia)," in *Proc. Interspeech*, 2005, pp. 149–152, 2183.

[9] J.-F. Bonastre, C. Fredouille, A. Ghio, A. Giovanni, G. Pouchoulin, J. Revis, B. Teston, and P. Yu, "Complementary approaches for voice disorder assessment," in *Proc. Interspeech*, 2007, pp. 1194–1197.

[10] F. Hönig, G. Stemmer, C. Hacker, and F. Brugnara, "Revising Perceptual Linear Prediction (PLP)," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, ISCA, Ed., Bonn, Germany, 2005, pp. 2997–3000.

[11] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustic Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, pp. 19–41, 2000.

[13] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[14] J.L. Gauvain and C.H. Lee, "Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[15] A.J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[16] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.