Biomedical Signal Processing and Control 23 (2016) 104–114

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Automatic classification of speech dysfluencies in continuous speech based on similarity measures and morphological image processing tools

Iman Esmaili [a,*], Nader Jafarnia Dabanloo [a], Mansour Vali [b]

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

ARTICLE INFO

Article history:
Received 8 May 2015
Received in revised form 26 August 2015
Accepted 27 August 2015
Available online 11 September 2015

Keywords:
Automatic dysfluency classification
Similarity measures
Morphological image processing
Stuttering severity measurement

ABSTRACT

Speech-language pathologists, traditionally, count the number of speech dysfluencies to measure the rate of stuttering severity. Subjective stuttering assessment is time consuming and highly dependent on clinician's experiences. The present study proposes an objective evaluation of speech dysfluencies (sounds prolongation, syllables\words\phrases repetition) in continuous speech signals. The proposed method is based on finding similarity in successive frames of speech features for sounds prolongation detection and in close segments of speech for repetition detection. Speech signals are initially parameterized to MFCC, PLP or filter bank energy feature sets. Then, similarity matrix is calculated based on similarities of all pairs of frames using cross-correlation or Euclidean criterion. Similarity matrix is considered as an image and highly similar components are extracted using proper threshold. By employing morphological image processing tools, irrelevant parts of similarity matrix are removed and dysfluent parts are detected. The effects of different feature sets and similarity measures on classification results were examined. The promising classification accuracy of 99.84%, 98.07% and 99.87% were achieved for detection of prolongation, syllable/word repetition and phrase repetition, respectively.

1. Introduction

Speech is the most important form of human communication. The efficiency of speech in transferring information depends on fluency. Fluency is characterized by simplicity with which sounds, syllables, word and phrases are connected to each other to form a message. Any disruption in fluency or forward flow of speech is called dysfluency. Table 1 provides the most commonly used classification of dysfluencies. This classification was first proposed by Johnson [1] in late 1950s and has been used by clinicians and researchers ever since. All people are dysfluent to some degree in speaking, but about 5% of children and 1% of adults have more dysfluencies in their speech, which is perceived as stuttering to listeners [2].

Diagnosis of individuals who stutter is carried out by Speech-Language Pathologist (SLP). To measure the severity of stuttering, SLPs traditionally counts the number of dysfluencies and divide it by total number of spoken words (or syllables) [3]. Clinicians need to measure fluency to decide if client is stutterer and to evaluate the response of stutterer during the treatment process. However, due to slightly different definitions of stuttering proposed by clinicians and their mistake in counting the dysfluencies, there is low agreement on the true number of dysfluencies among different SLPs (inter-rater agreement) and even when the process is repeated by the same SLP (intra-rater agreement) [4,5]. Such low agreement could be alleviated by proposing an accurate system of detecting and counting the number of dysfluencies automatically.

Due to existing complications in types and patterns of dysfluencies, automatic classification of dysfluencies is not a straightforward task. In general, researchers have used some simplifications to overcome the problem. The first simplification is to reduce the number of dysfluency classes that should be considered in automatic dysfluency classification. Some researchers believe that although there are many types of dysfluencies, people who stutter more likely produce some types of dysfluencies (e.g., repetition of sounds, repetition of syllables, repetition of monosyllabic words and sound prolongation) [3]. The other types of dysfluencies are more or less the same between people who stutter and those who do not. Almost all previous studies in automatic classification of dysfluencies have considered part-word repetition and sound prolongation classes among dysfluency classes of Table 1.

* Corresponding author. Tel.: +98 9364179987.
E-mail address: iman.esmaili@srbiau.ac.ir (I. Esmaili).

http://dx.doi.org/10.1016/j.bspc.2015.08.006
1746-8094/

Table 1
Description of different types of dysfluencies.

Type of disfluency	Description	Example
Part-word repetition	To repeat a sound or syllable	W-W-where is she going?
Word repetition	To repeat a whole word	WHERE-WHERE-where is she going?
Phrase repetition	To repeat a phrase	WHERE IS where is she going?
Prolongation	Sustain a sound for a long duration	Where is SHSHSHshe going?
Revision	To change the word	Where is HE she going?
Interjection	To add meaningless words irrelevant to the message	Where is UMMM she going?
Broken word	Pause within word	Where is she GO-(pause)-ING?
Incomplete phrase	To change the words	SHE MUST. . .where is she going?

Unlike most of speech pathology classification tasks [6], it is not possible to find the correct target class with a few number of speech frames in the dysfluent speech signals. In fact, we need whole repeated part or prolonged part to decide about the target class. Thus, the second simplification is applied on the task itself. Some researchers have manually segmented the speech in such a way that each segment contains whole dysfluent or fluent part. The general framework of this category is to train a set of classifiers, such as Artificial Neural Network (ANN) [7,8], Support Vector Machine (SVM) [9,10] and K-Nearest Neighbor (K-NN) [11,12], with speech features, such as Mel Frequency Cepstral Coefficient (MFCC) [9,12,13], Linear Predictive Cepstral Coefficient (LPCC) [11] and Perceptual linear predictive (PLP) analysis [14]. After training, input segments are classified into fluent and dysfluent classes.

Instead of considering manually segmented speech, some other studies have employed more suitable solution that segments the speech into windows of fixed length and tries to label each window as fluent or dysfluency classes. In [15], a feature set of 1st to 3rd formant frequencies, their amplitudes and two classifiers (ANN and rough set) was used to label input windows as repetition, prolongation or normal classes. In [16], a hierarchy of Kohonen network for feature reduction and three Multilayer neural networks were used to classify the input feature vectors (amplitudes of 1/3 octave weight filters on Fourier transform of 23 ms speech frames) as block, repetition and prolongation classes. They suggested a solution to extend the work to continuous speech. However, no tests and results were reported.

Although fixed window procedure does not need an SLP to segment the speech, the length of window is an important issue. If we choose a long duration window, it may contain more than one dysfluency while most of the fixed window approaches can only distinguish one dysfluency per window. On other hand, if we choose a short duration window, it probably cannot cover the whole dysfluency duration and the classifier will fail to find the dysfluency class. Additionally, these approaches do not provide any information regarding the exact location and duration of dysfluent parts.

Among numerous studies on automatic classification of dysfluencies in speech, only a few studies have considered the continuous speech. The problem of extending the work to continuous speech is mainly due to unknown location of dysfluent parts. Therefore, speech signal initially must be divided to segments which probably contain dysfluent segments then a classification task is performed to detect the dysfluency classes. In [17], sequence of phonemes first recognized employing an automatic speech recognition system. Then, dysfluent repetitions were extracted based on output phonemes. This method is based on the assumption of existence a speech recognition system which is robust to variety of speech and speaker variations. In spite of great improvements in recent automatic speech recognition systems, these systems are still far from recognizing the conversational speech. In [18], a syllable segmentation procedure is employed by considering neighbor local maxima and minima of FFT measures. Then, dysfluent syllables are detected by correlation analysis of adjacent syllables. Although the method of segmentation is appropriate for monosyllabic words, in multisyllabic words it will be difficult to find syllables boundaries. Employing symbolic representation of amplitude [19] and energy [20] of speech signal, the authors attempted to remove redundant parts of speech and to find disfluent segments. Although their idea is interesting, we believe that time domain features are not suitable criterion to search the similarities of speech patterns. There are also other studies that have investigated the relation between specific parameters, such as silence [21] and envelope [22] of speech signals with stuttering severity.

A practical automatic dysfluency classification system needs to deal with continuous speech without manual segmentation of dysfluent moments. Moreover, SLPs believe that considering more dysfluency classes provides more information about the diagnosis of stuttering [3]. Therefore, in this study, we focused on recognizing the dysfluent parts of continuous speech without segmenting the signal to limited windows. We also considered a broader range of dysfluency classes (i.e., syllable repetition, word repetition, phrase repetition, prolongation). To this end, we used the strength of speech recognition feature space. However, instead of using ordinary classification techniques, feature space was treated as time series in which repeated patterns were considered as dysfluencies of the original speech signal.

In this paper, short speech frames were initially represented as MFCCs, PLP or filter bank energies (FBE). Then, the silence parts of speech were removed using a voiced/unvoiced detection approach. Next, a similarity matrix was developed based on the similarities of each frame with the rest of the frames. We found that distinguishing the dysfluencies is easier in visual manner. Therefore, we treated the similarity matrix as a picture and tried to find target patterns using morphological image processing tools. To our knowledge, it is the first study in which morphological image processing tools have employed for speech pattern extraction. This paper is organized as follows: Section 2 describes methodology of the system and database used in the experiment. Experimental results and discussions about dysfluency classification for MFCC, PLP and FBE features using different distance measures (i.e., Euclidean and cross-correlation) are presented in Section 3. Finally, conclusions are given in Section 4.

2. Material and methods

Fig. 1 presents a schematic representation of the proposed dysfluency classification process. A voiced/unvoiced detection procedure is performed in the first block. In the second block, speech samples are parameterized to feature sets. Visualization of speech features is performed through developing a similarity matrix based on the distance of all pairs of frames in the third block. After elimination of silences, detection of sound prolongation, syllables/word repetition and phrase repetition is carried out in the next blocks. Finally, rate of dysfluency severity is calculated according to the detected dysfluencies. Details of the proposed method are presented in the following sections.