



# Learning and Detecting Stuttering Disorders

Fabio Fassetti, Ilaria Fassetti, Simona Nisticò

## ► To cite this version:

Fabio Fassetti, Ilaria Fassetti, Simona Nisticò. Learning and Detecting Stuttering Disorders. 15th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI), May 2019, Hersonissos, Greece. pp.319-330, 10.1007/978-3-030-19823-7\_26 . hal-02331315

**HAL Id: hal-02331315**

**<https://inria.hal.science/hal-02331315>**

Submitted on 24 Oct 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Learning and Detecting Stuttering Disorders

Fabio Fassetti<sup>1</sup>, Ilaria Fassetti<sup>2</sup>, and Simona Nisticò<sup>1</sup>

<sup>1</sup> DIMES Department, University of Calabria, Italy  
`{f.fassetti,s.nistico}@dimes.unical.it`

<sup>2</sup> Therapiea Rehabilitation Center, Italy  
`ilaria.fassetti@gmail.com`

**Abstract.** Stuttering is a widespread speech disorder involving about the 5% of the population and the 2.5% of children under the age of 5. Much work in literature studies causes, mechanisms and epidemiology and much work is devoted to illustrate treatments, prognosis and how to diagnose stutter. Relevantly, a stuttering evaluation requires the skills of a multi-dimensional team. An expert speech-language therapist conduct a precise evaluation with a series of tests, observations, and interviews. During an evaluation, a speech language therapist perceive, record and transcribe the number and types of speech disfluencies that a person produces in different situations. Stuttering is very variable in the number of repeated syllables/words and in the secondary aspects that alter the clinical picture. This work wants to help in the difficult task of evaluating the stuttering and recognize the occurrences of disfluency episodes like repetitions and prolongations of sounds, syllables, words or phrases silent pauses, hesitations or blocks before speech. In particular, we propose a deep-learning based approach able at automatically detecting disfluent production point in the speech helping in early classification of the problems providing the number of disfluencies and time intervals where the disfluencies occur. A deep learner is built to preliminarily valuate audio fragments. However, the scenario at hand contains some peculiarities making the detection challenging. Indeed, *(i)* fragments too short lead to uneffective classification since a too short audio fragment is not able to capture the stuttering episode; and *(ii)* fragments too long lead to uneffective classification since stuttering episode can have a very small duration and, then, the much fluent speaking contained in the fragment masks the disfluence. So, we design an ad-hoc segment classifier that, exploiting the output of a deep learner working with non too short fragments, classifies each small segment composing an audio fragment by estimating the probability of containing a disfluence.

**Keywords:** Deep learning · audio classification · stuttering.

## 1 Introduction

Stuttering is a communication disorder where the smooth flow of speech is disrupted. It begins during childhood and, in some cases, lasts throughout life. This dysfluency may interfere with the ability to be clear and understood. The effort

to learn to speak and the normal stress of the evolutionary growth can trigger in the child language manifestations characterized by brief repetitions, hesitations and prolongations of sounds that characterize both the early stuttering and the normal disfluency. About 5% of the child population experience a period of stuttering that lasts 6 months or more. A lot of Children of those who start stuttering will have a remission of the disorder in late childhood. Most of the risk for stuttering onset is over by age 5, earlier than has been previously thought, with a male-to-female ratio near onset smaller than what has been thought [10].

There is strong clinical evidence that more of 60% of the treated stuttering children have a stuttering relative in the family. Children who start stuttering before 42 months have a greater chance of overcoming and solving the problem. Physiological and normal developmental disfluencies are difficult to differentiate from the first signs of effective stuttering. But if the subject stutters for more than 6 months, it is difficult to solve the problem spontaneously. Signs of chronicity in older children (e.g., 6- or 7-year-olds) who had stuttered for two years may not be quite the same as those in 2- to 4-year-olds who have short stuttering histories [11]. It is important to remember that stuttering is not caused by nervousness nor is it related to personality or intellectual capabilities. Despite not demonstrating more severe stuttering, socially anxious adults who stutter demonstrate more psychological difficulties and have a more negative view of their speech [4]. Parents have not done anything that may have caused their son's stuttering even if they feel responsible in some way! Exist also an idiopathic stuttering that is caused by a possible deficiency in motor inhibition in children who stutter [6].

Stuttering is characterized by an abnormally high number of disfluencies, abnormally long disfluencies, and physical tension that is often evident during speech [8]. Some signs of stuttering are [9]:

- repetitions of whole words (e.g., “We, we, we went”)
- repetitions of parts of words (e.g., “Be-be-because”)
- prolongation or stretching of sounds (e.g., “Ssssssee”)
- silent blocks (getting stuck on a word or tense hesitations)

The child with severe stuttering often shows physical symptoms of stress, especially the increase in muscle tension, and tries to hide his stuttering and avoids speaking and exposing himself to linguistic situations. Although severe stuttering is more common in older children, it may, nevertheless, start at any age, between 1.5 years and 7 years. This person can exhibit behaviors associated with stuttering: blinking, looking away, or muscle tension in the buccal or other parts of the face. Moreover, part of the tension and of the impact can be perceived in a strong increase of the vocal tone or of the intonation (increase of the vocal frequency) during the repetitions or during the extensions. The subject with severe stuttering can resort to extraverbal sounds, interjections, such as “um, uh, well...” at the beginning of a word in which he expects to stutter. Especially moderate to severe stuttering had a negative impact on overall quality of life [5].

This work aims at contributing in stuttering recognition by helping therapists and patients in detecting episodes of disfluency. Technically speaking, we

propose a system that fed with an input audio outputs the time interval related to stuttering phenomena. The system consists in several phases, the two main ones are devoted to classify.

The rest of the paper is organized as follows. Section 2 presents the basic notions exploited in this work; Section 3 introduces the architecture of our technique; Section 4 details the proposed technique and the main phases it requires; Section 5 describes the experimental campaign we perform to validate our technique; Section 6 draws the conclusions.

## 2 Preliminaries

In this section we report some preliminary notions. The input audio file is quite clean since we assume that the therapist acquire the registration of the patient in a safe environment. From the input wav file we obtain feature vectors by considering spectrograms [3] and Mel frequency cepstral coefficients [1].

## 3 The proposed Architecture

The proposed architecture consists in several phases. For the sake of clarity, we introduce them next to provide a general overview. Each phase is detailed in the following section. The main flow is reported in Figure 1.

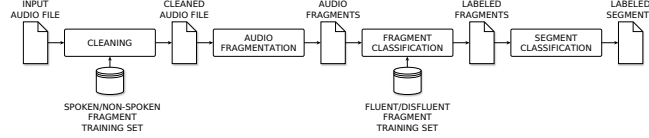


Fig. 1: Methodology Flow

**CLEANING PHASE.** This phase consists in cleaning the input audio file. Even if, as before stated, we assume that the input audio file is quite clean, without background sounds that could compromise quality, this phase is needed to clean the input audio file from intervals during which the patient does not speak. Also, during this phase the input file is normalized in terms of volume, and sampling frequency. See Section 4.1 for details.

**AUDIO FRAGMENTATION PHASE.** This phase consists in splitting the input file in fragments having length  $f_{len}$  and overlapped of  $\varepsilon$  seconds as detailed in Section 4.2.

**FEATURE EXTRACTION PHASE.** This phase consists in extracting features from raw audio fragments and represent a critical part of the architecture. Each fragment is transformed in a numeric vector as detailed in Section 4.3.

**FRAGMENT CLASSIFICATION PHASE.** This phase represents, with the succeeding *classification phase*, the core of our architecture. Here, a trained deep learner assigns each fragments to the class of *fluent* or *disfluent* with a certain probability. Details on this are reported in Section 4.4.

**SEGMENT CLASSIFICATION PHASE.** This phase is, with the previous *learning phase*, the core of the architecture. Here, a probabilistic model allows the classification of each segment composing a fragment, exploiting the overlapping of the fragments and taking as input the probabilities of belonging to the *fluent* or *disfluent* class computed by the previous phase. Details are provided in Section 4.5.

## 4 Detection Technique

In this Section we describe the proposed technique providing details about all the phases above introduced.

### 4.1 Noise Removal

The input audio file is assumed to be clean from background sounds that could compromise audio quality. However, there can be several time intervals during which the patient does not speak. These intervals cannot be completely removed since, in many cases, the silence is symptomatic of a disfluency. Thus, in this phase the system recognizes these intervals and reduces each of them so that its duration is large enough to guarantee that at least one fragment captures the segment just before the noise interval and the segment just after the noise interval. Note that, this operation is performed also for the audio files employed to train the system. To perform this operation, we trained a learner able to discriminate between spoken and no-spoken fragments. This learner is simple and highly accurated since the classes are well-separated.

### 4.2 Audio File Preparation

With the aim of obtaining classifiable fragments of the input audio file we recover to a fixed-size sliding window approach. Indeed, in order for the learner to correctly work, fragments cannot be too long, otherwise stuttering phenomena would be obfuscated by fluent speech, and cannot be too short since we need fragments of at least some seconds, how can be intuitively understood. Indeed, also a human, to recognize a stuttered phoneme, needs to hear at least for the interval covering the stuttering phenomenon that has a great variability but is, in general, wider than the duration of a segment. On the other hand, if there were not overlapping, the stuttering phenomenon could be split in two adjacent intervals and, then, not recognized. In other words, we need that at least one fragment contains all the stuttering phenomenon if this is shorter than the duration of a fragment, that a fragment and its adjacent ones cover all the stuttering phenomenon otherwise.

Let  $\mathcal{S}$  be the input audio stream and let  $d$  be its duration in seconds. Choose a fragment length in seconds, denoted as  $f_{\text{len}}$ , and an overlap size in seconds, denoted as  $\varepsilon$ , and letting

$$\begin{aligned} n &= \frac{f_{\text{len}}}{\varepsilon} && \text{be the number of segments per fragment,} \\ n_s &= \frac{d}{\varepsilon} && \text{be the total number of segments,} \\ n_f &= n_s - (n - 1) && \text{be the total number of fragments,} \end{aligned}$$

then,  $\mathcal{S}$  is sectioned in  $n_f$  fragments of equal size  $f_{\text{len}}$  and overlapped of  $\varepsilon$  seconds. Each fragment is composed by  $n$  segments and, thus,  $\mathcal{S}$  can be considered as partitioned in  $n_s$  segments and each segment, due to the overlap, belongs to  $n$  distinct fragments as illustrated in Figure 2.

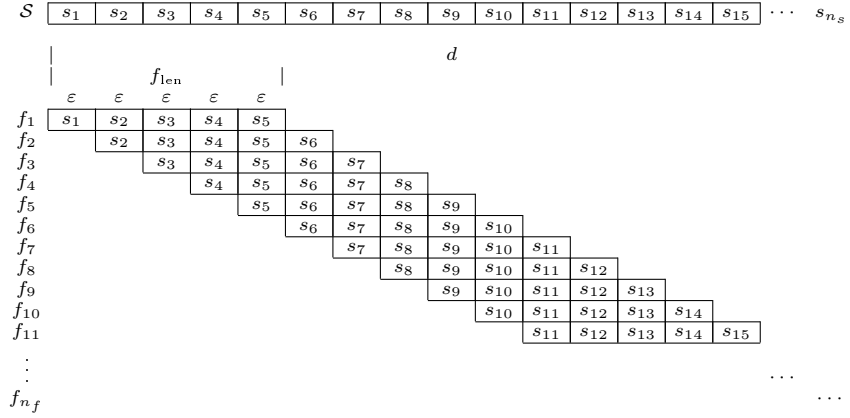


Fig. 2: Example of Audio Fragmentation with  $n = 5$ .

### 4.3 Feature Extraction

From each audio fragment, we need to build a numeric vector representing audio features. In particular, we compute spectrograms [3] and Mel frequency cepstral coefficients [1].

### 4.4 Fragment Classification

As for the fragments classification, we adopt a deep-learning based classifier.

The learning phase through a deep learner provides the fragment  $f_i, i \in [0 \dots n_f]$  (see Figure 2) with a classification  $\pi_i^\ell$  stating for the probability that the fragment  $f_i$  belongs to the class labeled  $\ell$ , with  $\ell \in \{fluent, disfluent\}$ .

#### 4.5 Segment Classification

The classification described in the previous section provides for each fragment  $f_i, i \in [0 \dots n_f]$  the probability  $\pi_i^\ell$  to belong to a certain class labeled  $\ell$ . Nevertheless, each segment  $s_i, i \in [0 \dots n_s]$  (see Figure 2) has to be classified in order to detect the time intervals of fluent speaking and the time intervals of disfluent speaking.

Consider, for example, the scenario where each fragment consists in 5 segments and a time interval  $\mathcal{I}$  with a  $c_2$  voice has a duration of 3 segments. There are, then, 7 overlapped fragments covering  $\mathcal{I}$  as illustrated in Figure 3.

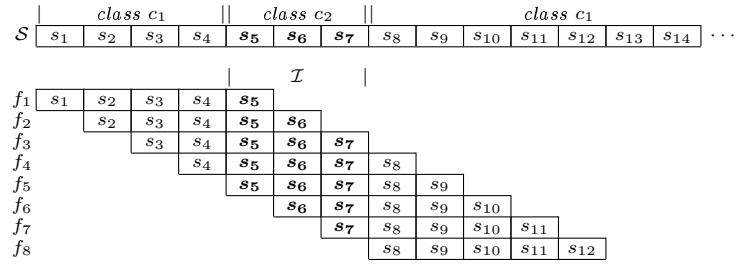


Fig. 3: Example of challenging segment classification.

Obviously, no fragment  $f_i$  has a probability  $\pi_i^{c_2}$  close to 1 since no fragment fully contains a female voice and the aim of the segment classification phase is to correctly individuate the 3 segments where the class label changes.

Let  $f_i$  with  $i \in \left[1 \dots \left(\frac{d-f_{\text{len}}}{\varepsilon} + 1\right)\right]$  denote the fragment starting from the segment  $s_i$ , then  $s_i$  belongs to the set of fragments

$$\{f_{\max(1, i-n+1)}, \dots, f_i\},$$

which are employed to evaluate the trend of classification when  $s_i$  appears, and does not belong to the succeeding set of fragments

$$\{f_{\min(i+1, n_s)}, \dots, f_{\min(i+n-1, n_s)}\},$$

containing some segments of fragment  $f_i$ . Thus, these fragments share with  $f_i$  some segments excepting  $s_i$  and are, roughly speaking, employed to evaluate the trend of classification when  $s_i$  disappears.

As described in Section 4.4, the learning phase through a deep learner provides the fragment  $f_i$  with a classification  $\pi_i^\ell$  stating for the probability that the fragment  $f_i$  belongs to the class labeled  $\ell$ .

In order to exploit the overlap to improve segment classification, we aim at evaluating the contribution that the  $i$ -th segment gives to the classification.

*Case 1: valuating the effect of the  $i$ -th segment when it appears.* The segment  $s_i$  firstly appears as  $n$ -th segment of the fragment  $f_{i-n+1}$ . Then, consider fragment  $f_{i-n}$  as referring fragment ( $s_i$  has not yet been seen) and let

$$\varphi_i^\ell(j) = \frac{\pi_{i-n+j}^\ell}{\pi_{i-n}^\ell}, \quad \forall j \in [1, n'), \quad \text{for each label } \ell$$

with  $n' \leq n$  is a parameter representing how many contributions are to be taken into account. In this case, we denote  $\pi_{ref}^\ell = \pi_{i-n}^\ell$  and  $\pi_{curr}^\ell = \pi_{i-n+j}^\ell$ .

*Case 2: valuating the effect of the  $i$ -th segment when it disappears.* The segment  $s_i$  firstly disappears for the fragment  $f_{i+1}$ . Then, consider fragment  $f_{i+j}$  as referring fragment ( $s_i$  is not seen) and let

$$\varphi_i^\ell(j) = \frac{\pi_i^\ell}{\pi_{i+j}^\ell}, \quad \forall j \in [1, n'), \quad \forall k \in \{0, 1\}$$

with  $n' \leq n$  is a parameter representing how many contributions are to be taken into account. In this case, we denote  $\pi_{curr}^\ell = \pi_i^\ell$  and  $\pi_{ref}^\ell = \pi_{i+j}^\ell$ .

We, firstly, compute the probability to observe a ratio smaller than  $\varphi_i^\ell(j)$

$$F(\varphi_i^\ell(j), \lambda) = 1 - e^{-\lambda \cdot \varphi_i^\ell(j)}, \quad \text{with } \lambda = \frac{\varepsilon \cdot j}{f_{len}}.$$

We use an exponential distribution since when  $f_i$  belongs to a class labeled  $\ell$ ,  $\pi_i^\ell$  is high, then we want to capture that the ratio is lowered with high probability and further raised with low probability. Also, the dependence of  $\lambda$  from  $j$  allows us to alleviate the exponential trend. The idea is that the more  $j$  is high, the more the change in the ratio can be high. In other words, it is quite improbable that a single segment can drastically change the ratio. Whereas, when  $j$  is high, more segments are taken into accounts and then the change can be high.

How much this value exceeds the no-change case, namely  $\varphi_i^k(j) = 1$ , is employed as weight for  $\pi_{curr}^\ell$

$$g_i^\ell(j) = \frac{F(\varphi_i^\ell(j), \lambda)}{F(1, \lambda)} \cdot \pi_{curr}^\ell.$$

In order to normalize this value in a vote ranging from 0 to 1, we apply an exponential kernel function to it:

$$h_i^\ell(j) = 1 - e^{-\lambda \cdot g_i^\ell(j)}, \quad \text{with } \lambda = -\frac{\log(1 - \pi_{curr}^\ell)}{\pi_{curr}^\ell}$$

so that, if  $\varphi_i^\ell(j) = 1$  then  $h_i^\ell(j)$  coincides with  $\pi_{curr}^\ell$ . To combine votes  $h_i^j$ , we compute the weighted *arithmetic* mean, where the weights are related to the



probability that observing  $\pi_{curr}^\ell$  given  $\pi_{ref}^\ell$  is due to chance. This probability follows a gamma distribution with parameters  $k$  and  $\theta$  and

$$f^\Gamma(x) = \frac{1}{\Gamma(k) \cdot \theta^k} \cdot x^{k-1} \cdot e^{-\frac{x}{\theta}}, \quad F^\Gamma(x) = \frac{1}{\Gamma(k)} \cdot \gamma\left(k, \frac{x}{\theta}\right)$$

are, respectively, the probability density function and the cumulative density function. Since the Gamma distribution is asymmetric and since both  $\pi_{ref}^\ell \in [0, 1]$  and  $\pi_{curr}^\ell \in [0, 1]$ , in comparing  $\pi_{curr}^\ell$  and  $\pi_{ref}^\ell$  we adopt the following strategy: if  $\pi_{ref}^\ell$  is smaller than 0.5 then we compute the probability of observing a value more extreme than  $\pi_{curr}^\ell$  given  $\pi_{ref}^\ell$ , otherwise, if  $\pi_{ref}^\ell$  is greater than 0.5 then we compute the probability of observing a value more extreme than  $(1 - \pi_{curr}^\ell)$  given  $(1 - \pi_{ref}^\ell)$ .

To compute  $F^\Gamma$ , we valuate  $k$  and  $\theta$  by fixing the value  $x$  (say  $t$  this value) maximizing the gamma probability density function and the value  $x$  such that the probability of observing  $x$  is equal to the probability of observing a value distant 4 standard deviations from the mean. In particular, we determine  $k$  and  $\theta$  with the following constraints: (i) the maximum is when  $\pi_{curr}^\ell = \pi_{ref}^\ell$  then  $t = \pi_{ref}^\ell$ ; and (ii), due to the fact that  $\pi_{curr}^\ell$  is at most 1, we want that 1 is at 4 standard deviations from the mean. The following theorem accounts for the computation of these parameters.

**Theorem 1 (Parameters  $k$  and  $\theta$  of  $F^\Gamma$ ).** *The value of parameters  $\theta$  and  $k$  of  $F^\Gamma$  such that the maximum is in a generic point  $t < 1$  and that 1 is at 4 standard deviation from the mean are*

$$k = \left( \frac{2 + \sqrt{4 + w \cdot (w - 1)}}{w - 1} \right)^2 \quad \text{with } w = -W_{-1} \left( -\frac{1}{t} \cdot e^{\frac{1}{t}} \right) \quad (1)$$

where  $W$  is the Lambert function and

$$\theta = \frac{t}{k - 1}. \quad (2)$$

*Proof.* The first constraint can be imposed by computing the first derivative of  $\hat{f}(x)$  and evaluating it in  $t$ .

$$\frac{\partial \hat{f}}{\partial x} = \frac{1}{\Gamma(k) \cdot \theta^k} \cdot (k - 1) \cdot x^{k-2} \cdot e^{-\frac{x}{\theta}} + \frac{1}{\Gamma(k) \cdot \theta^k} \cdot x^{k-1} \cdot \frac{-1}{\theta} \cdot e^{-\frac{x}{\theta}}$$

which is equal to 0 in  $t$  when

$$\begin{aligned} \left. \frac{\partial \hat{f}}{\partial x} \right|_t = 0 &\Rightarrow \\ \frac{1}{\Gamma(k) \cdot \theta^k} \cdot (k - 1) \cdot t^{k-2} \cdot e^{-\frac{t}{\theta}} + \frac{1}{\Gamma(k) \cdot \theta^k} \cdot t^{k-1} \cdot \frac{-1}{\theta} \cdot e^{-\frac{t}{\theta}} &= 0 \Rightarrow \\ (k - 1) - \frac{t}{\theta} &\Rightarrow \theta = \frac{t}{k - 1}. \end{aligned}$$

As for parameter  $k$ , since the mean is  $k\theta$  and the standard deviation is  $\theta\sqrt{k}$  we have to solve the following equation:

$$\widehat{f}(1) = \widehat{f}(k\theta + 4\theta\sqrt{k})$$

and, thus,

$$\begin{aligned} \frac{1}{\Gamma(k) \cdot \theta^k} \cdot e^{-\frac{1}{\theta}} &= \frac{1}{\Gamma(k) \cdot \theta^k} \cdot (k\theta + 4\theta\sqrt{k})^{k-1} \cdot e^{-\frac{k\theta + 4\theta\sqrt{k}}{\theta}} \Rightarrow \\ e^{-\frac{1}{\theta}} &= (k\theta + 4\theta\sqrt{k})^{k-1} \cdot e^{-\frac{k\theta + 4\theta\sqrt{k}}{\theta}} \Rightarrow \\ -\frac{1}{\theta} &= (k-1) \log(k\theta + 4\theta\sqrt{k}) - \frac{k\theta + 4\theta\sqrt{k}}{\theta} \Rightarrow \\ -1 &= \theta(k-1) \log(k\theta + 4\theta\sqrt{k}) - (k\theta + 4\theta\sqrt{k}). \end{aligned}$$

By substituting Equation (2) we obtain

$$-1 = t \log \frac{kt + 4t\sqrt{k}}{(k-1)} - \frac{kt + 4t\sqrt{k}}{k-1}$$

by setting

$$w = \frac{k + 4\sqrt{k}}{k-1} \quad (3)$$

we obtain

$$\frac{wt - 1}{t} = \log wt \Rightarrow e^{w - \frac{1}{t}} = wt$$

which can be solved by exploiting the Lambert function  $W$ , thus obtaining

$$w = -W_{-1} \left( \frac{1}{t} \cdot e^{-\frac{1}{t}} \right).$$

From Equation (3), we have

$$k = \left( \frac{2 + \sqrt{4 + w \cdot (w-1)}}{w-1} \right)^2$$

and, then, the theorem is proved.

Once  $F^\Gamma$  is fully determined, the weight of each vote  $h_i^j$  is

$$\nu_i^\ell(j) = \begin{cases} 1 - \widehat{F}(h_i^\ell(j)) & \text{if } \pi_{curr}^\ell \leq \pi_{ref}^\ell \\ \widehat{F}(h_i^\ell(j)) & \text{otherwise} \end{cases} \quad (4)$$

and, then, we estimate the probability of each segment to belong to a class through the following equation:

$$p_i^\ell = \frac{\sum_j \nu_i^\ell(j) \cdot h_i^\ell(j)}{\sum_j \nu_i^\ell(j)}. \quad (5)$$

## 5 Experiments

In this section we present test conducted to prove the effectiveness of our method. In particular, in order to validate the accuracy of the technique, we perform two families of experiments aimed at proving (i) the effectiveness of the segment classification phase in correctly detect time intervals where the class label changes and (ii) the effectiveness of the method in detecting stuttering phenomena.

**Segment classification accuracy.** First, we aim at proving that the segment classification phase is effective in distinguishing between classes. Specifically, we want to show how the length of the fragment and the length of the overlap are involved in classification performances and how the segment classification phase is able to correctly assign time intervals to class labels. In this scenario, we are not interested in the ability of the classification phase of correctly classifying a fragment fully belonging to a class while we are focused on the single segment classification. So we consider two well-separated classes, *male* voice and *female* voice. We built a synthetic test stream by randomly mix pieces of sentences pronounced by males or females and evaluate the capability of the method of correctly classifying single segments.

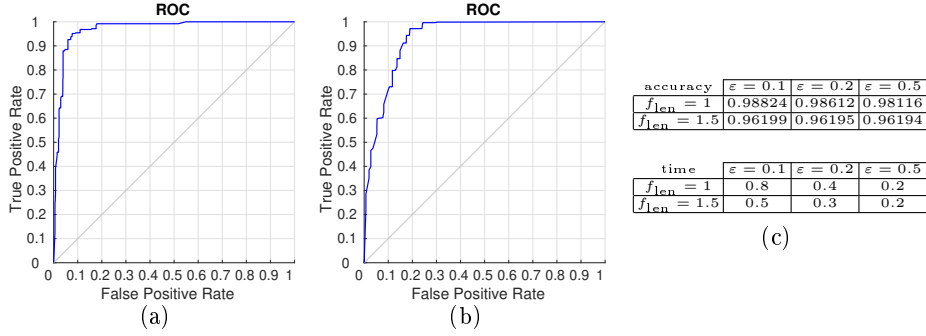


Fig. 4: Accuracy of the technique

In particular, for each combination, we considered 50 random male voice audio files of 5 seconds. For each of them, say  $f_m$ , we randomly select a female voice audio file  $f_f$ , we randomly picked a fragment from  $f_f$  having random duration  $d \in [0.5, 2]$  seconds, we randomly chose a starting point  $x$  in  $f_m$  and substituted  $d$  seconds of  $f_m$  starting from  $x$  with the fragment extracted from  $f_f$ . Figures 4 report the receiver operating characteristic (ROC) curves [7] for the case  $\varepsilon = 0.2$ , obtained by cumulatively considering all actual labels and all the scores assigned to each segment by the proposed technique. Conversely, Tables 4c report the mean area under the ROC curve (AUC) [7] (on the left) and the execution time (on the right) for each combination.

The accuracy of the proposed technique is very high. As expected, the shorter is the overlap the higher is the accuracy but, also, the higher is the execution time. In particular, the execution time is inversely proportional to the overlap since halving the overlap implies doubling the number of fragments to be analyzed. Also, note that shorter fragments performs better.

Figure 5 report the results associated with the test of robustness to the noise. In particular, for each combination of those previously depicted, we add to the input file *white noise* and *pink noise*. Tables on the right report the AUC associated with white and pink noise. The results are encouraging since they show a good robustness to the noise of our technique. We do not report execution time since they are almost the same as the previous case.

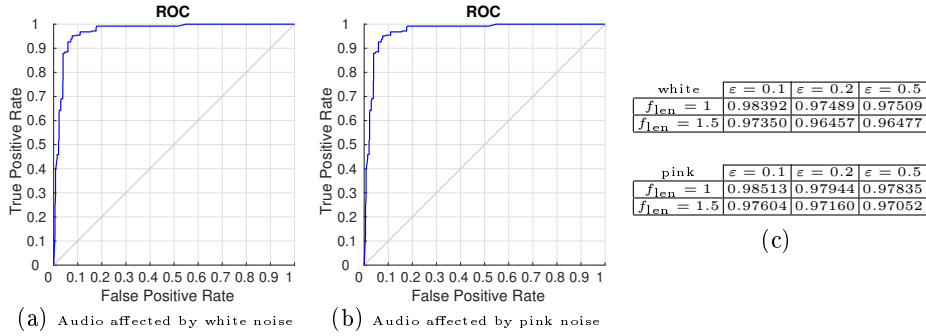


Fig. 5: Robustness of the technique

**Stuttering episode detection.** In this Section we discuss the effectiveness of the technique as far as the capability of recognizing disfluencies is concerned. In order to train the network we exploit the publicly available set of recordings of stuttering speakers, aged from 5 years to 47 years, reported in [2]. The dataset consists in 152 audio files in *wav* format with a length ranging from  $\approx 65$  sec to  $\approx 1028$  sec with a mean length of  $\approx 166$  sec. Domain experts prepare the datasets by listening and manually selecting time intervals where stuttering episodes occur. Results are accounted for in Figure 6.

## 6 Conclusions

The paper presents a technique to detect stuttering phenomena in audio files. The method employs a deep learning based classifier together with an ad-hoc segment classifier managing the output of the former classifier. The technique is able to effectively individuate stuttering episodes also having short duration and the experiments confirm that the approach is promising.

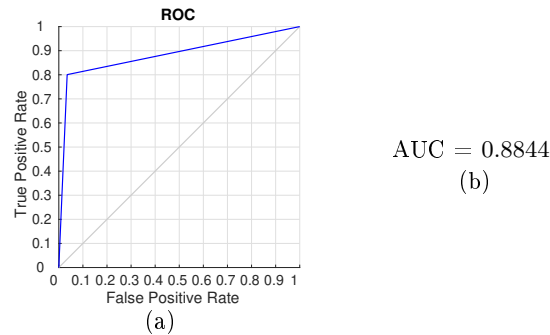


Fig. 6: Effectiveness of the technique

## References

1. Davis, S.B., Mermelstein, P.: Readings in speech recognition. chap. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, pp. 65–74. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1990), <http://dl.acm.org/citation.cfm?id=108235.108239>
2. Howell, P., Davis, S., Bartrip, J., Wormald, L.: Effectiveness of frequency shifted feedback at reducing disfluency for linguistically easy, and difficult, sections of speech (original audio recordings included). *Stammering Research* **1**(3) (2004)
3. Ingemann, F., Mermelstein, P.: Speech recognition through spectrogram matching. *Journal of The Acoustical Society of America - J ACOUST SOC AMER* **57**, 253–255 (01 1975)
4. Iverach, L., Jones, M., Lowe, R., O'Brian, S., Menzies, R.G., Packman, A., Onslow, M.: Comparison of adults who stutter with and without social anxiety disorder. *Journal of Fluency Disorders* **56**, 55 – 68 (2018)
5. Koedoot, C., Bouwmans, C., Franken, M.C., Stolk, E.: Quality of life in adults who stutter. *Journal of Communication Disorders* **44**(4), 429 – 443 (2011)
6. Markett, S., Bleek, B., Reuter, M., Prüss, H., Richardt, K., Müller, T., Yaruss, J., Montag, C.: Impaired motor inhibition in adults who stutter - evidence from speech-free stop-signal reaction time tasks. *Neuropsychologia* **91** (09 2016)
7. Sammut, C., Webb, G.I. (eds.): *Encyclopedia of Machine Learning*. Springer (2010)
8. Tichenor, S., Leslie, P., Shaiman, S., Yaruss, J.: Speaker and observer perceptions of physical tension during stuttering. *Folia Phoniatrica et Logopaedica* **69**, 180–189 (01 2017)
9. Weir, E., Bianchet, S.: Developmental dysfluency: Early intervention is key. *CMAJ: Canadian Medical Association Journal* **170**, 1790–1 (07 2004)
10. Yairi, E., Ambrose, N.: Epidemiology of stuttering: 21st century advances. *Journal of Fluency Disorders* **38**(2), 66 – 87 (2013)
11. Yairi, E., Ambrose, N.G., Paden, E.P., Throneburg, R.N.: Predictive factors of persistence and recovery: Pathways of childhood stuttering. *Journal of Communication Disorders* **29**(1), 51 – 77 (1996)