INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Objective Severity Assessment From Disordered Voice Using Estimated Glottal Airflow

Yu-Ren Chien, Michal Borský, Jón Guðnason

Center for Analysis and Design of Intelligent Agents, Reykjavik University, Iceland
yrchien@ntu.edu.tw, michalb@ru.is, jg@ru.is

Abstract

In clinical practice, the severity of disordered voice is typically rated by a professional with auditory-perceptual judgment. The present study aims to automate this assessment procedure, in an attempt to make the assessment objective and less labor-intensive. In the automated analysis, glottal airflow is estimated from the analyzed voice signal with an inverse filtering algorithm. Automatic assessment is realized by a regressor that predicts from temporal and spectral features of the glottal airflow. A regressor trained on overtone amplitudes and harmonic richness factors extracted from a set of continuous-speech utterances was applied to a set of sustained-vowel utterances, giving severity predictions (on a scale of ratings from 0 to 100) with an average error magnitude of 14.

Index Terms: voice quality, voice production, speech analysis, voice disorders, glottal airflow

1. Introduction

In the diagnosis and management of voice disorders, the voice quality of a patient is typically assessed personally by a voice clinician. Since the auditory-perceptual judgment used in the assessment is subjective by nature, ratings resulting from the assessment can vary substantially among raters. If multiple raters are recruited for objective assessment, the collective efforts involved would also be prohibitive. This study aims for the objective assessment of voice severity with a speech analysis algorithm, which estimates the severity of a disordered voice from its signal. Here the severity of a voice is defined as the median severity rating determined from a group of experienced raters assessing the voice. As a definition adopted from the CAPE-V protocol, a severity rating refers to a global, integrated impression of voice deviance [1].

The objective assessment of voice quality has been pursued prior to this study. Prosek *et al.* [2] used linear multiple regression analysis to show that residue features are highly correlated with severity ratings. A small number of acoustic parameters have been identified [3, 4] in the Multi-Dimensional Voice Program scale that correlate significantly with perceptual ratings. Wolfe *et al.* [5] used a combination of acoustic measures, including measures of signal periodicity, high-frequency noise, and fundamental frequency, to evaluate voice severity. For the measurement of overall voice quality, Maryn *et al.* [6] presented an approach where stepwise multiple regression analysis is applied to a set of features extracted from the speech signal, including a cepstral feature in particular. With a multifactor severity model incorporating cepstral and spectral speech features, Awan *et al.* [7] showed a strong relationship between acoustic severity predictions and auditory-perceptual severity ratings. See [8] for a survey of acoustic measures for voice quality assessment.

Since voice disorders result from physiological abnormalities occurring at one's glottis, an estimate of the glottal airflow would arguably provide vital clues on the severity of voice disorder. However, only a relatively small number of approaches took advantage of glottal flow estimation—Prosek *et al.* [2] is in our opinion the only existing approach that falls into this category. In their approach, glottal airflow is modeled by an impulse train and estimated by linear predictive autocorrelation analysis, from which six periodicity features are extracted as correlates of voice severity. In contrast, glottal airflow is estimated here with a closed-phase model [9], from which features are extracted to capture the overall waveform shape. With these features, the assessment of voice severity is learned by a regression algorithm from a training set of speech recordings and ratings.

2. Algorithm for Severity Assessment

To assess voice severity for a speech utterance, a set of features are extracted from the acoustic signal of the speech utterance. An assessment is produced by a regressor according to the extracted features.

2.1. Feature Extraction

To produce accurate voice severity assessments, it is imperative to base each assessment on particular speech features that are highly relevant to voice severity. Assuming that the physiological condition of vocal folds is well reflected in the glottal airflow, we let all the acoustic features be extracted from an estimated glottal airflow. In this study, a glottal flow signal is estimated from the speech signal with the sparse linear prediction (SLP) algorithm [9], which is based on a weighted model of the glottal-flow closed phase. SLP depends on a sequence of glottal closure instants (GCIs) that mark the beginning of each closed phase. These are estimated from the speech signal with the YAGA algorithm [10]. According to evaluation results presented by Chien *et al.* [11] for inverse filtering algorithms, SLP outperforms several alternative algorithms in terms of continuous speech analysis.

The GCIs detected from the speech signal constitute a cycle-wise segmentation of the estimated glottal flow signal. From each cycle of the glottal flow estimate, 4 types of features are extracted. First, the waveform shape of glottal airflow can be represented by a sequence of N_o *overtone amplitudes* (OA), which refer to those of the successive harmonics from the second to the $(N_o + 1)$th. Harmonic amplitudes of a periodic signal can be calculated from a single cycle by taking the absolute values of successive frequency components in the discrete-Fourier spectrum that correspond to the harmonics. Each OA is normalized by the fundamental amplitude before a conversion to dB. The other 3 types of features, i.e., *harmonic richness factor* (HRF), *normalized amplitude quotient* (NAQ), and *closed quotient* (CQ), are traditionally important for describing

http://dx.doi.org/10.21437/Interspeech.2017-138

voice quality. HRF [12] refers to the total power of overtones (up to 3 kHz in our implementation) normalized by the fundamental power and converted to dB. NAQ [13] is defined as the peak-to-peak amplitude of glottal airflow divided by the product of maximum flow declination rate and fundamental period. Finally, CQ is defined here as the quotient of the duration of closed phase divided by that of the cycle. The duration of open phase is defined as the longest high-flow duration, delimited by a pair of rising and falling edges detected from the glottal flow waveform, within the glottal-flow cycle. The duration of closed phase is calculated by subtracting that of the open phase from that of the cycle.

Some of the *GCI intervals* (i.e., intervals between contiguous GCIs) may actually be non-voiced, from which no feature should be extracted. To identify these non-voiced intervals, a voice activity detector (VAD) is applied to the analyzed speech signal, where voice activities are detected from a sequence of time frames that are spaced with a constant hop size. For a GCI interval longer than the hop size (which can occur when a small number of GCIs are spuriously detected from a non-voiced segment of speech signal), a total of M VAD frames ($M \geq 1$) can be found within the interval (in terms of the center position of each frame), which allows the voicing decision to be made from all the likelihood scores associated with this interval:

$$\frac{\left(\prod_{m=1}^{M} f_v(\mathbf{x}_m)\right)^{1/M}}{\left(\prod_{m=1}^{M} f_n(\mathbf{x}_m)\right)^{1/M}} \underset{\text{non-voiced}}{\overset{\text{voiced}}{\gtrless}} \eta. \quad (1)$$

Here $\mathbf{x}_m$ denotes a vector of MFCCs for the mth VAD frame in the interval, $f_v(\cdot)$ denotes the voiced GMM, and $f_n(\cdot)$ denotes the non-voiced GMM. For a GCI interval shorter than the VAD hop size, a voicing decision is taken from the nearest VAD frame.

Now, each of the $N_o + 3$ scalar feature types has as many cycle-specific instances as there are cycles (voiced GCI intervals) in the utterance. To create a fixed-size feature format for all speech utterances, moments of orders 1 to N_m are calculated from all the cycle-specific feature values for each feature type to give a $[(N_o + 3) \cdot N_m]$-dimensional feature (row) vector $\mathbf{y}$ for one utterance. This converts a variable-length representation of feature values into a fixed-length one that describes the distribution of values.

A final step of dimensionality reduction and whitening is applied to the feature vector $\mathbf{y}$ to give a transformed feature vector $\tilde{\mathbf{y}}$:

$$\tilde{\mathbf{y}} = \bar{\mathbf{y}}\mathbf{V}\mathbf{D}^{-1/2}, \quad (2)$$

where $\bar{\mathbf{y}}$ is a standardized version of $\mathbf{y}$, $\mathbf{V}$ is an $[(N_o + 3) \cdot N_m] \times N_p$ matrix whose columns specify a basis for an N_p-dimensional subspace, and $\mathbf{D}$ is an $N_p \times N_p$ diagonal scaling matrix. This transformation is determined from 128 continuous-speech utterances, which consist of 8 normal and 24 disordered voices each reading 4 sentences [7]. Specifically, analyses conducted on this data set include the estimation of mean and standard deviation for standardizing $\mathbf{y}$, and a principal component analysis [14] for determining the subspace and the variances along the subspace dimensions (which define the diagonal elements of $\mathbf{D}$).

2.2. Regression

Automated severity assessment can be realized by a regressor when a training set of speech utterances, each rated by multiple listeners, is available. To determine a ground-truth severity for each training utterance, a median rating is taken over all the listeners that rated the utterance. We apply a regression algorithm to such a training set in order to construct a regressor that predicts the median severity rating from any speech utterance represented by the features described in Section 2.1. For the regression, two alternative algorithms are adopted.

2.2.1. Algorithms

One of the two alternative algorithms for the regression is support vector regression (SVR) [15], which finds a prediction function by minimizing its deviations from a subset of the training data, such that only those exceeding an insensitivity threshold ϵ are minimized. The minimization is formulated as a convex optimization problem, to which a solution at the global optimum is guaranteed. The implementation used here is from LIBSVM [16], which minimizes

$$H(\mathbf{w}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N_s} V(z_i - \mathbf{w}^T\phi(\tilde{\mathbf{y}}_i) - b), \quad (3)$$

where N_s denotes the number of training instances, $V(\cdot)$ denotes the ϵ-insensitive deviation measure, z_i denotes the ith median rating in the training data, $\tilde{\mathbf{y}}_i$ denotes the features extracted from the ith training utterance, and $\phi(\cdot)$ denotes a nonlinear mapping defined here via the RBF kernel:

$$K(\tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j) = \phi(\tilde{\mathbf{y}}_i)^T\phi(\tilde{\mathbf{y}}_j) = \exp(-\gamma||\tilde{\mathbf{y}}_i - \tilde{\mathbf{y}}_j||^2). \quad (4)$$

The other regression algorithm is random forest regression [17], which is based on an array of (N_t) regression trees that are fitted to different resampled versions of the training data. These regression trees are randomized in terms of the resampling, and of the order in which the features are used to partition the training data. Thus, the final regressor is generalized by averaging the predictions over all the decision trees in the random forest. The implementation used here is from the (Matlab) Statistics and Machine Learning Toolbox. The number of trees N_t is set to 1,000, so as to give stable results across repeated runs of the same randomized training task. Two other parameters are the number of features included (by random selection) in the candidacy for the single feature used to determine each partition, which is denoted N_f, and the minimum number of training instances associated with each leaf node, denoted N_i.

2.2.2. Training Data

The training data is the same 128 continuous-speech utterances as used in determining the dimensionality reducing and whitening transformation (8 normal and 24 disordered voices each reading 4 sentences [7]), along with the corresponding median severity ratings. In addition to these 128 continuous-speech utterances, the original data set used in [7] also includes 32 sustained-vowel utterances, which are not included in this training set. Each median severity rating is calculated for the corresponding utterance from 125 severity ratings (continuous-valued in the range 0–100) produced by 25 listeners each responding to 5 unidentified repetitions of the utterance.

3. Experimental Procedure

3.1. Data Set

To evaluate the effectiveness of the severity assessment algorithm presented in Section 2, we conduct experiments on 32 sustained-vowel utterances which were previously used in [7],

consisting of 8 normal and 24 disordered voices pronouncing the vowel /ä/.[1] As with the training data for regression, a ground-truth severity is available for each utterance as the median over all 125 severity ratings produced by 25 listeners to 5 repetitions of the utterance. The mean and standard deviation of the 32 ground-truth severities are 36 and 26, respectively.

3.2. Performance Measure

Consider a speech utterance for which a ground-truth severity is available. After applying a severity assessment algorithm to this speech utterance, we evaluate the accuracy of assessment by calculating a *severity error magnitude* (SEM), which is defined as the absolute value of the difference between the assessed and ground-truth severities. For instance, if an utterance with a ground-truth severity of 100 receives an assessed severity of 0, the resulting SEM will be 100, indicating an utter misevaluation of voice severity. On the contrary, an SEM of 0 would indicate a perfect consistency between acoustic and auditory-perceptual assessments.

To obtain from the data set a baseline value for this performance measure, consider a constant "prior guess" applied to all the 32 sustained-vowel utterances. The guess is 22, calculated as the median over the 128 ground-truth severities (for continuous-speech utterances) from the training data for regression. The resulting average SEM over the 32 sustained-vowel utterances is 22, which (approximately) equals the guess value by coincidence. Since the guess completely disregards any feature extracted from an utterance, any value of SEM calculated on the data set that is close to or higher than 22 should be regarded as an indication of ineffective severity assessment. Moreover, while the 125 ratings available for each sustained-vowel utterance have been used to define the ground truth, they can also be used separately as a set of subjective assessments, from which human performance can be measured. The average SEM of the 4,000 subjective assessments is 36.

3.3. Algorithm Variants

In the experiments, several algorithm variants are considered. When SVR is used for regression, 4 types of features are tested alternatively, i.e., OA (OA-SVR), HRF (HRF-SVR), NAQ (NAQ-SVR), and CQ (CQ-SVR). The combination of OA, HRF, and NAQ is tested both with SVR (OA-HRF-NAQ-SVR) and with random forest regression (OA-HRF-NAQ-RFR). In addition, a decision-level fusion is tested among the single-feature-type regressors based on OA, HRF, and NAQ, respectively (OA-HRF-NAQ-DF), where a weighted average is used among the 3 predictions with a weight of W_h for the HRF regressor, a weight of W_n for the NAQ regressor, and a weight of $1 - W_h - W_n$ for the OA regressor. CQ is excluded from any feature- or decision-level fusion described above because of a relatively high SEM associated with it, as will be shown in the results.

3.4. Tested Parameter Settings

To see the effect of parameter settings on the performance of severity assessment, we test the following sets of alternative settings in the experiments:

$$N_m \in \{1, 2, 3, 4\} \quad (5)$$
$$N_o \in \{1, 2, ..., 9\} \quad (6)$$
$$\gamma \in \{0.0005, 0.001, 0.002, 0.005, ..., 50000\} \quad (7)$$
$$C \in \{0.2, 0.5, 1, 2, ..., 50000\} \quad (8)$$
$$\epsilon \in \{0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20\} \quad (9)$$
$$\eta \in \{e^{-14}, e^{-13.5}, ..., e^{4}\} \quad (10)$$
$$N_p \in \{1, 2, ..., 9\} \quad (11)$$
$$N_f \in \{1, 2, ..., 6\} \quad (12)$$
$$N_i \in \{1, 2, ..., 9\} \quad (13)$$
$$W_h \in \{0, 0.1, ..., 0.8\} \quad (14)$$
$$W_n \in \{0, 0.1, ..., 0.8\} \quad (15)$$

Instead of testing all the combinations of these settings across different parameters, we test only the combinations used in the course of a coordinate-descent procedure that minimizes the data-set average of SEM with respect to the parameters. This procedure is applied to each algorithm variant separately, where the parameters are updated one at a time and cyclically, each time with the lowest-error setting selected for the updated parameter. When this procedure is applied to the variant OA-HRF-NAQ-RFR, settings of the parameters N_m, N_o, η, and N_p are fixed at the optimal values from OA-HRF-NAQ-SVR, so that the two variants share the same set of features and the two regression algorithms can be compared. Note that whereas these experiments evaluate the generalization of regression with a data set separate from the training data for regression, they do not evaluate the generalization of parameter selection. The listener-rated voice data available in this study has not been sufficiently large for a division into training, validation, and test sets. With a new data set separate from the one used here, further experiments will be carried out in the future to evaluate the generalization of parameter selection.

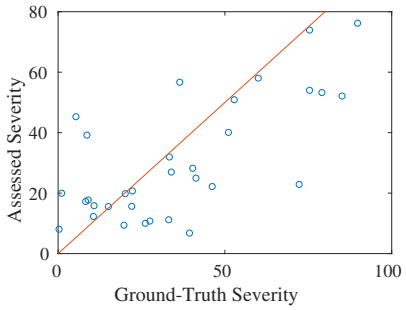

Figure 1: *Scatter plot (with an identity line) of ground-truth and assessed severities for the algorithm variant OA-SVR. The average SEM over these assessments is 15. A tendency of underestimation is observed.*

4. Results

Experimental results are presented in Table 1, which includes, for each algorithm variant, only the experiment that produced the lowest average SEM, out of all the experiments using different parameter settings for the same algorithm variant.

The lowest average SEM from OA-SVR is 15, which is a 32% reduction from the "prior-guess" SEM of 22. This accuracy of assessment is graphically demonstrated with a scatter

[1] If available, another source of data instead of this data set will be ideal for the evaluation of generalization in regression.

Table 1: *Best parameter settings and corresponding average SEMs (a lowest average SEM for each algorithm variant) from the results of severity assessment experiments. In case that a parameter is not applicable to a particular algorithm variant, the algorithm-parameter combination is identified with a dash.*

Algorithm	**SEM**	N_m	N_o	γ	C	ϵ	η	N_p	N_f	N_i	W_h	W_n
OA-SVR	15	3	8	0.1	200	10	e^{-12}	2	–	–	–	–
HRF-SVR	17	3	–	0.005	20000	10	e^{-12}	3	–	–	–	–
NAQ-SVR	18	4	–	2000	20	10	e^{-1}	3	–	–	–	–
CQ-SVR	20	2	–	0.2	20	10	e^{2}	2	–	–	–	–
OA-HRF-NAQ-SVR	15	4	9	0.02	200	10	$e^{-8.5}$	6	–	–	–	–
OA-HRF-NAQ-RFR	16	–	–	–	–	–	–	–	3	1	–	–
OA-HRF-NAQ-DF	14	–	–	–	–	–	–	–	–	–	0.3	0

plot shown in Fig. 1, which depicts the assessed and ground-truth severities from which this average SEM is calculated. The specific mechanisms in this algorithm for feature extraction and regression can be revealed by examining several parameter settings that resulted in this average SEM. First, the best setting for N_m is greater than 2, which suggests that higher-order, non-Gaussian feature statistics across the cycles in a glottal flow signal are useful. Secondly, the best setting of 8 for N_o suggests that the overall waveform shape of the glottal airflow, represented by multiple harmonic amplitudes, provides important information about voice severity. Thirdly, according to the optimal setting of 10 for ϵ, SVR is using training severity labels that are more than 10 units of rating away from the fitted prediction model. Fourthly, the optimal setting of e^{-12} for η suggests a relative tendency for the severity assessment algorithm to include a possibly non-voiced GCI interval for feature extraction. Extracting features from all GCI intervals can be feasible because only a small number of detected GCIs would actually occur at non-voiced time positions. Lastly, the best result was obtained by taking 2 principal components from 24 features, which confirmed the advantage of dimensionality reduction in the case of small training data.

A performance gain relative to OA-SVR is possible with alternative feature extraction or regression techniques. To investigate this possibility, we compare results between OA-SVR and other algorithm variants. A spectral feature that represents only the balance between fundamental and overtone energy, HRF did not give a lower average SEM than OA. Similarly, a lower average SEM did not result from the use of NAQ (in place of OA), which is limited to characterization of the maximum negative slope in the glottal-flow waveform. The utility of CQ in severity assessment is particularly limited by the difficulty in properly defining and estimating a glottal opening instant, which typically involves temporally locating a gradual ramp-up in the glottal airflow. This could explain a lowest average SEM from CQ-SVR that is the largest among all the variants. In relation to OA-SVR, none of the 3 fusion variants reduced the lowest average SEM by an amount greater than 1, which suggests that the information carried by HRF and NAQ on voice severity could be alternatively available from OA. Differences in lowest average SEM among the 3 fusion variants include a 1-unit increase resulting from substituting random forest regression for SVR, and a 1-unit decrease resulting from using decision-level fusion instead of feature-level fusion. The best weighting for decision-level fusion is 0.7 for OA and 0.3 for HRF. The contribution of HRF in this optimal fusion could result from the use of some higher harmonics beyond those used in OA.

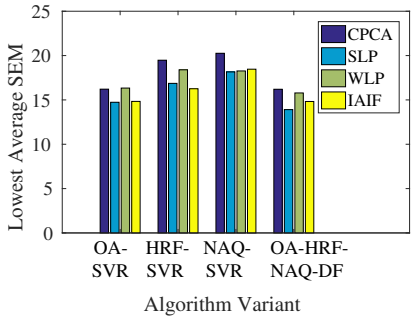

Figure 2: *Lowest average SEMs of several further algorithm variants that differ in the underlying inverse filtering algorithm.*

Best severity assessment results for several inverse filtering algorithms are presented in Fig. 2, where we have 3 algorithms in addition to SLP, i.e., closed-phase covariance analysis (CPCA) [18], weighted linear prediction (WLP) [19], and iterative adaptive inverse filtering (IAIF) [20]. For HRF-SVR, the lowest error was given by IAIF; in the other 3 cases, the lowest error was given by SLP.

5. Conclusions

An algorithm for voice severity assessment has been presented, which is a regressor that uses spectral and temporal features extracted from an estimate of glottal airflow. Results showed a better potential for harmonic features to generate accurate assessments than for temporal features. Future work includes comparison to existing acoustic measures of voice severity, and evaluation on a larger data set.

6. Acknowledgements

This work is sponsored in part by The Icelandic Centre for Research under Grant No 152705-051. The authors wish to thank Dr. Robert Hillman of the Center for Laryngeal Surgery and Voice Rehabilitation at Massachusetts General Hospital for making listener-rated voice data available for this study. The authors are also grateful to Dr. Daryush D. Mehta and Dr. Jarrad H. Van Stan of Massachusetts General Hospital for proofreading this paper.

7. References

[1] G. B. Kempster, B. R. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol," *American Journal of Speech-Language Pathology*, vol. 18, pp. 124–132, 2009.

[2] R. A. Prosek, A. A. Montgomery, B. E. Walden, and D. B. Hawkins, "An evaluation of residue features as correlates of voice disorders," *Journal of Communication Disorders*, vol. 20, no. 2, pp. 105–117, 1987.

[3] P. H. Dejonckere, M. Remacle, E. Fresnel-Elbaz, V. Woisard, L. Crevier-Buchman, and B. Millet, "Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements," *Revue de Laryngologie-Otologie-Rhinologie*, vol. 117, no. 3, pp. 219–224, 1996.

[4] T. Bhuta, L. Patrick, and J. D. Garnett, "Perceptual evaluation of voice quality and its correlation with acoustic measurements," *Journal of Voice*, vol. 18, no. 3, pp. 299–304, 2004.

[5] V. I. Wolfe, D. P. Martin, and C. I. Palmer, "Perception of dysphonic voice quality by naive listeners," *Journal of Speech, Language, and Hearing Research*, vol. 43, pp. 697–705, 2000.

[6] Y. Maryn, P. Corthals, P. V. Cauwenberge, N. Roy, and M. D. Bodt, "Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels," *Journal of Voice*, vol. 24, no. 5, pp. 540–555, 2010.

[7] S. N. Awan, N. Roy, M. E. Jetté, G. S. Meltzner, and R. E. Hillman, "Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the CAPE-V," *Clinical Linguistics & Phonetics*, vol. 24, no. 9, pp. 742–758, Sep. 2010.

[8] Y. Maryn, N. Roy, M. D. Bodt, P. V. Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality: A meta-analysis," *J. Acoust. Soc. Am.*, vol. 126, no. 5, pp. 2619–2634, 2009.

[9] V. Khanagha and K. Daoudi, "An efficient solution to sparse linear prediction analysis of speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 3, 2013.

[10] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012.

[11] Y.-R. Chien, D. D. Mehta, J. Guðnason, M. Zañartu, and T. F. Quatieri, "Performance evaluation of glottal inverse filtering algorithms using a physiologically based articulatory speech synthesizer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, submitted.

[12] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[13] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Am.*, vol. 112, no. 2, pp. 701–710, 2002.

[14] I. T. Jolliffe, *Principal component analysis*. Springer, 1986.

[15] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapoik, "Support vector regression machines," in *Advances in Neural Information Processing Systems 9*, 1996.

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[17] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 4, pp. 350–355, 1979.

[19] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *J. Acoust. Soc. Am.*, vol. 134, no. 2, pp. 1295–1313, 2013.

[20] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, pp. 109–118, 1992.