Annales UMCS Informatica AI 3 (2005) 131-138

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI

http://www.annales.umcs.lublin.pl/

Speech disfluency detection with the correlative method

Waldemar Suszyński[a*], Wiesława Kuniszyk-Jóźkowiak[a], Elżbieta Smołka[a], Mariusz Dzieńkowski[b]

[a]*Maria Curie-Skłodowska University,*
pl. M. Curie-Skłodowskiej 1, 20-031 Lublin, Poland
[b]*Department of Informatics System, Management Department, Technical University of Lublin,*
Nadbystrzycka 38, 20-618 Lublin, Poland

Abstract

The presented work constitutes a continuation of research on automatic disfluency recognition in utterances by stuttering people. One of the most frequently occurring episodes are syllable repetitions. The repeated fragments have a similar spectral structure, but they differ in their duration times. In order to detect them, correlation of 1/3 octave spectra was applied in connection with the procedures analysing the amplitude-time structure of sound files. The elaborated computer programme allows for recognition of that type of difluency in continuous speech and for exact, graphically illustrated location of the detected episodes in a sound file. It has been verified on the basis of over a hundred 4-second non-fluent utterances. Its functioning has been examined at various border values of the correlation co-efficient and various widths of the time window. Over 70% efficiency of the automatic detection of the episodes has been achieved. The result is comparable to those achieved with the use of the audio monitoring method.

1. Introduction

The measurements of disfluency episodes in the speech of stuttering people is vital in diagnosing, monitoring of the course and assessment of the final results of therapy. Currently, they are carried out with the use of audio monitoring, which is related to a significant effort of the logopaedist. Due to the subjective nature of such measurements and, sometimes, little accord among the listeners, it is difficult to compare various therapeutic methods which are in use. Thus, it would be desirable to elaborate an objective method of detection and duration time measurement of particular types of disfluencies, which would be based on acoustical characteristics of a speech signal.

The aim of the research carried out by the authors of the present work is to elaborate a programme algorithm for automatic recognition of various types of speech disfluencies in stuttering people's utterances, on the basis of their

[*]Corresponding author: *e-mail address*: biker@tytan.umcs.lublin.pl

parameterised characteristics in the amplitude-frequency space. It was for this aim that the disfluencies were classified according to similarities of acoustic features, which, after proper parameterisation, constitute the basis for their automatic recognition. In utterances of stuttering people there occur many disfluencies which vary in terms of their characteristics and duration times. Among them, some groups may be singled out which have similar acoustic features. For instance, the frequently occurring prolongation of fricatives are characterised by the spectrum maximum remaining in the range of high frequencies, prolongation of other sounds, e.g. nasals and vowels – by little variability of the average sound level and spectrum maximum location, syllable repetitions – by similarity of the spectra, and the disfluencies during the articulation of plosive consonants they take the form of short courses separated with long pauses. The procedures using fuzzy logic [1] have been elaborated for detection of sound prolongation [2-4] and of disfluencies related to the articulation of plosive consonants as well as blockades [5]. They allow for over 90% accurancy of recognition of these episodes.

The presented work concerns automatic detection of syllable repetitions. This type of difluency is the one to occur most frequently in the speech of stuttering people. The elaborated programme allows for recognition of these episodes in continuous speech.

2. Computer programme

The programme algorithm contains the following procedures:

1. FFT analysis of sound files of arbitrary duration time (with the use of the Hamming window of appr. 20 msec).
2. Calculation of amplitude values in 1/3-octave bands and their correction with an A filter. 21 filters were applied within frequency range from 100 to 10,000 Hz. The procedure reflects closely the perception process of the human ear. The choice of the 1/3-octave scale is a compromise between the width of Zwicker critical band frequency and Moore's measurements [6-8]. The A correction filter approximates the dependence of subjective sound loudness on its frequency.
3. Projection of the spectrum obtained in point 2, arbitrary stretching of it on the time scale, choice of any given fragment for visualisation and paying back of the sound.
4. Calculation and projection in any given window of the courses of averaged sound level in the function of time x(t) (Eq. 1).

$$x(t) = \frac{1}{21}\sum_{i=1}^{21} x_i(t)\,, \tag{1}$$

where xi (t) – sound level in the 1/3-octave band number i in the t moment.

5. Calculation of correlation coefficients Rt(T, τ) for subsequent time moments as the function of the window width T and time relocation τ according to the formula (Eq.2-3).

$$R_t(\tau,T)=\frac{\sum_{i=1}^{21}\sum_{t}^{t+T}[x_i(t)-\mu(t)][x_i(t+\tau)-\mu(t+\tau)]}{\sqrt{\left(\sum_{i=1}^{21}\sum_{t}^{t+T}[x_i(t)-\mu(t)]^2\right)\left(\sum_{i=1}^{21}\sum_{t}^{t+T}[x_i(t+\tau)-\mu(t+\tau)]^2\right)}}, \quad (2)$$

$$\mu(t)=\frac{1}{21T}\sum_{i=1}^{21}\sum_{t}^{t+T}x_i(t), \quad (3)$$

where $x_i(t)$, $x_i(t+\tau)$ – the corrected sound levels in the band number i for the time samples t, $t+\tau$, $\mu(t)$, $\mu(t+\tau)$ – the average values of all the values in the *21* bands in the T window, t – the current number of the time sample, T – the width of the window (number of samples in the time window).

The time window in the working version of the programme is optionally chosen. Thanks to that, the assessment of the efficiency of automatic detection as the function of the window width and choosing the proper setting of the parameter were possible.

6. Visualisation of the values of correlation coefficients in the function of time in a separate time window, with the opportunity of setting freely the cursor on the spectrum picture and with a synchronically moving marker of the correlation coefficient. A three-dimensional picture was applied based on a coloured spectrum.
7. Detection of neighbouring syllables by finding two subsequent maxima separated with a significant decrease of the sound level. The aim of this procedure was to eliminate correlated identical syllables occurring in fluent speech, which, however, are not repetitions. The condition of qualification of two fragments (marked as ***1*** and **n**) as belonging to the set of repetitions was the following inequality:

$$\sum_{i=1}^{n}Max_i-\sum_{i=1}^{n-1}Min_i\le 3\left[\frac{Max_1+Max_n}{2}-Min_1\right], \quad (4)$$

Max_i, Min_i – the mean the subsequent values of local maxima and minima.

8. Comparison of average levels of neighbouring maxima and minima. The earlier research showed that the decreases in average level (differences between maximum and minimum) in syllable repetitions exceed 10 dB (see Fig. 2).
9. Comparison of average levels of neighbouring maxima and elimination of correlation of these fragments, for which the difference is larger than 10 dB.

10. Measurement of time, for which the correlation coefficient exceeds the given constant parameter of relocation τ. In order to isolate syllable repetitions it was assumed that the time exceeds 100 msec. This way, accidental correlations of short fragments and correlations in sound prolongations are eliminated. The border value of the correlation coefficient is also a selectable parameter in order to optimise repetition recognition.

3. Optimisation of the programme for automatic repetition detection

Proper functioning of the programme, i.e. high efficiency of repetition detection with a small number of false alarms depended on the following adjustable parameters:

1) decrease of the average level (difference between the maximum and the minimum) in the repeated syllable (see point 8),
2) border value of the correlation coefficient (see point 9),
3) width of the time window T.

The parameters described in points 1 and 2 were selected on the basis of earlier research of the authors of the present work, and the optimum width of the time window T was set during the programme verification.

Re point 1) Figure 1 shows the histogram of the decreases in average levels in syllable repetitions measured in 126 4-second utterances of 5 stuttering people. It turned out that for 98% of the episodes the difference exceeded 10 dB.

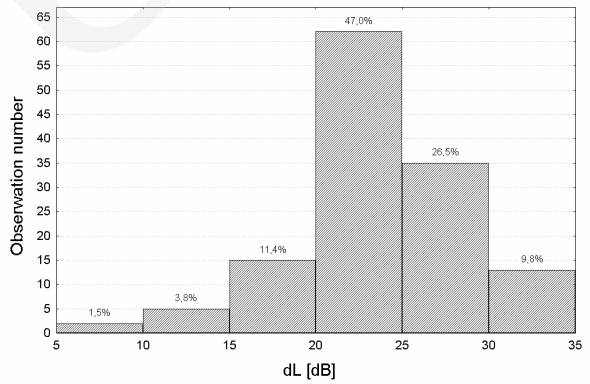

Fig. 1. Histogram of decreases in average levels in syllable repetitions

Re 2) In the same utterances maximum values of correlation coefficients for the repeated syllables were determined. They were contained within the range from 0.7 to 1.0, and in over 80% of the analysed episodes they exceeded 0.85. The graph of the cumulated values is presented in Fig. 2.

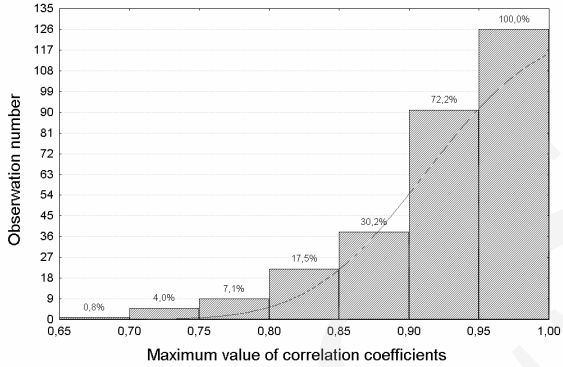

Fig. 2. Cumulated distribution of maximum values of correlation coefficients observed in syllable repetitions

4. Programme verification

In order to verify the programme for automatic detection of repetitions, 116 4-second sound files were chosen by five stuttering people. In each file there occurred a syllable repeated once or multiply in the environment of fluently uttered words. The choice of fragments of such length is related to the optimum for which the listeners' assessments are the most coincident, set experimentally by speech disfluency diagnosticians. The chosen fragments were independently listened to by two people. Their task was to answer the question whether in the chosen utterance a repetition occurs and to determine the beginnings of the repeated syllables. The coincidence of their assessments equalled 77%.

In the final version of the programme containing all the procedures described earlier in the work, the ability of the programme to detect disfluencies dependent on time window width T. The aim of the analysis was to select the window width in the way which would assure high detection sensitivity (percent of the detected disfluencies) with the maximum reduction of incorrect recognitions.

For various windows sensitivity was determined (the relation of the sum of correctly identified repetitions to the sum of all repetitions) and predictability (the relation of the sum of correct recognitions to the sum of all the recognitions) [9] (Eq. 5-6)

$$sensitivity = \frac{\sum true\ \ detected\ \ repetitions}{\sum true\ \ detected\ \ repetitions + \sum non\text{-}detected\ \ repetitions} \quad (5)$$

$$predictability = \frac{\sum true\ \ detected\ \ repetitions}{\sum true\ \ detected\ \ repetitions + \sum false\ \ detected\ \ repetitions} \quad (6)$$

With the optimum choice of the parameters (the border correlation coefficient of 0.85, time window of over 60 msec) almost 0.7 sensitivity and 0.7 predictability of the automatic repetition detection method was obtained.

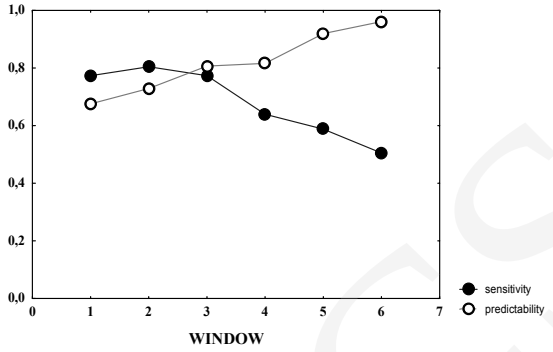

Fig. 3. The relation of sensitivity and predictability of automatic detection method to the time window width T

In addition, starting times of the disfluencies measured on the basis of audio monitoring and simultaneous observation were compared with those obtained from the correlative analysis and they proved to be coincident. With the optimum selection of parameters an accurate determination of time moments where these episodes began was obtained. That is illustrated in Fig. 4, where on the X axis the starting time *tn* measured from the image of time course of the signal on the Y axis *tn* determined on the basis of the maximum of the correlation coefficient.

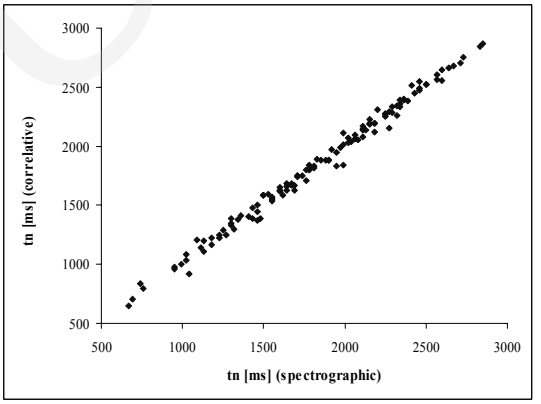

Fig. 4. Coincidence of disfluency outsets measured directly on the basis of time courses and determined with the use of correlative method in 4-second sound files

5. Conclusions

The correlative procedure is part of the computer programme for automatic detection of disfluencies in the speech of stuttering people. Recognition of speech pathologies on the basis of acoustic analysis of an utterance creates the opportunity of simple, non-invasive diagnostics. Stuttering is a very complex,

difficult in diagnosing and therapy articulation fault, which has not been thoroughly studied in spite of the intensive research carried out in this field. It may take various forms in different people and change under the influence of external factors. Acoustic analysis of disfluency episodes, their classification and recognition is of great significance for optimum diagnosis of that speech disturbance. The issue of speech and speaker recognition is currently the subject of intensive research carried out all over the world, with a vast area of possible applications. In the research, digital transfer and signal analysis methods prevail. Recognition of pathological speech, and, in particular, stuttering requires a different approach from the applied techniques of recognition of undisturbed speech. This problem has not been solved in the world research, although there exists high demand for such work among logopaedists. An attempt at automatic detection of non-fluent speech with the use of neural networks was made by Howell's team (London) [10-11], as well as with the use of Czyżewski's rough sets (Gdańsk) [12].

In the automatic detection method whose fragments are presented in this paper, the following types of disfluencies are detected: sound prolongations, disfluencies in articulation of plosives, repetitions and insertions. In the detection of repetitions and non-fluent articulations of plosive consonants, features characteristic of these episodes and their fuzziness in the time-amplitude space were used. They distinguish the decidedly non-fluent episodes from fluent speech, that is why their detection efficiency is high with a neglectable number of false alarms.

Recognition of repetitions is much more difficult, due to their variety and complexity and to the fact that in fluent utterances the same fragments (syllables, sounds) are often repeated. The authors applied the correlative method in the recognition of this type of disfluencies, with additional conditions determined on the basis of many observations. The selection of parameters resulted from a compromise between the detection efficiency of non-fluent realisations and its predictability, which became greater with the decrease in the number of false alarms resulting from the correlation of fluent fragments. The correlation procedure without specific conditions for syllable repetitions is also used to detect insertions. As it follows from the authors' research, particular speakers have their own specific type of insertions. Marking one of them and application of the correlative method allows for automatic detection of similar episodes in the analysed utterance.

Acknowledgements

Scientific work financed from the funds of the Scientific Research Committee in the years 2004-2005 as a research project.

The authors wish to thank Natalia Fedan for translating the text into English.

References

[1] Zadeh L.A., Fu K.S., Tanaka K., Shimura M., *Fuzzy sets and their applications to cognitive and decision processes*, New York: Academic Press (1975).
[2] Suszyński W., Kuniszyk-Jóźkowiak W., Smołka E., Dzieńkowski M., *Prolongation detection with application of fuzzy logic*, Annales Informatica UMCS, Lublin, 2 (2004) 183.
[3] Suszyński W., Kuniszyk-Jóźkowiak W., Smołka E., Dzieńkowski M., *Automatic Recognition of Nasals Prolongations in the Speech of Person Who Stutter*, Structures-Waves – Human Health, Kraków, XII(2) (2003) 175.
[4] Suszyński W., *Automatic recognition of the speech disfluencies*, 50 Open Seminar in Acoustic, Szczyrk-Gliwice, (2003) 386.
[5] Suszyński W., Kuniszyk-Jóźkowiak W., Smołka E., Dzieńkowski M., *Automatic recognition of non-fluent stops*, Annales Informatica UMCS, Lublin, 1 (2003) 133.
[6] Gold B., Morgan N., *Speech and audio signal processing*, John Wiley & Sons, Inc. New York (1999).
[7] Moore B.C.J., Peters, R.W., Glasberg B.R., *Auditory filter shapes at low center frequencies*, The Journal of the Acoustical Society of America, 88 (1) (1990) 132.
[8] Zwicker E., *Subdivision of the audible frequency range into critical bands (Frequenzgruppen)*, The Journal of the Acoustical Society of America, 33 (1961) 248.
[9] Jungk A., Thull B., Ran G., *Intelligent Alarms for Anaesthesia Monitoring Based on Fuzzy Logic Approach*, In: Fuzzy Logic in Medicine, Physica-Verlag, Heidelberg New York (2002) 113.
[10] Howell P., Sackin S., Glenn K., *Development of 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*, Journal of Speech, Hearing and Language Research, 40 (1997) 1085.
[11] Howell P., Sackin S., Glenn K., Au-Yeung, J., *Automatic stuttering frequency counts*, In Speech Motor Production and Fluency Disorders, edited by H. Peters, W. Hulstijn, and P. van Lieshout (Elsevier, Amsterdam) (1997).
[12] Czyżewski A., Kaczmarek, A., Kostek, B., *Intelligent processing of stuttered speech*, Journal of Intelligent Information Systems, 21(2) (2003) 143.