Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Technology 5 (2012) 812 – 819

Procedia Technology

CENTERIS 2012 - Conference on ENTERprise Information Systems / HCIST 2012 - International Conference on Health and Social Care Information Systems and Technologies

Measure and Comparison of Speech Pause Duration in Subjects with Disfluency Speech

João Paulo Teixeira*, Maria Goreti Fernandes, Rita Alexandra Costa

Politechnic Institute of Bragança, Apartado 1134, Portugal

Abstract

This work has the goal of comparing the pause duration in the disfluency speech and normal speech. Disfluency and normal spontaneous speech was recorded in a context were the subjects had to describe a scene from each other. The pause determination algorithm was developed. The automatic pause determinations allowed the measure of percentage of silence along the record of several minutes of speech. As expected these parameters is rather different in subjects with and without disfluency speech, but it do not seems that it is proportional to the severity of the disfluency.

Keywords: Speech Pause Duration Measurement, Disfluency Speech.

* Corresponding author. Tel.: +351 273 303129; fax: +351 273 313051.
E-mail address: joaopt@ipb.pt.

2212-0173 © 2012 Published by Elsevier Ltd. Selection and/or peer review under responsibility of CENTERIS/SCIKA - Association for Promotion and Dissemination of Scientific Knowledge
doi:10.1016/j.protcy.2012.09.090

1. Introduction

The speech is essential to communication between human beings, enabling the exchange of different cultures and allowing personal development and cultural of man.

To produce the human voice it is necessary that the air is pulled out of the lungs, through the throat and vocal cords being projected by the mouth.

Three states may be referred for the representation of speech, among them the silence, i.e. when no speech is produced and the vocal cords are relaxed, the unvoiced speech, where the vocal cords do not vibrate and the glottis is still open and the voiced speech, where there is vibration of the vocal cords that do vary the degree of opening of the glottis and consequently the volume of air that passes through it, Tate et al. [1].

Within the speech disorders, the dysfluencies is one of the areas of study most fascinating. This disturbance of fluency is found in all parts of the world regardless of race, socioeconomic levels and degrees of schooling.

The dysfluencies, also known as stutter, is seen as a disorder in which the fluency of speech is hampered by blockades, interjections, involuntary repetitions, prolongations of sounds and/or silent pauses, Jakubovicz [2]. In addition to the changes in the rhythm of speech, also may occur associated with movements, such as wink, twisting of the head, foot tapping, among others. These movements may arise spontaneously, or as a way to be free of the blockade that prevents him from speaking, Ribeiro [3].

Jakubovicz [2] divided the dysfluencies into four types, and they were not normal fluency, the stutter, disprosodia organic and taquilalia. According to the author Koogan [4] the stutter is classified into only three types: tonic, clonic seizures and mixed. Finally, according to Wendell Johnson (cit in Jakubovicz, [5]) the dysfluencies are classified into eight categories: interjections, repetition of sounds, repetition of syllables, repetition of words, repetition of phrases, review of the sentence, incomplete phrases, words and broken long sounds.

There are some classification instruments specific to reach the degree of severity of dysfluencies, being the scale "Stuttering Severity Instrument for Children and adults", proposed by Riley [6], one of them.

On the basis of a study carried out previously by the authors [7], it was assigned a degree of severity of dysfluencies to 6 subjects. The typology followed was similar to that of author Wendell Johnson, since it shows episodes of stutter that more fit with the scale of measurement of the degree of severity of dysfluencies.

Such tests were performed in spontaneous speech and reading speech and made the visual observation and later analysis of the behaviors of the subjects during the tests.

The evidence allowed the obtaining of frequency, and is the number of syllables disfluent divided by the total number of syllables spoken, expressed as a percentage and the duration, which is the average of the three events disfluent longer of both tests. These were converted to a score, using the standard scales proposed by Silva [8].

By using the Matlab, it was made the automatic counting of parameter for further evaluation of the degree of severity of stutter, as well as obtaining the means of the three longer disfluent events.

Finally, statements were added to the scores of the three stages of evaluation (frequency, duration of the three longer disfluent events and behaviors associated with). The final value was converted into percentage, allowing this, when compared to legal values, the classification of the degree of severity of dysfluencies in very low, low, moderate, severe and very severe.

Table 1 shows the results obtained in the study mentioned above.

Table 1. Scores and degrees of severity for each of the subjects, Costa and Fernandes [7]

Score									
Subject nº	Gender	Age	Frequency			Duration of the 3 longer events	Behaviors Associated	Final Score	Degree of Severity
			Spontaneous Speech	Reading	Total				
1	M	23	9	5	14	8	10	32	Serious
2	M	12	9	9	18	14	14	46	Very Serious
3	M	22	6	4	10	8	10	28	Moderate
4	F	18	5	4	9	10	6	25	Moderate
5	F	21	2	0	2	4	0	6	Not Disfluente
6	F	25	0	2	2	4	0	6	Not Disfluente

2. Speech Signal Record

To ensure that the information obtained is credible, it is necessary that the collection of the signal is made under certain conditions.

There are several processes that enable the acquisition of speech signal for later analysis. One which is commonly used is a process where the acquisition of the acoustic signal produced by the speaker and made in a soundproof room, using a unidirectional microphone. This signal is amplified by a pre-amplifier of linear signal and is stored on a magnetic tape of good quality, being later or immediately held their conversion to digital signal using an anti-aliasing filter. The storage of the signal is then done in digital format, Teixeira [9] and McAulay & Quatieri [10].

In the present work the realization of the recordings was done in a quiet room, trying to close a soundproof room. For the acquisition of speech signals it was used: unidirectional microphones Best e840, with a frequency response of 40Hz to 18kHz. It was also used two personal computers with the Praat software installed, Boersma P, Weenink [11], allowing this to obtain recordings in the form of Wave file. This software also allowed a segmentation of the speech signal for a later use of the same. Since short recordings were used the Record parameters were mono sound, a 16-bit resolution and a sampling frequency of 11025 Hz.

3. Methodology

In this work participated 8 subjects of both genders (5 male and 3 female), among which 7 are adults and 1 child, with ages ranging between eleven and twenty-six years. Of the eight subjects, four are considered disfluent aged between 11 and 23 years and the rest are individuals of control (with normal speech), with ages between 22 and 26 years.

All subjects were fully informed about the study and consented to the collection of the signal for the purposes of research, and, in the case of the minor subject it was needed a permission of their legal representatives to be able to carry out the recordings.

It is important to note that none of the subjects had any other type of disturbance of speech or language and only the child was in therapeutic treatment.

With the aim to cause a spontaneous conversation sought to establish a dialog between an individual with dysfluencies and another with normal speech, without visual contact between them. Resorted to the use of pairs of similar images, in which each individual had one in their possession. The participants were told that they would establish a conversation in which discuss the images. The duration of each was approximately 15 minutes.

Finally, the pause measurement algorithm was applied to the speech of individuals with dysfluencies and individuals without this disturbance.

4. Automatic Determination of the Pauses in Speech

For the automatic determination of pauses some tools were used, among them the moving average, the moving energy and the zero crossing rate. These tools perform their processing only in the temporal domain.

The algorithm was originally developed by Teixeira [9] for the classification of the signal in the zones of silence, voiced speech and unvoiced speech. The algorithm is based on the three mentioned tools (moving average, the moving energy and the zero crossing rate) and in one decision area. This decision being based on the result of two vectors obtained by the moving energy and zero crossing rate.

Initially the algorithm reads the signal. This signal is squared. Then, the moving average is applied using a window of length 20 and spaced with 10 samples allowing smoothing of the signal. This new signal will have a length 10 times lower than the original. Subsequently, the zero crossing rate is applied to the original signal derivate with a window of 20 samples and a spacing of 10. The zero crossing rate is applied to the signal derivate and not directly to the original signal because original signal can have an offset eliminating the zero crossing.

The output signal is smoothed out by applying the moving average with a window length of 100 samples and unitary spacing. Also this signal will have a length ten times less than the original. In Figure 1, is presented the speech signal corresponding to the “Eu tenho” (I have), as well as the application of moving average and the zero crossing rate of the derived signal. The derivative consists in application of eq. 1:

$$d(n) = x(n) - x(n-1) \tag{1}$$

The reason to use the derivative is because in the acquisition of speech signal, sometimes there is a small offset that makes changing the zero crossing rate of the signal.

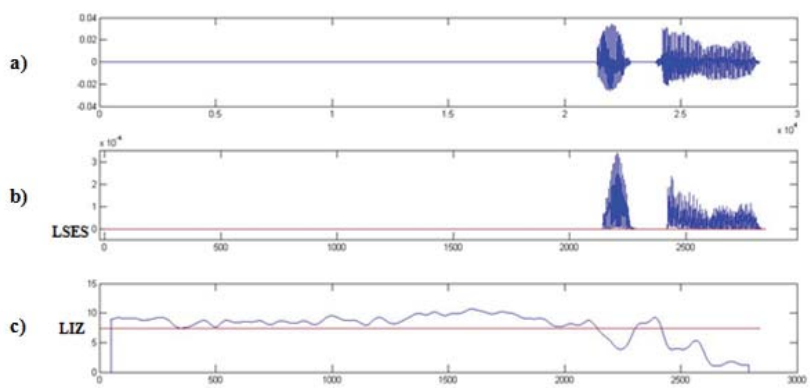

Fig. 1. a) Speech signal corresponding to the “Eu tenho” (I have); b) Moving energy in relation to the LSES; c) Zero crossing rate of the derivative with respect to LIZ

In Figure 1, secured silence is in the first instants of time. This signal of silence is a noise signal with low energy and a high rate of zero crossing. In Figure 1 (b) there is a visible line in red that relates to the maximum level of energy corresponding to the silence, called LSES (upper limit of the energy signal), and this is the maximum signal energy average sliding after smoothed. In turn, it is also established in the zero crossing rate signal after smoothed a minimum level of zero-crossings, called LIZ (Lower Limit of Zeros) ,which is visible in sub-figure 1 (c) in red. It should be noted that a speech signal contains lower zero crossing rate and greater energy that a signal of silence.

Continuing the follow-up of the algorithm, were declared the variables *e* and *z*, which count the number of elements below the LSES limit and above the limit LIZ respectively in a window of length 10. High values of *e* represents low-energy, low values of z imply low zero crossing rate. Based on the variables *e* and *z* a decision is made to classify the original signal in 1-Silence or 2-Speech.

On the basis of variables *e* and *z* has established a "field of decision", having the *z* variable in the abscissa, between 0 and 10 and the e variable *e* in the ordinate axes, also between 0 and 10. In the "field of decision" the areas were settled for the silence and speech, as can be seen in Figure 2.

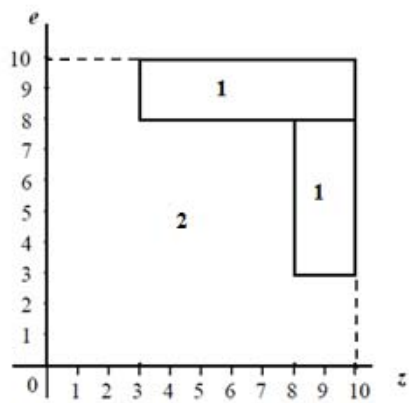

Fig. 2. “Field of decision” for classification of speech signals based on e and z variables. 1- Silence and 2 – Speech

In figure 2 the areas of classification of speech signal can be observed, in which the silence corresponds to : $z \geq 8 \wedge e \geq 3 \vee z \geq 3 \wedge e \geq 8$.

In figure 3, the zones of silence can be identify easily. These areas are represented by continuous trace of the magenta color with 0.5 magnitude. In the zones of speech, this same continuous line is replaced by 1. The green are represented all of the above elements of LIZ and the red all elements below of LSES.

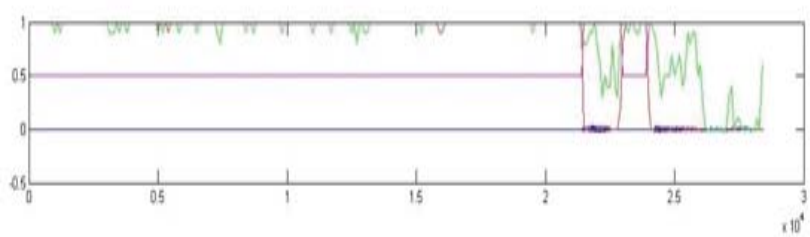

Fig. 3. Classification of the signal “Eu Tenho” (I have) according to the field of decision established

4.1. Program Code

Figure 4 presents the main algorithm to make the decision of silence or speech for each segment of speech.

```
Detrend filter aplied to the speech signal
Moving average of the energy of the signal
Determination of the treshould of energy in silent speech (LSES)
Signal is derivate
The zero crossing rate (ZCR)is applied to the signal derivate
A moving average is applied to the ZCR
Determination of the treshould of ZCR in silent speech (LIZ)
Cicle 1,
    z=0; % number of elements above LIZ
    e=0; % number of elements below LSES
    Cicly 2 (for j=1:10)
        if ZCR>LIZ,
           z=z+1;
        end
        if energy<LSES,
            e=e+1;
        end
    end cycle 2
    Energy(i)=e; %energy
    CrossingRate(i)=z; %zero crossing rate
end of cycle 1
Cycle 3
    if ((car(i,1)>=8 & car(i,2)>=3) | (car(i,1)>=3 & car(i,2)>=8)),
      decisao(i)=1; % silence
 else
      decisao (i)=2; % Speech
 end
end of cycle 3
```

Fig. 4. Mains Algorithm of the silent/speech decision for the speech segments

5. Analysis of Results

With the implementation of the algorithm it was possible to make the collection of the times of silence and times of speech of all the segments useful for later retrieval of the percentage of silence in each of the 8 subjects.

Table 2 shows the values obtained for the percentage of silence in the disfluent not control subjects. To obtain these results, were selected 5 segments of the signal of each of the subjects, which ensured equal conditions of measurement. The segments for analysis were obtained by leaving an initial portion of speech signal that would guarantee at least 1 second of silence and a start and end time that corresponded to the beginning and end of the speech, respectively. The percentage of silence was obtained based on the following equation:

$$\%\ of\ \ silence = \frac{Time\ of\ \ Silence}{Time\ of\ \ Speech} \times 100 \quad (2$$

Table 2. Results obtained for the percentage of silence in all subjects

	Subject	% of Silence
Disfluent	1	19.99
	2	41.77
	3	23.23
	4	20.19
	Average	26.30
Not Disfluent	5	6.22
	6	16.24
	7	6.75
	8	14.78
	Average	11.00

As shown in table 2, the subjects with dysfluencies present higher percentage values of silence, with an average of 26.30 %. In this group the percentage of silence varies from 19.99 to 41.77 %. Subject number 2, presenting 41.77 % of silence will be the individual from among the 4 disfluent belonging to the sample under study that will present a dysfluencies more accentuated as recorded in table 1, on the previous study, Costa and Fernandes [7]. Based on the results in tables 1 and 2 it can be alleged that it is not a viable degree classification of dysfluencies solely based on the percentages of silence. For instance, the subject 1, ranked with a degree of severity serious has the lowest percentage of silence than the subjects 3 and 4, classified with a degree of moderate severity. This can be explained by the fact that there are a number of parameters, in addition to the silence that influence the classification of the degree of severity.

The percentage of silence in the control group (not disfluent) is minor than the ones of the disfluent subjects. For this group the percentage of silence varies from 6.22 to 16.24, and has an average of 11.00 %. All the controls have lower values than any disfluent.

One aspect that may alter results is the emotional state of the subject, because when it is under pressure of an evaluation it may become nervous, causing a possible increase of disfluent moments throughout the speech. In addition to being under pressure for an assessment, the fact of having to run the tests with strange people nearby can also cause changes in the occurrence of disfluent moments.

6. Conclusion

In this work an algorithm to measure the silences in speech was developed. This algorithm was applied to determine the percentage of silence in spontaneous disfluent speech and in a normal spontaneous speech.

For the purpose of the work a simple program was developed to determine the segments of silent in opposition t the segments of speech in the speech signal.

In the course of this study were found some difficulties, particularly in finding individuals with dysfluencies which accept to be part of the study sample. A larger sample would allow results more credible, and that would be an added value for the study in question.

Based on the results obtained, it is possible to conclude that individuals with dysfluencies show percentages of silence much higher compared with individuals without any kind of disturbance in speech. Moreover this measure alone is not able to determine the degree of severity of the disfluent speech.

References

[1] Tate P, Stephens T, Seeley R. Anatomia & Fisiologia. 6ª Edição. 2003; 825-831.

[2] Jakubovicz R. A Gagueira: Teoria e Tratamento de Adultos e Crianças. Rio de Janeiro, Revinter. 2009.

[3] Ribeiro I. Conhecimentos Essenciais para Atender Bem a Pessoa que Gagueja. Pulso Editorial, 2003.

[4] Koogan. Peña-Casanova, et al. Manual de Fonoaudiologia. 2ª Edição. Porto Alegre, artes Médicas.1997.

[5] Jakubovicz R. Gagueira. 6ª Edição. Rio de Janeiro, Revinter. 2009.

[6]Riley G. A Stuttering Severity Instrument for Children and Adults. Journal of Speech and Hearing Disorders. 1972; 37: 314-322.

[7] Costa RA, Fernandes MA. Classificação do Grau de Disfluência em Indivíduos com Gaguez. Projeto Fim de Curso, IPB. 2011.

[8] Silva S. Classificação do grau de disfluência com e sem o uso de feedback acústico modificado em adolescentes e adultos gagos portugueses. Porto. 2009.

[9] Teixeira JP. Dissertação de Mestrado: Modelação Paramétrica de Sinais Para a Aplicação em Sistemas de Conversão Texto-Fala. FEUP. 1995.

[10] McAulay R, Quatieri T. Speech Analysis/Synthesis Based on a Sinusoidal Representation. IEEE Transactions on Acoustics Speech, and Signal processing. Vol. ASSP- 34, Nº4. 1986.

[11] Boersma P, Weenink D. Praat: doing phonetics by computer, http://www.fon.hum.uva.nl/praat/, University of Amsterdam.