Voice Pathology Detection and Classification Using MPEG-7 Audio Low-Level Features

Ghulam Muhammad, Moutasem Melhem

Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

ghulam@ksu.edu.sa, moutasem10@hotmail.com

Abstract

In this paper, a new pathological voice detection and pathology classification method based on MPEG-7 audio low-level features is proposed. MPEG-7 features are originally used for multimedia indexing, which includes both video and audio. Indexing is related to event detection, and as pathological voice is a separate event than normal voice, we show that MPEG-7 audio low-level features can do very well in detecting pathological voices, as well as classifying the pathologies. The experiments are done on a subset of sustained vowel (namely, "AH") recordings from healthy and voice pathological subjects, from the MEEI database. For classification, support vector machine (SVM) with 10-fold cross-validation is applied. The proposed method with MPEG-7 audio features and SVM classification is evaluated on voice pathology detection, as well as pathology classification. The experiment results show that the proposed method outperforms some recent methods in the literature both in detection and in classification. The proposed method is able to achieve an accuracy of 99.994 ± 0.0105% for detecting pathological voices and an accuracy of 100% for binary pathologies classifying.

Index Terms: MPEG-7 audio features, dysphonia recognition, support vector machines, pathology binary classification, Fisher discrimination ratio

1. Introduction

Analysis of acoustic signals of human voice has many purposes and can be seen from different points. From a technological point of view, there is a quick and huge growth for the need to store, code, transmit, recognize and synthesize voice signals. From a health science point of view, it was proved that the human health condition and the pathological status do affect the human voice [1]. The quality of voice can be heavily affected, and is related to the health conditions of vocal folds and their functionality. If the vocal folds become inflamed, some growth may develop on them or they become not functioning effectively, and hence, the speech production process will differ from normal conditions. In cases with disordered voice, speech samples carry symptoms of disorder from their origin. As a result any abnormality in larynx most likely will affect the quality of voice signal and its characteristics. Some of the most spreading disorders are vocal fold paralysis, vocal fold nodules, adductor spasmodic dysphonia, keratosis and others [2, 3, 4]. Many techniques like stroboscopy, laryngoscopy, and endoscopy are widely used by physicians to diagnose voice disorders, the aim of pathology detection is to be able to diagnose at early stages of occurrence before they lead to critical conditions. However, using these techniques can cause discomfort to the patients and need costly tools. Suggesting a new automated method will reduce time, cost, and making inspection easier and lead to patient's comfort during inspection.

Voice perturbation and quality measures, such as jitter and shimmer, depend on accurate extraction of fundamental frequency and the amplitude of various waveform types. The extraction method directly affects the accuracy of the measure, particularly if several waveform types (with or without formant structure) are under construction and if noise and modulation are present in the signal [5]. Lieberman proposed the first acoustic voice parameter in pathological voice analysis in 1961 [6]. In [7], a method was developed for short-time jitter and the AUC (area under curve) reached 94.82%. In [8], detection of pathological voice was done by means of Gaussian mixture models and short-term mel cepstral parameters complemented by frame energy together with first and second derivatives. The method was applied on a subset samples from the MEEI database, and a detection accuracy of 94.07% was achieved. In [9], modulation spectra method was suggested to identify pathological voices, and this method was applied on the same subset used in [8], and a detection accuracy of 94.1% was achieved. In [10], a new method which is based on extracting eleven features using nonlinear analysis of time series was introduced; the samples which were used in the experiment are the same samples used in [8]. A detection accuracy of 98.23% was obtained.

In this paper we explain and evaluate detection and classification of pathological voices using MPEG-7 low-level audio features. In the MPEG-7 part 4, the audio framework contains low-level tools designed to provide a basis for construction of higher-level audio applications. Low-level audio descriptors consist of a collection of simple, low complexity features [11]. These features are extracted from a subset of the MEEI database [12] with sustained vowel samples. Ten-fold cross validation using support vector machines (SVM) is used and results are expressed in terms of accuracy, sensitivity, specificity and AUC. The contribution of the paper is to evaluate the MPEG-7 audio low-level features, which were tested in many other detection applications, on voice pathology detection and classification.

The rest of the paper is organized as follows. Section 2 presents the proposed voice pathology detection method, Section 3 gives experimental setup, Section 4 presents results, and finally, Section 5 draws some conclusions.

2. Proposed method

This section describes the proposed method of voice pathology detection. In the proposed method, MPEG-7 audio low level features are extracted from sustained-vowel input signals. An optional feature selection module based on Fisher discrimination ratio (FDR) is applied on the extracted descriptors. SVM with radial basis function (RBF) is used for detection and classification.

2.1. Features: MPEG-7 audio low-level features

MPEG-7 Part 4 audio features can be applied to all forms of audio content. It has restriction neither on the encoding format or medium of the audio, nor on the audio content whether it is with or without music, speech, sound effects or others. The MPEG-7 features are originally proposed for multimedia indexing, which contains both video and audio parts. Since the invention of MPEG-7 features, they have been used in many applications such as speaker recognition, environment recognition [13], audio sports event detection [14], musical instrument classification [15] and musical onset detection [16], but never tested on pathology detection and classification, as a result, we use it in this paper for voice pathology detection and classification.

The MPEG-7 audio features that we extract here are low-level features which are of two types: scalar and vector. There are 45 features in total; four vector-type features: Audio Spectrum Envelope (3 features), Audio Spectrum Flatness (ASF) (22 features), Audio Spectrum Basis (2 features) and Audio Spectrum Projection (2 features). The others are scalar-type features: Audio Wave Form (2 features; Min and Max), Audio Power, Audio Spectrum Centroid (ASC), Audio Spectrum Spread (ASS), Audio Harmonicity (2 features; Harmonic Ratio and Upper Limit of Harmonicity), Audio Fundamental Frequency (2 features; Raw and Weight), Log Attack Time, Temporal Centroid, Spectral Centroid, Harmonic Spectral Centroid, Harmonic Spectral Deviation, Harmonic Spectral Spread (HSS) and Harmonic Spectral Variation [11].

2.2. Feature selection: Fisher discrimination ratio (FDR)

After obtaining the MPEG-7 features, feature selection in the form of FDR is used to find which features contribute higher to detection of pathologies. FDR for each feature (denoted *i*) is calculated as shown in (1)

$$FDR(i) = (\mu_1 - \mu_2)^2 / (\sigma_1^2 + \sigma_2^2) \quad (1)$$

where μ_1 and μ_2 represent the means for classes normal and pathological, respectively, while σ_1 and σ_2 represent the same but as the variances, and where *i* represents the feature number.

In the experiments on MEEI database with ten different training sets (in 10-fold cross validation), five features are constantly found having their FDR >1; these are: ASC, ASS, HSS and two features from ASF (9^{th} and 10^{th} bands). These highly discriminative features are described below:

(i) ASC describes the center of gravity of the log-frequency power spectrum and is calculated as shown in (2), where p_i is the power associated with frequency f_i; (ii) ASS describes the second moment of the log-frequency power spectrum and is calculated as shown in (3); (iii) HSS is computed as the average over the sound segment duration of the instantaneous Harmonic Spectral Spread, which is in turn computed as the amplitude weighted standard deviation of the harmonic peaks of the spectrum, normalized by the instantaneous Harmonic Spectral Centroid; (iv, v) ASF describes the flatness properties of the spectrum of an audio signal within a given number of frequency bands and is calculated as shown in (4), where *N* is the number of coefficient within a sub-band and c_n is the n-th spectral power coefficient of the sub-band [11, 17].

$$ASC = \sum_i \log_2(f_i/1000)\, p_i \Big/ \sum_i p_i \quad (2)$$

$$ASS = \sqrt{\sum_i ((\log_2(f_i/1000) - ASC)^2\, p_i) \Big/ \sum_i p_i} \quad (3)$$

$$ASF = \sqrt[N]{\prod_{n=1}^{N} c_n} \Big/ \frac{1}{N} \sum_{n=1}^{N} c_n \quad (4)$$

2.3. Classification: support vector machine (SVM)

SVM is widely used for data classification and it is known for its high prediction capabilities in many speech recognition applications. Many kernels can be used in SVM. For RBF kernel, an optimal combination values for two parameters has to be found, these parameters are called 'c' (optimization parameter) and 'gamma' (kernel function parameter), ensuring finding the optimal values for these will improve the prediction accuracy for the SVM classifier, this is accomplished in the experiments by doing grid search for these two parameters.

3. Experiment

This section describes the database, experimental setup used in the experiments.

3.1. Database

In [18], it is mentioned that "the main benefit of using standard corpora is that it allows researchers to compare performance of different techniques on common data, thus making it easier to determine which approaches are most promising to pursue. In addition, standard corpora also can be used to measure current state-of-the-art performance in research areas for particular tasks and highlight deficiencies that require further research". Thus, the database used in the experiments is the one developed by the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Labs [12]. The MEEI database contains sustained vowel /AH/ recordings by 53 normal speakers with duration of around 3 seconds and by 657 pathological speakers, with a wide variety of diseases, with duration of around 1 second. Also, it contains voice recordings with duration of around 12 seconds of a reading text from "Rainbow passage". Since we are interested in vocal pathologies we used sustained vowel recordings only to ensure that the vocal folds remain in motion during the entire utterance [8]. We used the same subset of the MEEI database which was used in [8,9,10,19], denoted as $MEEI_{subset}$; this subset has sustained vowel /AH/ recordings for 53 normal and 173 pathological speakers. The criteria in selecting this subset took into consideration a wide variety of voice disorders and a uniformly distributed gender and age between both normal and pathological classes [19], statistical description of the $MEEI_{subset}$ is shown in Table 1. The MEEI database has been thoroughly tested in numerous research works since its development and it is the most widespread and available of all the voice quality databases [20].

3.2. Setup

All the sound files in the $MEEI_{subset}$ have a sampling rate of either 50 kHz or 25 kHz. Therefore, all the files having a sampling rate of 50 kHz were down-sampled to half the

original sampling rate so that all the files in the $MEEI_{subset}$ will have the same sampling rate (which is 25 kHz). We extracted the MPEG-7 audio features for each file in the $MEEI_{subset}$ using TU-Berlin MPEG-7 audio analyzer [21]. All features were considered except the three features: Log Attack Time, Temporal Centroid and Spectral Centroid, as they all give a single value across the whole file, and therefore irrelevant to our analysis. The parameters set to extract MPEG-7 audio features are as follows: hop size=10 ms, frame size=20 ms, low edge=250 Hz, high edge=12.5 kHz, resolution=4 Octave/band, low limit=40 Hz, high limit=500 Hz, Mode= Instantaneous, these values were found to the optimum during our experiments.

The classification process was done using a ten-fold cross-validation SVM. RBF kernel was chosen for SVM as it is more general than other kernels (especially linear one) and usually it produces better accuracy and has less restriction than other kernels. In ten-fold cross-validation, the normal files and the pathological files' features values are randomly divided into ten equal groups each. In each iteration, nine groups each from the normal and the pathological are used for training, while the remaining are for testing. Therefore, at the end of ten iterations, all the ten groups are tested. There is no overlapping between the training set and the testing set in one iteration. Feature selection and the optimization of the SVM parameters are done with the training set. Grid search was conducted to obtain the optimal parameters values for 'c' and 'gamma' (in our experiments, the average optimum was found with c = 2 and gamma = 0.177). The final accuracy is obtained by averaging ten accuracies of the folds. We used LIBSVM, which is a library for SVM [22].

Two different types of experiments are conducted: (i) detection of pathology voices in the $MEEI_{subset}$, and (ii) classification of the pathologies. For pathologies classification, a full pair-wise classification was performed on four types of diseases: vocal nodules, vocal fold polyp, keratosis and adductor. Pathologies classification was done on exactly the same samples and the same permutations as done in [9].

4. Results and discussion

Results for pathology detection and classification are expressed in terms of accuracy, sensitivity (*Sn*: the likelihood that an event will be detected given that it is present), specificity (*Sp*: the likelihood that the absence of an event will be detected given that it is absent) and Area under the Receiver Operating Characteristic (ROC) curve, called AUC. All the results are calculated frame-by-frame basis, where the frame size is 20 ms. As the $MEEI_{subset}$ contains 173 recordings of pathological and 53 recordings of normal speakers, the number of pathological is almost three times the number of normal speakers while the duration of normal recordings is almost three times (3 sec) the pathological ones (1 sec) the total number of frames for both classes are almost close to each other.

4.1. Pathology detection

The results of pathology detection using the proposed MPEG-7 based method are shown in Table 2. The best performance was with all the features while as can be seen most of the contribution is by the five features explained earlier (Section 2.2). With all features (total 42 features), the proposed method achieves average accuracy of 99.994% with 0.011 standard deviation (confidence interval 95%), and AUC = 0.999. The accuracies of seven selected features (FDR ≥ 0.8; Section 2.2 plus ASF^{8th} and ASF^{13th}), five selected features (FDR ≥ 1; Section 2.2), and three selected features (FDR ≥ 1.2; ASC, ASS, ASF^{10th}) are 99.412%, 99.4%, and 92.535%, respectively (confidence interval 95%), Figure 1 shows the ROC curves for all features and for the top three features. Table 3 shows a comparison of our proposed method with three other methods ([8], [9], [10]) on the same subset of MEEI database ($MEEI_{subset}$), we can clearly see that our proposed method outperforms the other methods in terms of accuracy. Figure 2 shows a 3D scatter plot of the values for the top 3 features (ASC vs ASS vs HSS), where each marker represents the average value per file in the $MEEI_{subset}$ (total of 226 = 173+53); we can see that the normal samples can be visually easily separated from pathological ones, this can give an idea on how much these MPEG-7 features discriminate normal from pathological voices.

4.2. Pathology classification

Results for pathology classification; among the four diseases: vocal nodules, vocal fold polyp, keratosis and adductor, are shown in Table 4. With all features we achieved an accuracy of 100% for all diseases combinations except for "nodules vs keratosis" where an accuracy of 99.97% was achieved. The last column of the table shows AUC obtained in method [9]. As we see, the proposed method even with top five features has a competitive AUC compared to AUC in [9] and with using all features higher AUC was achieved than achieved in

Table 1. *Statistics of $MEEI_{subset}$ (♂ for male, ♀ for female).*

	Number		Mean age (years)		Age range (years)		Standard Deviation (years)	
	♂	♀	♂	♀	♂	♀	♂	♀
Normal	21	32	38.8	34.2	26 to 59	22 to 52	8.5	7.9
Pathological	70	103	41.7	37.6	26 to 58	21 to 51	9.4	8.2

Table 2. *Results of the proposed method on pathology detection.*

# of features	%Accuracy	Sn	Sp	AUC
All features	99.994±0.011	1	0.999	0.999
Top 7 features	99.412±0.066	0.996	0.986	0.991
Top 5 features	99.400±0.375	0.996	0.985	0.988
Top 3 features	92.535±0.382	0.965	0.797	0.880

Table 3. *Comparison of different methods on voice pathology detection.*

Method	%Accuracy
Proposed Method (all features)	99.994 ± 0.0105
Method [9]	94.1 ± 0.28
Method [8]	94.07 ± 0.0033
Method [10]	98.23 ± 0.001

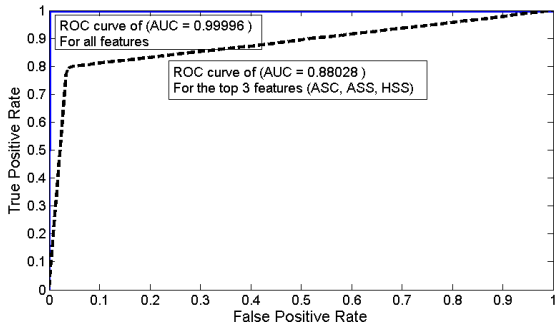

Figure 1: *ROC curve with AUC for: all features and top 3 features.*

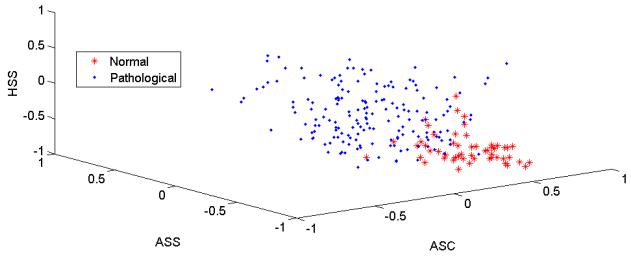

Figure 2: *3D Plot of (ASC vs ASS vs HSS) averaged values for each file in MEEI$_{subset}$ (53 normal and 173 pathological).*

[9]. In Figure 3, we can see ASC, which has the highest FDR, distribution along time (for 99 frames) for random samples of the four diseases and the normal voice. We can see that no intersection among all the samples. This can give an idea on how much MPEG-7 features discriminate between normal and pathological diseases and between the pathologies themselves.

5. Conclusion

MPEG-7 audio features based voice pathology detection and classification is proposed in this paper. The proposed method

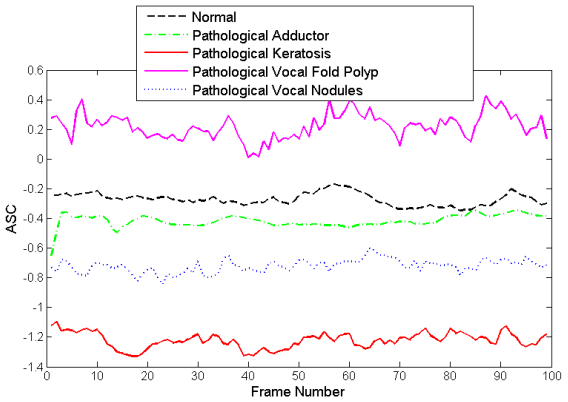

Figure 3: *ASC along time of different voice samples.*

achieves 99.994% accuracy for detection and 100% accuracy for classification. It also outperforms some state-of-the-art methods in terms of accuracy. We can justify MPEG-7 features powerful distinguishing pathology from normal ability by noticing that the power spectrum of normal voices are dominated by high frequencies as opposite to pathological ones where low frequencies are dominating and MPEG-7 features describe power spectrum; also, MPEG-7 features can differentiate noise-like sounds (more like pathological voices) from tone-like sounds (more like normal ones) by indicating whether the power spectrum is concentrated in the vicinity of its centroid or else spread out over the spectrum. Also, various types of pathological samples have different spectral centroids, distinct spectral flatness, etc. due to different types of vocal fold opening and closing behavior and MPEG-7 features can show such differences. A future work will be to apply the proposed method on continuous speech.

6. Acknowledgement

This work has been supported by the Research Center, College of Computer and Information Sciences, King Saud University, Saudi Arabia under the project RC120910.

Table 4. *Results of the proposed method on pathology classification (95% confidence interval). The last column gives AUC in method [9].*

Diseases Types	**# of features**	**%Acc**	**Sn**	**Sp**	**AUC**	**AUC in [9]**
Nodules vs Adductor	All Features	100.0 ± 0	1	1	1 ± 0	0.9578 ± 0.0064
	Top 5 Features	97.47 ± 0.365	0.9829	0.9674	0.9753 ± 0.004	
Nodules vs Keratosis	All Features	99.97 ± 0.003	1	0.9995	0.9998 ± 0.001	0.9527 ± 0.0053
	Top 5 Features	95.75 ± 0.7	0.9766	0.9429	0.9602	
Nodules vs Polyp	All Features	100.0 ± 0	1	1	1 ± 0	0.9428 ± 0.0073
	Top 5 Features	97.62 ± 0.362	0.9797	0.9728	0.976 ± 0.004	
Adductor vs Keratosis	All Features	100.0 ± 0	1	1	1 ± 0	0.9949 ± 0.0017
	Top 5 Features	93.19 ± 0.393	0.9466	0.9147	0.9321 ± 0.008	
Adductor vs Polyp	All Features	100.0 ± 0	1	1	1 ± 0	0.9585 ± 0.0087
	Top 5 Features	94.55 ± 0.701	0.9523	0.9394	0.9468 ± 0.009	
Keratosis vs Polyp	All Features	100.0 ± 0	1	1	1 ± 0	0.9359 ± 0.0058
	Top 5 Features	94.91 ± 0.857	0.9562	0.9437	0.9494 ± 0.011	

7. References

[1] I.R. Titze, “Workshop on Acoustic Voice Analysis: summary statement,” National Center for Voice and Speech, 1994.

[2] MHL. Hecker and EJ. Kreul, “Description of the speech of patients with cancer of the vocal folds—part I: measures of fundamental frequency,” J Acoust Soc Am, 49, pp. 1275–1282, 1970.

[3] R.J. Baken and R. Orlikoff, Clinical Measurement of Speech and Voice, 2nd ed. San Diego, CA: Singular, 2000.

[4] G. Muhammad, T. Mesallam, K. Almalki, M. Farahat, A. Mahmood, and M. Alsulaiman, “Multi Directional Regression (MDR) Based Features for Automatic Voice Disorder Detection,” Journal of Voice, vol. 26, no. 6, pp.817.e19-27, 2012.

[5] I.R. Titze and H. Liang, “Comparison of F0 extraction methods for high-precision voice perturbation measurements,” J Speech Hear Res, 36, pp. 1120-1133, 1993.

[6] P. Lieberman, “Perturbation in vocal pitch,” J. Acoust Soc Am, 33, pp. 597-603, 1961.

[7] M. Vasilakis and Y. Stylianou, “Sepctral jitter modeling and estimation” Biomed. Signal Process Control, pp.183-193, 2009.

[8] J.I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, “Dimensionally reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters,” IEEE Trans. Biomed. Eng., vol. 53, no. 10, pp. 1943-1953, Oct. 2006.

[9] M. Markaki and Y. Stylianou, “Voice Pathology Detection and Discrimination Based on Modulation Spectral Features” IEEE Trans. Speech Audio Process., vol. 19, no. 7, pp. 1938-1948, Sep. 2011.

[10] J.D. Arias-Londono, J.I Godino-Llorente, N. Saenz-Lechon, V. Osma-Ruiz and G. Castellanos-Dominguez, “Automatic Detection of Pathological Voices using Complexity Measures, Noise Parameters and Mel-Cepstral Coefficients” IEEE Trans. Biomed. Eng., vol. 58, no. 2, pp. 370-379, Feb. 2011.

[11] Information Technology-Multimedia Content Description Interface-Part 4: Audio, ISO/IEC CD 15938-4, 2001.

[12] Massachusetts Eye and Ear Infirmary, Elemetrics Disordered Voice Database (Version 1.03). Boston, MA: Voice and Speech Lab., 1994.

[13] G. Muhammad and K. Alghathbar, "Environment recognition for digital audio forensics using mpeg–7 and mel cepstral features," Journal of Electrical Engineering, Vol. 62, No. 4, pp.199–205, August 2011.

[14] Z. Xiong, R. Radhakrishnan, A. Divakaran and T. Huang, “Audio-based highlights extraction from baseball, golf and soccer games in a unified framework.” In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 5, pp. 628-631, 2003.

[15] P. Szczuko, P. Dalka, M. Dabrowski and B. Kostek, “MPEG-7-based Low-Level Descriptor Effectiveness in the Automatic Musical Sound Classification,” Proc. AES 116th Convention, 2004.

[16] D. Smith, E. Cheng and I. S. Burnett, “Musical Onset Detection using MPEG-7 Audio Descriptors,” In Proc. of the 20th Int. Congress on Acoustic (ICA 2010) Sydney, Australia, Aug. 23-27 2010.

[17] J. C. Wang, J. F. Wang, K. W. He, and C. S. Hsu, “Environmental sound classification using hybrid SVM/KNN classifier and MPEG-7 audio low-level descriptor”, Proc. IEEE International Joint Conference on Neural Networks, Canada, July 2006, pp. 1731-1735.

[18] J.P. Campbell, D.A. Reynolds, “Corpora for the evaluation of speaker recognition systems.” In Proc. of ICASSP ’99, vol. 2, Phoenix, AZ, USA, (March 1999), pp. 829–832.

[19] V. Parsa and D. Jamieson, “Identification of pathological voices using glottal noise measures,” J.Speech, Lang., Hear. Res., vol. 43, no. 2, pp.469-485, Apr. 2000.

[20] N. Saenz-Lechon, J.I. Godino-Llorente, V. Osma-Ruiz, and P. Gomez-Vilda, “Methodological issues in the development of automatic systems for voice pathology detection,” Biomed. Signal Process. Control, vol. 1, no. 2, pp. 120–128, 2006.

[21] TU-Berlin MPEG-7 Audio Analyzer. http://mpeg7lld.nue.tu-berlin.de/.

[22] Chih-Chung Chang and Chih-Jen Lin, “LIBSVM: a library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, 2:27:1—27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.