Expert Systems with Applications 39 (2012) 2157–2165

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Classification of speech dysfluencies with MFCC and LPCC features

Ooi Chia Ai, M. Hariharan *, Sazali Yaacob, Lim Sin Chee

Biomedical Electronic Engineering Programme, School of Mechatronic Engineering, Universiti Malaysia Perlis(UniMAP), 02600 Arau, Ulu Pauh, Perlis, Malaysia

ARTICLE INFO

Keywords:
Stuttering
MFCC
LPCC
*k*NN
LDA

ABSTRACT

The goal of this paper is to discuss comparison of speech parameterization methods: Mel-Frequency Cepstrum Coefficients (MFCC) and Linear Prediction Cepstrum Coefficients (LPCC) for recognizing the stuttered events. Speech samples from UCLASS are used for our analysis. The stuttered events are identified through manual segmentation and used for feature extraction. Two simple classifiers are used for testing the proposed features. Conventional validation method is used for testing the reliability of the classifier. The experimental investigation elucidates MFCC and LPCC features which can be used for identifying the stuttered events and LPCC features were slightly outperformed than MFCC features.

1. Introduction

Speech is a verbal means used by humans to express their feelings, ideas and thoughts in communication. Speech consists of articulation, voice and fluency. However, 1% of the population has noticeable speech stuttering problem and it has found to affect female to male with ratio 1:3 or 4 times (Awad, 1997; Chia Ai & Yunus, 2006; Van Borsel, Achten, Santens, Lahorte, & Voet, 2003). Stuttering is defined as a normal flow of speech disrupted by unintentionally of dysfluencies such as repetition, prolongation, interjection of syllables, sounds, words or phrases and involuntary silent pauses or blocks in communication (Awad, 1997; Chia Ai & Yunus, 2006; Sin Chee, Chia Ai, Hariharan, & Yaacob, 2009; Tian-Swee, Helbin, Ariff, Chee-Ming, & Salleh, 2007).

Stuttering cannot be completely cured; it may go into remission for a time (Awad, 1997). Stutterers can learn to shape their speech into fluent speech with appropriate speech pathology treatments. Therefore a stuttering assessment is needed to evaluate performance of stutterers before and after therapy.

Traditionally, speech language pathologist (SLP) count and classify occurrence of dysfluencies such as repetition and prolongation in stuttered speech manually. However, these types of stuttering assessments are subjective, inconsistent, time consuming and prone to error (Awad, 1997; Howell, Sackin, & Glenn, 1997a, 1997b; Nöth et al., 2000; Ravikumar, Rajagopal, & Nagaraj, 2009; Ravikumar, Reddy, Rajagopal, & Nagaraj, 2008). Therefore, it might be good if stuttering assessment can be done automatically and thus having more time for the treatment session between stutterer and SLP.

1.1. Related work

Researchers have focused on developing objective methods to facilitate the SLP during stuttering assessment. Table 1 depicts some of the significant research works that have been conducted in the last two decades chronologically.

In Howell and Sackin (1995) located stuttered speech events namely repetition and prolongation. They extracted totally 39 acoustic parameters, 20 vector based on autocorrelation function plus spectral coefficient based on a 19 channel vocoder. Furthermore, envelope of speech waveform was obtained by filtering the signal using 10 Hz low-pass filter. Artificial neural network was employed to discriminate between the stuttered events. The best hit/miss rate was 0.82 and 0.77 for prolongations and repetitions respectively using autocorrelation function plus spectral coefficient.

In 1997, an automatic dysfluency count was presented by Howell, Sackin and Glenn (1997a, 1997b). They employed 12 children who speak stuttered English. The speech samples can be obtained from UCLASS database. They approached the recognition task based on nine parameters, for instances, whole word and part word duration; whole word, first part and second part fragmentation; whole word, first part and second part spectral measure; and part word energy. Before extracting the parameters, speech signals were segmented manually into word unit. At the same time, supra-lexical dysfluency like interjection, revisions, incomplete phrase and phrase repetitions were eliminated. Artificial neural networks (ANN) were employed to classify a word either as repetition, prolongation or fluent. The nine parameters were input to the networks. The system yielded 95% accuracy for fluent words, 78% of the overall dysfluent (combining prolongations and repetitions) accuracy with only 58% and 43% of prolongations and repetitions respectively.

* Corresponding author. Tel.: +0060 4 9885237; fax: +0060 4 9885167.
E-mail addresses: hari@unimap.edu.my (M. Hariharan), sclim3@gmail.com (L. Sin Chee).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.07.065

Table 1
Summary of several research works on stuttering recognition, detailing the number of subjects, the features used and the classifiers employed.

First author	Database	Features	Classifiers	Best results (%)
Howell (Howell & Sackin, 1995)	–	Autocorrelation function and envelope parameters	ANNs	≈80%
Howell (Howell et al., 1997a, 1997b)	12 Speakers (UCLASS)	Duration, energy peaks, spectral of word based and part word based	ANNs	78.01%
Geetha (Geetha et al., 2000)	51 Speakers	Age, sex, type of dysfluency, frequency of dysfluency, duration, physical concomitant, Rate of speech, historical, attitudinal, and behavioral scores, family history.	ANNs	92%
Nöth (Nöth et al., 2000)	37 speakers	Duration and frequency of dysfluent portions, speaking rate	HMMs	–
Czyzewski (Czyzewski et al., 2003)	6 normal speech samples + 6 stop-gaps speech samples	Frequency, 1st to 3rd formant's frequencies and its amplitude	ANNs & rough set	73.25% & ⩾90.0%
Prakash (Prakash, January 9-11, 2003)	10 normal + 10 stuttering children	Formant patterns, speed of transitions, F2 transition duration and F2 transition range	–	–
Szczurowska (Szczurowska et al., 2006)	8 speakers	Spectral measure (FFT 512)	Multilayer Perceptron (MLP), Kohonen	76.67%
Wiśniewski (Wiśniewski etal., October 18, 2007)	38 samples for prolongation of fricatives + 30 samples for stop blockade + 30 free of silence samples	Mel Frequency Cepstral Coefficients (MFCC)	HMMs	70%
Wiśniewski (Wiśniewski et al., 2007)	–	Mel Frequency Cepstral Coefficients (MFCC)	HMMs	Approximately 80%
Tian-Swee (Tian-Swee et al., 2007)	15 normal speakers + 10 artificial stuttered speech	Mel Frequency Cepstral Coefficients (MFCC)	HMMs	96%
Ravikumar (Ravikumar et al., 2008)	10 speakers	Mel Frequency Cepstral Coefficients (MFCC)	Perceptron	83%
Ravikumar (Ravikumar et al., 2009)	15 speakers	Mel Frequency Cepstral Coefficients (MFCC)	SVM	94.35%
Świetlicka (Świetlicka et al., May 12, 2009)	8 stuttering speakers + 4 normal speakers (yields 59 fluent speech samples + 59 non-fluent speech samples)	Spectral measure (FFT 512)	Kohonen , Multilayer Perceptron(MLP), Radial Basis Function(RBF)	88.1%–94.9%

Geetha, Pratibha, Ashok, and Ravindra (2000) presented a research on classification of childhood dysfluencies using ANNs in year 2000. Fifty one children were employed, 25 were used to train ANN and 26 children were used to test the ANN. Ten variables viz. age, sex, type of dysfluency, frequency of dysfluency, duration, physical concomitant, rate of speech, historical, attitudinal and behavioral scores, family history were used as input to the network. They achieved 92% accuracy of predicting normal non-fluency and stuttering.

Nöth et al. (2000) implemented a system which combined the work of SLP and speech recognition system to evaluate the degree of stuttering during therapy session. Thirty seven patients with stuttering symptoms were employed to read an English passage. They used frequency of dysfluent portions in the speech, duration of dysfluency and speaking rate to classify the degree of stuttering. They employed HMM as the classifier and the system achieved high correlation coefficient of 0.99 from the average actual dysfluencies per word to the average detected dysfluencies.

In year 2003, Czyzewski, Kaczmarek, and Kostek (2003) approached the recognition task based on detection of stop-gaps, discerning vowel prolongation, detection of syllable repetition. Six fluent speech samples and six stop-gaps speech samples in Polish were used in the experiment. Two classifiers, namely ANNs and rough set were used to detect stuttering events. Results were favourable to rough set-based system yielding best results more than 90% of classification accuracy compared to ANNs with accuracy equal to 73.25%.

Prakash (2003) presented a study to evaluate speech of 10 normal and 10 stuttering children speaking Kannada (a south Indian language). They proposed acoustic parameters such as formant patterns, speed of F2 transition, F2 transition duration, F2 transition range. Some statistical analysis such as mean and standard deviation of the acoustic feature were computed. Walsh Test was applied to find out the significant differences between the two groups and results prove that acoustic parameters are useful for differential diagnosis of children with stuttering and normal non fluency.

In 2006, Szczurowska, Kuniszyk-Jozkowiak, and Smolka (2006) described the neural networks tests on ability of recognition and categorizing the non-fluent and fluent speech samples. Recordings that taken from eight stuttering Polish speakers were used as research material. The recordings were analyzed based on FFT512 with the use of 21 digital 1/3-octave filters of centre frequencies between 100 Hz and 10 kHz. Kohonen and Multilayer Perceptron Networks were applied to recognize and classify between fluent and dysfluent. Best result of 76.67% achieved neural network architectures with 171 input neurons, 53 neurons in hidden layer and 1 output neuron.

In year 2007, three papers related to automatic stuttering detection system were presented. Two of the papers were published by (Wiśniewski, Kuniszyk-Jóźkowiak, Smołka, & Suszyński, October 18, 2007b). In (Wiśniewski et al., October 18, 2007b), 38 samples were employed for prolongation of fricatives recognition model, 30 samples for stops blockade of recognition model and 30 samples