World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:2, No:10, 2008

Automatic Detection of Syllable Repetition in Read Speech for Objective Assessment of Stuttered Disfluencies

K. M. Ravikumar, Balakrishna Reddy, R. Rajagopal, and H. C. Nagaraj

***Abstract*—**Automatic detection of syllable repetition is one of the important parameter in assessing the stuttered speech objectively. The existing method which uses artificial neural network (ANN) requires high levels of agreement as prerequisite before attempting to train and test ANNs to separate fluent and nonfluent. We propose automatic detection method for syllable repetition in read speech for objective assessment of stuttered disfluencies which uses a novel approach and has four stages comprising of segmentation, feature extraction, score matching and decision logic. Feature extraction is implemented using well know Mel frequency Cepstra coefficient (MFCC). Score matching is done using Dynamic Time Warping (DTW) between the syllables. The Decision logic is implemented by Perceptron based on the score given by score matching. Although many methods are available for segmentation, in this paper it is done manually. Here the assessment by human judges on the read speech of 10 adults who stutter are described using corresponding method and the result was 83%.

***Keywords*—**Assessment, DTW, MFCC, Objective, Perceptron, Stuttering.

I. Introduction

STUTTERING, also known as stammering in the United Kingdom is a speech disorder. The type of disfluencies that employed are: 1.Interjections (extraneous sounds and words such as “uh” and “well”); 2.Revisions (the change in content or grammatical structure of a phrase or pronunciation of a word as in “there was a young dog, no, a young rat named Arthur”); 3. Incomplete Phrases (the content not completed); 4. Phrase–repetitions; 5.Word-repetitions; 6.Part-word-repetitions; 7.Prolonged sounds (sounds judged to be unduly prolonged); 8.Broken words (words not completely pronounced) [11].

Stuttering is often associated with “Repetitions.” As described above, part-word or syllabic repetitions are one of the defining elements of stuttering. The dominant features of normal nonfluent (NNF) speech reported are: 1.Word Repetitions, but not part-word Repetition is a prevalent feature of early stuttering [5]. 2. In early stuttering, there is a high proportion of Repetition in general, as opposed to other types of disfluency like prolongation [3].

Conventional way of making stuttering assessment are to count the occurrence of these types of disfluencies and express them either as the number of disfluent words as a proportion of all words in a passage or measure the time the disfluencies take compared with the duration of the entire passage. The main difficulties in making such counts are:

1. They are time consuming to make and
2. There are poor agreements when different judges make counts on the same material [1].

The following area of research concerns finding the repetitions automatically, which is one of the key parameters in assessing the stuttering events objectively, thus making the work of the speech-language pathologist easier and also improve interjudge agreements about stuttered events.

A Standard English passage of 150 words was selected for preparing the database. All the ten clients with mean age group of 25 were made to read the passage and these speeches were recorded using cool edit version 2 at sampling rate of 16000 samples per second with number of bits to represent as 16-bits.

II. Automatic Detection Method

The detection scheme that is used for assessment is divided into four steps:

A. Segmentation

Phonetics gives no exact specification of syllables. The characteristic feature of the syllable is the dynamical transient part consonant-vowel or consonant –vowel –consonant. The feeling of syllable boundaries, although usually very strong, is subjective and often not unique. For Automatic segmentations of syllable many methods are available, which uses signal extremes, first Autoregressive (AR) coefficient, etc [12]. The speech samples collected in the databases are segmented manually to obtain the syllable. The segmented speech syllables are subjected to Feature Extraction.

K. M. Ravikumar is with Ghousia college of Engg., Ramanagara, Dept. of ECE (phone:9180-7271353, e-mail: kmravikumar@rediffmail.com).

Balakrishna Reddy is with centiti, Bangalore as Tech. Director (e-mail: balu@centiti.in).

R. Rajagopal was with Central Research Laborotary, Bangalore. Now he is with L&T as CTO (e-mail: rrcrl@yahoo.co.in).

H. C. Nagaraj is with NMIT, Bangalore as Principal (e-mail: principal@nmit.co.in).

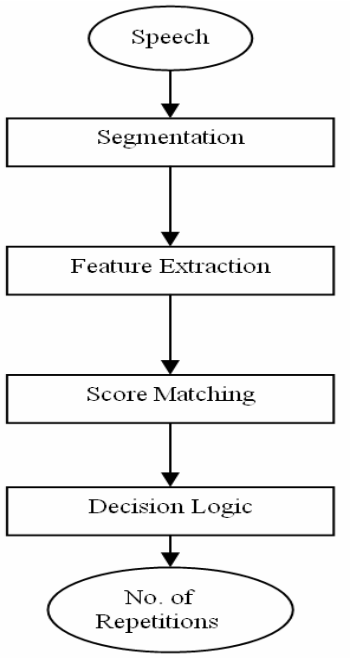

Fig. 1 Block diagram of Automatic detection method

B. Feature Extraction

A common first step in feature extraction is frequency or spectral analysis. The signal processing techniques aim to extract features that are related to identify the characteristics. The speech signal is analyzed in successive narrow time windows of 10msec width, for its frequency content with 2msec offset. For each and every window we obtain the intensity of several bands on the frequency scale using feature extraction algorithm.

Several different feature extraction algorithms exist:
1. Linear Prediction Coefficient (LPC) Cepstra
2. Mel frequency Cepstra coefficient (MFCC)
3. Perceptual linear prediction (PLP) Cepstra.

Most feature extraction package produce a multi-dimensional feature vector for every frame of speech. This study considers 12MFCC.The Cepstral coefficient are a set of features reported to be robust in some different pattern recognition tasks concerning human voice. They are widely used in speech recognition and also in speaker identification. The human voice is very well adapted to the ear sensitivity, most of the energy developed in speech being comprised in the lower frequency energy spectrum, below 4 kHz. In speech recognition tasks, usually the 12 coefficients are retained, that they represent the slow variations of the spectrum of the signal, characterizing the vocal tract shape, the spectrum of the uttered words [7].

The Mel-scale equivalent value for frequency f expressed in Hz is:

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

The MFCCs are computed by redistributing the linearly spaced bins of the log-magnitude Fast Fourier Transform (FFT) into Mel-spaced bins according to above equation and applying Discrete Cosine Transform (DCT) on the redistributed spectrum. A relatively small number of coefficients (typically 12) provide a smoothed version of the spectral envelope, leading to the isolation of the vocal tract response by the simple retention of the desired amount of information. An additional advantage in using MFCC is that they have a decorrelating effect on the spectral data, maximizing the variance of the coefficients, similar to the effect to Principal Component Analysis.

The Each dimension is a floating point value. Feature extraction modules are also called front-end or just signal processing modules.

C. Score Matching

In this paper the DTW based score matching is done. The DTW procedure combines alignment and distance computation in one dynamic programming procedure. Basic DTW assumes that:

i) Global variation in speaking rate for a person uttering the same word at different times can be handled by linear time normalization;

ii) Local rate variations within each utterance are small and can be handled using distances penalties;

iii) Each frame of test utterance contributes equally to recognition;

iv) A single distance measure applied uniformly across all frames is adequate.

These give intuitive distance measurements between time series by ignoring both global and local shifts in the time dimension. The 12 dimensional MFCC obtain for each syllable are used to compute the angle between them (normalized inner product) which serve as local-distance and represent in the form of matrix. Using Dynamic Programming (DP) the min-cost path through matrix is found [4, 6]. These values were given to decision logic to identify whether the syllable were repeated or not.

D. Decision Logic

In this paper the Decision logic uses the Perceptron to take a decision whether a syllable is repeated or not. Perceptron was the first iterative algorithm for learning linear classification and was proposed by Rosenblatt in 1956. It is a single –layer network with threshold activation function:

$$y = \mathrm{sgn}\left(w^T x + b\right)$$

The weight vector w^T is updated each time a training point is misclassified. The algorithm is guaranteed to converge when data are linearly separable. Two classes of pattern are "linearly separable" if they can be separated by a linear hyperplane. The example for a pattern of linearly separable is shown in Fig. 2 by one of the samples of speech collected which has such pattern.

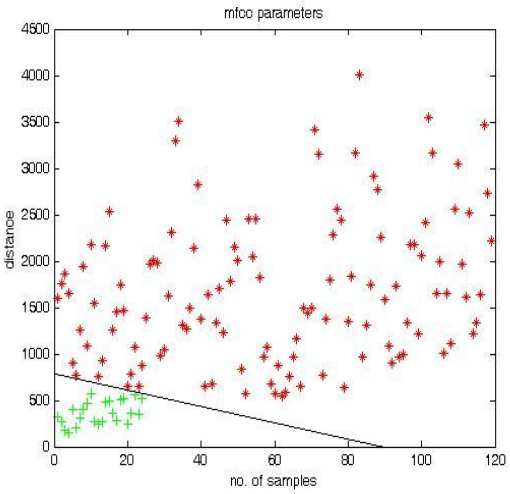

Fig. 2 Hyperplane which separate two classes

Suppose that target values (d_t) take either 1 or -1:

$$d_t = \begin{cases} 1 \ if & x \in c1 \\ -1 if & x \in c2 \end{cases}$$

Here we find w such that

$$w^T x > 0 \qquad for \qquad x \in c1$$
$$w^T x < 0 \qquad for \qquad x \in c2$$

This implies that

$$w^T xd > 0 \quad \forall x$$

The Perceptron criterion leads to the following objective function

$$\varepsilon(w) = -\sum_{x_t \in m} w^T x_t d_t$$

where m is the set of vectors, x_t which are misclassified by the current weight vector.

The gradient of $\varepsilon(w)$ is:

$$\frac{\partial \varepsilon}{\partial w} = -\sum_{x_t \in m} x_t d_t$$

The basic idea behind Perceptron is shown in Fig. 3

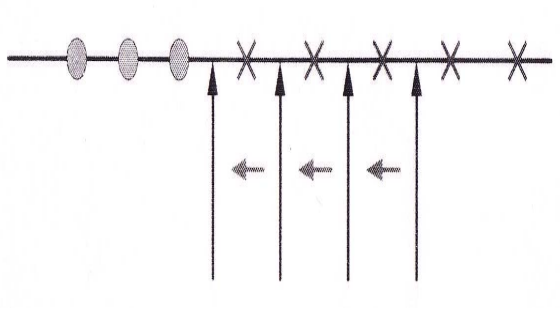

Fig. 3 Basic idea of Perceptron

If distinct parameters are separated, do not move. If not, move it to the left. If the pattern is correctly classified, do nothing, else:

$$\Delta w = \eta \sum_{x_t \in m} x_t d_t$$

The Perceptron classifier minimizes the error probability much better then Minimum Mean Square Error (MMSE) classifier [10].

The Perceptron learning algorithm is given below:

i) Get a training sample.

ii) Check to see if it is misclassified.

a) If classified correctly, do nothing.

b) If classified incorrectly, update w by

$$\Delta w = \eta x_t d_t$$

iii) Repeat steps (i)and (ii) until convergence

III. Results

Ten samples of speech were collected, out of which eight samples were used for training and remaining two samples for testing. The percentage of accuracy was 83% with testing data, where as the existing method which uses the ANN, has an accuracy of 78.01% [8, 9]. The results for two test data are listed in the Table I.

TABLE I
PERCENTAGE ACCURACY FOR TEST DATA

Feature Extraction algorithm	Test data1	Test data2	Average accuracy
MFCC	82%	84%	83%

The syllables per minute (SPM) and percent disfluency (PD) were calculated using following formulae:

$$SPM = \frac{Total \quad number \quad of \quad syllables \quad read}{Total \quad time \quad in \quad \sec onds} X \quad 60$$

$$PD = \frac{Total \quad number \quad of \quad disfluent \quad syllable}{Total \quad number \quad of \quad syllable} X \quad 100$$

The results are tabled in Table II for two testing data.

TABLE II
PERCENT DISFLUENCY (PD)

Parameters	Testdata1	Testdata2
No. of syllable	171	147
Time in Secs	68.4	62.4
Fluent syllable	130	121
NON-fluent syllable	41	26
SPM	150	141
PD (%)	23.97	17.68

The Table II helps the speech-language pathologist to assess the client and also improve interjudge agreements about stuttered events.

IV. DISCUSSION

In this paper a novel approach for automatic detection of syllable Repetition was presented for objective assessment of stuttered disfluencies. The novel approach present in this paper has a better performance than the existing method which uses ANN. In the present paper we discussed the different steps involved in finding the number of repetitions from the speech samples using MFCC feature extraction algorithm. Some of the main areas for future work are:

1.The number of Training data can be increased and checked with testing data to improve the accuracy.

2.Lot more training algorithm can be tried to improve the results.

3.Different feature extraction algorithm including LPC, PLP and others may be verified [2].

4.An alternate method for score matching may be designed.

ACKNOWLEDGMENT

The Authors would like to thank L&T and itie Knowledge Solutions for providing technical support and timely guidance.

REFERENCES

[1] D.Kully and E.Boerg, "An investigation of inter_clinic agreemet in the identification of fluent and stuttered syllables," Journal of fluency disorders, vol.13, pp.309-318, 1988.
[2] Dalouglas O'Shaughnessy, "Speech Communication," Human and Machine, Universities press, second edition, 2001.
[3] E.G.Conture. "Englewood cliffs,new jersey:Prentice-Hall," 2nd edition,1990.
[4] E.Keogh, "Exact indexing of dynamic time warping," .In VLDB.pp.406-417.Hong Kong, China, 2002.
[5] E.Yairi & B.Lewis, "Disfluencies at the onset of stuttering," Journal of speech & Hearing Research, vol.27, pp.154-159, 1984.
[6] H. Silverman & D.morgan, "the application of dynamic programming to connected speech segmentation," IEEE ASSP Mag.7, no.3, 7-25, 1990.
[7] L. Rabiner and B.H. Juang. "Fundamental of speech recognition," PTR Prentice Hall, Englewood Cliffs, New Jersey, 1993.
[8] Peter Howell, Stevie Sackin, and Kazan Glen, "Development of a Two-stage procedure for the Automatic Recognition of Dysfluencies in the speech of children who stutter: I. Psychometric Procedure Appropriate for selection of Training Material for Lexical Dysfluency Classifiers," JSLHR, vol.40, pp.1073-1084, October 1997.
[9] Peter Howell, Stevie Sackin, and Kazan Glen, "Development of a Two-stage procedure for the Automatic Recognition of Dysfluencies in the speech of children who stutter: II. ANN Recognition of Repetitions and Prolongations with supplied word segment markers," JSLHR, vol. 40, pp.1085-1096, October 1997.
[10] Tack Mu Kuson and Michael E. Zervakis, "Gaussian Perceptron: Learning Algorithms," IEEE International Conference on Systems, Man, and Cybernetics, vol. 1 pp. 105-110, Oct 1992.
[11] W.Johnson et al., "The onset of stuttering, minneapolies university of minnesata press," 1959.
[12] W.Reichl and G.Ruske, "syllable segmentation of continuous speech with Artificial Neural Networks," In Processing of Eurospeech, Berlin, vol.3, pp.1771-1774, 1993.