Interspeech 2018
2-6 September 2018, Hyderabad

Automatic Evaluation of Soft Articulatory Contact for Stuttering Treatment

Keiko Ochi[1] , *Koichi Mori*[2], *Naomi Sakai*[2]

[1] Tokyo University of Technology, Hachioji, Japan
[2]National Rehabilitation Center for Persons with Disabilities, Tokorozawa, Japan
[3]Research Institute, National Rehabilitation Center for Persons with Disabilities, Tokorozawa, Japan

ochikk@stf.teu.ac.jp, mori-koichi@umin.ac.jp, sakai-naomi@rehab.go.jp

Abstract

We describe a new method for the automatic discrimination and evaluation of phonation beginning with a consonant with soft articulatory contact, which is used in the treatment of stuttering, and normal phonation. Soft articulatory contact is trained to relax articulators and remove hard contacts that occur during stuttering. We use features related to the changes in acoustic characteristics and the voice quality under the hypothesis that the slowing down of articulatory movement of the initial consonant and the relaxing of phonatory muscles co-occur with soft articulatory contact. The results of an experimental evaluation showed that high accuracy was obtained when acoustic features were related to the peaks of the first derivative of the mel frequency cepstral coefficients (MFCCs) corresponded to the slowing down of the movement of the articulators. The features of vocal quality only slightly contributed to the classification.

Index Terms: Articulation, speech treatment, stuttering

1. Introduction

Stuttering is a common speech disorder in the general population with an incidence of 5% in children and 1% in adults [1]. It is characterized by the core symptoms of the repetition and prolongation of sounds and "blocks" (abnormal stoppages of phonation, airflow and/or gestures of articulation). As stuttering becomes persistent, the duration of the blocks tends to increase and their tension becomes stronger, sometimes accompanied by a tremor. People who stutter (PWS) have difficulty in oral communication. Some of them, especially those who choose to seek therapy, often have difficulties in social activities and negative attitudes toward communication [2]; 50% of them meet the diagnostic criteria of social anxiety disorder [3]. The quality of life (QOL) of PWS in terms of the vitality, social functioning, emotional functioning and mental health status has been reported to be negatively affected by stuttering [4].

Fluency shaping is one of the most widely used treatment methods for stuttering [1]. It changes the speech pattern to one that can reduce stuttering. It includes soft voice onset for the training of words beginning with a vowel and soft articulatory contact for the training of words beginning with consonants. Soft articulatory contact is used to mitigate stutterers' too strong contact between the lips or between the tongue and other articulation organs in producing plosives and affricates (e. g., the contact between the upper anterior teeth and the tip of the tongue when producing the plosive /t/). The muscle tension caused by blocks can be prevented by soft articulatory contact.

Computer-assisted fluency shaping therapies have been carried out to facilitate self-practice in intensive training and follow-up training after hospitalization, and the effect was demonstrated in a large participant study [5]. Because the time that a patient can receive face-to-face treatment by a clinician is limited, repeated self-practice is crucial to enable the patient to use the new speech pattern in everyday life. We previously proposed a method for discriminating soft and hard voice onsets, at the beginning of the initial vowel of words that used acoustic parameters related to whether the glottis closes or opens before the oscillation begins [6].

Soft articulatory contact has conventionally been evaluated by measuring the onset of the vowel following the initial consonant of words by either visually or using the amplitude contour [7][8]. While the correlation between the perceived softness of voice onset of vowels and the rise time (RT) for phonation is reasonably high [9], the previous training system [7] uses the RT for words which begins with both vowels and consonants. Thus, a gradual increase in volume is scored highly, though too long phonation is penalized.

There have been many studies on the automatic measurement of characteristics of consonants, such as voice onset time (VOT) [10]-[12], or the use of acoustic features of consonants to detect or assess disordered speech [13]-[16]. Regarding stuttering, methods for detecting repetition and/or prolongation have been proposed [17]-[19]. However, methods for automatically evaluating the consonant with soft articulatory contact itself have not been developed.

In this study, we propose a method for the automatic discrimination of soft and hard articulatory contacts with the aim of developing a computer-based speech training system for stuttering, specifically for modifying consonant production, that uses acoustic features that do not depend on the RT. We focus on the change in acoustic characteristics caused by reducing the speed of articulatory motion. The first set of features relates to the decrease in the speed of plosions and the slow transition of the formants. The second set relates to the soft voice quality resulting from reducing the tension of the muscles involved in the adjustment of the vocal folds. Using these acoustic features, it is expected that the patient can learn to eliminate excessively abrupt articulatory motions, preventing the acquisition of an unnatural phonation pattern that takes time too much to increase the volume, without the need for special equipment such as the electromyogram (EMG) used in a biofeedback method [20].

2. Conventional Acoustic Features

One of the computer-based scoring systems uses the average magnitude profile (AMP), defined as the sum of the absolute amplitude of speech signal samples within a windowed frame [7]. The score of a given patient's utterance is calculated on

10.21437/Interspeech.2018-2544

Table 1: *Proposed acoustic features*

	Feature	Dimensions
AM features	$\boldsymbol{g}_p$	2
	T_p	1
	$\boldsymbol{d}_M$	13
	$\boldsymbol{d}_m$	13
	$\boldsymbol{p}_M$	13
VQ features	H1-H2	1
	H1-H3	1
	CPP	1

the basis of the time until the AMP value increases to the upper threshold. It has been reported that the time until the AMP value exceeds an upper threshold is up to 600 ms in an easy onset of phonation [7]. The patient receives a score of 100% if the difference between the therapist's rise time and the patient's rise time is within 5 ms, and the score is reduced as the time difference increases.

Whereas the rise time described above is used to score the patient's phonation, it was compared with the proposed features in a simple discrimination task in this study. The RT feature is defined as the time in which the root mean square (RMS) of the sound signal increases from 10% to 90% of the maximum value to simplify the calculation.

The RMS of the sound signal is calculated with a window length of 20 ms and a sliding window step of 10 ms. Although the maximum magnitude of the AMP and the duration of the phonation were also used for scoring in [7], we do not use them because our aim is to discriminate the phonation regardless of the uttered word or sentence, in which case these parameters are not well defined or useful.

3. Proposed Acoustic Features

As it is assumed that a plosion is smaller and/or slower when a consonant is softly produced, the resultant formant transition will be smaller and/or slower. Accordingly, we tested eight acoustic features with a total of 45 dimensions shown in Table 1. The features consist of articulatory motivated (AM) and voice quality (VQ) subsets.

3.1. Articulatory Motivated Features

To capture the differences in the power and in the rate of change of the vocal tract shape, we used features related to the dynamic features of the mel frequency cepstral coefficients (ΔMFCCs). Let the ΔMFCC vector at the kth frame be defined as

$$\boldsymbol{d}^{[k]} = \left(d_1^{[k]}, d_2^{[k]}, \cdots, d_n^{[k]}\right)^{\mathrm{T}}, \tag{1}$$

where n is the number of dimensions.

The n-dimensional vectors $\boldsymbol{d}_M$ and $\boldsymbol{d}_m$ have components that respectively represent the maximum and minimum values of the components of $\boldsymbol{d}^{[k]}$ within the interval from the K_sth to K_eth frame at the beggining of the phonation. $\boldsymbol{p}_M$ is defined as the vector of the maximum values of the squared components of $\boldsymbol{d}^{[k]}$. The ith components of $\boldsymbol{d}_M$, $\boldsymbol{d}_m$, and $\boldsymbol{p}_M$ are given by

$$d_{M,i} = \max_{K_s \le k \le K_e} d_i^{[k]}, \tag{2}$$

$$d_{m,i} = \min_{K_s \le k \le K_e} d_i^{[k]}, \tag{3}$$

$$p_{M,i} = \max_{K_s \le k \le K_e} \left(d_i^{[k]}\right)^2, \tag{4}$$

respectively. The acoustic features $\boldsymbol{g}_p$ and T_p are defined as the vector of the first two peak values of the sequence of the norm of the ΔMFCC vector and the time interval between the two peaks, respectively. When the squared norm of $\boldsymbol{d}$ is defined as

$$\|\boldsymbol{d}^{[k]}\|^2 = \boldsymbol{d}^{[k]^{\mathrm{T}}} \boldsymbol{d}^{[k]}, \tag{5}$$

$\boldsymbol{g}_p$ and T_p are described by

$$\boldsymbol{g}_p = \left(\|\boldsymbol{d}^{[k_{p1}]}\|^2, \|\boldsymbol{d}^{[k_{p2}]}\|^2\right)^{\mathrm{T}}, \tag{6}$$

$$T_p = (k_{p2} - k_{p1})\, s, \tag{7}$$

where the sequence $\{\|\boldsymbol{d}^{[k]}\|^2; k = 1, 2, \cdots, K\}$, which is defined as the "delta curve" g, has its first and second peaks at the k_{p1}th and k_{p2}th frames, respectively. The peaks are between the K_s'th and K_e'th frames. A peak appears when the spectral characteristics change. We extracted the first and second peaks to capture the change in acoustic characteristics at the boundaries before and after the first consonant. Because we observed that adding Δ^2MFCC features did not improve the performance in a preliminary experiment, we omitted them in this experiment.

3.2. Voice Quality Features

Because the voice quality is assumed to become relaxed when speaking with soft articulatory contact, we used three acoustic features related to voice quality: H1-H2, H1-H3 and cepstral peak prominent (CPP) [21]. The relative amplitude of the fundamental frequency component to the upper harmonics increases with the phonation with "relaxed" separation of the arytenouds because the open quotient becomes longer [22]. This difference in the acoustic characteristics can be quantified by the difference in the amplitude of the first and second harmonics (H1-H2) [22], of the first and third harmonics (H1-H3) or the CPP.

H1-H2 and H1-H3 are features used to measure the slope of the amplitude spectrum in the region relatively unaffected by the formants. They are defined as the difference between the amplitudes of the first harmonic and second or third harmonic in the log domain. They are defined as follows:

$$\text{H1-H2} = 20\log_{10} a(2F_0) - 20\log_{10} a(F_0)\ [\text{dB}], \tag{8}$$

$$\text{H1-H3} = 20\log_{10} a(3F_0) - 20\log_{10} a(F_0)\ [\text{dB}], \tag{9}$$

where $a(f)$ is the amplitude spectrum of the waveform and F_0 is the fundamental frequency.

In a periodic signal, the peak of the cepstrum corresponds to the fundamental frequency (F_0). The amplitude of the cepstral peak has a larger value when the harmonic structure is prominent because the amplitudes of the high-order harmonic components are large. The CPP is defined by

$$\text{CPP} = \max_k c(T_0) - a_c T_0 - b_c, \tag{10}$$

where $c(T_0)$ is the value of the cepstrum at the quefrency of the fundamental period T_0 $(= 1/F_0)$ s, and a_c and b_c are the slope and intercept of the regression line of the cepstrum, respectively.

4. Experimental Evaluation

4.1. Speech Samples

Eleven male PWS and three speech therapists (STs) (two males, one female), who were Japanese native speakers, uttered 36–56 Japanese words or single morae, beginning with a consonant, with instructions to make the utterance with normal articulatory contact (NAC) or soft articulatory contact (SAC). SAC was used only at the first consonant, since stuttering occurs most often in the beginning of a sentence or word [23][24].

The PWS were partly recruited from the patients of the hospital of the National Rehabilitation Center for Persons with Disabilities (NRCD) in Japan, and recruited from the Research Institute of NRCD. This study has been approved by the ethics board of NRCD. Five out of the eleven PWS received training in SAC by an ST. All PWS were instructed how to produce phonation with SAC by an experimenter using the instruction manual created by the STs in NRCD. All of the STs had at least one year of clinical experience with stuttering.

Each word was uttered one to three times. The first consonant of the word was /k/, /g/, /t/, /d/, /p/ or /b/. The total number of speech samples was 2400. Recording was conducted in a sound-attenuated chamber using a headset microphone (AKG C420) connected to a personal computer through a USB audio interface (MOTU 828). The sampling rate and the number of quantization bits were 48 kHz and 24, respectively. The audio data were downsampled to 16 kHz, requantized to 16 bits and analyzed.

4.2. Methods

4.2.1. Listening Evaluation

To label whether or not words were phonated with SAC, an ST not included in the speakers evaluated the weakness of the articulatory contact of the initial consonants (1 = 'very strong'; 5 = 'very weak'). The ST was asked to score 336 speech samples on a five-point scale, which consisted of 12 utterances randomly selected from each speaker's utterances. Three samples uttered by each speaker were presented twice at random to confirm the consistency of the scoring; the ST listened a total of 420 samples. The speech samples were normalized by the mean RMS value. The Pearson's correlation coefficient between two evaluations of the same voice was 0.81.

Figure 1 shows histograms of the scores of utterances by the PWS group and ST group with intended SAC and NAC. All STs utterances except for one had a score of four or five. Thus, we assumed that the utterances scoring four or five could be considered as SAC and the others could be considered as NAC, and we labeled the utterances accordingly. We used the mean score for the utterances that were presented twice. Table 2 presents the percentage of utterances with intended NAC and SAC that were scored as 4 or 5. Note that repeated practice of SAC is required to use it in daily life even if the patients can produce utterances with SAC in a laboratory or training room setting.

4.2.2. Feature Extraction

The speech signals were first normalized in amplitude and analyzed frame by frame to obtain the MFCCs, H1-H2, H1-H3, and CPP. Auditory Toolbox [25] for MATLAB was used to calculate the 13-order MFCCs, using a 16 ms frame length and 10 ms frame shift. We used a 25 ms frame length and 10 ms frame shift and applied a Hanning window for the calculation of the other features.

K_s and K_e were heuristically set to the frame numbers at the utterance start time and 100 ms later, respectively. The utterance start time was defined as the point when the RMS value exceeds a threshold. K'_s and K'_e were the frame numbers 100 ms before the utterance start time and 200 ms after the utterance start time, respectively. A voiced-unvoiced decision was made for each frame on the basis of whether the maximum magnitude of the autocorrelation function exceeded a threshold. These thresholds were determined experimentally.

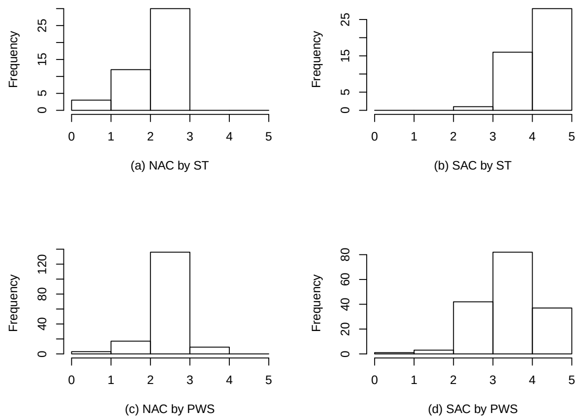


Figure 1: *Histograms of (a) NAC and (b) SAC intended utterances by STs and (c) NAC and (d) SAC intended utterances by PWS.*

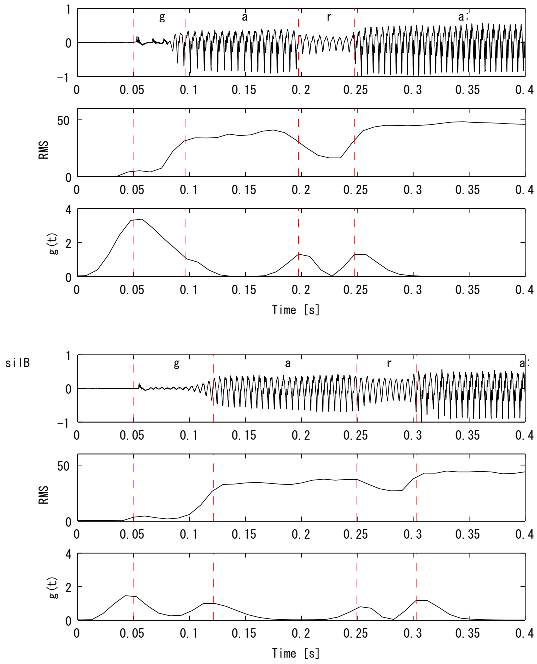
Table 2: *Percentage of the utterances with intended NAC and SAC by PWS scored as 4 or 5 by the ST*

Speaker	Treatment	NAC	SAC
S1	-	0.0	8.3
S2	-	0.0	25.0
S3	-	0.0	41.7
S4	-	8.3	75.0
S5	-	0.0	91.7
S6	-	0.0	100.0
S7	+	16.7	58.3
S8	+	8.3	75.0
S9	+	8.3	100.0
S10	+	0.0	100.0
S11	+	0.0	100.0

The amplitudes of the first to third harmonics were obtained by linearly interpolating between the neighboring data points of the spectrum. To calculate the CPP, it was assumed in this study that the fundamental frequencies were in the range of 50–400 Hz, and we measured the cepstral peak between 2.5 and 20 s. We used the average values of H1-H2, H1-H3 and CPP over the voiced frames between the utterance start time and 200 ms later.

We also discriminated between SAC and NAC and estimated scores of weakness of articulatory contact by using a support vector machine (SVM) and support vector regression. We used the kernlab [26] package implemented for the statistical environment R. We used the Gaussian radial basis kernel. The classifier was evaluated by leave-one-speaker-out cross-validation. The classification and estimation were conducted using the conventional feature (RT), the articulatory motivated features (subset AM), the voice quality features (subset VQ), subsets AM and VQ (AM+VQ), and all features (AM+VQ+RT).

The dimension of features used for classification and regression were selected by a forward stepwise selection algorithm. In this method, the set of features starts with an empty basis and then the best feature is added "greedily" in each step.



(b)

Figure 2: *Waveforms, RMS and delta curve of (a) normal and (b) soft contact utterance /gara:ki/ ("completely empty").*

We used the criterion of the highest accuracy.

5. Results and Discussion

Figure 2 shows examples of the waveforms, RMS and delta curve of utterances with NAC and SAC. In the utterance with SAC, the amplitude of the plosion of /g/ with SAC is weaker than that with NAC. The first peak of delta curve for NAC is larger than that for SAC. The magnitude of the delta curve for NAC remains large during the transition from /g/ to /a/, while the delta curve for SAC has two small peaks at the boundary between the silence and /g/ and at the boundary between /g/ and /a/.

Table 3 presents the maximum accuracy rates of the classification and the classification and correlation coefficients between the score estimated by the regressor and the score rated by the ST. Figure 3 show the accuracies of the classifiers trained by the features selected by stepwise feature selection for the classification. The accuracy was maximum when the classifier was trained with 13 selected dimensions. Even the accuracy of the classifier trained by one proposed feature, which was selected in the first step, was higher than that of the RT. Table 4 shows the selected dimensions with the highest accuracy of the classifier and the highest correlation coefficient of the regressor.

The accuracy did not decrease when the RT was removed from the candidates in the dimension selection. The accuracy was maximum when the classifier was trained by 13 selected dimensions. This indicates that the proposed method can discriminate SAC sufficiently well without using the slowness of the increase in volume. The accuracy was only slightly improved when subset VQ was added to subset AM. The accuracy was low with only subset VQ (66.1%). The results imply that the soft vocal quality does not always co-occur with soft articulatory contact.

The score estimated using subset AM was highly correlated

Table 3: *Accuracies of the classification and correlation coefficients between the estimated and rated weakness scores (r) when the best dimensions of features was selected.*

Method	Accuracy [%]	r
RT	61.3	0.23
AM	88.0	0.71
VQ	66.1	0.18
AM+VQ	89.3	0.72
AM+VQ+RT	88.3	0.72

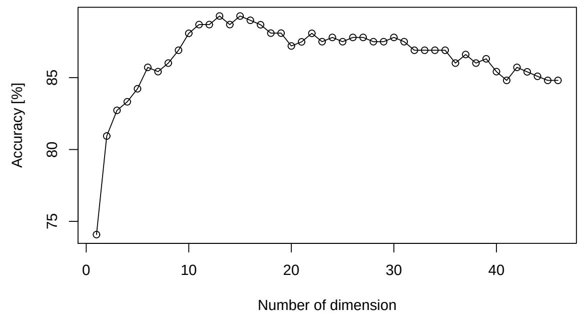


Figure 3: *Accuracy obtained by stepwise selection.*

with the rated score ($r = 0.71$), whereas that estimated using subset RT or VQ had a low correlation with the rated score. This indicates that the features of subset AM may be the descriptors of the weakness of the articulatory contact.

6. Conclusions and Future Work

In this study, we presented a new method for discriminating soft and normal articulatory contact of the initial consonants of words. By training a classifier using features related to the changes in acoustic characteristics, better discrimination can be obtained than when using the rise time of the amplitude. This indicates that PWS would be able to use the evaluation system for their training to correct or reduce their abrupt articulation pattern, which could be an essential training tool for reducing stuttering. In future, we will develop a speech training system for a PC or mobile phone by utilizing the proposed acoustic features in the present study.

7. Acknowledgements

This work was supported by a JSPS KAKENHI Grant-in-Aid for Young Scientists (B) (Grant Number: 20623713) and a Grant-in-Aid for Scientific Research (C) (Grant Number: 17K01495). The authors would like to thank Prof. N. Ono for useful discussions.

Table 4: *Selected dimensions of the acoustic features*

	Selected dimensions
Classification	$\boldsymbol{p}_{M,1}$ $\boldsymbol{d}_{M,10}$ $\boldsymbol{p}_{M,4}$ $\boldsymbol{g}_{p,2}$ $\boldsymbol{g}_{p,1}$ $\boldsymbol{d}_{m,5}$ $\boldsymbol{d}_{M,1}$ $\boldsymbol{p}_{M,13}$ $\boldsymbol{d}_{M,3}$ CPP $\boldsymbol{p}_{M,7}$ $\boldsymbol{d}_{m,8}$ $\boldsymbol{d}_{M,4}$
Estimation	$\boldsymbol{d}_{M,1}$ $\boldsymbol{g}_{p,2}$ $\boldsymbol{d}_{M,2}$ $\boldsymbol{d}_{m,13}$ $\boldsymbol{d}_{m,4}$ $\boldsymbol{p}_{M,13}$ H1-H2 $\boldsymbol{d}_{m,1}$ $\boldsymbol{d}_{M,13}$ $\boldsymbol{p}_{M,1}$ $\boldsymbol{p}_{M,5}$ $\boldsymbol{d}_{m,2}$ $\boldsymbol{p}_{M,6}$ $\boldsymbol{d}_{M,7}$ $\boldsymbol{d}_{M,9}$ $\boldsymbol{p}_{M,11}$ $\boldsymbol{g}_{p,1}$ $\boldsymbol{p}_{M,3}$ $\boldsymbol{d}_{M,8}$ T_p

8. References

[1] B. Guitar, *Stuttering: An integrated approach to its nature and treatment,* Lippincott Williams & Wilkins, 2013.

[2] M. Vanryckeghem and G. J. Brutten, "The BigCAT: A normative and comparative investigation of the communication attitude of nonstuttering and stuttering adults," *J. Commun. Disord.*, vol. 44, no. 2, pp. 200-206, 2011.

[3] F. W. Kraaimaat, M. Vanryckeghem and R. Van Dam-Baggen, "Stuttering and social anxiety", *J. Fluency Disord.*, vol. 27, no. 4, pp. 319-331, 2002.

[4] A. Craig, E. Blumgart and Y. Tran, "The impact of stuttering on the quality of life in adults who stutter," *J. Fluency Disord.*, vol. 34, no. 2, pp. 61-71, 2009.

[5] H. A. Euler, A. W. V. Gudenberg, K. Jung and K. Neumann, "Computer-assisted therapy for speech disfluency: The long-term effectiveness of the Kassel Stuttering Therapy (KST)," *Sprache, Stimme, Gehör*, vol. 33 no. 4, pp. 193-202, 2009.

[6] K. Ochi, K. Mori, N. Sakai and N. Ono, "Automatic discrimination of soft voice onset using acoustic features of breathy voicing," *Proc. INTERSPEECH 2016*, pp. 267-271.

[7] S. S. Awad, "The application of digital speech processing to stuttering therapy," *Proc. IMTC/97*, Vol. 2, pp. 1361-1367, 1997.

[8] S. S. Awad and C. Piechocki, "Speech therapy software on an open web platform," *Proc. ICENCO 2014*, pp. 53-56, 2014.

[9] H. F. Peters, L. Boves and I. C. van Dielen, "Perceptual judgment of abruptness of voice onset in vowels as a function of the amplitude envelope," *Journal of Speech and Hearing Disorders*, vol. 51, no. 4, pp. 299-308, 1986.

[10] M. Sonderegger and J. Keshet, "Automatic measurement of voice onset time using discriminative structured prediction," *Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3965–3979, 2012.

[11] A. P. Prathosh, A. G. Ramakrishnan and T. V. Ananthapadmanabha, "Estimation of voice-onset time in continuous speech using temporal measures," *Journal of the Acoustical Society of America*, vol. 136, no. 2, EL122128, 2014.

[12] Y. Adi, J. Keshet, O. Dmitrieva and M. Goldrick, "Automatic measurement of voice onset time and prevoicing using recurrent neural networks," *Proc. INTERSPEECH 2016*, pp. 3152–3155, 2016.

[13] A. Maier, F. Hönig, C. Hacker, M. Schuster and Nöth, "Automatic evaluation of characteristic speech disorders in children with cleft lip and palate," *In Ninth Annual Conference of the International Speech Communication Association*, 2014.

[14] M. Novotný, J. Rusz, R. Čmejla and E. Růžička, "Automatic evaluation of articulatory disorders in Parkinson's disease," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 2, no. 9, pp. 1366-1378, 2014.

[15] A. Maier, T. Haderlein, F. Stelzle, E. Nöth, E. Nkenke, F. Rosanowski and M. Schuster, "Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, 926951, 2009.

[16] L. He, J. Zhang, Q. Liu, H. Yin and M. Lech, "Automatic evaluation of hypernasality and consonant misarticulation in cleft palate speech," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1298–1301, 2014.

[17] J. Zhang, B. Dong and Y. Yang, "A computer-assist algorithm to detect repetitive stuttering automatically," *IALP 2013*, pp. 249–252, 2013.

[18] P. A. Heeman, A. McMillin and J. S. Yaruss, "Computer-assisted disfluency counts for stuttered speech," *In Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[19] P. S. Savin, P. B. Ramteke and S. G. Koolagudi, "Recognition of Repetition and Prolongation in Stuttered Speech Using ANN," *Proc. 3rd International Conference on Advanced Computing, Networking and Informatics,* pp. 65–71, 2016.

[20] A. R. Craig and P. J. Cleary, "Reduction of stuttering by young male stutterers using EMG feedback," *Appl. Psychophysiol Biofeedback*, vol. 7, no. 3, pp. 241-255, 1982.

[21] J. Hillenbrand, R. A. Cleveland and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *J. Speech Lang. Hear. Res.*, vol. 37, no. 4, pp. 769-778, 1994.

[22] D. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice quality variations among female and male talkers, *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820-857, 1990.

[23] U. Natke, P. Sandrieser, M. van Ark, R. Pietrowsky and K. T. Kalveram, "Linguistic stress, within-word position, and grammatical class in relation to early childhood stuttering," *J. Fluency Disord.*, vol. 29, no. 2, pp. 109-122, 2004.

[24] M. Jayaram, "Distribution of stuttering in sentences: relationship to sentence length and clause position," *J. Speech Lang. Hear. Res.*, vol. 27, no. 3, pp. 338-341, 1984.

[25] M. Slaney, "Auditory Toolbox", *Interval Research Corporation, Tech. Rep*, 10, pp. 29-32, 1998.

[26] A. Zeileis, K. Hornik, A. Smola andA. Karatzoglou, "kernlab - an S4 package for kernel methods in R," *J. Stat. Softw.*, vol. 11, no. 9, pp. 1-20, 2004.

[27] B. Caputo, K. Sim, F. Furesjo andA. Smola, "Appearance-based object recognition using SVMs: which kernel should I use?," *Proceedings of NIPS workshop on Statitsical Methods for Computational Experiments in Visual Processing and Computer Vision, 2002*, 2002.