Automatic Selection of Acoustic and Non-linear Dynamic Features in Voice Signals for Hypernasality Detection

J.R. Orozco-Arroyave.[1], *S. Murillo-Rendón*[2], *A.M. Álvarez-Meza*[2], *J.D. Arias-Londoño*[3]
E. Delgado-Trejos[4], *J.F. Vargas-Bonilla*[1] *and C.G. Castellanos-Domínguez*[2]

[1]Departamento de Ingeniería Electónica, Universidad de Antioquia.
GEPAR and GITA Research Groups, Medellín, Colombia
[2]Departamento de Ingeniería Eléctrica, Electrónica y Computación,
Universidad Nacional de Colombia, Manizales
[3]Facultad de Ingeniería Electrónica y Biomédica, Universidad Antonio Nariño, Bogotá DC, Colombia.
[4]MIRP - Research Center, Instituto Tecnológico Metropolitano, Medellín, Colombia.

rafael.orozco@udea.edu.co, smurillor@unal.edu.co, amalvarezme@unal.edu.co
julian.arias@uan.edu.co, edilsondelgado@itm.edu.co, jfvargas@udea.edu.co
cgcastellanosd@unal.edu.co

Abstract

Automatic detection of hypernasality in voices of children with Cleft Lip and Palate (CLP) is made considering two charcaterization techniques, one based on acoustic, noise and cepstral analysis and other based on nonlinear dynamic features. Besides characterization, two automatic feature selection techniques are implemented in order to find optimal sub-spaces to better discriminate between healthy and hypernasal voices. Results indicate that nonlinear dynamic features are valuable tool for automatic detection of hypernasality; addtionally both feature selection techniques show stable and consistent results, achieving accuracy levels of up to 93.73%.

Index Terms: Hypernasality, Cleft Lip and Palate, acoustic, cepstral, nonlinear dynamics.

1. Introduction

The Cleft Lip and Palate (CLP) is a congenital malformation which produces, in most of the cases, velopharingeal insufficiency inducing an incomplete closure of the velopharyngeal cavity produced by deficient control of articulatory muscles [1]. Commonly, that impairment is associated with the presence of several speech pathologies such as: hypernasality, hyponasality, glottal stop, among others. Moreover, the most commonly pathology presented in CLP patients is the hypernasality.

The development of automatic systems for the detection of hypernasality in CLP patients (must of them children) is important to be used as computer-aided medical diagnosis tool, allowing a better following of the clinical treatment given to each patient [2].

The objective assessment of hypernasal voices has been addressed using acoustic analysis mainly oriented to evaluate voice tone variation [3]. In [4], a detailed analysis of the spectral components in vowel /a/ is made by means of high order linear prediction spectrums; in the same way, the authors in [5] consider a spectral analysis in different frequency bands, finding differences between first and second formant amplitudes. In [6] the authors consider pronunciation features and twelve Mel-Frequency Cepstral Coeffcients (MFCC) to detect different articulation problems in 26 CLP patients, reporting accuracy labels of 71.1% for vowels, while Linear prediction (LP) spectrum analysis based on group delay functions is made in [7], reporting accuracies of 100%, 88.15% and 80.25% for automatic detection of hypernasality in vowels /a/, /i/ and /u/, respectively.

This work presents a methodology for automatic detection of hypernasal voices which includes measurements of conventional features in the pathological speech processing field such as: acoustic, cepstral and noise measures as well as features based on non-linear dynamic analysis, in order to take into account as much information as possible about the physical phenomena involved in the voice production process, which could be useful for the clinical evaluation of hypernalsal voices. Acoustic, spectral and noise measures have been widely used for the characterization of pathological speech, whilst, nonlinear dynamic analysis techniques are still not commonly employed for this task. Nevertheless, in the few last years there have been published several studies that have demonstrated the presence of nonlinear behavior in the production of speech [8], and also the usefulness of such information for the characterization of pathological voices. Taking into account that due to the viscoelasticity properties of the soft palate, it presents intrinsic nonlinear behavior related to the stress-strain relationship and the tissue deformations [9], the nonlinear analysis becomes a valuable tool for evaluating pathologies associeated with velopharingeal insufficiency.
According to the state of the art in hypernasality detection, this is the first work devoted to comparing the performance of acoustic features with non-linear dynamics ones.
Additionaly, since the accuracy of the computerized system depends on the discriminant power of the features used for characterizing the voice signals, and that there is not a clear knowledge about which features provide a better characterization of the hypernasal voices, this work also explores the use of two different feature selection techniques, the first one based on a Principal Component Analysis and the second one based on a Heuristic Floating Search, in order to find the most important features for the detection of hypernasality. The rest of the paper is organized as follows: section 2 contains the description of voice characterization, section 3 presents the feature selection techniques

28 – 31 August 2011, Florence, Italy

considered in the work, section 4 includes details about the experiments, section 5 contains the results, and section 6 includes the conclusions

2. Characterization of Voice

Two groups of features are considered. Subsections 2.1 and 2.2 indicate some details of acoustic and nonlinear characteristics of voice, respectively.

2.1. Acoustic, Noise, and Cepstral Features

Acoustic measures of perturbations in fundamental frequency have already been employed for the evaluation of hypernasality [3]. In this study, the variation of the fundamental frequency also called *Jitter* is considered because it is affected mainly by the lack control of vocal fold vibration. The variation of the maximum amplitude in this fundamenta frequency, which is called *Shimmer*, allow to detect general problems with vocal folds movement due to improper or incomplete closure of the pharyngeal velum. Jitter indicates variations on the frequency vibration in vocal folds due to the lack control of vocal fold muscles, while shimmer represents variations of glottic resistance and mass lesions in vocal folds [10].

Considering the influence of noise in hypernasal voices [11], and their demonstrated usefulness for the characterization of pathological speech, the following noise measures are included in this work: Harmonics to Noise Ratio (HNR), Cepstral-HNR (CHNR), Normalized Noise Energy (NNE) and Glottal to Noise Excitation Ratio (GNE).

According to [12], velopharyngeal insufficiency or incompetence suffered by CLP patients, leads them to the need for compensatory movements in vocal tract, producing glottal stops and general problems with glottal articulation. Therefore, eleven Mel-Frequency Cepstral Coefficients (MFCC) are considered suitable for the characterization of hypernasal voices, because in presence of voice disorders they show an inherent ability to model either an irregular movement of the vocal folds, or a lack of closure induced by an increase of mass or due to changes in the properties of the tissue covering the vocal folds.

2.2. Non-linear Dynamics Features

According to evidences of non-linearities in vocal fold vibration [13], and considering that CLP patients suffer problems with their vocal tract articulation and with movements on their vocal folds, non-linear dynamic analysis can be considered as an alternative for the assessment of pathological voices [8].

The research community has shown interest for comparing classical acoustic analysis with non-linear dynamics, as presented in [14], where accuracy levels obtained with jitter, shimmer, NNE, CHNR and GNE are compared to recurrence and fractal scaling in the detection of voice disorders different to hypernasality. More recently in [15], the accuracy of jitter and shimmer is compared to the accuracy of five non linear dynamic features in the detection of different voice pathologies. Recent works that tray to correlate voice disorders and abnormal vocal tract movements, are mainly oriented on the analysis of the state space. For a sinosoidal signal, its corresponding attractor is a perfect circle, for a healthy voice, the attractor is not a perfect circle, but the trayectories can converge. In the case of a pathological voice, its attractor doesn't have convergent trayectories, because it is produced by more complex dynamical system.

In the presented work, Correlation Dymension (D_c) is implemented following the proposed method in [16], and considering that this feature is related with the number of independent variables necessary for describing the corresponding process it can be a good complexity measure. Additionally, chaos in voice indicates exponential growth due to infinitesimal perturbations. This exponential instability can be measured using the Largest Lyapunov Exponent (λ_1) which can be calculated according to the algorithm presented in [17]. For characterizing randomness of voice signals, Lempel-Ziv Complexity (LZ) is measured, showing an idea of the rate of new pattern occurrences in the time series. Finally, Hurst Exponent (H) is calculated to consider possible relationships between future and past samples in voice recordings, which is implemented according to an algorithm based on the scaling range method proposed in [18].

3. Automatic Feature Selection

The aim of automatic feature selection is to find out the m most relevant characteristics of the original feature space $\mathbf{X} \in \mathbb{R}^{n \times p}$ (n: number of observations, p: number of original features), which make possible to build the subspace representation $\mathbf{Y} \in \mathbb{R}^{n \times m}$, ($m < p$). Relevant features contained in $\mathbf{Y}$ allow to diminish irrelevant and/or redundant information and the computational load in classification stages. Here, two different algorithms of feature selection are considered. The first one is based on Principal Component Analysis (PCA), and the second one is based on a Heuristic technique called Sequential floating forward selection (SFFS). Both methods look for a subset of features that best discriminates healthy and hypernasal voices.

3.1. Feature Selection based on Relevance Analysis

PCA is a statistical technique applied here to find out a low-dimensional representation of the original feature space, searching for directions with greater variance to project the data.

Although, PCA is commonly used as a feature extraction method, it can be useful to properly select a relevant subset of original features that better represent the studied process [19]. In this sense, given a set of features ($\boldsymbol{\xi}_k : k = 1, \ldots, p$) corresponding to each column of the input data matrix $\mathbf{X}$, the relevance of each $\boldsymbol{\xi}_k$ can be analyzed for finding the resulting subspace $\mathbf{Y}$. More precisely, relevance of $\boldsymbol{\xi}_k$ can be identified looking at $\boldsymbol{\rho} = \begin{bmatrix} \rho_1 & \rho_2 & \cdots & \rho_p \end{bmatrix}^\top$, where $\boldsymbol{\rho}$ is defined as $\boldsymbol{\rho} = \sum_{j=1}^{m} |\lambda_j \mathbf{v}_j|$. ($\lambda_j$ and $\mathbf{v}_j$ are the eigenvalues and eigenvectors of the initial matrix, respectively). Therefore, the main assumption is that the largest values of ρ_k point out to the best input attributes, since they exhibit higher overall correlations with principal components.

3.2. Feature Selection based on Heuristic Floating Search

SFFS is a heuristic algorithm that finds the best subset of features of the original set through the iterative inclusion and exclusion of features. In this procedure after each forward step a number of backward steps are applied as long as the resulting subsets are better than the previous ones. Sorting features according to their discriminant capacity are necessary to get stable and consistent results, which is reflected in the performance of the system; for more details about the implemented algotithm see [20].

4. Experimental Setup

The proposed methodology is tested on a Hypernasality dataset provided by *Grupo de Control y Procesamiento Digital de*

Señales-(GCyPDS) of the Universidad Nacional de Colombia, Manizales. The database contains 266 voice registers from children between 5 and 15 years old, who uttered the vowels of Spanish. There are 110 children labeled by phoniatry expert as healthy, and 156 labeled as hypernasal.

To compare the discriminant performance of both kind of characterizations: acoustic, noise, and cepstral features, against nonlinear dynamics, two different input spaces are calculated for each vowel: $\mathbf{X}_A$ and $\mathbf{X}_{NLD}$.

More precisely, $\mathbf{X}_A$ and $\mathbf{X}_{NLD}$ are calculated according to sections 2.1 and 2.2, using temporal windowing 40ms and 55ms, respectively as in [21]. Thence, the mean value, standard deviation, and variance of each feature vector are calculated, building the input data matrix $\mathbf{X}_A \in \mathbb{R}^{266\times 51}$. For nonlinear dynamic feature space, only the mean values and standard deviation of each feature vector are calculated, building the input space $\mathbf{X}_{DNL} \in \mathbb{R}^{266\times 8}$.

Furthermore, two feature selection methods are compared (section 3), generating two different subspace matrices: $\mathbf{Y}_{PCA}$ and $\mathbf{Y}_{SFFS}$. The $\mathbf{Y}_{PCA}$ matrix is calculated based on a relevance analysis in the original input space according to section 3.1. 10-fold cross validation analysis is made to generate a mean weighted vector $\bar{\boldsymbol{\rho}}$, with $\bar{\boldsymbol{\rho}}_k = {}^1\!/_{10} \sum_{i=1}^{10} \boldsymbol{\rho}_{ki}$, being $\boldsymbol{\rho}_{ki}$ the weight of relevance of the k-th feature in the i-th fold. Therefore, the original features are sorted according to $\bar{\boldsymbol{\rho}}$, and a redundancy analysis is used to eliminate correlated features. We eliminate features with a correlation greater than 80%, but considering the relevance order given by $\bar{\boldsymbol{\rho}}$. On the other hand, the $\mathbf{Y}_{SFFS}$ matrix is calculated using SFFS (section 3.2). In order to obtain $\mathbf{Y}_{SFFS}$, a $10-$fold cross validation is performed. For each fold a sorted sub set of features is obtained, forming 10 feature vectors. This process is repeated 10 times, building a total of 100 feature vectors. Then, the maximum size among these 100 vectors is chosen, and the mode per position in all 100 vectors are obtained to form the output matrix.

Finally, given a subset of features for each $\mathbf{Y}_{PCA}$ and $\mathbf{Y}_{SFFS}$, a soft-margin Support Vector Machines - (SVM) classifier is trained using a gaussian kernel with band-width σ [22]. We generate a curve of performance adding one by one the characteristics obtained in each subspace representation. For a given sub set, the optimum working point has been searched using a 10-fold cross validation to fix the C and σ values. The proposed methodology can be summarized as in Figure 1.

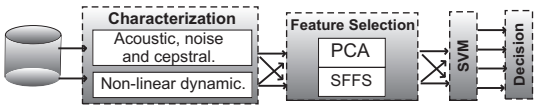

Figure 1: General Scheme

5. Results

Table 1 shows the overall results obtained for acoustic, noise and cepstral features. Note that results are better after either automatic feature selection process, PCA or SFFS. For each vowel, representation space is reduced in up to 23 dymensions, achieving not only less complexity but also higher success rates.

Table 2 shows results using nonlinear dynamic features. In general, accuracy, sensitivity and specificity values are lower than those showed in table 1; however, the representation space in that case has a lower dimension.

Last rows in tables 1 and 2 show results of combining best features set per vowel, achieving higher accuracies with its respective sensitivity and specificity.

Table 1: Classification Results for Acoustic Features.

Vowel	SM	NF	Accuracy	Sensitivity	Specificity
/a/	WS	51	86.11±8.88	81.90±8.30	92.58±12.71
	PCA	21	88.28±8.38	90.77±12.06	83.31±10.90
	SFFS	23	86.03±6.27	92.93±6.26	76.48±9.31
/e/	WS	51	89.87±10.06	89.54±12.30	90.46±8.86
	PCA	23	92.45±4.06	95.51±4.46	89.00±8.94
	SFFS	27	93.28±4.54	97.76±5.62	87.17±10.33
/i/	WS	51	87.58±6.27	86.66±6.11	88.85±8.90
	PCA	32	91.28±5.77	99.33±2.11	80.69±9.45
	SFFS	18	92.45±6.83	97.45±4.55	86.35±11.95
/o/	WS	51	89.87±3.15	88.30±4.39	91.35±8.48
	PCA	26	89.13±4.86	88.41±7.98	90.35±10.27
	SFFS	20	89.49±9.62	87.11±14.78	93.82±9.34
/u/	WS	51	86.44±6.77	84.05±7.53	88.66±9.26
	PCA	28	87.22±6.70	95.45±4.64	76.01±12.76
	SFFS	20	89.83±4.74	92.01±7.17	83.25±19.20
Union	WS	147	93.28±4.12	92.99±6.41	93.30±7.72
	PCA	97	93.73±5.28	92.86±9.19	93.71±11.69
	SFFS	99	92.86±5.63	92.28±9.96	94.11±9.96
SM: Selection Method, WS: Without Selection, NF: Number of Features					

Table 2: Classification Results for Non-linear Dynamic Features.

Vowel	SM	NF	Accuracy	Sensitivity	Specificity
/a/	WS	8	86.78±5.10	86.77±7.44	87.90±9.32
	PCA	8	86.01±5.73	83.43±9.06	89.73±8.57
	SFFS	3	87.16±4.37	87.01±6.70	88.57±8.92
/e/	WS	8	87.19±6.05	85.62±9.06	89.86±5.24
	PCA	6	87.96±7.21	86.40±9.13	90.67±6.69
	SFFS	6	87.57±6.56	85.60±9.06	90.86±6.14
/i/	WS	8	87.98±3.40	87.54±8.92	88.98±9.94
	PCA	2	86.42±8.33	84.64±12.32	87.69±7.02
	SFFS	5	86.86±5.62	82.99±8.45	93.09±5.18
/o/	WS	8	86.48±8.65	84.85±12.49	88.71±10.02
	PCA	8	86.14±4.31	86.85±5.27	87.07±11.23
	SFFS	5	85.73±5.96	83.49±8.90	88.34±8.96
/u/	WS	8	86.11±6.42	87.84±6.90	84.42±13.45
	PCA	4	86.15±8.23	87.26±8.40	83.95±14.21
	SFFS	7	86.85±8.20	85.83±9.61	87.84±10.60
Union	WS	36	91.16±7.24	90.84±11.00	91.28±10.52
	PCA	23	92.08±8.21	95.49±7.21	88.05±12.73
	SFFS	16	92.05±5.71	93.58±7.39	90.06±9.78
SM: Selection Method, WS: Without Selection, NF: Number of Features					

Although the spaces built with PCA and SFFS include differents features, for all five Spanish Vowels, is possible to set that mean values of Cepstral Coefficients (MFCC) are most consistent features in conjunction with the means of HNR, CHNR, GNE, thus, these features set up an optimal space to perform hypernasality detection regardless the pronounced vowel.

Figure 2 shows Receiver Operating Characteristic (ROC) curves for acoustic, noise and cepstral analysis and figure 3 shows the case of non linear dynamic analysis. Results for SFFS, PCA and without feature selection techniques are provided on each figure. In the case of acoustic analysis, best result is with PCA, where the Area Under the Curve (AUC) is 0.9616. For non linear dynamic analysis the best result is with SFFS where $AUC = 0.9578$.

6. Conclusions

Results in last rows of tables 1 and 2 indicate that accuracy level increases when best feature spaces for each vowel are joined, building general representation spaces useful to perform automatic detection of hypernasality in voice independent of the pronounced vowel and with higher accurary rates.

Non linear dynamic features are presented as a promising alternative for the automatic detection of hypernasality in voice. Besides both characterization methods show to be stable and robust, for non linear dynamics the number of features is lower than in acoustic case.

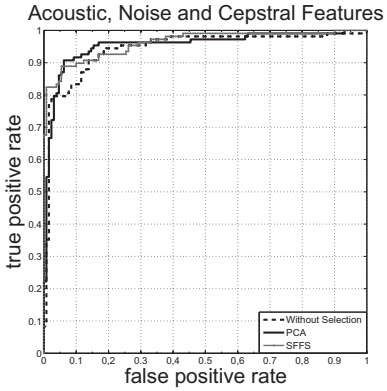

Figure 2: ROC curves for two different features sets

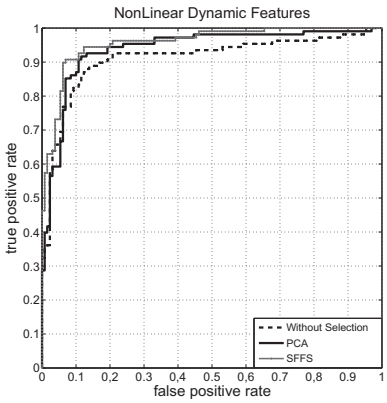

Figure 3: ROC curves for two different features sets

Both techniques for selection of features are able to perform significant reduction of the dimensionality in representation spaces; however, technique based on PCA showed to be more stable along the experiments. Additionally, PCA yields a weight vector that allows to organize automatically the features without any statistical posterior analysis, unlike the case of floating search, where is necessary to perform a mode analysis to chose the representation vectors.

Cepstral Coefficients are important for the detection of hypernasality, due to their ability to model voice signals in spatio-temporal terms. For the case of noise features that were consistent in the experiments (HNR, CHNR and GNE), they are important because of their ability to detect irregular movements in the vocal tract.

7. Acknowledgements

This work was carried out in association with Clinica Noel in Medellín, and under grants of ARTICA, funded by COLCIENCIAS and Ministerio de TIC in Colombia, project N°1115-470-22055 and through young researchers program by resolution DFIA-0265 in Universidad Nacional de Colombia, Manizales.

8. References

[1] Henningsson, G.E. and Isberg, A.M., "Velopharyngeal movement patterns in patients alternating between oral and glottal articulation: a clinical and cineradiographical study citation", *The Cleft Palate Journal*, 23(1):1–9, 1986.

[2] Godino-Llorente, J.I., Sáenz-Lechón, N., Osma-Ruiz, V., Aguilera-Navarro, S. and Gómez-Vilda, P., "An integrated tool for the diagnosis of voice disorders", *Medical Engineering & Physics*, 28(3):276-289, 2006.

[3] Lee, G.S., Wang C.P. and Fu S., "Evaluation of hypernasality in vowels using voice low tone to high tone ratio", *The Cleft Palate Journal*, 46(1):47–52, 2009.

[4] Rah, D.K., Ko, Y.I., Lee, C. and Kim, D.W., "A Noninvasive Estimation of Hypernasality Using a Linear Predictive Model", *Annals of Biomedical Engineering*, 29(7):587-594, 2001.

[5] Kataoka, R. Warren, D.W., Zajac, D.J., Mayo R. and Lutz R.W., "The relationship between spectral characteristics and perceived hypernasality in children", *Journal of Acoustic Society of America*. 109(5):2181-9, 2001.

[6] Maier, A., Hönig, F., Hacker C., Shuster M. and Nöth E., "Automatic Evaluation of Characteristic Speech Disorders in Children with Cleft Lip and Palate", Proc. of *11th Int. Conf. on Spoken Language Processing*, Brisbane, Australia, pp. 1757–1760, 2008.

[7] Vijayalakshmi, P., Nagarajan, T. and Jayanthan Ra, V., "Selective pole modification-based technique for the analysis and detection of hypernasality", Proc. of *IEEE Conference - TENCON*, pp. 1-5, 2009.

[8] Jiang, J.J., Zhang, Y. and Mcgilligan, C., "Chaos in voice, from modeling to measurement", *Journal of Voice*, 20(1):2–17, 2006.

[9] Birch, M.J. and Srodon, P.D., "Biomechanical Properties of the Human Soft Palate", *The Cleft Palate-Craniofacial Journal*, 46(3):268-274, 2009.

[10] Wertzner H.F., Schreiber S., and Amaro L., "Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders", *Rev. Bras. Otorrinolaringol*, 71(5):582–588, 2005.

[11] Setsuko, I., "Effects of Breathy Voice Source on Ratings of Hypernasality", *The Cleft Palate Journal*, 42(6):641–648, 2005.

[12] Golding, K.K., *Therapy Techniques for Cleft Palate Speech and Related Disorders*, Singular Thomson Learning [Ed], 2001.

[13] Giovanni, A., Ouaknine, M., Guelfucci, R., Yu, T., Zanaret, M. and Triglia, J.M., "Nonlinear behavior of vocal fold vibration: the role of coupling between the vocal folds", *Journal of Voice*, 13(4):456–476, 1999.

[14] Little, M.A., McSharry, P.E., Roberts, S.J., Costello, D.E. and Moroz, I.M., "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection", *Biomedical Engineering Online*, 6(23), 2007.

[15] Vaziri, G. Almasganj F., and Behroozmand, R., "Pathological assessment of patients' speech signals using nonlinear dynamical analysis", *Computers in Biology and Medicine*, 40(1):54–63, 2010.

[16] Grassberger, P. and Procaccia, I., "Measuring the strangeness of strange attractors", *Physica D*, 9:189–208, 1983.

[17] Rosenstein, M.T., Collins, J.J. and De Luca C.J., "A practical method for calculating largest Lyapunov exponents from small data sets", *Physica D*, 65:117–134, 1993.

[18] Hurst, H.E., Black, R.P. and Simaika, Y.M., "Long-term storage: an experimental study", London, 1965.

[19] G. Daza-Santacoloma, J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruíz, and G. Castellanos-Domínguez, "Dynamic feature extraction: An application to voice pathology detection," *Intelligent Automation and Soft Computing*, 15(4):665-680, 2009.

[20] P. Pudil, J. Novovicova and J. Kittler, "Floating Search Methods in Feature Selection", *Pattern Recognition Letters*, 15(11):1119–1125, 1994.

[21] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruíz and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE transactions on bio-medical engineering*, 58(2):370-379, 2011.

[22] Scholköpf, B. and Smola, A.J., *Learning with Kernels*, The MIT Press, 2002.