Stream analysis for detecting stuttering episodes

Fabio Fassetti[1], Ilaria Fassetti[2], Simona Nisticò[1]
[1]DIMES, University of Calabria, Italy
[2]Therapeia Rehabilitation Center, Italy
https://doi.org/10.36505/ExLing-2019/10/0019/000381

Abstract

Stuttering is a communication disorder where a person is not able to speak fluently. A fluency disorder causes problems with the flow, rhythm and speed of speech. If one stutters, its speech may sound interrupted or blocked, as though the individual is trying to say a sound but it doesn't come out. This work is designed to help specialists in the evaluation of stuttering and recognize occurrences of disfluency episodes such as repetitions of sounds, syllables or words, silent pauses, hesitations or blocks during the speech.

Key words: Stuttering, stream detection, machine learning

Introduction

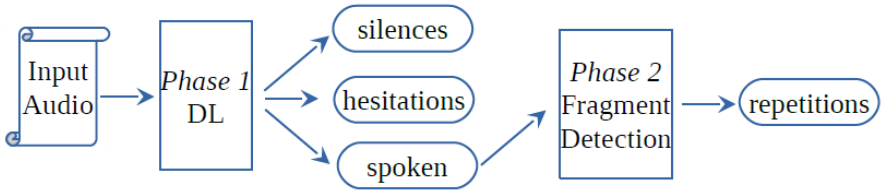
Stuttering is a disorder of speech motor control, but the disorder of stuttering is more than just the stuttering it also involves a lifetime of dealing with the anxiety and avoidances caused by the stuttering. Typical disfluencies are: a whole word or phrase repeated three time or less; interjections and/or revisions in speech. Conversely, less-typical disfluencies are: any word that is repeated as part of a word or an initial sound, whole words or phrases repeated four or more times, prolonging one sound of a word, "blocking", and any associated tension or effort observed in speech production. A fluency disorder can be evaluated by a speech-language pathologist which will conduct a precise evaluation with a series of tests, observations, interviews and listening. This work aims at helping the evaluation of stuttering and several types of disfluency. The proposed approach is based on deep learning and stream analysis and its goal is to automatically detect the points where verbal output is influential, thus helping in the early classification of stuttering or cluttering problems, providing the number of disfluencies and time intervals in which the disfluencies occur.

Preliminaries

In the proposed work, an audio stream S in *wav* format and in mono mode is assumed as input, if the audio is in stereo mode the mean value between the two channels is considered. A window W in a stream S is the portion $S(t_1, t_2)$ of S between instants t_1 and t_2. The *energy* of a window W is the inner product $<W,$

W>. The proposed framework makes use of a deep learner. Due to lack of space, details about deep learning theory are not provided.

Framework architecture:



The framework consists in two main phases. The first one aims at extracting relevant fragments from the input audio file. The goal is to remove intervals of silence and of hesitations. Hesitations are very difficult to distinguish from silence since they consist in spluttered letters/parts of words of very low energy.

The spoken fragments are then concatenated to obtain a clean stream provided as input of the second phase. During this phase, the goal is to individuate repetitions in form of very similar subsequent fragments in the clean stream. Next details about the detection technique are provided.

Detection technique

Let F be a fragment of a stream S. Three features are extracted from F: (f_S) portion of S associated with F, (f_F) the spectrum of F, (f_C) the MFCC of F. The similarity $\Sigma(F_i, F_j)$ between two fragments F_i and F_j of a stream S is

$$\Sigma(F_i, F_j) = 1 - (w_S \cdot d_S(F_i, F_j) + w_F \cdot d_F(F_i, F_j) + w_C \cdot d_C(F_i, F_j)),$$

where w_S, w_F and w_C are weights to be tuned and d_S, d_F and d_C are three distance measures based on the three features as detailed in the following.

Stream distance (d_s)

The distance $d_S(F_i, F_j)$ between F_i and F_j according to feature f_S is obtained by a novel notion of distance based on the *Levenshtein* distance. This is due to the fact that, in order to make two windows comparable, the associated signal has to be aligned in a way such that the similarity is maximized. The *Levenshtein* distance is defined on strings and evaluates the number of edit operations to make the two strings equals. It uses three kinds of operations: *insert*, *delete* and *substitute* and each of them has a cost of 1. For example, the distance between "*sitting*" and "*kitten*" is 3 since the Stream analysis for detecting stuttering episodes substitution of *s* in *k*, of *i* in *e* and the removal of *n* have to be performed. In other words, the two strings are aligned for minimizing the cost. Following the same approach, the aim is to align the fragments so that the

Euclidean distance between them is minimized. Thus, in order to compute the distance between F_i and F_j, the insertion operation of element h is $F_i(h)^2$ the deletion operation of element h costs $F_j(h)^2$ the substitution operation of element h_x of F_i and the element h_y F_j costs $(F_j(h_x) - F_j(h_y))^2$.

Spectrum-based distance (d_F)

The distance $d_F(F_i, F_j)$ between F_i and F_j according to feature f_F is obtained by computing the *Fast Fourier Transform* (FFT) of the signal associated with F_i and F_j, by normalizing the power and, then, by computing the Euclidean distance between the two spectra.

MFCC-based distance (d_C)

The distance $d_C(F_i, F_j)$ between F_i and F_j according to feature f_C is obtained by computing the vectors composed by the *Mel-frequency cepstral coefficients* (MFCCs) and, then, by computing the Euclidean distance between these two vectors.

Detection algorithm

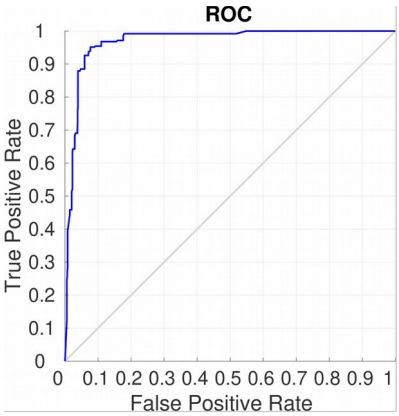
The *first phase* consists in recognizing and classifying spoken, noise, silence and hesitation fragments exploiting a multi-class deep-learner. The input stream is split in fragments of length $\lambda = 0.5$ sec which are overlapped of $\varepsilon = 0.25$ sec. With each fragment, a class is associated by the classifier and *hesitation fragments* are returned as *output*, *noisy* and *silence fragments* are removed and the *spoken fragments* compose the input of the second phase. Let S be the stream returned by the first phase, the goal of the *second phase* is to find windows $W = S(t_w, t_w+ws)$ starting at time t_w of size ws, such that

1. $\Sigma(W, S(t'_w, t'_w+ws)) < \alpha$ with $t'_w > t_w$ and α a threshold, and
2. Energy in $S(t_w, t'_w)$ is negligible with respect to $S(t_w, t_w+ws)$.

Conditions are due to the fact that a repetition is characterized by a fragment F very similar to a subsequent fragment Fj (condition 1) interleaved by a fragment absent, namely $t'_w = t_w+ws$, or with spluttered letters, namely low energy signal (condition 2) which have not be filtered by phase 1.

Experimental results

Performed experiments show that the approach is significantly accurated. The following figure reports the ROC curve associated with the output of the experiments.



The dataset employed for this preliminary campaign is publicly available [Howell et al. 2004] and it is composed by 152 audio wav files of stuttering English-speaking people of ages from 5 years to 47 years. Recordings have a length ranging from ≈65 s to ≈1028 s with a mean length of ≈166 s. Domain experts prepared the datasets by listening and manually selecting intervals where silences, hesitations and stuttering episodes occur.

References

Davis, S.B., Mermelstein, P. 1990. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In: Readings in Speech Recognition, pp. 65–74. Morgan Kaufmann Publishers Inc.

Howell, P., Davis, S., Bartrip, J., Wormald, L. 2004. Effectiveness of frequency shifted feedback at reducing disfluency for linguistically easy, and difficult, sections of speech. Stammering Res. 1(3), 309–315.

Ingemann, F., Mermelstein, P. 1975. Speech recognition through spectrogram matching. J. Acoust. Soc. Am. 57, 253–255.

Iverach, L., et al. 2018. Comparison of adults who stutter with and without social anxiety disorder. J. Fluency Disord. 56, 55–68.

Koedoot, C., Bouwmans, C., Franken, M.C., Stolk, E. 2011. Quality of life in adults who stutter. J. Commun. Disord. 44(4), 429–443.

Markett, S., et al. 2016. Impaired motor inhibition in adults who stutter - evidence from speech-free stop-signal reaction time tasks. Neuropsychologia 91, 444–450.

Sammut, C., Webb, G.I. (eds.) 2010. Encyclopedia of Machine Learning. Springer.

Weir, E., Bianchet, S. 2004. Developmental dysfluency: early intervention is key. CMAJ Can. Med. Assoc. J. 170, 1790–1791.

Yairi, E., Ambrose, N. 2013. Epidemiology of stuttering: 21st century advances. J. Fluen. Disord. 38(2), 66–87.

Yairi, E., Ambrose, N.G., Paden, E.P., Throneburg, R.N. 1996. Predictive factors of persistence and recovery: pathways of childhood stuttering. J. Commun. Disord. 29(1), 51–77