Original Article

An Automatic Prolongation Detection Approach in Continuous Speech With Robustness Against Speaking Rate Variations

Abstract

In recent years, many methods have been introduced for supporting the diagnosis of stuttering for automatic detection of prolongation in the speech of people who stutter. However, less attention has been paid to treatment processes in which clients learn to speak more slowly. The aim of this study was to develop a method to help speech-language pathologists (SLPs) during diagnosis and treatment sessions. To this end, speech signals were initially parameterized to perceptual linear predictive (PLP) features. To detect the prolonged segments, the similarities between successive frames of speech signals were calculated based on correlation similarity measures. The segments were labeled as prolongation when the duration of highly similar successive frames exceeded a threshold specified by the speaking rate. The proposed method was evaluated by UCLASS and self-recorded Persian speech databases. The results were also compared with three high-performance studies in automatic prolongation detection. The best accuracies of prolongation detection were 99 and 97.1% for UCLASS and Persian databases, respectively. The proposed method also indicated promising robustness against artificial variation of speaking rate from 70 to 130% of normal speaking rate.

Keywords: *Attention, language, learning, pathologists, speech, speech-language pathology, stuttering*

Iman Esmaili, Nader Jafarnia Dabanloo, Mansour Vali[1]

Biomedical Engineering Department, Science and Research Branch, Islamic Azad University, Tehran, Iran, [1]Electrical and Computer Engineering Department, K.N. Toosi University of Technology, Tehran, Iran

Introduction

Fluent speech is characterized by smoothness (lack of interruptions), normal speaking rate (not too fast or too slow), prosody (emotional intonation), and minimum mental effort (effortless speech).[1] Dysfluency, on the other hand, refers to any disorder in fluency parameters. People who have more than 10% dysfluency in their speech are classified as people who stutter. Approximately 80% of preschool children recover with or without treatment, while others develop a chronic stuttering.[2] Prolongation, which is defined as involuntary lengthening of speech sounds (e.g., MMMMother or SSSSave), is one of the important types of speech dysfluencies. In addition to its importance in measurement of stuttering severity, this type of dysfluency can be utilized in prediction of stuttering development. In fact, in case early stuttering is dominated by prolongation, the chance of recovery is lower than domination by other dysfluencies.[3]

Speech-language pathologists (SLPs) listen to the client's speech samples and calculate the number of speech dysfluencies to diagnose stuttering and trace the clients' response to treatments.[4] In addition to the fact that these methods are time consuming, the evaluation results are inconsistence because of the dependency of the methods on clinicians' experiences.[5] To support objective classification of speech dysfluencies, several researches have been performed on the classification of prolongation either individually[6,7] or in combination with other dysfluencies.[8-10]

The general framework of previous studies was to segment the speech signals manually into fluent or prolongation classes. Then, different sets of features and classifiers were used to detect prolongations among normal samples. Mel frequency cepstral coefficient (MFCC) and hidden Markov model (HMM),[6] entropy of wavelet packet transforms (WPT) and support vector machine (SVM),[8] MFCC and Gaussian mixture model (GMM),[10] speech envelop

How to cite this article: Esmaili I, Dabanloo NJ, Vali M. An automatic prolongation detection approach in continuous speech with robustness against speaking rate variations. J Med Sign Sence 2017;7:1-7.

Address for correspondence:
Dr. Iman Esmaili, Biomedical Engineering Department, Science and Research Branch, Islamic Azad University, Tehran, Iran.
E-mail:
iman.esmaili@srbiau.ac.ir

Website: www.jmss.mui.ac.ir

and artificial neural network (ANN),[11] and linear predictive cepstral coefficient (LPCC) and K-nearest neighbor (K-NN)[12] are a few examples of such procedures.

In addition to the problem of extending the researches to continuous speech, none of the previous researches have considered the classification of dysfluencies during the treatment process. As a matter of fact, errors made by SLPs in detecting and counting the dysfluencies will not lead to different judgments regarding the diagnosis of stuttering.[4] Indeed, the real problem occurs during the treatment process, wherein the success of treatments is measured by decrease in dysfluency occurrence in the current treatment session compared to the previous one. Treatment of stuttering is usually performed by the so-called fluency-shaping methods in which clients learn to speak more slowly.[13] Although much effort is put into selecting classifiers and feature sets in automatic dysfluency detection,[14] less attention has been devoted thus far to the existing variations of speaker rate during the treatment process.

In this study, a fast and accurate system with robustness against speaking rate variation is introduced to support detection of prolongation dysfluencies in continuous speech. For this purpose, the speech signals were initially parameterized to perceptual linear predictive (PLP)[15] features. To detect the prolonged segments, the similarities between successive frames of speech signals were calculated based on autocorrelation similarity measures. The segments were then labeled as prolongation if the duration of highly similar successive frames exceeded a threshold specified by the speaking rate. All previous studies have employed either a fixed threshold for minimum sustaining duration in sounds[6,7] or classifiers that were trained on predefined speaking rates,[8-10,12] although it must actually be judged according to how slow or fast the individual's speech is. The performance of the proposed method was appraised using English and Persian language databases by comparing our results with those of three accurate methods[7-9] from the previous researches. This paper is organized as follows: the proposed method and databases are explained in the next section; results of experiments and its comparison with other researches are given in the following section; and finally, the last section contains the conclusion of this study.

Materials and Methods

Figure 1 shows the block diagram of the proposed prolon gation detection method. First, the speaking rates were determined for further use in prolongation detection process. Next, the speech samples were parameterized to PLP feature sets. To detect the prolongation dysfluencies, a sequence of highly similar frames was initially recognized and then compared with the speaking rate to check whether or not the segment was long enough to be considered as prolongation. The following are the details of the proposed method.

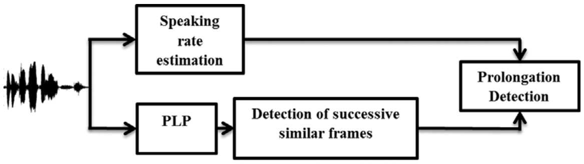


Figure 1: The block diagram of the proposed system

Databases

UCLASS

UCLASS[16] is an archive of audio recordings from stuttering people in the following three forms: monolog, conversation, and reading. In this study, 39 audio recordings of individuals (i.e., 37 males and 2 females) with age ranging from 8 years and 4 months to 20 years and 1 month were taken from UCLASS database. As suggested by other researchers,[8,10,12] the samples were taken from "reading" category that mostly contains monosyllabic words. Because most of the recordings do not have labels, two well-trained SLPs were employed to label the 39 samples of our choice into fluent or prolongation classes. Finally, a total of 105 speech samples of sound prolongations and 14,200 speech samples of fluent words were obtained.

Persian database

Persian language database was collected in a clinic from people who stutter. The speech samples were recorded in the SLPs' office to reflect the real test conditions. It contained 20 recordings (18 males and 2 females). The recordings covered a wide range of ages (ranging from 5 years 1 month to 30 years 6 months). Speech samples were collected during treatment sessions in the form of monolog when the clients were speaking about their vacations. Audio recordings were judged by SLPs and a total of 122 speech samples of sound prolongations and 5400 speech samples of fluent words were obtained.

Speaking rate estimation

People who stutter usually have a lower speaking rate.[17] Furthermore, the treatment processes encompass methods in which the clients learn to speak slowly. Therefore, it is necessary to estimate the speaking rate of each client for further employment as a threshold in the prolongation detection algorithm. As all syllables contain a vowel, the general idea in speaking rate estimation is based on counting the vowels in speech signals. Signal energy and zero-crossing rate,[18] signal energy and fundamental frequency,[19] and vowel classifier[20] are examples of speaking rate estimation techniques. In this study, we employed signal energy and zero-crossing rate as suggested in Pfau and Ruske[18] to estimate the number of syllables. There is a peak in the energy level of voiced phonemes, and similar peaks can be found in the zero-crossing rate of unvoiced phonemes. Accordingly, the number of syllables can be estimated by counting the local maxima of energy signal, wherein there is no local maximum of zero-crossing rate. For a reliable estimation of speaking rate, the local maxima with a long distance from the adjacent maxima are excluded. These maxima result from the existing prolongations that cause overestimation of the final speaking rate. Figure 2 is an example of syllable counting method in which vertical lines, solid line, and dashed line represent actual syllable boundaries, energy signal, and zero-crossing rate, respectively. The peaks with circle marks are the voiced phonemes considered in the syllable counting process. Ultimately, the speaking rate was calculated by counting the number of syllables in a second.

Feature extraction and similarity measure

Cross-correlation similarity measure and PLP feature were demonstrated in Esmaili *et al.*[21] as the best choices of classifier and feature for prolongation detection. Figure 3 shows the block diagram of the PLP feature extraction method. Initially, speech is divided into short duration frames. Hamming windows of length 30 ms, which shifted every 10 ms, were employed in this study. In the first block, Fast Fourier Transform converted the speech frames into the frequency domain. In the next three blocks, the frequency domain speech frames were subjected to three filters according to Eqs. (1)–(3),[15] where w is the angular frequency in rad/s. The mentioned filters were as follow: filter banks of Bark scale named as critical band analysis (i.e., Eq. (1)), a function that approximated the sensitivity of human hearing in different frequencies, which was named as equal-loudness pre-emphasis (i.e., Eq. (2)), and another function that simulated the nonlinear relation between intensity of sound and its perceived loudness, which was named as intensity-loudness conversion (i.e., Eq. (3)).

$$\Omega(w) = 6\ln\left(\frac{w}{1200\pi} + \left[\left(\frac{w}{1200\pi}\right)^2 + 1\right]^{0.5}\right) \qquad (1)$$

$$E(w) = \frac{\left[\left(w^2 + 56.8\times 10^6\right)w^4\right]}{\left[\left(w^2 + 6.3\times 10^6\right)^2 \times \left(w^2 + 0.38\times 10^9\right)\right]} \qquad (2)$$

$$\Xi(w) = \left[\Omega(w)E(w)\right]^{0.33} \qquad (3)$$

In the next block, inverse FFT was performed, and in the final block, the PLP features were approximated by spectrum of an all-pole model using the autocorrelation method. In this study, 12 coefficients of PLP features were employed for automatic detection of prolongation sounds.

In this study, we used cross-correlation to check the similarity of speech frames. The cross-correlation between the samples of two feature sets of $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ of length n is defined in Eq. (4).

$$r_{XY} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2}} \qquad (4)$$

where $\overline{X}$ and $\overline{Y}$ denote the mean values for X and Y, respectively.

Prolongation detection

Because prolongation occurs in a single sound, the frames of the prolonged segments have the same frequency structure. Thus, long speech segments whose frames were highly similar could be considered as the reason for the prolongation dysfluency. Two frames were regarded as highly similar if, on the basis of correlation measure, their similarity was higher than a specified threshold (i.e., T1). Another threshold (i.e., T2) was employed to check whether the segment was long enough to be regarded as prolongation.

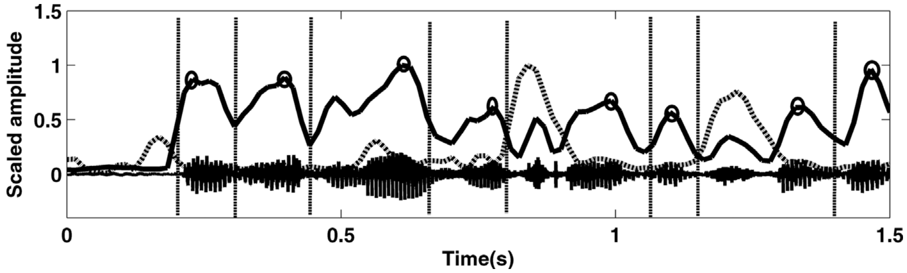


Figure 2: An example of syllable counting method. Vertical lines, solid line, and dashed line are actual syllable boundaries, energy signal, and zero-crossing rate, respectively. Circle marks are considered in the syllable counting process. [The amplitude of signals (i.e., *Y* axis) were normalized to fit all signals in proper view]

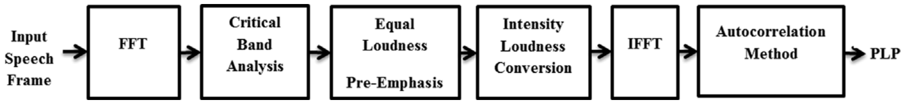


Figure 3: The block diagram of the PLP feature extraction method

Given $S = \{s_1, s_2, \ldots, s_n\}$ as a sequence of extracted features of speech signal S, where S_n is the feature vector of the nth frame (PLP features of length 12), the prolongation algorithm can be explained through the following steps:

The first step is to find highly similar successive frames by calculating the similarities between s_i and the subsequent frames based on correlation similarity measures. If the similarity between s_i and s_{i+1} be more than a threshold (i.e., T1), the location of s_{i+1} is recorded and the algorithm proceeds by calculating the similarity between s_i and s_{i+2}. This step ends when the similarity between s_i and $s_{i+k}(i \leq k \leq n)$ drops under T1.

The second step involves counting the number of highly similar successive frames and labeling them as prolongation if their duration is more than T2.

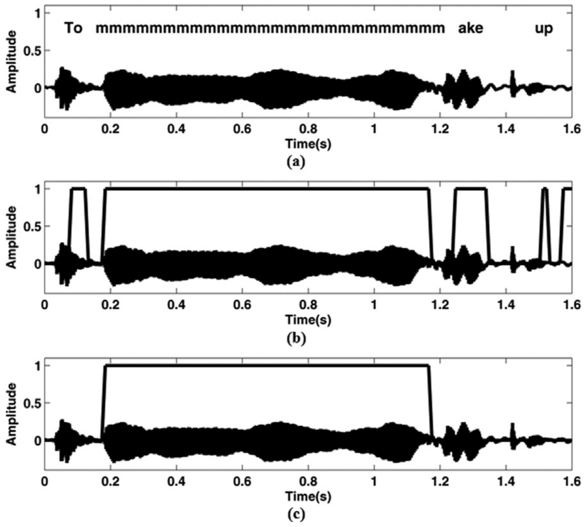
The third step is a return to the first step by starting from s_{i+k} until $i+k \leq n$.

The similarity values between pairs of speech frames map to [0:1] where 1 and 0 values correspond to maximum and minimum similarity, respectively. Two frames were considered as highly similar frames provided that their similarity exceeded the value of T1. Our experiments revealed that the best choice of T1 for signal-to-noise ratios (SNRs) higher than 25 dB was 0.9. The other threshold (i.e., T2) was used to determine prolongation segments. In other words, if the duration of highly similar frames exceeded T2, the sequence of frames would be considered as prolongation segment. T2 was automatically calculated based on speaker rate estimation. In fact, we initially found the average syllable length by inversing the speaking rate. Subsequently, a prolongation label was further assigned to the segments, which were two times longer than the average syllable length. The threshold values were achieved by detailed examination of UCLASS and Persian databases. Even if one did not accept the proposed threshold values, it was required to adjust these values just once for the operating environment, and this did not seriously affect the automatic framework of the proposed method.

Figure 4 indicates the prolongation detection algorithm in a speech signal. Figure 4a depicts the UCLASS speech sample ("to make up") with prolongation in the initial phoneme of the word "make." Figure 4b illustrates the highly similar successive frames detected through the recursive execution of the algorithm. In this example, the average syllable length was 200 ms. Therefore, the segment whose length was more than 400 ms (i.e., twice the average syllable length) was considered as prolongation and is presented in Figure 4c.

Evaluation methods

The performance of the proposed method was assessed by the sensitivity, specificity, and accuracy parameters. These



(a)

(b)

(c)

Figure 4: Prolongation detection. (a) Speech sample "to make up" with prolongation in word "make," (b) highly similar segments, and (c) detected prolonged segment which is longer than threshold (here 400 ms)

parameters were defined based on true positive (TP), false positive (FP), true negative (TN), and false negative (FN) terms as follows:[22]

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{TP} + \text{FN}}$$

Herein, prolongation and fluent words were considered as positive and negative classes, respectively. Therefore, sensitivity and specificity referred to probability of correctly detected prolongation and fluent words, respectively, while accuracy showed the probability of a correct classification (i.e., both prolongation and fluent).

Selected studies

The results of our proposed method were compared with those of three high-performance researches[7-9] in automatic dysfluency classification. Hariharan *et al.*[8] achieved the best results by employing the entropy of WPT as the feature set and the SVM as the classifier. As recommended by these researchers, three levels of decomposition and "coiflets" wavelet family were employed in the detection process. Świetlicka *et al.*[9] utilized amplitudes of weight filters on Fourier transform of speech frames (i.e., FFT measures) as the feature set, as well as an order of Kohonen and ANN for feature selection and classification, respectively. In these studies, the prolongations were segmented manually.

Suszyński *et al.*[7] applied the same features, but the prolongations were detected by means of fuzzy sets. Two fuzzy membership functions were suggested by these researchers for fricative and nasal sounds. Another membership function was also offered for the duration of prolongation. In fact, they considered a ramp between 200 ms and 400 ms for the prolongation duration. Indeed, all sounds above 400 ms were also considered as prolongation. In our research, unlike the first two studies, the features were directly extracted from the continuous speech. As mentioned before, the first two studies used manual segmentation of prolongation parts, not being practical for detection of dysfluencies in continuous speech. Then, we first employed a sliding window on speech signals, and each window was subjected to feature extraction and classification by following the original methods of the selected studies. As described in the section "Prolongation detection," our proposed method does not need segmentation and training stage. The procedure of the proposed method consisted of the following two main steps: detection of highly similar frames and comparison with automatically calculated threshold.

Results and Discussion

The evaluation was performed by comparing the results of our proposed method (i.e., similarity-based method) with those of three high-quality researches in this field. In Hariharan's and Świetlicka's procedures, 75% of the data were utilized at the training stage and the remaining 25% for the evaluation stage. Suszyński's procedures and the proposed methods do not need the training stage; however, the same amounts of data (i.e., 25%) were employed for the evaluation.

Table 1 shows the accuracy, sensitivity, and specificity of similarity-based method along with selected studies on UCLASS and Persian databases. The best prolongation recognition rates were achieved by the similarity-based method with 99.0, 99.3, and 99.4% for accuracy, sensitivity, and specificity, respectively. The accuracy of similarity-based method is 2.2, 3.2, and 3.8% higher than the methods based on SVM, ANN, and fuzzy sets, respectively. The same experiments were conducted on the Persian database to assess the reliability of the results. As demonstrated in Table 1, an accuracy of 97.1% was achieved by employing similarity-based method that is 3.6, 5, and 7.2% higher than that of the methods based on SVM, ANN, and fuzzy sets, respectively.

There are logical explanations on how a simple correlation measure can reach better prolongation recognition rates compared to the methods based on classifiers. This is mostly as a result of false alarms brought by the sliding window to the systems. In fact, whole prolongation frames must be exposed to the classifiers to have a correct decision. In case of employing the sliding window, it is quite possible that the window covers just a part of the prolongation segment and not all of it. The problem is not to be solved even if we increase the overlap time between the sliding windows. Indeed, in this case, a single prolongation may be recognized twice. Then, employing a long duration window with a small overlap time seems to be a suitable solution. In prolongation detection, the classifiers show high performance when the input just contains the prolongation parts or the fluent parts, not their combination. On the other hand, increasing the sliding window duration increases the probability of the occurrence of both classes in a single window. That is why authors of selected studies in their original methods preferred to use the manual segmentation instead of the sliding window.

As mentioned previously, the speaking rates in normal speaking differ from people who stutter especially during treatment periods. To evaluate the methods in this situation, the speaking rates of both databases artificially altered from 60 to 140% of its original speaking rate, and all of the experiments were performed again. Table 2 presents the accuracy of similarity-based method along with that of three other studies on UCLASS and Persian databases. Obviously, the accuracy of similarity-based method has not changed by variation of the speaking rates from 70 to 130%, while the accuracy of the methods based on SVM, ANN, and fuzzy sets has decreased by 19.7, 22.7, and 30.5%, respectively. The same trends can be seen in the Persian database. There is no change in the accuracy of similarity-based method, while the accuracy of the methods based on

Table 1: The recognition rates for prolongation detection on UCLASS and Persian database

Method	Database	Feature set	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)
Similarity-based method	UCLASS	PLP	Correlation	99.0	93.3	99.4
Hariharan *et al.*[8]	UCLASS	WPT entropy	SVM	96.8	78.1	98.1
Świetlicka *et al.*[9]	UCLASS	FFT measure	ANN	95.8	68.6	97.7
Suszyński *et al.*[7]	UCLASS	FFT measure	Fuzzy sets	95.2	64.7	97.4
Similarity-based method	Persian	PLP	Correlation	97.1	90.1	98.1
Hariharan *et al.*[8]	Persian	WPT entropy	SVM	93.5	71.3	96.9
Świetlicka *et al.*[9]	Persian	FFT measure	ANN	92.1	65.6	96.1
Suszyński *et al.*[7]	Persian	FFT measure	Fuzzy sets	89.9	52.3	95.6

Table 2: The accuracy of prolongation detection methods for artificial variations of speaking rate from 60 to 140% of normal speaking

Method	Database	Speaking rate								
		60%	70%	80%	90%	100%	110%	120%	130%	140%
Similarity-based method	UCLASS	91.3	99.0	99.0	99.0	99.0	99.0	99.0	99.0	89.1
Hariharan *et al.*[8]	UCLASS	72.9	77.1	87.4	94.6	96.8	96.6	96.4	96.0	95.7
Świetlicka *et al.*[9]	UCLASS	70.5	73.1	79.0	92.8	95.8	95.6	95.1	94.7	93.9
Suszyński *et al.*[7]	UCLASS	58.8	64.7	76.1	91.3	95.2	95.0	94.3	93.8	93.2
Similarity-based method	Persian	87.3	97.1	97.1	97.1	97.1	97.1	97.1	97.1	82.2
Hariharan *et al.*[8]	Persian	69.4	76.6	81.8	91.1	93.5	93.2	92.7	92.4	91.9
Świetlicka *et al.*[9]	Persian	66.7	70.2	76.1	88.3	92.1	91.9	91.5	91.1	90.4
Suszyński *et al.*[7]	Persian	64.4	61.2	72.8	84.4	89.9	89.6	88.7	88.1	87.5

SVM, ANN, and fuzzy sets has decreased by 14.5, 18.1, and 20%, respectively. The change in the accuracy of the proposed method starts from 140% variation in speaking rate; however, these amounts of changes in speech variation rarely occur in human speech.

In Hariharan's and Świetlicka's methods, classifiers were trained on specific speaking rates. Thus, encountering different speaking rates decreases the performance of their methods. It persists for the method based on the fuzzy logic, because it uses a specific membership function for predefined speaking rates. However, similarity-based method consists of an automatic speaking rate estimator with which a suitable threshold for prolongation detection can be achieved. It must be noted that methods based on the classifier or the fuzzy set cannot be easily equipped with the automatic speaking rate estimator. In fact, several trained classifiers or fuzzy membership functions corresponding to different speaking rates are required in this case.

Conclusion

In this study, a fast and accurate method with robustness against speaking rate variation based on PLP features and cross-correlation was introduced for automatic detection of prolongation in continuous speech. By employing a speaking rate estimator, the proposed method can cope well with a wide range of speaking rates (i.e., 70–130% of the original speaking rate). Moreover, the method is quite fast, as in each frame, we must just perform the similarity calculation with the adjacent frame and the comparison with the threshold. The proposed method indicated promising results compared to the three high-quality researches conducted in the same field. The reliability of our findings was also evaluated by a self-recorded Persian database. This method can be applied by a SLP for the diagnosis of stuttering or during the treatment sessions.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Starkweather C. Fluency and Stuttering. Englewood Cliffs, NJ: Prentice-Hall; 1987.
2. Adams MR. A clinical strategy for differentiating the normally nonfluent child and the incipient stutterer. J Fluen Disord 1977;2:141-8.
3. Conture EG. Stuttering. 2nd ed. Englewood Cliffs, NJ: Prentice Hall; 1990.
4. Yaruss JS. Clinical measurement of stuttering behaviors. Contemp Issues Commun Sci Disord 1997;24:33-44.
5. Curlee RF. Observer agreement on disfluency and stuttering. J Speech Lang Hear Res 1981;24:595-600.
6. Wiś niewski M, Kuniszyk-Jó kowiak W, Smołka E, Suszyński W. Automatic detection of prolonged fricative phonemes with the hidden Markov models approach. J Med Inform Technol 2007;11:293-8.
7. Suszyński W, Kuniszyk-Jó kowiak W, Smołka E, Dzieńkowski M. Prolongation detection with application of fuzzy logic. Ann UMCS Inform 2003;1:133-40.
8. Hariharan M, Fook CY, Sindhu R, Adoma AH, Yaacob S. Objective evaluation of speech dysfluencies using wavelet packet transform with sample entropy. Digit Signal Process 2013;23:952-9.
9. Świetlicka I, Kuniszyk-Jó kowiak W, Smołka E. Hierarchical ANN system for stuttering identification. Comput Speech Lang 2013;27:228-42.
10. Mahesha P, Vinod DS. Gaussian mixture model based classification of stuttering dysfluencies. J Intell Syst 2015;25: 387-99.
11. Howell P, Sackin S. Automatic recognition of repetitions and prolongations in stuttered speech. Proceedings of the First World Congress on Fluency Disorders, 1995. p. 372-4.
12. Ai OC, Hariharan M, Yaacob SB, Chee LS. Classification of speech dysfluencies with MFCC and LPCC features. Expert Syst Appl 2012;39:2157-65.
13. Roth FP, Worthington CK. Treatment Resource Manual for Speech-Language Pathology. 4th ed. Clifton Park, NY: Delmar; 2011.
14. Fook CY, Hariharan M, Chee LS, Yaacob SB, Adom AH. Comparison of speech parameterization techniques for classification of speech dysfluencies. Turk J Electr Eng Comput Sci 2013;21:1983-94.
15. Hermansky H. Perceptual linear predictive (PLP) analysis of speech. J Acoust Soc Am 1990;87:1738-52.

16. Howell P, Davis S, Bartrip J. The UCLASS archive of stuttered speech. J Speech Lang Hear Res 2009;52:556-69.
17. de Andrade CR, Cervone LM, Sassi FC. Relationship between the stuttering severity index and speech rate. Sao Paulo Med J 2003;121:81-4.
18. Pfau T, Ruske G. Estimating the speaking rate by vowel detection. Acoust Speech Signal Process 1998;2:945-8.
19. de Jong NH, Wempe T. Automatic measurement of speech rate in spoken Dutch. ACLC Work Pap 2007;2:49-58.
20. Yuan J, Liberman M. Robust speaking rate estimation using broad phonetic class recognition. 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010. p. 4222-5.
21. Esmaili I, Jafarnia Dabanloo N, Vali M. Automatic classification of speech dysfluencies in continuous speech based on similarity measures and morphological image processing tools. Biomed Signal Process 2016;23:104-14.
22. Duda RO, Hart PE, Stork DG. Pattern Classification. 2nd ed. New York, NY: John Wiley and Sons; 2001.