Voice Pathology Detection on the Saarbrücken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit

David Martínez, Eduardo Lleida, Alfonso Ortega, Antonio Miguel, and Jesús Villalba

Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
{david,lleida,ortega,amiguel,villalba}@unizar.es

Abstract. The paper presents a set of experiments on pathological voice detection over the Saarbrücken Voice Database (SVD) by using the MultiFocal toolkit for a discriminative calibration and fusion. The SVD is freely available online containing a collection of voice recordings of different pathologies, including both functional and organic. A generative Gaussian mixture model trained with mel-frequency cepstral coefficients, harmonics-to-noise ratio, normalized noise energy and glottal-to-noise excitation ratio, is used as classifier. Scores are calibrated to increase performance at the desired operating point. Finally, the fusion of different recordings for each speaker, in which vowels /a/, /i/ and /u/ are pronounced with normal, low, high, and low-high-low intonations, offers a great increase in the performance. Results are compared with the Massachusetts Eye and Ear Infirmary (MEEI) database, which makes possible to see that SVD is much more challenging.

Keywords: Pathological Voice Detection, Saarbrücken Voice Database, GMM, Fusion, MultiFocal toolkit.

The detection of laryngeal pathologies through an automatic voice analysis is one of the most promising tools for speech therapists, mainly due to its noninvasive nature and its objectivity for making decisions. The performance of these systems is nevertheless not perfect, and nowadays it is used as an additional source of information for other laryngoscopial exams [1].

Researchers have focused their efforts on finding new features that could discriminate between normal and pathological voices or even assess their quality, but also on finding different approaches for classification. Some of the most useful features are considered to be acoustic parameters such as mel-frequency cepstral coefficients (MFCC) [17, 1], amplitude and frequency perturbation parameters [9], and noise related parameters [2, 6], but there are different alternatives like nonlinear analysis [3, 4]. Regarding to the classifiers, well-known approaches in speech processing like hidden Markov models (HMM) [7], Gaussian mixture models (GMM) [1], multilayer perceptrons (MLP) [8], or support vector machines (SVM) [9], have been studied.

D.T. Toledano et al. (Eds.): IberSPEECH 2012, CCIS 328, pp. 99–109, 2012.

Most of the works in the literature make use of the MEEI database, openly commercialized by *Kay Elemetrics* [12]. In other cases, private databases collected in local hospitals are the alternative. However, recently a new open and freely downloadable database, the SVD [13], has been recorded by the Institute of Phonetics of Saarland University. On it, sustained /a/, /i/ and /u/ vowels, pronounced with normal, low, high and low-high-low intonations, and a spoken sentence in German, are found, what make of this database a very complete set to conduct experiments, and easy to reach by all the community. No previous results for voice pathology detection have been found on it, and with this work we also aim at proposing a baseline for future research.

Experiments related with pathological voices can be focused on three main tasks. While the most simple and direct idea is to classify voices as pathological or normal, like in [2, 6, 7, 10, 11], another goal is to assess voice quality according to a perceptual scale, like GRBAS [24, 25], DSI [27, 28] or VPA [26], among others. A third typical problem is to identify a specific pathology, like for example, functional dysphonia in [28], nodules and other laryngeal injuries in [9], or polyps, keratosis leukoplakia, adductor spasmodic dysphonia and vocal nodules in [5].

The work developed in [11] explores different configurations for a GMM classifier to detect pathological voices, fed with MFCC, harmonics-to-noise ratio (HNR) [14], normalized noise energy (NNE) [15] and glottal-to-noise excitation ratio (GNE) [16]. The performance is tested on the MEEI database, and only files with recordings of vowel /ah/ sustained are used. A 30-fold strategy is followed and several random partitions are created to average results. The best performance is obtained with 3 Gaussians and 16 MFCCs, and the area under the curve (AUC) [18] is 0.98459, with the 95.49% of the pathological files and 90.70% of the normal files correctly classified. In addition, the same study is performed in different environments, such as MP3 compression and telephone channel distortion.

In this work, we have tried to follow the guidelines marked in [11] to discriminate between normal and pathological voices, extrapolating the techniques to the SVD, and taking benefit of the MultiFocal toolkit [19] for fusing different subsystems. MultiFocal toolkit is a toolkit developed by Niko Brümmer, and is widely used among the speaker and language recognition community.

The rest of the paper is organized as follows: in Section 1, the MEEI database and the SVD are presented; in Section 2, the features extracted from the audio are described; in Section 3, it is detailed how a GMM classifier works; in Section 4, a brief description of MultiFocal toolkit is given: in Section 5, the experiments that have been performed are presented and analyzed, for the two databases mentioned above; and in Section 6, the conclusions of this work are drawn.

1 Databases

1.1 MEEI, *Kay Elemetrics*

The same configuration adopted in [2, 6, 11] has been taken for the present work. There, 226 recordings of the whole database were used, corresponding to vowel /ah/ sustained. From this subset, 173 files belong to pathological patients and 53 to normal speakers. Male and female speakers covering ages from 21 to 59 are uniformly distributed in both groups. The mean length of pathological recordings is 1 second and the one of normal recordings is 3 seconds. All files are converted to a common sampling frequency of 25 kHz and 16-bit resolution.

1.2 SVD, Saarland University

This database has been recently made freely available online [13]. It is a collection of voice recordings from more than 2000 persons, where a session is defined as a collection of:

- recordings of vowels /a/, /i/, /u/ produced at normal, high, low and low-high-low pitch.
- recording of sentence "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?").

That makes a total of 13 files per session. In addition, the electroglottogram (EGG) signal is also stored for each case in a separate file. The length of the files with sustained vowels is between 1 and 3 seconds. All recordings are sampled at 50 kHz and their resolution is 16-bit. 71 different pathologies are contained, including both functional and organic. For our experiments only files with sustained vowels and people older than 18 are used. A total of 1970 sessions are kept, after discarding those where some of the recordings were missing or damaged. 1320 (609 males and 711 females) sessions belong to pathological speakers and 650 (400 males and 250 females) to normal speakers.

2 Features

The features used in this work are divided in two groups, according to their nature: acoustic features, represented by the MFCC, where the aim is to characterize the frequency content of the signal; and noise related features, represented by HNR, NNE and GNE, where the aim is to measure how good the quality of the signal is, or simply, how noisy it is.

2.1 Acoustic Features

MFCC are a family of parameters widely used for many tasks related with speech processing. It makes a frequency analysis of the signal based on the human perception of the sounds. This idea matches well with the fact that an experienced speech therapist can detect the presence of a disorder just by listening to the signal [10].

In the extraction procedure, after downsampling to 25 kHz, a 40 ms window with 50% overlap has been used, with a bank of 30 Mel filters, to obtain 15 MFCC plus log-energy. The first two and last two frames have been discarded to avoid possible errors in the edges of the recordings, like peaks due to the on and off switches. Finally, the coefficients are mean and variance normalized within each file.

2.2 Noise Related Features

Harmonics-to-Noise Ratio. HNR was introduced to measure in an objective manner the perceptual feeling of hoarseness in the voice [14]. To calculate it, the signal is firstly downsampled to 16 kHz, and split into 25 ms length frames, with 10 ms shift. In each frame, a comb filter is applied to the signal to compute the energy in the harmonic components. To the logarithm of this quantity, the log-energy of the noise is substracted to get the HNR.

Normalized Noise Energy. In a similar process to the calculation of the HNR, and also with the signal downsampled to 16 kHz and with 25 ms length frames and 10 ms shift, the noise estimation is calculated and normalized by the total energy of the signal. This was first used in [15] and it assumes that pathological voices are noisier than normal voices.

Glottal-to-Noise Excitation Ratio. The goal of this parameter is to compare the amount of signal due to vocal folds vibration with the amount of signal due to noise produced by air turbulences produced during phonation [16]. It is a good measurement of breathiness, although not the only factor that can cause it. To compute it, the signal is first downsampled to 10 kHz, and frames of 40 ms length with 20 ms shift are taken. For each frame, the spectrum is divided into bands of 2000 Hz with centers separated 500 Hz. For each of these bands, the Hilbert envelope in time domain is calculated and the correlation of this envelope with the envelopes of the bands separated more than half of the bandwidth (in this case, bands must be at least 1000 Hz) is computed. The GNE is the maximum of all correlations. If the voice is not pathological, the correlation should be high, because all bands should be excited at the same time when the glottis is closed.

3 GMM Classifier

The features extracted from the signal are used to train a generative GMM model [22] for each class. This model is the basis for many speech processing tasks, like speech, speaker, or language recognition. It is a generalization of the Gaussian model, and it permits to generate much more complicated likelihood functions.

For D-dimension features $\mathbf{x}$ calculated in a frame-by-frame basis, a GMM probability density function has the form

$$p(\mathbf{x}|\omega, \mu, \mathbf{\Sigma}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k), \tag{1}$$

where K is the number of Gaussians in the model, ω_k is the weight of the k*th* Gaussian, and $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ is the Gaussian function with mean μ_k and covariance Σ_k.

To train this model the expectation-maximization (EM) algorithm [23] has been used. K random Gaussians are generated for initialization and 10 iterations of the algorithm are performed. Full-covariance matrices are used. Then for each test file y, the likelihoods for pathological and normal classes are calculated, calibrated as explained in next section, and the log-likelihood ratio between them is obtained as

$$LLK(y) = \log p(y|pathological) - \log p(y|normal), \tag{2}$$

which will decide to which class the file belongs.

4 Calibration and Fusion with MultiFocal Toolkit

In pathology detection, the traditional metrics to evaluate the performance of the classifiers are:

- the area under the receiver operating characteristic (ROC) curve (AUC) [18]
- equal-error-rate (EER), point in ROC where the probability of miss is equal to the probability of false alarm
- correct classification rate (CCR), percentage of trials correctly classified
- error rate (ER), percentage of trials wrong classified, complementary to CCR
- sensitivity (S), ability of the classifier to detect the target class
- specifity (E), ability of the classifier to detect the impostor class

AUC and EER are calibration-insensitive evaluation metrics, and the last four are calibration-sensitive metrics. Since we are interested in a specific operating point, that is, in the hard decisions made by our classifier, we find more interesting to evaluate the real performance of our classifier with a calibration-sensitive metric than with AUC and EER. AUC and EER can be useful in early stages of our system development, when hard decisions are not of immediate interest and we are only interested in the goodness of uncalibrated scores [20].

In addition, two more metrics that we consider meaningful are the detection cost function (DCF) or empirical Bayes risk, and its minimum value for the selected operating point (minDCF) [19]. DCF is defined as

$$DCF = \pi C_{miss} P_{miss} + (1 - \pi) C_{fa} P_{fa}, \tag{3}$$

where π is the prior probability of the target class, in our case the pathological voice, C_{miss} and C_{fa} are the costs of a miss, that is, a pathological voice classified as normal, and a false alarm, that is, a normal voice classified as pathological, respectively, and P_{miss} and P_{fa} are the probabilities of a miss and a false alarm. It is a calibration-sensitive metric, since it depends on the current threshold. However, minDCF is calibration-insensitive, and it gives the minimum cost that could have been obtained with optimal calibration, at every operating point. It is calculated by varying the threshold from $-\infty$ to ∞ for each operating point, and then picking the minimum.

MultiFocal is a toolkit developed in Matlab primarily designed for calibrating and fusing scores of a language recognition task [19]. The aim of using this toolkit is twofold: *i)* to calibrate scores so cost effective Bayes decisions can be made, by setting the threshold to the *Bayes decision threshold*, η,

$$\eta = \log \frac{C_{fa}}{C_{miss}} - logit(\pi); \tag{4}$$

ii) to fuse scores coming from different recognizers to obtain a better recognizer.

The idea behind calibration is that scores are converted in such a way that the Bayes decision threshold can be used for making the best possible decisions. Equivalently, the user could tune the threshold manually to minimize the error metric.

To calculate calibrated log-likelihoods, MultiFocal optimizes another calibration-sensitive metric, C_{llr} [21]. C_{llr} is defined as

$$C_{llr} = -\frac{1}{T}\sum_{t=1}^{T} \omega_t \log_2 P_t, \tag{5}$$

where T is the number of trials, ω_t is a weight to normalize the class proportions in the evaluation trials,

$$\omega_t = \frac{\pi_{c(t)}}{Q_{c(t)}} \quad (6)\ , \qquad\qquad Q_i = \frac{nr.\,of\,trials\,of\,class\,H_i}{T}, \quad (7)$$

c(t) is the true class of trial t, π_i the prior of class i, P_t is the posterior probability of hypothesis $H_{c(t)}$ of true class given the vector of calibrated log-likelihoods, $\boldsymbol{l'}(x_t)$, at trial t,

$$P_t = P(H_{c(t)}|\boldsymbol{l'}(x_t)) = \frac{\pi_{c(t)}e^{l'_{c(t)}(x_t)}}{\pi_{c(t)}e^{l'_{c(t)}(x_t)} + \pi_{\bar{c}(t)}e^{l'_{\bar{c}(t)}(x_t)}}, \tag{8}$$

being $\bar{c}(t)$ the impostor class at trial t, and x_t the observation at t. As we are in a 2-class problem, $c(t)$ ∈{'0','1'}, being '0' the label for the target class, and '1' the label for the impostor class.

C_{llr} has the sense of a cost and it is measured in terms of bits of information. $0 \leq C_{llr} \leq \infty$, being 0 for perfect recognition.

Well-calibrated log-likelihoods, $\boldsymbol{l'}(x_t)$, are the final output of our calibration procedure. They are obtained as,

$$\boldsymbol{l'}(x_t) = \alpha \boldsymbol{l}(x_t) + \boldsymbol{\beta}, \tag{9}$$

where $\boldsymbol{l}(x_t)$ is the uncalibrated log-likelihood obtained from the classifier. Then, through the mimization of C_{llr}, we obtain the scalar α, that scales our outputs, and the vector β, that shifts our outputs. The optimization is made via a discriminative logistic regression.

More generally, to fuse K systems what we want is our calibrated log-likelihoods to be a linear combination of the uncalibrated log-likelihoods of the K systems,

$$\boldsymbol{l'}(x_t) = \sum_{k=1}^{K} \alpha_k \boldsymbol{l_k}(x_t) + \boldsymbol{\beta}. \tag{10}$$

As we can check, the fusion is a generalization of the calibration of a single system (where K=1), and since the fusion is also a calibration, because of the linearity of the operation, there is no need to pre-calibrate each input system, or to post-calibrate the fusion [19].

5 Experiments

The experiments conducted in this work are divided in two, according to the database tested. First, results for the MEEI datatabase will be shown. We have followed a similar procedure to [11]. This will be useful to compare our classifier with a state of the art system. After this check, our classifier will be used to test the SVD. This database will make possible to show how the fusion of different sources of information increase the performance of the system. The results will be given in terms of the traditional metrics described in Section 4, AUC, EER, CCR, ER, S and E, and also in terms of DCF and minDCF as additional information to check how good the calibration is. Finally, a confidence interval (CI) at 95% confidence ($\alpha = 0.05$) will be given for each experiment. Note that it is computed over the CCR.

For all our experiments, $C_{fa} = C_{miss} = 1$, $\pi_0 = \pi_1 = 0.5$, and threshold equal to the Bayes decision threshold, in our case η=0.

5.1 Results on MEEI

The database is divided in 30 folds, in the same manner as in [11]. For every test fold, the remaining 29 are used for training. Then, an average performance measure is extracted from the 30. GMMs with 3 components are trained. This is the optimal number found in [11]. One difference with [11] is that all recordings of the same speaker are grouped into the same fold, in such a way that one speaker is not in the training and test subsets at the same time. This will avoid recognizing the speaker instead of the pathology. Note that a slight drop in performance could be seen as a consequence, compared to the experiments in [11].

In table 1, results between normal and pathological classes are shown, where every recording is considered as a trial. The features are 19 dimensional, including 15 MFCCs plus log-energy, HNR, NNE and GNE, previously normalized in mean and variance. A comparison between results with and without calibration is done. As we can see, the calibration-insensitive metrics, AUC and EER, do not change. However, S, E, CCR and ER, indicate that something has changed. With calibration our system detects better both normal and pathological classes. The reason is that the scores have been transformed in such a way that the posterior probabilities of the true class are maximized, what at the same time minimizes the Bayes risk, because now the Bayes decision threshold will be the optimal one to make decisions. Note that we have trained the calibration with the data under test, and this gives the optimal values for α and β. In a real system, these values should be trained before any clinical evaluation, and the performance would not probably be optimal. However, in this way an upper bound of the results is obtained, and this is more reliable for comparison with other systems tested over

Table 1. Evaluation metrics for the experiments on MEEI database. Averages over 30 folds.

Metric	AUC	EER	CCR	ER	S	E	CI	DCF	minDCF
Calibrated	0.943	**0.048**	**0.948**	**0.052**	0.949	**0.950**	0.071	**0.050**	**0.033**
Uncalibrated	0.943	0.048	0.923	0.077	0.950	0.850	0.123	0.099	0.033
Work in [11]	**0.985**	-	0.943	0.057	**0.955**	0.907	**0.034**	0.069	-

the same data, since there is no dependence on any development data used for training the calibration.

We consider DCF a reasonable criterion to choose one classifier or another, because it weights both kind of errors, misses and false alarms, at the desired operating point. minDCF will tell us how good our system could have been with a perfect calibration. In turn, AUC and EER are evaluation metrics of the performance of our system considering all operating points. Also in table 1, DCF and minDCF can be found for the experiments made with MEEI. In [11] they have actually worked at the operating point given by EER, but we do not know the values for C_{fa}, C_{miss} and π. Assuming they are the same as ours (which can be an unfair but reasonable assumption since they work at EER and their effective prior is 0.5, what would give those values of C_{fa}, C_{miss} and π, for η =0), in terms of DCF, our system performs better. However, their AUC is very good, and it would probably give better estimated DCF if a calibration had been performed.

5.2 Results on SVD

In this case, 12 subsets of data are created by grouping separately the recordings belonging to /a/, /i/, and /u/, pronounced with normal, low, high, and low-high-low intonation. Then, for each one of these subsets a 30-fold strategy is followed. Also 3 components are trained in the GMM. In this case, grouping of the same speaker into the same subset is not guaranteed, what could give optimistic results. In short, the same procedure as in Section 5.1 is followed for each subset. In table 2, results for the same 19 dimension features as above are shown in terms of AUC, EER, CCR, ER, S, E, CI, DCF, and minDCF. Only calibrated results are shown. The behavior of the classifier for each vowel and intonation is interesting. It seems that the recognition rate is slightly better for /a/, but the differences are small. It can be checked that for /a/, normal and low intonations help the most, for /i/ all intonations behave similar, and for /u/, the normal intonation is the least discriminative.

Next, a partial fusion is made for each vowel, where the 4 intonations of each one are fused. This is made for every fold and then all folds are averaged. In table 2 this is in the last line of each vowel. It can be seen that the results are improved with regard to the case with every vowel and intonation tested individually, and that the classifier built with /a/ outperforms the ones with /i/ and /u/. However, looking at the confidence interval, no definitive conclusions should be made.

Table 2. Evaluation metrics for the experiments on SVD. Averages over 30 folds. Intonations are N: normal; L: low; H: high; and LHL: low-high-low.

Metric	AUC	EER	CCR	ER	S	E	CI	DCF	minDCF
Vowel /a/									
N	**0.747**	**0.321**	**0.670**	**0.330**	0.636	**0.739**	**0.112**	**0.313**	**0.270**
L	0.743	0.335	0.656	0.340	0.650	0.680	0.114	0.334	0.286
H	0.722	0.336	0.666	0.334	**0.655**	0.687	0.112	0.328	0.285
LHL	0.702	0.353	0.645	0.355	0.640	0.655	0.114	0.352	0.304
Fusion /a/	**0.804**	**0.277**	**0.718**	**0.282**	**0.701**	**0.752**	**0.108**	**0.273**	**0.234**
Vowel /i/									
N	0.702	0.350	**0.645**	**0.355**	0.627	0.682	**0.114**	**0.345**	0.305
L	**0.705**	**0.348**	0.642	0.358	0.620	**0.687**	0.115	0.347	**0.303**
H	0.700	0.354	0.640	0.359	0.629	0.664	0.115	0.352	0.305
LHL	0.679	0.373	0.639	0.361	**0.652**	0.612	0.116	0.368	0.322
Fusion /i/	**0.783**	**0.283**	**0.710**	**0.290**	**0.694**	**0.741**	**0.110**	**0.282**	**0.247**
Vowel /u/									
N	0.706	0.354	0.634	0.366	0.615	0.671	0.116	0.357	0.307
L	0.712	**0.342**	0.646	0.354	0.624	**0.692**	0.115	0.342	0.301
H	0.713	0.348	0.640	0.356	0.619	0.684	0.116	0.348	0.293
LHL	**0.715**	0.344	**0.666**	**0.334**	**0.678**	0.642	**0.114**	**0.340**	**0.293**
Fusion /u/	**0.797**	**0.282**	**0.715**	**0.284**	**0.702**	**0.741**	**0.108**	**0.278**	**0.242**
Fusion	**0.879**	**0.206**	**0.794**	**0.206**	**0.778**	**0.826**	**0.095**	**0.198**	**0.165**

Finally, a global fusion with all vowels and intonations is made for each fold, and as before, all folds are averaged to obtain a single figure for each of the metrics. This is in the last line of table 2. Actually, we do not fuse different systems, but the same system trained on different data. A huge increase in performance is obtained. Comparing with the best result without fusion (/a/ normal), the increase in perfomance is 17.67% for the AUC, 35.83% for the EER, and 36.74% for the DCF. We believe that the main reason is the fact of having much more data, and containing different information, because they come from different vowels and intonations. Again, these results are optimal in terms of the fusion, since the fusion parameters have been trained on the test data. If we compare this fusion with the partial fusion of each vowel, it can be seen that all vowels contribute to the improvement, because the global fusion outperforms the partial ones.

6 Conclusions

A new voice database, the SVD, has been evaluated for the task of pathology detection. The amount of recordings of different sounds and intonations included in this database makes possible to conduct different and interesting experiments. This is an open and free database available online. A robust GMM with 3 components, trained on MFCC, HNR, NNE and GNE, has been used as classifier, and the effect of calibration has been shown. Finally, a fusion of the classifiers trained on /a/, /i/ and /u/, pronounced with normal, low, high and low-high-low intonations, has been performed, showing that every sound gives different information to the system and their combination offers a huge improvement: 17.67% for the AUC and 36.75 % for the DCF. As future work we plan to test the effect of the number of Gaussians on the performance and other methods for fusing, such as a simple concatenation of files of the same session. In addition, this work is thought to be a starting point for a further research with the SVD database, which is currently being perceptually classified according to the GRABS scale by a speech therapist of the *Bioingeniería y Optoelectrónica* (ByO) group at the Universidad Politécnica de Madrid.

Acknowledgements. Thanks to Manfred Pützer and his team for the great work of recording the SVD.

This work was funded by the Spanish Ministry of Science and Innovation under projects TIN2011-28169-C05-02 and INNPACTO IPT-2011-1696-390000.

References

[1] Godino Llorente, J.I., et al.: Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters. IEEE Tr. Biomed. Eng. 53(10) (2006)

[2] Sáenz-Lechón, N., et al.: Methodological Issues in the Development of Automatic Systems for Voice Pathology Detection. Biomed. Signal Proc. and Control 1(2) (2006)

[3] Jiang, J.J., Zhang, Y.: Nonlinear Dynamic Analysis of Speech from Pathological Subjects. Electron. Lett. 38(6) (2002)

[4] Zhang, Y., Jiang, J.J.: Nonlinear Dynamic Analysis in Signals Typing of Pathological Human Voices. Electron. Lett. 39(13) (2003)

[5] Markaki, M., Stylianou, Y.: Using Modulation Spectra for Voice Pathology Detection and Classification. In: Proc. IEEE EMBS Annual Intern. Conf., Minneapolis, MN (2009)

[6] Parsa, V., Jamieson, D.G.: Identification of Pathological Voices Using Glottal Noise Measures. J. Speech, Lang. and Hearing Res. 43(2) (2000)

[7] Gavidia-Ceballos, L., Hansen, J.H.L.: Direct Speech Feature Estimation Using an Iterative EM Algorithm for Vocal Fold Pathology Detection. IEEE Tr. Biomed. Eng. 43(4) (1996)

[8] Tadeusiewicz, R., et al.: The Evaluation of Speech Deformation Treated for Larynx Cancer Using Neural Network and Pattern Recognition Methods. In: Proc. EANN 1998 (1998)

[9] Gelzinis, A., et al.: Automated Speech Analysis Applied to Laryngeal Disease Categorization. Comput. Methods Programs Biomed. 91 (2008)

[10] Arias-Londoño, J.D., et al.: On Combining Information from Modulation Spectra and Mel-Frequency Cepstral Coefficients for Automatic Detection of Pathological Voices. Logop. Phoniatrics Vocology (2010)

[11] Sáenz Lechón, N.: Contribuciones Metodológicas para la Evaluación Objetiva de Patologías Laríngeas a partir del Ánalisis Acústico de la Voz en Diferentes Escenarios de Producción. PhD Thesis (2010)

[12] Kay Elemetrics Corp., Disordered Voice Database, Version 1.03 (CD-ROM), MEEI, Voice and Speech Lab, Boston, MA (October 1994)

[13] Barry, W.J., Pützer, M.: Saarbrücken Voice Database, Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.uni-saarland.de/

[14] Yumoto, E., et al.: Harmonics-To-Noise Ratio as an Index of the Degree of Hoarseness. J. Acoust. Soc. Am. 71 (1982)

[15] Kasuya, H., et al.: Normalized Noise Energy as an Acoustic Measure to Evaluate Pathologic Voice. J. Acoust. Soc. Am. 80(5) (1986)

[16] Michaelis, D., et al.: Glottal-to-Noise Excitation Ratio. A New Measure for Describing Pathological Voices. Acustica/Acta Acustica 83 (1997)

[17] Davis, S.B., Mermelstein, P.: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Tr. Acoust. 28(4) (1980)

[18] Hanley, J.A., McNell, B.J.: The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. Radiology 143 (1982)

[19] Brümmer, N.: FoCal Multi-class: Toolkit for Evaluation, Fusion and Calibration of Multi-class Recognition Scores - Tutorial and User Manual, `http://sites.google.com/site/nikobrummer/focalmulticlass`

[20] Brümmer, N.: The BOSARIS ToolkitUser Guide: Theory, Algorithms and Code for Binary Classifier Score Processing, `http://sites.google.com/site/bosaristoolkit`

[21] Brümmer, N., du Preez, J.A.: Application-Independent Evaluation of Speaker Detection. Computer Speech and Language 20(2-3) (2006)

[22] Reynolds, D.A., Rose, R.C.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Models. IEEE Tr. on Speech and Audio Proc. 3 (1995)

[23] Dempster, A.P., et al.: Maximum Likelihood from Incomplete Data Via the EM Algorithm. J. of the Royal Statistical Society 39, Series B (1977)

[24] Hirano, M.: Clinical Examination of Voice. Springer, New York (1981)

[25] Sáenz-Lechón, N., et al.: Automatic Assessment of Voice Quality According to the GRBAS scale. In: Proc. 28th IEEE EMBS Annual Intern. Conf. (2006)

[26] Carding, P., et al.: Formal Perceptual Evaluation of Voice Quality in the United Kingdom. Logop. Phoniatrics Vocology 25 (2000)

[27] Wuyts, F., et al.: The Dysphonia Severity Index: An Objective Measure of Vocal Quality Based on a Multiparameter Approach. J. Speech, Lang. and Hearing Res. 43 (2000)

[28] Hakkesteegt, M.M., et al.: The Relationship between Perceptual Evaluation and Objective Multiparametric Evaluation of Dysphonia Severity. J. of Voice 22 (2008)