INVESTIGATING THE SPEECH CHARACTERISTICS OF SUICIDAL ADOLESCENTS

Stefan Scherer[1], *John Pestian*[2], *Louis-Philippe Morency*[1]

[1] University of Southern California, Institute for Creative Technologies, Playa Vista, CA
[2] Department of Pediatrics, Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati Cincinnati, OH

ABSTRACT

Suicide is a very serious problem. In the United states it ranks as the second most frequent cause of death among teenagers between the ages of 12 and 17. In this work, we investigate speech characteristics of prosody as well as voice quality in a dyadic interview corpus with suicidal and non-suicidal adolescents. In these interviews the adolescents answer specifically designed questions. Based on this limited dataset, we reveal statistically significant differences in the speech patterns of suicidal adolescents within the investigated interview corpus. Further, we investigate the classification capabilities of machine learning approaches both on an utterance as well as an interview level. The work shows promising results in a speaker-independent classification experiment based on only a dozen speech features. We believe that once the algorithms are refined and integrated with other methods, they may be of value to the clinician.

Index Terms— Suicide prevention, speech characteristics, voice source model, voice quality, classification

1. INTRODUCTION

Suicide among adolescents, which for the past 10 years has been the second most frequent cause of death among 12- to 17-year-olds, is a serious issue.[1]

Suicide risk factors include family history, demographics, mental illness co-morbidities, and nonverbal behavior and cues [1, 2, 3, 4, 5, 6]. There are, however, no standardized approaches for analyzing these nonverbal behaviors, which traditionally include gestures, facial expressions, and voice characteristics. This work investigates a method for classifying one of these nonverbal behaviors: acoustic characteristics of speech. In particular, we investigate prosodic and voice quality-related features from 16 suicidal and non-suicidal patients. These data are from a larger interview corpus that includes the speech of 30 suicidal and 30 non-suicidal adolescents. The data for these 16 patients are balanced according to gender and condition. We compare hidden Markov models (HMMs) and support vector machines (SVMs) to classify the speech of the subjects in the two categories. We compare the classification results on both a segment-based level of analysis as well as on an interview-based analysis in an entirely speaker-independent setup.

The remainder of the paper is organized as follows: In Section 2, we position this work within the current body of research and state the goals of this research. In Section 3, we introduce the dyadic interview dataset. Section 4 introduces the acoustic features utilized in the analysis of the speech. In Section 5, we detail the experimental setup to classify the speech of the adolescents on both an utterance-level and an interview-level analysis. We then statistically analyze the features in Section 6 and find significant differences between the groups. We further report the results of the speaker-independent classification task. In Section 7, we discuss the results, and in Section 8 we conclude the work and discuss future avenues of research.

2. RELATED WORK

Several researchers have investigated the correlates between severe depression, suicide, and the characteristics of speech. In [2], for example, the speech of 10 suicidal, 10 depressed, and 10 control subjects was analyzed in great detail. All subjects were males between the ages of 25 and 65. The data for the suicidal subjects were obtained from a large spectrum of recording setups comprising, for example, suicide notes recorded on tape. The other two groups were recorded under more controlled conditions at Vanderbilt University. For each subject the researchers concatenated speech to clips of 30 seconds of uninterrupted speech (i.e., removing pauses larger than 500ms). Then they analyzed jitter in the voiced parts of the signal as well as glottal flow spectral slope estimates. Both features helped to discern the classes in binary problems with high above-chance accuracies by utilizing simple Gaussian mixture model-based classifiers (e.g., control vs. suicidal 85% correct, depressed vs. suicidal 75% correct, control vs. depressed 90% correct). A holdout validation was employed. However, the fact that the recordings were done over such a large variance of recording setups, as acknowledged by the authors themselves, makes it difficult to assess "the accuracy about the extracted speech features and, therefore, the meaningfulness of the classification results." Nevertheless, the fact that the researchers have analyzed real-world data with speech recorded from subjects shortly before they attempted suicide is remarkable and needs to be acknowledged.

Further, in [7] a similar approach was utilized to assess the suicide risk of subjects with the same categories as in [2]. In [7], spectral density features were again used to classify the three classes in three separate binary problems. The data utilized comprised both interview data and read speech. It seems that the authors utilized a cross-validation approach for which it is not clear if the analysis was entirely speaker-independent, as they claim to have used randomized sets of 75% of the data for training and 25% of the data for testing. The observed accuracies are quite high: control vs. suicidal 90.25%, depressed vs. suicidal 88.5%, and control vs. depressed 92.0%.

The study in [4] involved the analysis of glottal flow features as well as prosodic features for the discrimination of depressed read speech of 15 male (nine controls and six depressed subjects, ages 33-50) and 18 female (nine controls and nine subjects, ages 19-57) speakers. In total, 65 sentences were recorded per speaker. The

[1] http://webappa.cdc.gov/sasweb/ncipc/leadcaus10_us.html

extracted glottal flow features are closely related to the Liljencrants-Fant model parameters used in the present work and comprised instances such as the minimal point in glottal derivative, maximum glottal opening, start point of glottal opening, and start point of glottal closing. The prosodic features extracted consist of fundamental frequency, energy, and speaking rate. The classification was performed on a leave-one-observation-out paradigm, which renders the analysis highly speaker-dependent. Hence, strong classification results were observed, well above 85% accuracy for male speakers and above 90% for female speakers. However, the main focus of the paper was not to find great classification results but rather to identify features that are often chosen to be relevant in the feature selection used. The authors identified glottal flow features to be selected for the majority of the classifiers as well as energy-based features for female speakers.

2.1. Prior Work Statement

The present work differs from that described above in several aspects, the most apparent being the analysis of the speech of adolescents between the ages of 13 and 17. In addition, we investigate novel acoustic features—voice source parameters and features relevant for the identification of voice qualities—that have to date not been utilized to characterize suicidal speech.

Lastly, we would like to emphasize that the present investigations are all based on a strict speaker-independent setup that has not been adequately studied in previous work. The partly speaker-dependent analysis in prior research might have led to overestimations of the classification accuracies. The present approach thus sacrifices percentage points of accuracy in order to reveal a more realistic performance.

2.2. Research Goals

(1) Based on the extracted features, we investigate the performance of HMMs and SVMs to classify each subject's voice into the categories suicidal and non-suicidal. We anticipate that we are capable of correctly classifying the suicidal adolescents on a temporally integrated interview level.

(2) We will investigate the performance of the classifiers on an utterance or segment level of speech.

(3) We will investigate which features contributed the most to the observed performances.

3. DATASET

From March 2011 through August 2011, 60 patients were enrolled in a prospective, controlled trial at the Cincinnati Children's Hospital Medical Center (CCHMC) ED (IRB#2008-1421). Eligible patients were between the ages of 13 and 17 and had come to the ED with suicidal ideation, gestures, attempts, or orthopedic injuries. Patients with orthopedic injuries were enrolled as controls because they are seen as having the fewest biological and neurological perturbations of all of the ED patients. Potential controls were excluded if they had a history of major mood disorder or if first-degree family members had a history of suicidal behavior. The parent(s) or legal guardian(s) had to consent to the study, the patients had to consent, and the physician(s) had to agree that the patients were appropriate for inclusion. Each patient received $75USD compensation for participation.

Data were collected by a trained social worker. Each subject completed the Columbia Suicide Severity Rating Scale (C-SSRS version 1/14/2009) [8], Suicidal Ideation Questionnaire - Junior (SIQ-Jr version 1987) [9], and the Ubiquitous Questionnaire (UQ version 2011) [10]. The UQ consists of five open-ended questions selected to elicit conversational responses: Does it hurt emotionally? Do you have any fear? Are you angry? Do you have any secrets? Do you have hope?

Potential subject and control patients were identified from the hospitals' electronic medical records. The attending physician was asked to determine whether the patient was appropriate for the study. If so, the parent(s) or legal guardian(s) were approached for consent. After that consent was obtained, the patient was then asked to consent. The same social worker interviewed all subjects.

All interviews were audio recorded and transcribed on a question-response level. The recordings were conducted in a private examination room using a tabletop microphone. The audio is sampled at 16 kHz with an average signal-to-noise ratio of 17.2 dB. Due to the fact that the interview was recorded with one single microphone, the speech utterances of both the interviewer and the interviewee are present on the single mono channel of the recordings. Hence, we manually annotated the speech segments using WaveSurfer.[2] The speech turns of a single interlocutor were segmented based on pauses greater than or equal to 300 ms. The resulting speech segments form the basis of the utterance-level analysis in Section 5. Overlapping speech was annotated separately.

For this work, we analyzed the interviews of 16 adolescents (eight female, eight male) with an average age of 15.53 years (σ = 1.5).[3] Eight of those had attempted suicide in the past; eight had not. All of those who had attempted suicide stated that they had wished to end their lives within the previous six months. The average length of the interviews with suicidal adolescents was 778.27 s (σ = 161.21), with 249.96 s (97.95 standard deviation) time spoken by the participant and 332.91 s (σ = 123.65) of pauses on average. The interviews with non-suicidal adolescents lasted for 451.55 s (σ = 107.01) on average, with 123.66 s (σ = 56.95) time spoken by the participant and 170.46 s (σ = 44.43) of pauses on average. The average length of a speech segment is 1.75 s (σ = 0.34) for suicidal adolescents and 1.66 s (σ = 0.52) for non-suicidal adolescents.

4. ACOUSTIC MEASUREMENTS

We analyzed the participants' prosody and voice quality using several acoustic measures, described below. The automatically extracted features were chosen based on previous findings in the literature (cf. Section 2). The abbreviations for the features used throughout the paper are shown italicized in parentheses after the paragraph headers. All features are sampled at 100 Hz.

Energy in dB (en, en_{slope}): The energy of each speech frame is calculated on 32 ms windows with a shift of 10 ms. Further, we calculate the slope, i.e., the first derivative, of the energy signal as a measurement of the change of speech intensity.

Fundamental frequency (f_0): We utilized the method in [11] for f_0 tracking based on residual harmonics, which is especially suitable in noisy conditions.

Peak slope ($peak$): This voice quality parameter is based on features derived following a wavelet-based decomposition of the speech signal [12]. The parameter, named *peak*, is designed to identify glottal closure instances from glottal pulses with different

[2] http://www.speech.kth.se/wavesurfer/

[3] This limited number is due to the time-consuming manual segmentation of the speech. In the future we anticipate utilizing the full corpus.

closure characteristics. It was used to differentiate between breathy, modal, and tense voice qualities in [13].

Spectral stationarity (ss): To characterize the range of the prosodic inventory used over utterances and the monotonicity of the speech, we make use of the so-called *spectral stationarity* measure ss. This measurement was previously used in [14] as a way of modulating the transition cost used in a dynamic programming method used for f_0 tracking.

LF model parameters from time domain estimation methods (R_a,R_k,R_g,EE,OQ,R_d): The most commonly used acoustic voice source model is the Liljencrants-Fant (LF) model [15]. It is a five-parameter (including f_0) model of differentiated glottal flow.

The model has two segments. The first segment, the open phase, is a sinusoid function. The second segment, which models the return phase, is an exponential function [16, 17].

The pulse shape of the LF model can be characterized using an amplitude parameter, EE (which is the negative amplitude corresponding to the main excitation), and three time-based parameters R_a, R_k, and R_g. These parameters have been shown to be suitable for characterizing a range of voice qualities, including breathiness and tenseness [18].

We extract the open quotient OQ with:

$$OQ = \frac{1 + R_k}{2 \cdot R_g}. \quad (1)$$

Further, R_d, which is characterizing the basic shape of the LF model [19], is extracted following [20].

Normalized amplitude quotient (NAQ): The normalized amplitude quotient parameter was introduced as a global voice source parameter capable of differentiating breathy to tense voice qualities [21] and is closely related to the R_d parameter described in [19].

Although NAQ is closely related to R_d, it is subtly but nevertheless significantly different: NAQ is a direct measure of the glottal flow and glottal flow derivative, whereas R_d is a measure derived from a fitted LF model pulse.

5. EXPERIMENTAL SETUP AND RESULTS

We investigate the capabilities of machine learning algorithms in a two-fold analysis: we investigate the accuracy of HMMs and SVMs both on an utterance level (cf. Section 3 for the definition of an utterance) and on an interview level. All experiments are conducted using a leave-one-speaker-out validation strategy. Hence, for the training of the classifiers in one fold, we leave out the speech samples of one speaker entirely from the training and test the classifiers on the speech of the left-out speaker. Note that due to the limited amount of data, we currently refrain from using parameter optimization and feature selection for the machine learning algorithms, leaving this for future analysis on the full dataset. We employ two separate classifiers, namely three-state HMMs with three mixtures for each state with full transition matrix and SVMs with radial basis function kernels. The HMMs can take advantage of the sequential and dynamic characteristics of the observations and classify each segment on the full 100Hz sampled feature vector. The SVMs, on the other hand, do not take previous observations into account and are trained on the median and standard deviations of the features over the single utterances.

5.1. Interview-Level Analysis

For the interview-level analysis, we integrate the decisions of the classifiers for each single speech segment and form an overall temporally integrated decision for each interview. Based on this we achieve an accuracy of 81.25% for the HMMs. The HMMs confused only three interviews; one of these misclassifications is a false negative (i.e., suicidal adolescent as non-suicidal). The accuracy of the SVMs was at 75%, with one additional confusion. Again, only one of them was a false negative classification, and it is the same interview as for the HMMs. The results are highlighted in Figure 1.

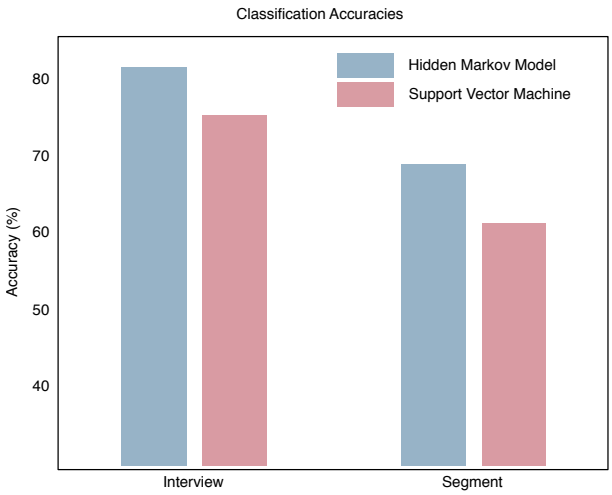

Fig. 1. Accuracies in % for the speaker-independent classification experiments on the interview and the utterance level. The performance of HMMs (blue) is compared to that of SVMs (red).

5.2. Utterance-Level Analysis

For the utterance-level analysis, we classify every single speech segment spoken by either a suicidal or non-suicidal adolescent. The overall classification accuracy for the HMMs is 69%. The overall accuracy for the SVMs is slightly lower at 61%.

6. STATISTICAL EVALUATION

In this section we report the results of our statistical investigations. We compare the observations of voice characteristics found in the speech of suicidal and non-suicidal adolescents. We conduct statistical tests on all 12 extracted parameters, as described in Section 4, and their standard deviations. Overall we conduct 24 independent t-tests with the very conservative Bonferroni correction for multiple testing [22]. Hence, the significance level p is adjusted to be at least $p < 0.002$. Additionally, we present Hedges' g value, as a measure of the effect size found in the data [23]. The g value denotes the required shift of the mean of one set to match the mean of the other in magnitudes of standard deviations [23]. Values of $g > 0.4$ are considered substantial effects.

The features with effect sizes close to the threshold of 0.4 for the suicidal vs. non-suicidal participants are summarized in Table 1. Unfortunately, due to space constraints, we are not able to show all the results. All the observed p-values for the listed features are smaller than the mentioned threshold of 0.002. To visualize the statistical results we plot the three strongest effects in Figure 2.

7. DISCUSSION

Based on our research goals outlined in Section 2, we discuss our findings in this section.

	Suicidal	Non-Suicidal	Hedges' g
$peak$	-.25 (.04)	-.23 (.05)	-.540
NAQ	.12 (.05)	.09 (.04)	.557
R_k	.36 (.12)	.30 (.10)	.495
R_g	1.43 (.58)	1.70 (.64)	-.450
OQ	.42 (.20)	.31 (.13)	.664
NAQ **Std.**	.08 (.03)	.06 (.02)	.618
EE **Std.**	.01 (.01)	.01 (.01)	-.448
R_k **Std.**	.12 (.07)	.10 (.05)	.396
R_g **Std.**	.61 (.24)	.74 (.34)	-.467
OQ **Std.**	.19 (.10)	.13 (.05)	.653

Table 1. Statistically significant acoustic measures discerning suicidal and non-suicidal adolescents. The mean and standard deviation (in parentheses) values as well as the effect sizes measured in magnitudes of standard deviations are summarized. All mentioned measures are statistically significantly different for the two groups with a p-value < 0.002.

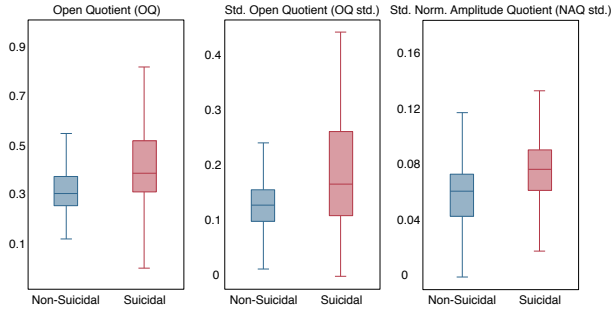

Fig. 2. Boxplot comparison of three significantly different acoustic measures for the two groups of suicidal and non-suicidal adolescents. All three measures indicate more a more breathy voice quality for suicidal subjects.

In the first research goal of the present work, we anticipate that the extracted features will be useful for classifying suicidal and non-suicidal adolescents' speech using standard machine learning algorithms. For the interview-level analysis, we could achieve classification accuracies of 81.25% for the HMMs and 75% for the SVMs after temporal integration. The HMMs confused only three interviews; one of these misclassifications is a false negative (i.e., suicidal adolescent as non-suicidal). In the future, we hope to further improve these results using more sophisticated temporal fusion algorithms and utilizing the full available dataset of 60 interviews.

As for our results on the second research goal, we were able to classify segments of speech with an average length of about 1.7 s with accuracies of 69% using HMMs and 61% using SVMs. The HMMs, which take advantage of the sequential and dynamic structure of the extracted features, are clearly outperforming the SVMs, which were trained on the median and standard deviation values of the features over each segment. These results are well above chance level and imply strong differences in the speech characteristics of suicidal and non-suicidal adolescents.

Based on our investigations on the third research goal, we want to identify the speech features that contributed the most to the classification. We could identify several statistically significant differences between the speech characteristics of suicidal and non-suicidal adolescents. The voice source and voice quality-related features show the strongest differences between the two groups. In particular, OQ, NAQ, and $peak$, features that have been associated with voice qualities on the breathy to tense dimension, reveal that suicidal adolescents' voices are often more breathy than the voice of non-suicidal subjects. Additionally, parameters of the LF model, such as R_k and R_g, reveal strong statistically significant differences. The larger R_k values suggest a more symmetric glottal pulse, which is again characteristic of a breathy phonation type. Similarly, a smaller R_g indicates a lower frequency of the glottal formant, which is also typical of a breathy voice.

Anecdotally, we would like to mention that not only do the speech characteristics of the suicidal and non-suicidal adolescents differ significantly, but also those of the interviewer himself. This phenomenon was observed while annotating the speech turns. In particular, the backchannels provided by the interviewer were significantly different in the two conditions. The observed significant differences include but are not limited to more breathy tones as observed with $peak$, lower speech intensity (en), higher monotonicity (ss), and larger OQ variations. This adaptation to the participant's voice and context implies that it is desirable to be able to adapt one's voice to the given situation [24] (e.g., for spoken human computer interaction systems or virtual agent systems as described in [25]). Current speech synthesis often lacks the capability to vary the produced voice along the voice quality domain. This finding of adaptation between interviewer and subject has also been confirmed by findings in the literature; for example, it was found that the clinician's behavior was strongly correlated with the patient's severity of depression [26].

8. CONCLUSIONS AND FUTURE WORK

Based on our research goals, the two major findings of this study are (1) based on the few extracted features, we could identify the speech of suicidal and non-suicidal adolescents with a high degree of accuracy in a speaker-independent analysis scenario, and (2) suicidal adolescents exhibit significantly more breathy voice qualities than non-suicidal subjects. We are confident that with some additional refinement we can provide professional healthcare providers with objective speech measures of suicidal patients to improve clinical assessments.

For future work, we are planning to first make use of the full corpus, including the speech of 60 suicidal and non-suicidal adolescents, and to start contextualizing responses of the subjects on a question-level basis. We believe that this more fine-grained analysis might reduce the number of errors significantly. Additionally, we seek to enrich the body of utilized features to incorporate more prosodic features such as the articulation rate as well as video-based features (e.g., gaze, smiles, gestures, and posture) in the interviews [6]. Lastly, we would like to mention that the investigated dataset is limited in its relevance with respect to everyday life conversations and general voice characteristics. We are planning to address this in the future and plan to assess the veracity and applicability of the reported results on a broader spectrum of interactions. With respect to this, we have already started a multi-center study designed, in part, to test the generalizability of our findings. With respect to this, we have already started a multi-center study designed, in part, to test the generalizability of our findings.

9. ACKNOWLEDGEMENTS

This work is supported by DARPA under contract (W911NF-04-D-0005) and U.S. Army Research, Development, and Engineering Command and. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

10. REFERENCES

[1] W. J. Fremouw, M. Perczel, and T. E. Ellis, *Suicide Risk: Assessment and Response Guidelines*, Pergamon, 1990.

[2] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 9, pp. 1530–1540, 2004.

[3] J. A. Hall, J. A. Harrigan, and R. Rosenthal, "Nonverbal behavior in clinician-patient interaction," *Applied and Preventive Psychology*, vol. 4, no. 1, pp. 21–37, 1995.

[4] M. Elliott, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 1, pp. 96–107, 2008.

[5] D. J. France, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[6] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *accepted for publication at IEEE Conference on Automatic Face and Gesture Recognition*, 2013.

[7] T. Yingthawornsuk, *Acoustic Analysis of Vocal Output Characteristics for Suicidal Risk Assessment*, Ph.D. thesis, Vanderbilt University, 2007.

[8] K. Posner, G. K. Brown, B. Stanley, D.A. Brent, K. V. Yershova, M. A. Oquendo, G. W. Currier, G. A. Melvin, L. Greenhill, S. Shen, and J. J. Mann, "The ColumbiaSuicide severity rating scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults," *The American Journal of Psychiatry*, vol. 168, no. 12, pp. 1266–1277, Dec 2011.

[9] W. M. Raynolds, *Suicidal Ideation Questionnaire - Junior*, Odessa, FL: Psychological Assessment Resources, 1987.

[10] J. P. Pestian, "A conversation with Edwin Shneidman," *Suicide and Life-Threatening Behavior*, vol. 40, no. 5, pp. 516–523, 2010.

[11] T. Drugman and A. Abeer, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proceedings of Interspeech 2011*. 2011, pp. 1973–1976, ISCA.

[12] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform," in *Proceedings of Interspeech 2011*. 2011, pp. 177–180, ISCA.

[13] S. Scherer, J. Kane, C. Gobl, and F. Schwenker, "Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification," *Computer Speech and Language*, vol. 27, no. 1, pp. 263–287, 2013.

[14] David Talkin, "A Robust Algorithm for Pitch Tracking," in *Speech coding and synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., pp. 495–517. Elsevier, 1995.

[15] G. Fant, J. Liljencrants, and Q. Lin, "A four parameter model of glottal flow," *KTH, Speech Transmission Laboratory, Quarterly Report*, vol. 4, pp. 1–13, 1985.

[16] C. Gobl and A. Ní Chasaide, "Amplitude-based source parameters for measuring voice quality," in *Proceedings of ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03)*. 2003, pp. 151–156, ISCA.

[17] C. Gobl, "The voice source in speech communication," *Ph. D. Thesis, KTH Speech Music and Hearing, Stockholm*, 2003.

[18] C. Gobl, "A preliminary study of acoustic voice quality correlates," *KTH, Speech Transmission Laboratory, Quarterly Report*, vol. 4, pp. 9–21, 1989.

[19] G. Fant, J. Liljencrants, and Q. Lin, "The LF-model revisited. transformations and frequency domain analysis," *KTH, Speech Transmission Laboratory, Quarterly Report*, vol. 2-3, pp. 119–156, 1995.

[20] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 19, no. 5, pp. 1080–1090, 2011.

[21] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parameterization of the glottal flow," *Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[22] O. J. Dunn, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, pp. 52–64, 1961.

[23] L. V. Hedges, "Distribution theory for glass's estimator of effect size and related estimators," *Journal of Educational Statistics*, vol. 6, no. 2, pp. 107–128, 1981.

[24] S. Kopp, "Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors," *Speech Communication*, vol. 52, no. 6, pp. 587–597, 2010.

[25] S. Scherer, S. Marsella, G. Stratou, Y. Xu, F. Morbini, A. Egan, A. Rizzo, and L.-P. Morency, "Perception markup language: Towards a standardized representation of perceived nonverbal behaviors," in *Proceedings of Intelligent Virtual Agents (IVA'12)*. 2012, LNAI 7502, pp. 455–463, Springer.

[26] A. L. Bouhuys and R. H. van den Hoofdakker, "The interrelatedness of observed behavior of depressed patients and of a psychiatrist: an ethological study on mutual influence," *Journal of Affective Disorders*, vol. 23, pp. 63–74, 1991.