Digital Signal Processing 104 (2020) 102779

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

# Automatic detection of Voice Onset Time in voiceless plosives using gated recurrent units

T. Arias-Vergara [a,c,d,*], P. Argüello-Vélez [b,e], J.C. Vásquez-Correa [a,c], E. Nöth [c], M. Schuster [d], M.C. González-Rátiva [b], J.R. Orozco-Arroyave [a,c]

[a] Faculty of Engineering, Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia
[b] Faculty of Communications, Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia
[c] Pattern Recognition Lab, Friedrich-Alexander University, Erlangen-Nürnberg, Germany
[d] Department of Otorhinolaryngology, Head and Neck Surgery, Ludwig-Maximilians University, Munich, Germany
[e] Facultad de Salud, Universidad Santiago de Cali, Colombia

ARTICLE INFO

*Article history:*
Available online 27 May 2020

*Keywords:*
Voice Onset Time
Voiceless stop consonants
Diadochokinesis
Recurrent neural network

ABSTRACT

Voice Onset Time (VOT) has been used by researchers as an acoustic measure in order to gain some understanding about the impact of different motor speech disorders in speech production. However, VOT values are usually obtained manually, which is expensive and time consuming. In this paper we proposed a method for the automatic detection of VOT based on pre-trained Recurrent Neural Networks with Gated Recurrent Units (GRUs). Speech recordings from 50 Spanish native speakers from Colombia (25 male) are considered for the experiments. The recordings include the utterance of the diadochokinesis task /pa-ta-ka/ which is typically used for the evaluation of motor speech disorders like those caused due to Parkinson's disease. Additionally, the diadochokinesis task allows us to train a system to detect the VOT of voiceless plosive sounds in intermediate positions. Acoustic analysis is performed by extracting different temporal and spectral features from the recordings. According to the results, it is possible to detect the VOT with F1-score values of 0.66 for /p/, 0.75 for /t/, and 0.78 for /k/ when the predicted values are compared with respect to the manual VOT labels.



## 1. Introduction

Voice Onset Time (VOT) is defined as the interval between the initial burst of a stop consonant and the onset of voicing for the following vowel. VOT was initially developed as a parameter to produce synthetic sounds of voiced and voiceless stop consonants in spoken English [1]. Few years later, it was introduced as an acoustic feature to model the perception of stop consonants in the syllable-initial position [2]. In the recent years, researchers have studied how to use the VOT as an acoustic cue to understand several aspects of speech production and language development [3–6]. Usually, VOT is labeled manually using a time-frequency representation of the speech signal. However, this is an expensive and time-consuming task, which has motivated the research community to develop methods for the automatic detection of VOT. Several works have addressed this task using different machine learning and signal processing techniques. For instance, the studies presented in [7–11] consider Automatic Speech Recognition (ASR) systems for phoneme detection in speech recordings containing stop consonants in word-initial positions. In these studies, forced alignment is performed in order to extract speech segments containing the stop-to-vowel transitions, then, further signal processing algorithms are applied to measure the VOT. The main limitation of this approach is that every sound file requires a phonetic transcription to perform the measurement, which leads to a significant amount of manual work or assuming error free ASR. Other approaches are based on energy content and zero-crossing rate to detect the initial burst and vowel onsets [7,12–14]. In general, these methods are based on the assumption that the energy content of the vowel is higher compared to the stop consonants in word-initial utterances. Thus, hand-crafted thresholds, which are commonly extracted from the amplitude envelope of the signal, are used to differentiate between vowels and voiced/voiceless stop. However, these methods fail when the energy of the vowel is lower than expected, e.g., in the case of unstressed vowels. In an attempt to perform the VOT detection automatically, pre-trained classifiers based on random forest and support vector machines have been also considered [9,15]. In those works, the classifier is trained with feature vectors extracted from stop-to-vowel transitions, which are detected using either phoneme forced alignment (which requires

* Corresponding author.
*E-mail address:* tomas.ariasvergara@lmu.de (T. Arias-Vergara).

https://doi.org/10.1016/j.dsp.2020.102779
1051-2004/

phonetic transcriptions) or a customized rule (which is not suitable for generalization).

In this paper we propose a deep learning-based approach for the automatic detection of VOT considering Recurrent Neural Networks (RNN). The proposed approach only considers VOT speech segments from the voiceless plosive sounds /p/, /t/, and /k/ produced during the rapid repetition of the syllables /pa-ta-ka/. This is called diadochokinesis (DDK) task and is typically used for the evaluation of motor speech disorders like those related to Parkinson's Disease (PD). Our final goal is to train a system that can be used later to evaluate motor speech disorders produced by neurological diseases such as PD [13,14,16,17]. Particularly, the articulation capability of the patients can be analyzed by considering the changes in the duration of the VOT with respect to a group of age-matched healthy speakers. For instance, producing a stop consonant requires the precise coordination of different articulators to produce the sound. Thus, if the disease affects the motor coordination of the muscles involved in the speech production, then, these problems can be detected by means of the VOT. In this paper, VOTs are manually labeled by an expert in linguistics by considering speech recordings with the rapid repetition of /pa-ta-ka/ uttered by 50 Spanish native speakers from Colombia (healthy speakers). The manual VOTs are used as targets to train a stacked Bidirectional Recurrent Neural Network with Gated Recurrent Units (BiGRU). The input to the network consists of sequences of feature vectors formed with temporal and spectral acoustic features extracted from the speech recordings. The GRUs were proposed as a modification of the Long Short-Term Memory (LSTM) recurrent network, replacing the separate input and forget gates with a reset gate to control the input information to the network. GRUs and LSTMs have provided similar results for several tasks, including speech and language modeling [18]; however, the GRUs are faster to train and require less parameters [19], which makes these units more suitable to be used when less training data are available. Additionally, white noise is added to the speech signals in order to evaluate the robustness of the model in different acoustic surroundings. The BiGRU is tested using the speech recordings with Signal-to-Noise Ratios (SNRs) of 50 dB, 40 dB, and 30 dB. It is expected for the performance of the system to decrease as the SNR of the recordings decreases. Thus, it is important to highlight that the purpose of this experiment is just to show how the system performs when speech signals, recorded in different acoustic conditions than those considered to train the model, are used to predict the VOT. There are, however, different strategies to cope with Gaussian and non-Gaussian noises. For instance, the studies proposed in [20–22] suggest to introduce nonlinear transformations to improve the robustness of the system, resulting in a different analysis than the source-filter theory adopted in this study [23]. A more suitable approach, would be to either include a pre-processing stage or to perform data augmentation to model the speech signals under different acoustic conditions [24].

### 1.1. Related work

Automatic detection using deep learning-based methods has been considered before. In [25] a multi-class bidirectional LSTM (BiLSTM) with two recurrent layers is used to predict whether a speech frame is silence, pre-voiced, burst, or vowel. These classes are used to determine whether a stop consonant is pre-voiced (negative VOT) or not (positive VOT). For positive VOT, the authors considered speech recordings from 9 speakers who were asked to read consecutive stop consonants in word-initials, e.g., *pin bin bin pin*. Seven acoustic features based on the energy of the signal computed from the Short-Time Fourier Transform (STFT) are computed from speech frames of 5 ms extracted every 1 ms from the recordings. The RNN is trained with feature vector sequences extracted from the speech recordings of 4 speakers. Then, 15% of the training data is considered for validation. The speech data of the remaining 5 speakers is used for testing the BiLSTM. The authors reported accuracies of up to 99% when the error between the manual and predicted labels is shorter than 50 ms. The main limitation of this study is that the VOT speech frames are predicted using isolated words where the stop consonant is produced in the initial position of the utterance. Furthermore, the results reported by the authors are optimistic due to the low amount of speakers used for training and testing. Later in [26], the same authors presented a methodology to extract VOT segments from speech recordings by means of a 2-stacked BiLSTM. Two different datasets were considered for the test. One consists of isolated words uttered by 48 English speakers (24 native speakers) and the other one contains spontaneous speech recordings from 4 English speakers from United Kingdom. Automatic word segmentation was applied and only the utterances with stop consonants were used for the prediction of VOT. The authors reported accuracies of up to 97% to detect VOT speech frames. It is not clear, however, how many utterances were considered for training/testing, which stop consonants were annotated for the automatic detection of VOT, or if the stop consonants were produced in different positions within syllables. In [27], a similar approach was presented with two differences: a BiRNN is used instead of a stacked BiLSTM and an adversarial network was included during the training process in order to make the system dataset invariant. The authors reported accuracies of up to 98 % when detecting the VOT from voiceless and voiced stop consonants. Similar to their previous work, the authors only considered isolated words with stop consonants in the initial position of the utterance.

### 1.2. Contributions of this work

Automatic detection of VOT with RNNs has been addressed before; however, according to the literature revision presented in this work, still there is space for more contributions as it is explained below:

- The studies in the literature only consider isolated words to measure VOT. In our work, we consider speech recordings with the rapid repetition of /pa-ta-ka/, which is a standardized speech test commonly used to evaluate articulation problems in people with motor speech disorders such as those caused by PD. Thus, in the future we plan to use the proposed approach to evaluate speech problems in PD patients.
- The systems described in the state-of-the-art only consider utterances with stop consonants produced in the initial position. In the current study, the BiGRU is trained to measure the VOT in stop consonants produced in initial and intermediate positions between syllables and utterances of /pa-ta-ka/.
- Previous works consider binary RNNs to detect whether a speech frame is VOT or non-VOT. In our approach, a multi-class BiGRU is trained in order to detect speech frames that are non-VOT or VOT produced within the stop consonants /p/, /t/, or /k/. This opens the option to perform further analysis with reference to the place of articulation: lips for /p/, alveolar ridge for /t/, and velum for /k/.

## 2. Materials and methods

Fig. 1 shows the stages of the proposed methodology. First, manual labels of VOT values are annotated by an expert in linguistics considering speech recordings with the rapid repetition of the syllables /pa-ta-ka/ uttered by Spanish native speakers from Colombia. Then, different acoustic features are extracted by performing temporal and spectral analyses of the speech signals. Next,

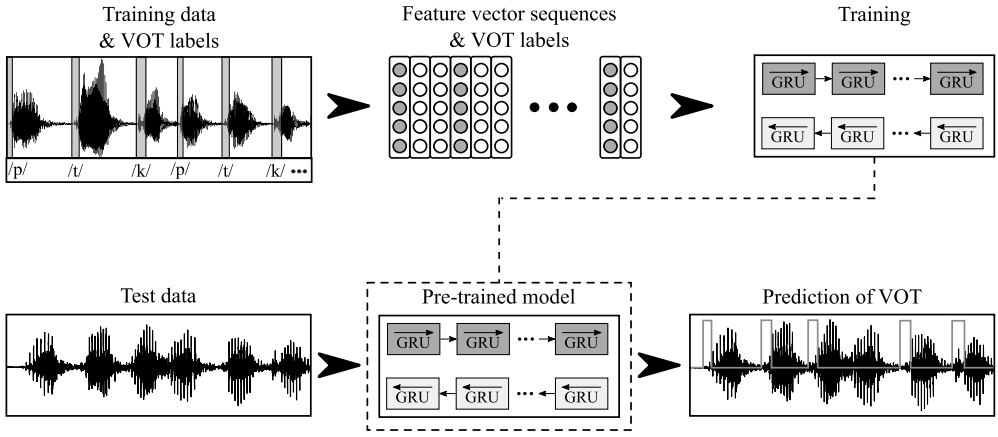


**Fig. 1.** Methodology implemented in this study. The VOTs of speech recordings with the rapid repetition of /pa-ta-ka/ are manually annotated by an expert in linguistics. The speech recordings (and their corresponding VOT labels) are divided into train and test sets. The speech signals are converted into sequences of feature vectors, which are time-aligned with their corresponding VOT labels in order to train the recurrent network. The trained model is used to predict the VOTs of the speech signals from the test set.

**Table 1**
Information about the speakers and the duration of VOT. $\mu$: Mean. $\sigma$: Standard deviation. Repetitions: Average repetitions of /pa-ta-ka/. $t_{/p/}, t_{/t/}, t_{/k/}$: Mean duration of the VOT for /p/, /t/, and /k/, respectively.

| | Male | Female |
|---|---|---|
| Number of speakers | 25 | 25 |
| Range of age [years] | 31–86 | 49–76 |
| Age [years] ($\mu \pm \sigma$) | 60 ± 11 | 61 ± 7 |
| Repetitions ($\mu \pm \sigma$) | 9 ± 3 | 9 ± 3 |
| Duration [secs] | 3.7 ± 1.1 | 4.4 ± 1.8 |
| $t_{/p/}$ [ms] ($\mu \pm \sigma$) | 13 ± 3 | 12 ± 3 |
| $t_{/t/}$ [ms] ($\mu \pm \sigma$) | 18 ± 7 | 14 ± 3 |
| $t_{/k/}$ [ms] ($\mu \pm \sigma$) | 27 ± 6 | 25 ± 6 |

sequences of feature vectors and their corresponding VOT labels (grey filled circles in Fig. 1) are used as inputs to a two-stacked BiGRU. The manual labels are used to train the BiGRU. In the test stage, the pre-trained BiGRU is used to predict the VOT values from the feature vector sequences extracted from the test data. During testing, the manual labels of the VOT are only considered to evaluate the performance of the network. The details of each stage are described in the following sections.

### 2.1. Data

The speech recordings of the 50 healthy speakers (25 male) from the PCGITA database are considered for the experiments [28]. The age of the male speakers ranges from 31 to 86 years old ($60 \pm 11$). In the case of the female, the age of the speakers ranges from 49 to 76 years old ($61 \pm 7$). The participants were asked to perform the rapid repetition of /pa-ta-ka/ for at least 3 seconds. The speech signals were captured in a sound-proof booth using a professional audio card and an omni-directional microphone. The speech signals were sampled at 16 kHz with 16-bit resolution. Table 1 summarizes demographic information of the speakers and includes details of the speech recordings considered in this paper.

Fig. 2 shows VOTs estimated with the manual labels. It can be observed that lower values are obtained for /p/ ($12 \pm 3$ ms), followed by /t/ ($16 \pm 6$ ms), and /k/ ($26 \pm 6$ ms). The Kruskal-Wallis test was applied and significant differences have been found between VOT values ($p < 0.001$). These differences can be explained considering that the place of articulation is one of the factors that affect the VOT, i.e., the values are longer as the place of articulation moves from the front to the back: short values for the bilabial stop (/p/), intermediate values for the alveolar stop (/t/), and long values for the velar stop (/k/) [29].

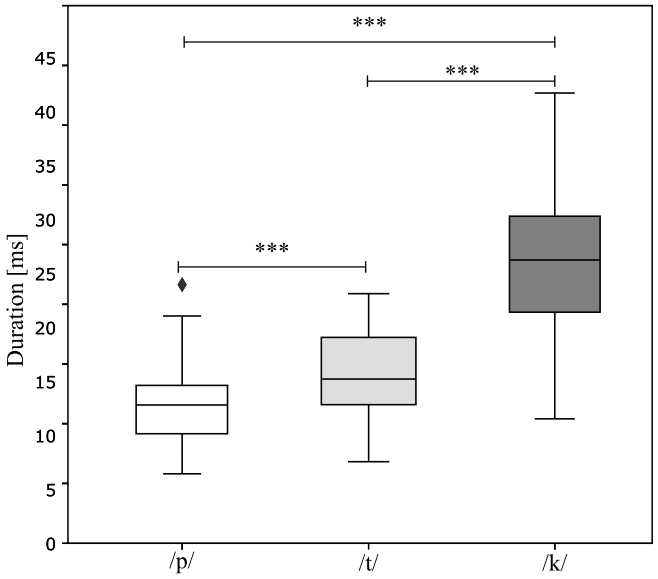


**Fig. 2.** VOT values of /p/ ($12 \pm 3$ ms), /t/ ($16 \pm 6$ ms), and /k/ ($26 \pm 6$ ms). Kruskal-Wallis $^{***}p < 0.001$. The diamond in the figure indicates outliers.

### 2.2. Manual labeling

The labeling procedure of the VOT was performed by an expert in linguistics. Manual labels are placed at the initial burst of the consonants and vowel onsets using the software Praat [30]. The complete procedure is as follows:

1. VOT in the absolute initial stop:
   Refers to the first /p/ in a repetition of /pa-ta-ka/.
   - The zero crossing points are computed in order to set the starting point of the initial burst visible in the time signal and the spectrogram (Fig. 3).
   - The vowel onset is set at the beginning of a periodic-like signal. Thus, the formant frequencies and the presence of pitch are used to mark the beginning of voicing in a stop-vowel transition. Fig. 4.A shows the position of the vowel onset in the time domain. The beginning of voicing

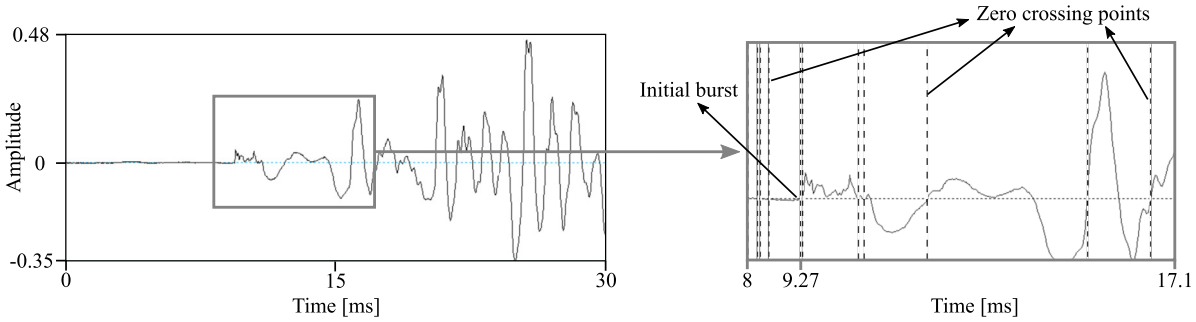


**Fig. 3.** Initial burst label for the stop consonant /p/. The vertical lines represent the zero crossing points. Only the positive zero crossing points are displayed.

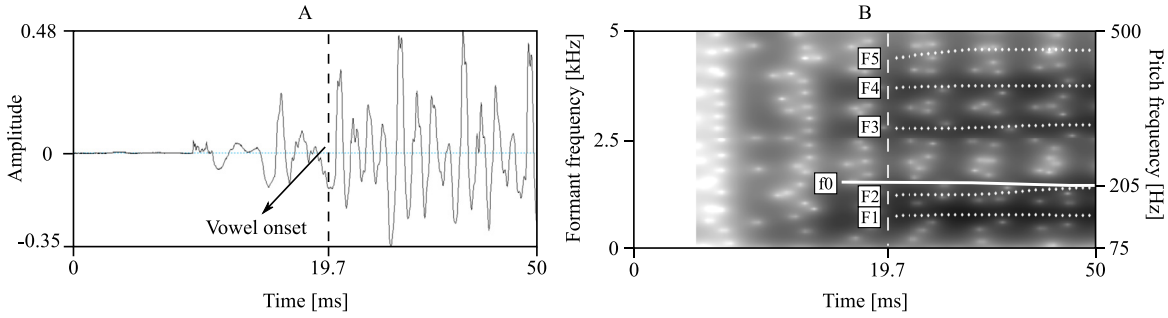


**Fig. 4.** Vowel onset label for the stop-vowel transition /pa/. The vertical dotted lines in Figs. 4.A and 4.B represent the vowel onset in time and frequency domain, respectively. The labels at the right and left of the y-axis in Fig. 4.B are the formant frequencies (F1, F2, F3, F4, and F5) and pitch (f0) frequency values, respectively.

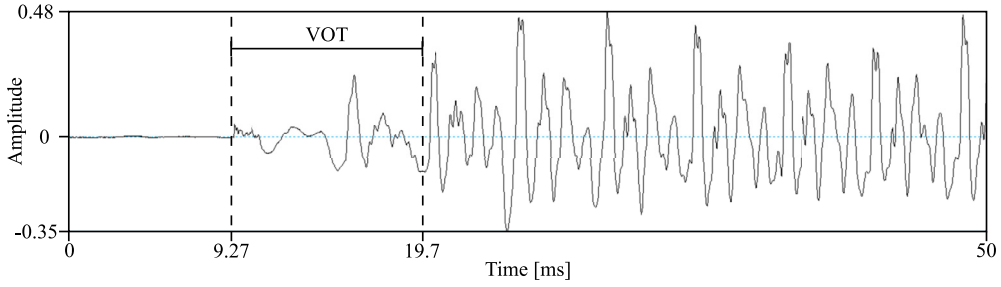


**Fig. 5.** Voice onset time at the initial stop consonant /p/.

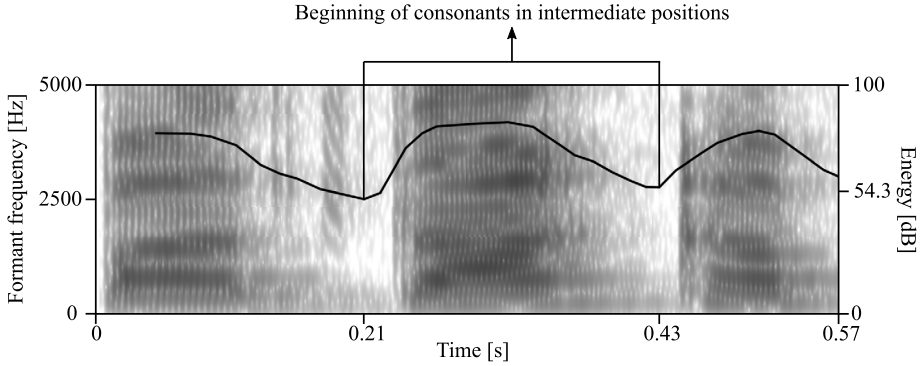


**Fig. 6.** Spectrogram of the first /pa-ta-ka/ uttered by one of the speakers considered in this study. The minimum energy value between vowels is used to identify the beginning of the consonants /t/ and /k/ at intermediate positions. The labels at the right and left of the y-axis are the frequency and the energy values, respectively.

is marked by looking at the formants (F1, F2, F3, F4, and F5) and pitch (f0) in the spectrum (Fig. 4.B).

- The VOT is labeled by the expert considering the time of the initial burst and the vowel onset. Fig. 5 shows the time stamps of the VOT at the absolute initial stop for one of the utterances.

2. VOT in intermediate positions:

- The beginning of the stop consonant (/p/, /t/, or /k/) is located by searching for the point of the lowest energy value between vowels (Fig. 6).

3. After locating the beginning of the consonant the zero crossing points, the frequency formants, and the presence of pitch are considered to measure the VOT. Fig. 7 shows the VOT for /t/ and /k/ in intermediate positions.

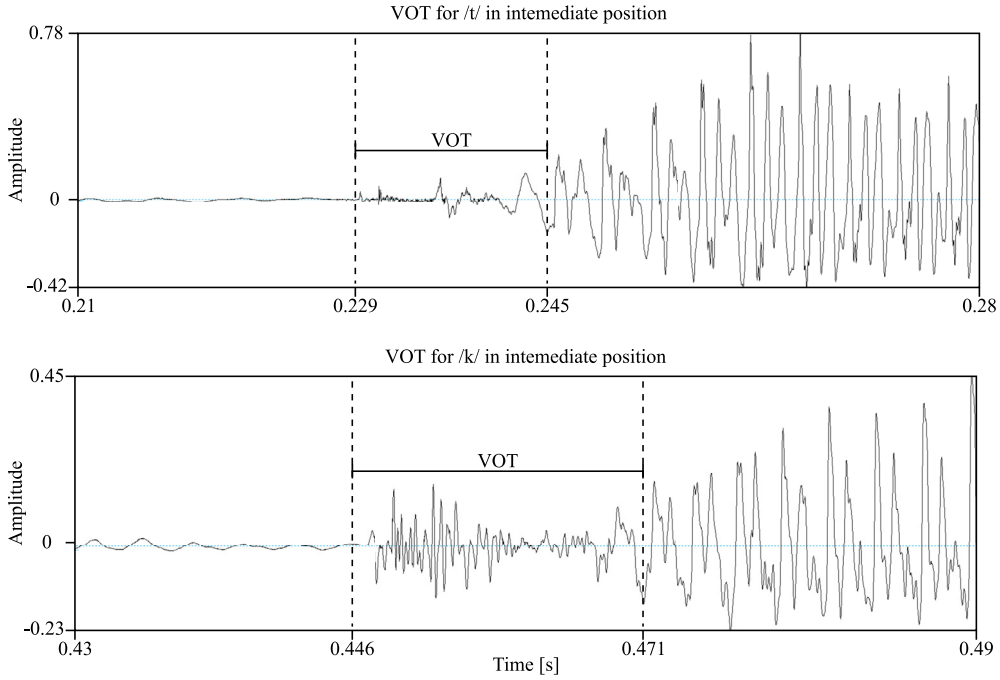


**Fig. 7.** VOT labels for the consonants /t/ (top) and /k/ (bottom) in one of the recordings.

### 2.3. Distinctive features of voiceless stops

In the Spanish language, voiceless consonants are characterized by three stages: closure, release, and aspiration. In the closure stage, an obstruction of airflow is created by the articulators resulting in a silence region in the speech signal. Then, the articulators move away from each other during the release stage producing an explosive burst of air with energy spread across the audible spectrum. After the burst, the air pressure (generated by obstruction of the articulators) is decreased, which results in turbulent airflow with energy values no longer spread across the spectrum. Fig. 8 shows the three stages involved in the production of /p/, /t/, and /k/. Visually, the voiceless plosives can be differentiated from each other by looking at the burst during the release stage. In the case of the /p/ sound, there is a single burst "bar" (Figs. 8.A.1 and 8.A.2), two "bars" for the /t/ (Figs. 8.B.1 and 8.B.2), and a longer burst "bar" for the /k/ (Figs. 8.C.1 and 8.C.2) with respect to the /p/. In other cases, however, the production of the voiceless stops is affected by different acoustic factors. Most of these phenomena occur at intermediate positions in an utterance, i.e., in the vowel-consonant-vowel context. In the case of our data, the following acoustic factors were found:

1. Voicing: Is characterized by the presence of glottal pulses during the closure and release stages, i.e., the glottal pulses are present during the silence region of the plosive and during the VOT. Fig. 9 shows an example of voicing in a speech segment with the transition from /ka/ to /pa/. Even though the voicing is present during the production of the /p/ sound, the burst can be visualized in the spectrogram (bottom picture), thus the VOT can be measured.
2. Partial voicing: It can be identified by the presence of glottal pulses during the closure stage of the plosive sound but not in release and aspiration stages, i.e., the glottal pulses are present in the silence region but not in the VOT. Fig. 10 shows an example of a partially-voiced /t/ sound. The glottal pulses are present during the closure stage, but not during the VOT.
3. Consonant weakening: This phenomenon occurs mainly in /p/ sounds. Is characterized by the absence of the burst. As a result, the /p/ sound is weaker and perceived as a /b/ sound. Fig. 11 shows an example of the weakening effect in a /p/ sound.

## 3. Feature extraction

Time and spectral acoustic features are extracted from the speech signals in order to capture similar information considered by the annotator to manually label the VOTs. Each speech signal is divided into frames of 40 ms (Section 3.1) taken every 1 ms, resulting in a (time) sequence $\vec{S} = \{\vec{s}_1, \vec{s}_2, ..., \vec{s}_T\}$. Then, 42 acoustic features (18 temporal, 24 spectral) are extracted by iterating $\vec{s}_t$, with $t = \{1, 2, \ldots, T\}$. The sequence of speech frames $\vec{S}$ is then converted into a sequence of feature vectors $\vec{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$, where $T$ is the number of frames extracted from the speech recording.

### 3.1. Temporal analysis

The temporal information of the speech signals is modeled by extracting the pitch, zero-crossing rate (ZCR), and four descriptors calculated from Intrinsic Mode Functions (IMFs): the Root Mean Square (RMS), the mean, the standard deviation, and the maximum amplitude. The IMFs are obtained by means of an iterative algorithm called Empirical Mode Decomposition (EMD). The aim of the EMD algorithm is to decompose a signal into a finite number of amplitude and frequency modulated time series that satisfy the following two conditions:

- The number of extrema and the number of zero-crossings are equal or differ at most by one.
- At any point, the mean defined by the upper and lower envelopes is zero.

The number of IMF components $N$ to be considered was determined by selecting the number of IMFs containing at least 80% of the total energy from all components [31]. The energy of each component is computed as

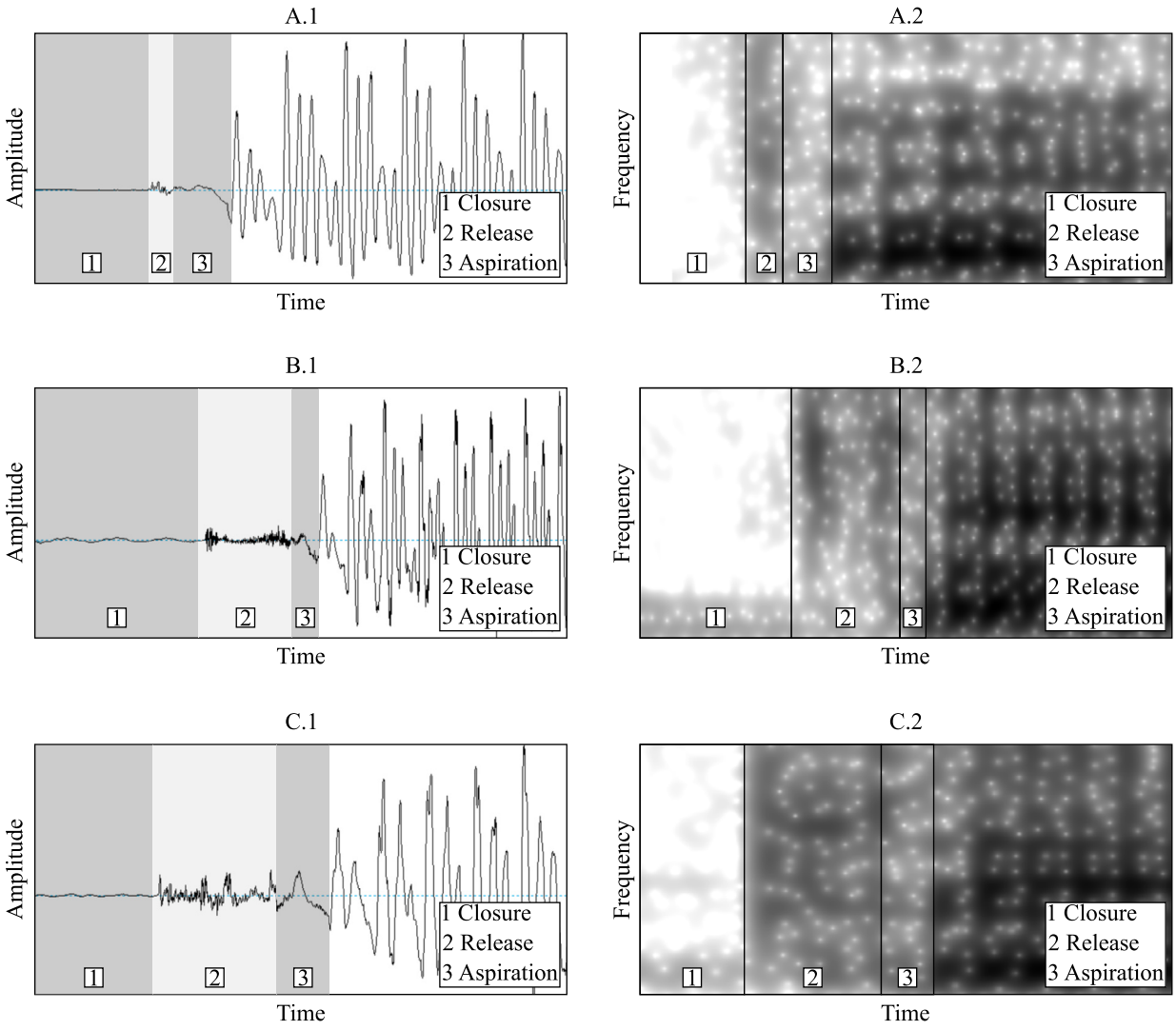


**Fig. 8.** Time and spectral representations of the voiceless plosive sounds /p/ (Figures A.1 and A.2), /t/ (Figures B.1 and B.2), and /k/ (Figures C.1 and C.2). In the Figures, the shaded regions represent the closure, release, and aspiration stages.

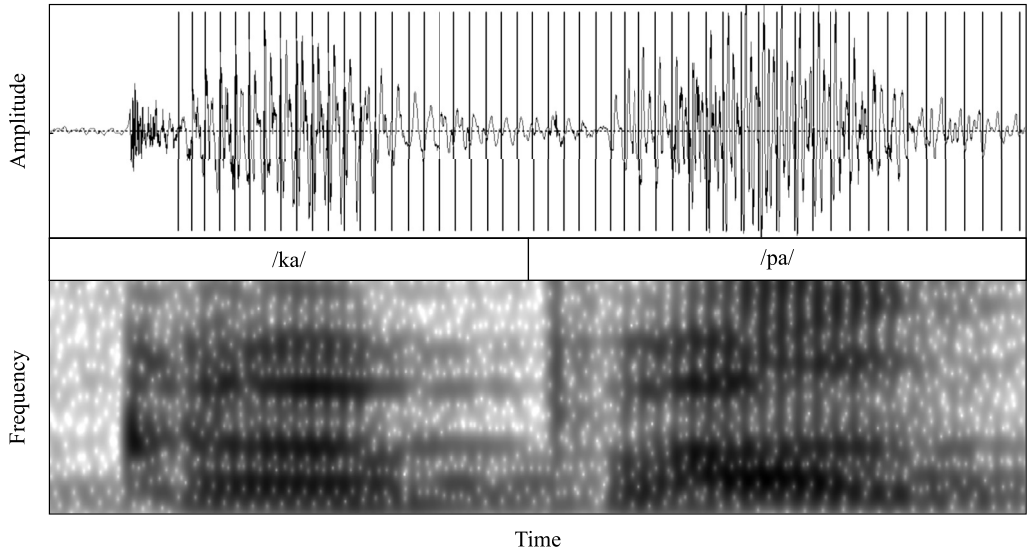


**Fig. 9.** Voicing effect present in the speech transition from /ka/ to /pa/. The vertical lines in the time domain signal (top) are the glottal pulses extracted with Praat.

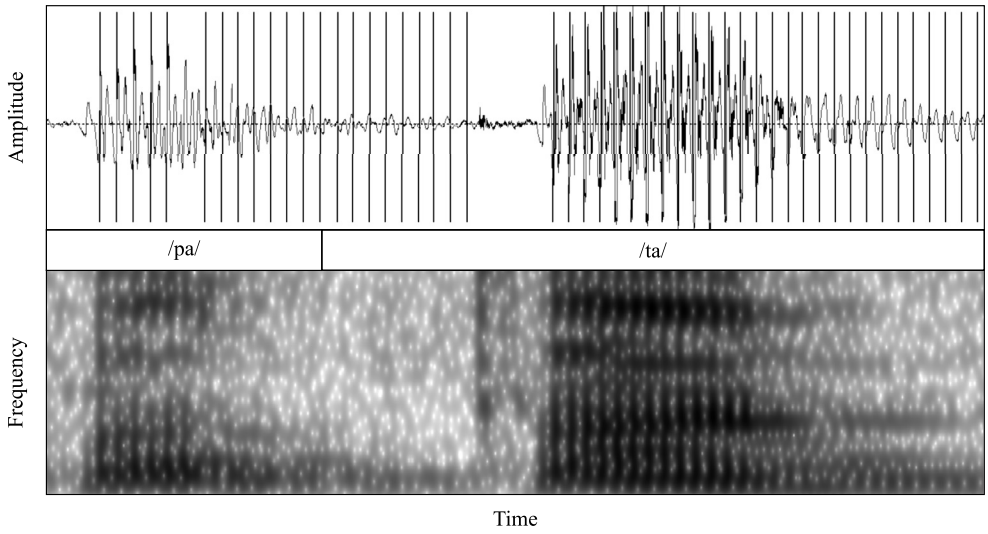


**Fig. 10.** Partially voiced /t/ sound extracted from a transition from /pa/ to /ta/. The vertical lines in the time domain signal (top) are the glottal pulses extracted with Praat.

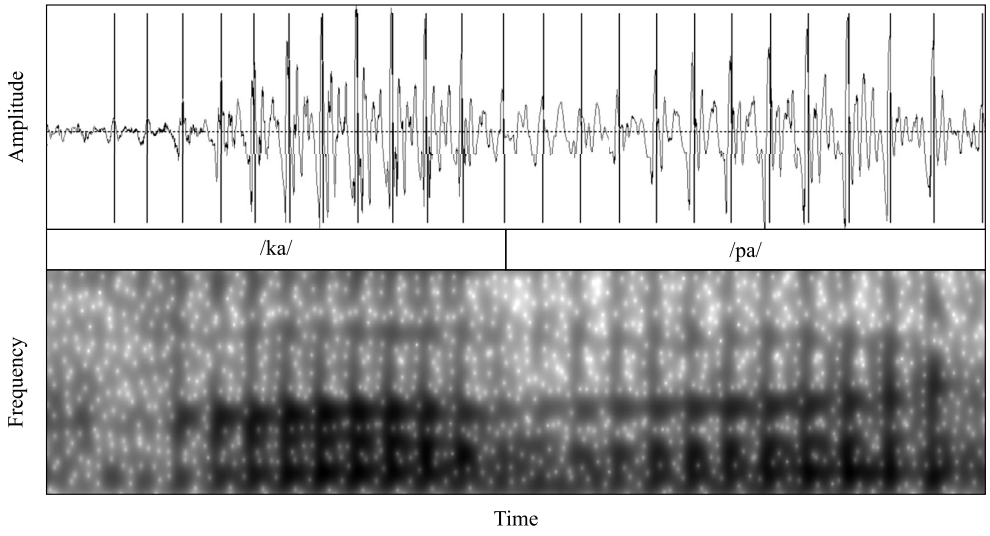


**Fig. 11.** Consonant weakening of the /p/ sound in an intermediate position. The vertical lines in the time domain signal (top) are the glottal pulses extracted with Praat. Note the absence of the burst in /p/.

**Table 2**
Relative energy of the IMF components extracted from the recordings. $E_{VOT}$: Relative energy of the speech segments manually labeled as VOT. $E_{RES}$: Relative energy of the non-VOT speech segments.

| Relative Energy | IMF Components | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $E_{VOT}$(%) | 6.15 | 21.98 | 31.23 | 24.34 | 11.65 | 4.15 | 0.49 | 0.01 | 0.00 | 0.00 |
| $E_{RES}$(%) | 6.32 | 22.14 | 29.64 | 23.17 | 13.52 | 4.62 | 0.57 | 0.02 | 0.00 | 0.00 |

$$E_n = \frac{1}{L}\sum_{l=0}^{L}(c_n)^2 \tag{1}$$

where $c_n$ is the $n$th IMF component extracted from each $\vec{s}_t$, $L$ is the number of samples in $\vec{s}_t$. The relative energy of each component is computed as:

$$E_n(\%) = \frac{E_n}{E} \tag{2}$$

where $E$ is the total energy of the IMFs. Table 2 shows the relative energy of all IMF components extracted from the speech recordings of the speakers described in Section 2.1. The relative energy is computed for the IMFs of the speech segments manually labeled as VOT ($E_{VOT}$) and the non-VOT signals ($E_{RES}$), individually. The results indicate that both $E_{VOT}$ and $E_{RES}$ have more than 80% of the energy is contained in the IMF components 1 to 4 thus $N = 4$. In this study, the IMF decomposition is performed in Python using the PyHHT module.[1] In the case of the pitch, the periodicity detector algorithm implemented in the software Praat is considered for extraction [32]. According to Praat's documentation, the analysis window should be long enough to contain at least three periods of the minimum pitch. By default Praat considers that the minimum pitch value to be detected is 75 Hz which leads to an analysis win-

[1] https://pyhht.readthedocs.io/en/latest/index.html.

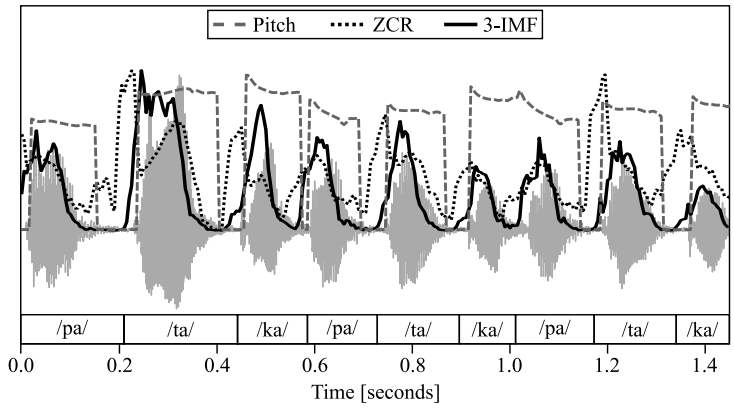


**Fig. 12.** Example of pitch (dashed gray lines), ZCR (dotted black lines), and 3-IMF values (straight black lines) extracted from three utterances of /pa-ta-ka/. The values of each signal (y-axis) have been re-scaled in order to have comparable curves in the same picture.

dow of 40 ms. In the case of ZCR, the number of zero crossings in each speech frame is calculated. Fig. 12 shows the pitch contour, ZCR, and RMS values of the third IMF component (3-IMF) extracted from three utterances of /pa-ta-ka/ from one of the recordings. The amplitude of the signal, pitch (dashed gray lines), ZCR (dotted black lines), and 3-IMF values (straight black lines) were re-scaled with respect to their respective maximum value in order to depict comparable curves in the same picture. In general, the voiced segments can be detected using the pitch information, the ZCR can be used to detect the stop consonants, and the RMS values of 3-IMF are suitable to detect the start and end of a syllable, i.e., /pa/, /ta/, or /ka/. However, in some cases these features may fail to correctly differentiate between voiceless and voiced sounds, which affects the automatic the detection of VOT. In Fig. 12, this situation can be observed in the transition between the second /ka/ to the third /pa/. In this case, pitch values are present during the utterance of /p/ due to the voicing effect presented in this consonant (Section 2.3).

### 3.2. Spectral analysis

The set of spectral features includes the first and second formant frequencies ($F_1$ and $F_2$), the ratio between formant frequencies ($F_1/F_2$), 13 Mel-Frequency Cepstral Coefficients (MFCCs), and two Hilbert spectrum-based features: the RMS value and centroid of the marginal spectral energy of each IMF component. The Hilbert spectrum $H(\omega, t)$ is obtained with the Hilbert-Huang Transform (HHT) by applying the Hilbert transform to each component [33]. Then, the Hilbert spectrum of each $n$th IMF is defined as

$$H_n(\omega, t) = \begin{cases} a_n(t) & \omega = \omega_n(t) \\ 0 & otherwise \end{cases} \tag{3}$$

where $a_n(t)$ is the amplitude envelope and $\omega_n(t)$ is the instantaneous frequency (in radians). The resolution of the Hilbert spectrum is then defined by equal-sized frequency bins [34]. Once we get $H_n(\omega, t)$, then marginal spectral energy is defined as

$$G(\omega) = \int_0^{T_s} \mathrm{H}(\omega, t)\, \mathrm{d}t \tag{4}$$

where $T_s$ is the duration of the signal. The RMS value of the marginal spectral energy of each IMF is defined as:

$$\mathrm{SE}_{\mathrm{RMS}} = \sqrt{\frac{1}{N_b}\sum_{i=1}^{N_b} G_n^2(i)} \tag{5}$$

where $N_b$ is the number of frequency bins in the Hilbert spectrum and $G_n(i)$ is the spectral energy of the $n$th IMF component at the frequency bin $i$. The spectral centroid is defined as

$$\mathrm{SE}_{\mathrm{CEN}} = \frac{\sum_{i=1}^{N_b} f(i) G_n(i)}{\sum_{i=1}^{N_b} G_n(i)} \tag{6}$$

where $f(i)$ is the value of the frequency at bin $i$. For the case of the formant frequencies, $F_1$ and $F_2$ are obtained by computing the Linear Prediction Coefficients using the Burg's algorithm, which is implemented in Praat. The MFCCs are simply obtained by applying a triangular filter bank (on the Mel scale) to each speech frame $\vec{s}_t$. Then, the discrete cosine transform is calculated upon the logarithm of the energy bands.

## 4. Model architecture

Fig. 13 shows the configuration of the recurrent network used in this study. As described in Section 3, each speech recording is converted into a sequence of feature vectors $\vec{X} = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_T\}$ formed with temporal and spectral acoustic features. The input sequence is then processed by two bidirectional recurrent layers (BiGRU-1 and BiGRU-2) with shared weights on each time frame $t$. Bidirectional recurrent layers are used in this study because of their ability to consider as much contextual information as needed from the sequence, i.e., for every input data $\vec{x}_t$ in the sequence, the network has sequential information about the data points before $(\vec{x}_1, \ldots, \vec{x}_{t-2}, \vec{x}_{t-1})$ and after $(\vec{x}_{t+1}, \vec{x}_{t+2}, \ldots, \vec{x}_T)$ [35]. As shown in Fig. 13, the sequence of feature vectors $\vec{X}$ is fed to the first recurrent layer (BiGRU-1), which computes the forward ($\overrightarrow{h}^1$) and backward ($\overleftarrow{h}^1$) hidden sequences. The super-index "1" denotes operations performed in the first recurrent layer. The sequence $\overrightarrow{h}^1$ is computed by iterating Equation (7) from $t = 1$ to $t = T$. In the case of $\overleftarrow{h}^1$, the hidden states are computed by iterating Equation (8) from $t = T$ to $t = 1$ [36]. The size of the hidden states is 512, which was chosen experimentally in a grid-search process.

$$\overrightarrow{h}^1_t = \tanh(\vec{W}_{\vec{X}\overrightarrow{h}^1}\vec{x}_t + \vec{W}_{\overrightarrow{h}^1\overrightarrow{h}^1}\overrightarrow{h}^1_{t-1} + b^1_{\overrightarrow{h}^1}) \tag{7}$$

$$\overleftarrow{h}^1_t = \tanh(\vec{W}_{\vec{X}\overleftarrow{h}^1}\vec{x}_t + \vec{W}_{\overleftarrow{h}^1\overleftarrow{h}^1}\overleftarrow{h}^1_{t+1} + b^1_{\overleftarrow{h}^1}) \tag{8}$$

where $\vec{W}$ and $b$ are the weight matrices and bias, respectively. The output of the first recurrent layer is the sequence $\vec{r}^1$ formed with the concatenation of $\overrightarrow{h}^1$ and $\overleftarrow{h}^1$. For each time frame, the output of the BiGRU-1 is defined as

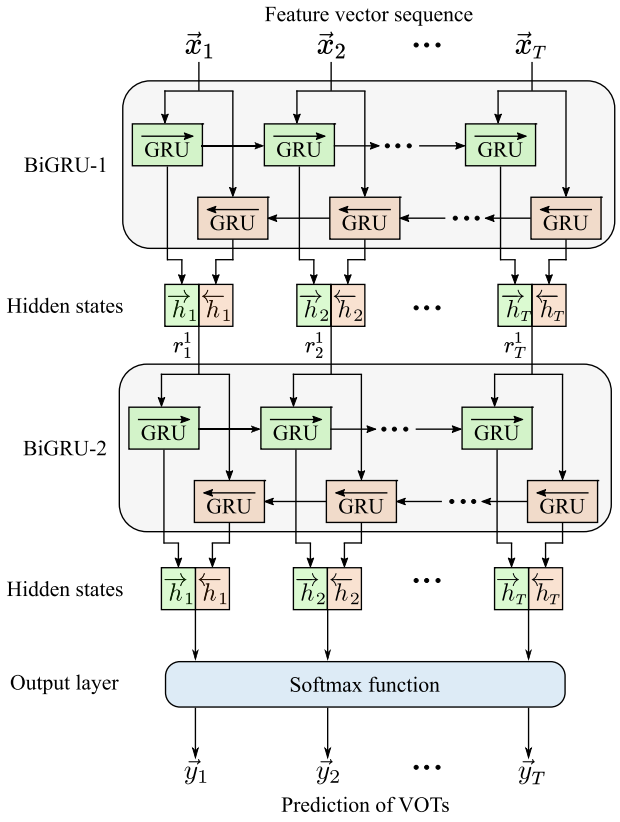


**Fig. 13.** Architecture of the network considered in this work. The sequence of feature vectors $\vec{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T\}$ is fed to the first recurrent layer (BiGRU-1), whose outputs are processed by the second recurrent layer (BiGRU-2). The resulting sequence of hidden states is processed by a softmax activation layer to predict the VOT labels for each time frame.

$$r_t^1 = (\vec{W}_{\overrightarrow{h}^1 r^1} \overrightarrow{h}_t^1 \oplus \vec{W}_{\overleftarrow{h}^1 r^1} \overleftarrow{h}_t^1) + b_r \tag{9}$$

where the symbol $\oplus$ denotes concatenation. The second recurrent layer is fed with the sequence $\vec{r}^1$, thus, the hidden states in the BiGRU-2 are computed as

$$\overrightarrow{h}_t^2 = \tanh(\vec{W}_{\vec{r}^1 \overrightarrow{h}^2} \vec{r}_t^1 + \vec{W}_{\vec{r}^1 \overrightarrow{h}^2} \overrightarrow{h}_{t-1}^2 + b_{\overrightarrow{h}^2}^2) \tag{10}$$

$$\overleftarrow{h}_t^2 = \tanh(\vec{W}_{\vec{r}^1 \overleftarrow{h}^2} \vec{r}_t^1 + \vec{W}_{\vec{h}^1 \overleftarrow{h}^2} \overleftarrow{h}_{t-1}^2 + b_{\overleftarrow{h}^2}^2) \tag{11}$$

The hidden states $\overrightarrow{h}^2$ and $\overleftarrow{h}^2$ are concatenated and the resulting sequence is processed by a softmax activation function to predict the VOT labels, thus, the output of the network at each time frame is defined as

$$y_t = (\vec{W}_{\overrightarrow{h}^2 y} \overrightarrow{h}_t^2 \oplus \vec{W}_{\overleftarrow{h}^2 y} \overleftarrow{h}_t^2) + b_y \tag{12}$$

The output of the network is a time sequence of VOT predictions $y = \{y_1, y_2, ..., y_T\}$.

### 4.1. Training of the stacked BiGRU

The network architecture considered in this study is implemented using PyTorch's deep learning modules [37]. In order to train the model, feature vector sequences $\vec{X}_j^i = \{\vec{x}_{1j}^i, \vec{x}_{2j}^i, ..., \vec{x}_{Tj}^i\}$ are computed considering speech segments of 500 ms extracted from every recording, where $j$ is the $j$th speech recording in the database, $i$ is the $i$th feature vector sequence extracted from $j$, and $T$ is the number of feature vector extracted from $j$. In this study, $T = 500$ since the acoustic features are extracted every 1 ms (Section 3). For every time frame $t$ in the sequence $\vec{X}_j^i$, there is a corresponding time sequence of targets $L_j^i = \{l_1, l_2, \ldots, l_T\}$, with $l_t = \{1, 2, 3, 4\}$. Each target label represents one of the following classes: (1) non-VOT frames, e.g., silence, vowels, (2) VOT frames of the /p/, (3) VOT frames of the /t/, and (4) VOT frames of the /k/. The Adam optimization algorithm with a learning rate of $\eta = 10^{-4}$ is considered for training [38]. The cross–entropy between the target labels $L_j^i = \{l_1, l_2, \ldots, l_T\}$ and the predictions of the network is used as the loss function. Additionally, class weights are computed for the targets $l = \{2, 3, 4\}$ (VOT frames from /p/,/t/,/k/ sounds, respectively) since those are the classes with the lowest count of samples (feature vectors) compared with respect to the class $l = 1$ (non-VOT frames). Thus, the loss function is described as[2]:

$$loss(p, l) = w[l](-p[l] + \log(\sum_k \exp(p[k]))) \tag{13}$$

where $p$ are the posterior probabilities of the sequences obtained from the output layer $y = \{y_1, y_2, \ldots, y_T\}$, $l$ are the target labels, and $w$ are the class weights. The optimization algorithm, learning rate, and the loss function were chosen based on a previous study where a set of parallel BiGRUs are trained to learn the representation of several phoneme classes grouped according to different phonological rules [39].

The performance of the multi-class BiGRU is measured at frame level by means of precision, recall, and F1-score. Precision measures the proportion of *predicted* VOT and non-VOT speech frames that are correctly classified. Recall measures the proportion of *actual* VOT and non-VOT speech frames that are correctly classified. The F1-score measures the performance of the BiGRU to classify all speech frames, which reaches its best value at 1 and worst score at 0. These three measures are computed as in [40].

### 4.2. Post-processing of predictions

As described in Section 4.1, the stacked BiGRU processes the speech signals in segments of 500 ms, which may produce discontinuities in the signal and causing errors in the predicted VOT. One solution is to apply a median filter mask of 5 ms to the sequence of predicted VOTs in order to interpolate the missing values. An example illustrating this situation is shown in Figs. 14.A and 14.B.

## 5. Experiments and results

The stacked BiGRU is trained and tested following a 5-fold cross-validation strategy. The average size of the training and test sets are 1434 and 358 feature vector sequences, respectively. Speaker independence is guaranteed during the training stage. Additionally, the network is trained considering three feature sets: temporal features, spectral features, and the combination of both. Table 3 shows the obtained results for the automatic detection of VOT feature vectors. The lowest performance is achieved when the BiGRU is trained only with the temporal features (Total F1-score = 0.65). This can be explained considering that temporal features are more sensible to acoustic phenomena such as voicing, as showed in Section 3.1. The highest performance is achieved when spectral and temporal features are combined (Total F1-score = 0.78), particularly, the precision of the system to detect VOT feature vectors increases for /p/, /t/, and /k/. Considering that the recall values are relatively high in all cases, an increase in the precision indicates that including the temporal features improves the performance of the system to correctly predict the time stamps of the initial burst and vowel onset. Regarding the detection of VOT segments, the lowest performance is obtained for /p/, then /t/ and then /k/. This

[2] https://pytorch.org/docs/stable/nn.html#crossentropyloss.

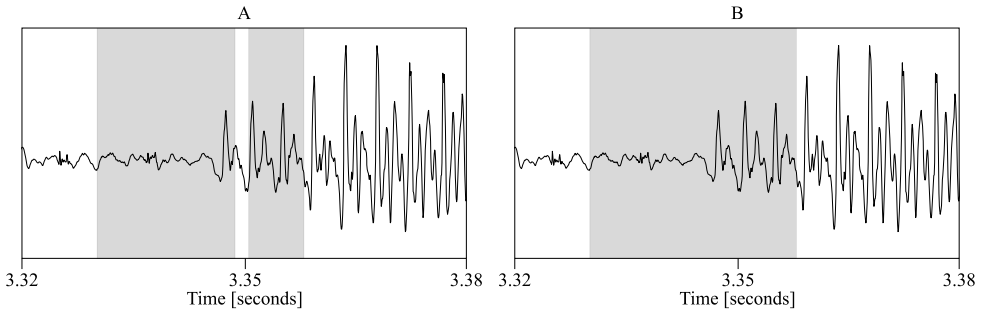


**Fig. 14.** Prediction of a VOT segment (shaded regions) before (Figure A) and after post-processing (Figure B). A median filter mask of 5 ms is applied to the sequence of predicted VOTs in order to interpolate the missing values.

**Table 3**
Performance of the stacked BiGRU for the detection of the VOT speech frames from /p/, /t/, and /k/. **Non-VOT:** Non-VOT speech frames. **Fusion:** Combination of temporal and spectral features. **Prec:** Precision. **Rec:** Recall. **F1:** F1 score.

| Feature | /p/ | | | /t/ | | | /k/ | | | Non-VOT | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Prec | Rec | F1 | Prec | Rec | F1 | Prec | Rec | F1 | Prec | Rec | F1 | F1-score |
| Temporal | 0.35 | 0.62 | 0.45 | 0.51 | 0.72 | 0.60 | 0.52 | 0.80 | 0.63 | 0.98 | 0.91 | 0.94 | 0.65 |
| Spectral | 0.47 | 0.88 | 0.61 | 0.57 | 0.90 | 0.70 | 0.62 | 0.92 | 0.74 | 0.99 | 0.91 | 0.95 | 0.75 |
| Fusion | 0.53 | 0.86 | 0.66 | 0.64 | 0.91 | 0.75 | 0.67 | 0.92 | 0.78 | 0.99 | 0.93 | 0.96 | 0.78 |

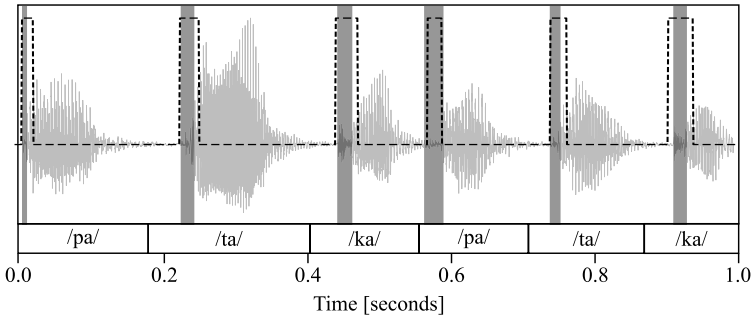


**Fig. 15.** Predicted VOT values and manual labels for the first two utterances of /pa-ta-ka/ from one of the speakers. The shaded regions represent the VOT measured by the expert and the dotted black lines are the predicted segments.

**Table 4**
Prediction error of time stamps of the initial burst and vowel onset for the VOT of /p/, /t/, and /k/. $\mathbf{E_{Burst}}$ : Time error between the predicted time and the manual label of initial burst. $\mathbf{E_{Vowel}}$ : Time error between the predicted time and the manual label of vowel onset. $\mu$: Mean. $\sigma$: Standard deviation.

| | /p/ | /t/ | /k/ | AVG |
|---|---|---|---|---|
| $E_{Burst}$ [ms] ($\mu \pm \sigma$) | $5.2 \pm 2.4$ | $4.1 \pm 1.8$ | $5.6 \pm 3.6$ | $5.0 \pm 2.8$ |
| $E_{Vowel}$ [ms] ($\mu \pm \sigma$) | $4.1 \pm 1.5$ | $4.6 \pm 1.6$ | $5.3 \pm 2.9$ | $4.7 \pm 2.2$ |

can be explained considering the acoustic phenomena described in Section 2.3, particularly, the voicing and consonant weakening phenomena that commonly occurs in the /p/ sounds produced in intermediate positions, e.g., in the transition from one utterance of /pa-ta-ka/ to another. Fig. 15 shows an example of the predicted VOT segments extracted from one of the recordings. The shaded regions represent the VOT segment manually labeled by the expert and the dotted black lines are the predictions of the system, after post-processing the posterior probabilities. Most of the mistakes of the network are made in the prediction of the time for the initial burst and the vowel onset, however, the system is able to correctly identify the occurrence of the VOT, which explains the performance obtained for the precision and recall reported in Table 3. The errors between the real and predicted time stamps of the initial burst ($E_{Burst}$) and the vowel onset ($E_{Vowel}$) are calculated and reported in Table 4. These results can be explained considering that each speech frame is 40 ms long, resulting in a poor time resolution to compute the spectral features, thus, future work should include narrowband and wideband spectral analysis in order to improve the precision in the prediction of VOT speech frames. Regarding the duration of the predicted VOT segments, Fig. 16 shows

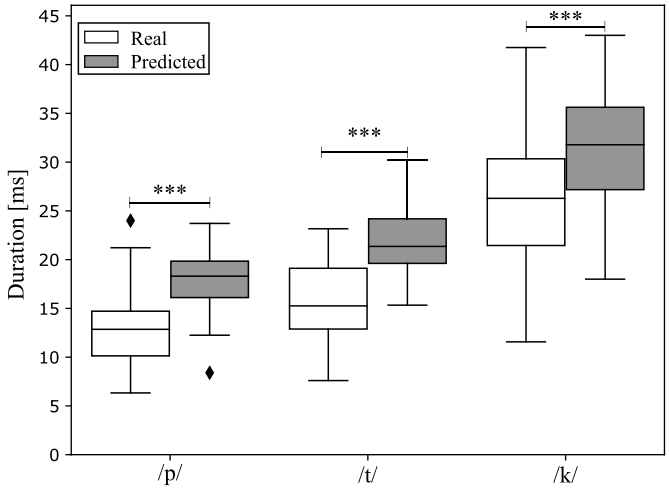


**Fig. 16.** The white boxes represent the real VOT values of /p/ ($12 \pm 3$ ms), /t/ ($16 \pm 6$ ms), and /k/ ($26 \pm 6$ ms). The gray boxes represent the predicted VOT values of /p/ ($18 \pm 3$ ms), /t/ ($22 \pm 4$ ms), and /k/ ($31 \pm 5$ ms). Kruskal-Wallis p-value: $^{***}p < 0.001$. The diamonds in the figure indicate outliers.

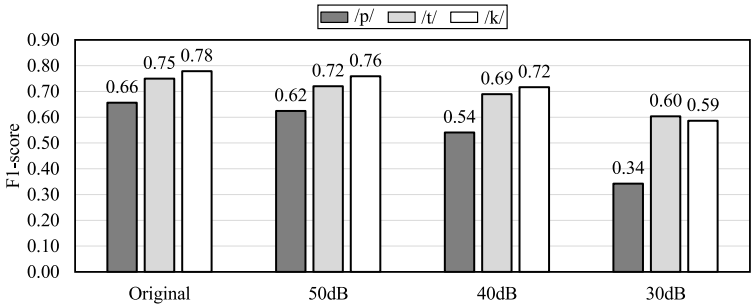


**Fig. 17.** F1-scores for the automatic detection of VOT speech frames from the consonants /p/ (dark grey bar), /t/ (light grey bar), and /k/ (white bar). The BiGRU is trained with the combination of temporal and spectral features extracted from the clean signals (original). The model is tested using clean signals (original) and noisy signals with SNRs of 50 dB, 40 dB, and 30 dB.

the box plots for the real and predicted VOT segments. Kruskal-Wallis test was applied and significant differences have been found between real and predicted VOT values ($p < 0.001$), which can be explained considering that there is a difference of approximately 5 ms between the predicted time stamps of the initial burst and the vowel onset (Table 4).

### 5.1. Experiments with white noise

The speech recordings considered in this study were captured in a soundproof booth. In order to simulate different acoustic settings, white noise is added to the speech signals. For the experiments, the BiGRU is trained with the original speech recordings (clean signals) and tested with clean and noisy signals. Furthermore, the model considered for testing is trained with the temporal and spectral features. Only the F1-score is reported in order to evaluate the general performance of the model when it is tested in different acoustic conditions. Fig. 17 shows the F1-scores for the automatic detection of VOT speech frames from the consonants /p/ (dark grey bar), /t/ (light grey bar), and /k/ (white bar). The BiGRU is tested with clean signals (original) and noisy signals with SNRs of 50 dB, 40 dB, and 30 dB. As expected, the performance of the BiGRU decreases as the noise level of the signal increases (lower SNR). Nevertheless, relatively good results are obtained when the model is tested with speech recordings with SNRs of 50 dB and 40 dB. These SNRs correspond to more realistic scenarios, for instance, when the speech recordings are captured in a quiet room and using a standard microphone. The performance of the model decreases considerably when the model is tested with signals upon 30 dB SNRs. Thus, future work should include training strategies to improve the robustness of the model again different acoustic conditions.

## 6. Discussion

Automatic detection of VOT speech frames was possible with F1-scores of up to 0.66 for /p/, 0.75 for /t/, and 0.78 for /k/. On average, the time error between the predicted VOT frames and the manual labels is 5 ms $\pm$ 2.8 ms for the initial burst and 4.7 ms $\pm$ 2.2 ms for the vowel onset. Previous works in the literature have reported accuracies of up to 99% detecting VOT segments with time errors of up to 15 ms. However, it is not possible to make fair comparisons between the systems described in Section 1.1 and our approach due to the following reasons: First, those systems are binary classifiers limited to differentiate between VOT and non-VOT speech segments. In our approach a multi-class system is trained in order to predict VOT labels from three different stop consonants. Second, the systems described in the literature only considered stop consonants produced in the initial position of isolated words. The speech recordings considered in our study include stop consonants produced in different positions between syllables and utterances. Third, the performance of the models it's measured in different ways. The studies in the literature only reported the proportion of correctly classified VOT segments when the time error between the prediction and the manual labels is less or equal to a predefined threshold (15 ms). In the current study, we measure the performance of the model taking into account the proportion of speech frames that are correctly classified among the predictions (precision) and the actual labels (recall).

Regarding the acoustic features, the results in Table 3 indicate that the performance of the stacked BiGRU mostly relies on the information provided by the spectral features. However, even if the network makes more mistakes when it's trained with the temporal features, it can be observed that VOT speech segments are modeled better when acoustic features from the time domain are combined with the spectral ones. These results show the importance of considering prior knowledge to process information from speech signals when using deep learning methods. The low performance obtained with the temporal features can be explained considering the acoustic factors that affect the production of voiceless stop consonants, for instance, the presence of glottal pulses during the release and aspiration stages indicate that the vocal folds are vibrating, resulting in the presence of pitch values during stages that normally are not periodic. One explanation for this phenomenon may be that the speakers perform rapid movements of the articulators during the DDK task, resulting in difficulties to control the movement of the vocal folds when alternating from vowels to voiceless stop sounds [41].

Regarding the detection of the VOT for each consonant, the highest performance of the network to predict VOT labels is achieved for the /k/, followed by the /t/, and the /p/. These results indicate that the BiGRU is able to better model VOT segments with longer durations. As described in Section 2.1, the duration of the VOT is related to the place of articulation, thus, a short duration is obtained for the bilabial stop /p/, intermediate values for the alveolar stop /t/, and long duration for the velar stop /k/. Additionally, it can be observed that the prediction of VOT frames from the consonant /p/ was considerably lower with respect to the /t/ and /k/. This can be explained considering acoustic phenomena such as the consonant weakening, which in our database was found to affect mainly the /p/ sounds located in intermediate positions. Such a phenomenon is characterized by the absence of the burst during the release stage resulting in a "weaker" production of the voiceless stop /p/. Future work will include labels of the different acoustic phenomena affecting the production of stop consonants in order to better model the speech signals.

The present study has some limitations. First, the detection of VOT segments is limited to the standardized DDK task, thus, the proposed approach is not suitable to automatically detect VOT segments from recordings containing continuous speech, e.g., monologues, picture descriptions, conversations. Second, the stacked BiGRU is trained only with speech recordings of elderly speakers,

which may present articulation disorders related to the aging process [42]. Nevertheless, our system was trained with a standardized speech task (repetition of /pa-ta-ka/) uttered by elderly speakers because in the future we plan to use the proposed approach to evaluate motor speech disorders in PD patients. Third, the accuracy of the model decreases when noisy speech signals are considered for the test. Thus, future work will include a data augmentation-based approach in order to improve the robustness of the model against signals recorded in non-controlled acoustic conditions.

Regarding the computational cost, most of the heavy work is performed during the training stage. For practical applications, only the pre-trained model is used to perform acoustic analysis, thus, most of the computational cost is produced during the feature extraction stage, particularly when computing the IMF-based features, due to the fact that such descriptors are computed after extracting all components by means of the EMD, which is an iterative algorithm.

## 7. Conclusions

In this study a methodology is presented to automatically detect VOT segments in voiceless stop consonants produced during the rapid repetitions of /pa-ta-ka/ by Spanish native speakers from Colombia. Temporal and spectral analyses are performed by an expert in linguistics in order to measure the VOT. For the automatic detection, we extract different acoustic features in the time and spectral domains in order to include similar information considered by the expert to label the VOT. As a result, the speech recordings are transformed into feature vector sequences that, together with the VOT labels, are used as input data to train a multi-class BiGRU to classify between non-VOT (silence, vowels) and VOT feature vectors of /p/, /t/, and /k/. According to the results, the spectral features proved to be better than temporal features for the detection of VOT segments, however, the highest performance was obtained when both spectral and temporal features were considered for automatic detection. Regarding the detection of individual VOT segments, the lowest performance was obtained for the /p/, which can be explained considering that there are different acoustic phenomena that affects the /p/ sounds produced in intermediated positions. Such phenomena include consonant weakening (weak production of a voiceless plosive) and voicing (vocal fold vibration during the release stage in a voiceless plosive). Additionally, the error between the predicted and real labels in the prediction of the initial burst and vowel onset was also calculated. According to the results, our system is able to predict the VOT segments with an average error of 5 ms for both the initial burst and vowel onset, with respect to the manual labels given by the expert. Currently, the VOT segments of the rapid repetition of /pa-ta-ka/ are being labeled considering speech recordings of Parkinson's disease patients, thus, in a future we expect to evaluate the proposed approach to analyze pathological speech.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

The authors acknowledge to the Training Network on Automatic Processing of PAthological Speech (TAPAS) funded by the European's Union Horizon 2020 programme under Marie Sklodowska-Curie grant agreement Number 766287. Tomás Arias-Vergara is under grants of Convocatoria Doctorado Nacional-785 financed by COLCIENCIAS. The authors also thanks to CODI from University of Antioquia (grant Numbers 2018-23541 and 2017-15530).

## References

[1] A.M. Liberman, P. Delattre, F.S. Cooper, Some cues for the distinction between voiced and voiceless stops in initial position, J. Acoust. Soc. Am. 29 (1957) 1254.
[2] L. Lisker, A.S. Abramson, A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements, vol. 20, Taylor & Francis, 1964, pp. 384–422.
[3] K.-Y. Chao, L.-M. Chen, A cross-linguistic study of voice onset time in stop consonant productions, Int. J. Comput. Linguist. Chin. Lang. Process. 13 (2008) 215–232.
[4] P.M. Sweeting, R.J. Baken, Voice onset time in a normal-aged population, J. Speech Hear. Res. 25 (1982) 129–134.
[5] V.Y. Yu, et al., Effects of Age, Sex and Syllable Number on Voice Onset Time: Evidence from Children's Voiceless Aspirated Stops, vol. 58, SAGE Publications Sage UK, London, England, 2015, pp. 152–167.
[6] G.S. Neiman, R.J. Klich, E.M. Shuey, Voice onset time in young and 70-year-old women, J. Speech Hear. Res. 26 (1983) 118–123.
[7] S. Das, J.H.L. Hansen, Detection of Voice Onset Time (VOT) for unvoiced stops (/p/, /t/, /k/) using teager energy operator for automatic detection of accented English, in: Proceedings of the 6th Nordic Signal Processing Symposium, 2004, pp. 344–347.
[8] V. Stouten, et al., Automatic Voice Onset Time Estimation from Reassignment Spectra, vol. 51, Elsevier, 2009, pp. 1194–1205.
[9] C.-Y. Lin, H.-C. Wang, Automatic estimation of voice onset time for word-initial stops by applying random forest to onset detection, J. Acoust. Soc. Am. 130 (2011) 514–525.
[10] C. Prakash, et al., Bessel features for detection of voice onset time using AM-FM signal, in: Eighteenth International Conference on Systems, Signals and Image Processing, IEEE, 2011, pp. 1–4.
[11] N. Ryant, J. Yuan, M. Liberman, Automating phonetic measurement: the case of voice onset time, in: Proceedings of Meetings on Acoustics ICA2013, vol. 19, ASA, 2013, pp. 1–10.
[12] A.P. Prathosh, A.G. Ramakrishnan, T.V. Ananthapadmanabha, Estimation of voice-onset time in continuous speech using temporal measures, J. Acoust. Soc. Am. 136 (2014) 122–128.
[13] M. Novotný, J. Pospíšil, R. Čmejla, J. Rusz, Automatic detection of voice onset time in dysarthric speech, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2015, pp. 4340–4344.
[14] D. Montaña, Y. Campos-Roca, C.J. Pérez, A Diadochokinesis-Based Expert System Considering Articulatory Features of Plosive Consonants for Early Detection of Parkinson's Disease, vol. 154, Elsevier, 2018, pp. 89–97.
[15] M. Sonderegger, J. Keshet, Automatic discriminative measurement of voice onset time, in: Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, 2010, pp. 2242–2245.
[16] E. Fischer, A.M. Goberman, Voice Onset Time in Parkinson's Disease, vol. 43, Elsevier, 2010, pp. 21–34.
[17] T. Tykalova, J. Rusz, J. Klempir, R. Cmejla, E. Ruzicka, Distinct patterns of imprecise consonant articulation among Parkinson's disease, progressive supranuclear palsy and multiple system atrophy, Brain Lang. 165 (2017) 1–9.
[18] K. Irie, Z. Tüske, T. Alkhouli, R. Schlüter, H. Ney LSTM GRU, Highway and a bit of attention: an empirical overview for language modeling in speech recognition, in: Proceedings of the Seventeenth Annual Conference of the International Speech Communication Association, 2016, pp. 3519–3523.
[19] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, in: NIPS 2014 Deep Learning and Representation Learning Workshop, 2014.
[20] V. Stojanovic, N. Nedic, Identification of time-varying OE models in presence of non-Gaussian noise: application to pneumatic servo drives, Int. J. Robust Nonlinear Control 26 (18) (2016) 3974–3995.
[21] V. Stojanovic, N. Nedic, Joint state and parameter robust estimation of stochastic nonlinear systems, Int. J. Robust Nonlinear Control 26 (14) (2016) 3058–3074.
[22] V. Filipovic, N. Nedic, V. Stojanovic, Robust identification of pneumatic servo actuators in the real situations, Forsch. Ingenieurwes. 75 (4) (2011) 183–196.
[23] M.A. Little, Mathematical foundations of nonlinear, non-Gaussian, and time-varying digital speech signal processing, in: International Conference on Nonlinear Speech Processing, Springer, 2011, pp. 9–16.
[24] T. Nguyen, S. Stueker, J. Niehues, A. Waibel, Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7689–7693.
[25] Y. Adi, J. Keshet, O. Dmitrieva, M. Goldrick, Automatic measurement of voice onset time and prevoicing using recurrent neural networks, in: Proceedings of the Seventeenth Annual Conference of the International Speech Communication Association, 2016, pp. 3152–3155.
[26] Y. Adi, J. Keshet, E. Cibelli, M. Goldrick, Sequence segmentation using joint RNN and structured prediction models, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2017, pp. 2422–2426.
[27] Y. Shrem, M. Goldrick, J. Keshet, Dr.VOT: measuring positive and negative voice onset time in the wild, in: Proc. Interspeech 2019, 2019, pp. 629–633.

[28] J.R. Orozco-Arroyave, et al., New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease, in: Language Resources and Evaluation Conference, LREC, 2014, pp. 342–347.
[29] D.H. Klatt, Aspiration and voice onset time in word-initial consonant clusters in English, J. Acoust. Soc. Am. 54 (1973) 319.
[30] P. Boersma, et al., Praat, a system for doing phonetics by computer, Glot Int. 5 (2002) 341–345.
[31] Z. Yang, Z. Yu, C. Xie, Y. Huang, Application of Hilbert–Huang Transform to Acoustic Emission Signal for Burn Feature Extraction in Surface Grinding Process, vol. 47, Elsevier, 2014, pp. 14–21.
[32] P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, in: Proceedings of the Institute of Phonetic Sciences, Amsterdam, vol. 17, 1993, pp. 97–110.
[33] N.E. Huang, et al., The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-stationary Time Series Analysis, vol. 454, The Royal Society, 1998, pp. 903–995.
[34] N.E. Huang, X. Chen, M.-T. Lo, Z. Wu, On Hilbert spectral representation: a true time-frequency representation for nonlinear and nonstationary data, Adv. Adapt. Data Anal. 3 (2011) 63–93.
[35] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (1997) 2673–2681.
[36] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5–6) (2005) 602–610.
[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: an imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019, pp. 8024–8035.
[38] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: 3rd International Conference on Learning Representations (ICLR), 2015, https://arxiv.org/abs/1412.6980.
[39] J.C. Vásquez-Correa, P. Klumpp, J.R. Orozco-Arroyave, E. Nöth, Phonet: a tool based on gated recurrent neural networks to extract phonological posteriors from speech, in: Proceedings of the 20th Annual Conference of the International Speech Communication Association, 2019, pp. 549–553.
[40] F. Pedregosa, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[41] T. Louzada, R. Beraldinelle, G. Berretin-Felix, A.G. Brasolotto, Oral and vocal fold diadochokinesis in dysphonic women, J. Appl. Oral Sci. 19 (6) (2011) 567–572.
[42] T. Arias-Vergara, J.C. Vásquez-Correa, J.R. Orozco-Arroyave, Parkinson's disease and aging: analysis of their effect in phonation and articulation of speech, Cogn. Comput. 9 (6) (2017) 731–748.

**Tomás Arias-Vergara** received the B.S. degree in Electronics Engineering from University of Antioquia (Medellin, Colombia) in 2014, and the M.Sc. degree at the same institution in 2017. Currently, he is a doctoral candidate at the GITA Lab from the University of Antioquia (Colombia) and at Pattern recognition Lab from the Friedrich-Alexander-Universität Erlangen-Nürnberg (Germany) in a joint degree program. Additionally, he is a guest researcher in the Department of Otorhinolaryngology, Head and Neck Surgery from the Ludwig-Maximilians-Universität München (Germany). His research interests include speech processing, signal processing, pattern recognition, machine learning and their applications in pathological speech.

**Patricia Argüello-Vélez** received the B.S. degree in speech therapist from the University of Santiago de Cali, Valle, Colombia, in 2007, and the M.Sc. degree in Linguistics from the Pereira technology university, Pereira, Colombia, in 2013. Currently, she is a PhD candidate in linguistics at the University of Antioquia. She is professor at the Faculty of Health at the University of Santiago de Cali, member of the research group in speech therapy and psychology of the university of Santiago de Cali and member of the Sociolinguistic Studies Group of University of Antioquia.

**Juan Camilo Vasquez-Correa** received his B.S. degree in electronics engineering and his M.Sc. in Telecommunication engineering both from the University of Antioquia (Colombia). He has performed research activities related to signal processing and machine learning for health-care and biometric applications for five years now, both in academic and industrial partners. Currently, he is a doctoral candidate and researcher at the Pattern recognition Lab from the Friedrich Alexander University (Germany) and at University of Antioquia (Colombia). His research interests include pathological speech processing, paralinguistics, machine learning, and deep learning for health-care data.

**Elmar Nöth** is a Professor for Applied Computer Science at the Friedrich-Alexander-University Erlangen-Nuremberg (FAU) in Germany. He studied in Erlangen and at M.I.T. and received the Dipl.-Inf. and the Dr.-Ing. from the FAU in 1985 and 1990, respectively. He received his Habilitation in 2006, also from FAU. Since 1990 he was an assistant professor at the Pattern Recognition Lab of the FAU. Since 2008 he is a full professor at the same institute and head of the speech group. From 2013 to 2014 he was Adjunct Professor at the King Abdulaziz University, Jeddah, Saudi Arabia. He is one of the founders of the Sympalog Company, which markets conversational dialogue systems. His current interests are prosody, analysis of pathologic speech, computer aided language learning and emotion analysis.

**Maria Schuster**, Prof. Dr. med., is a Medical Doctor specialized in phoniatrics and pediatric audiology. She studied medicine at the Ruprecht-Karls-Universität Heidelberg (Germany), at the University of Paris IV (France) and at the University of Lausanne (Switzerland). Afterwards she was specializing in otorhinolaryngology at the Technical University of Munich and then in phoniatrics and pediatric audiology at the university of Heidelberg and Erlangen-Nuremberg. She received her Ph.D. in 2005 from the Friedrich-Alexander-University Erlangen-Nuremberg. From 2011 she was head of the dept. of phoniatrics and pediatric audiology at the clinic of otorhinolaryngology, head and neck surgery, at the university hospital of the Ludwig-Maximilians-University Munich (LMU). She is currently working at the LMU and at the Metropol Medical Center in Nuremberg (Germany). Her main scientific interests lie in speech diagnostics and therapy and hearing assessment.

**María Claudia González-Rátiva** received the B.S. degree in Pedagogy in English and Spanish from National Pedagogical University, in 1988, and the M.Sc. degree in Hispanic Linguistics from the Caro and Cuervo Institute, Bogotá, Colombia, in 1995. In 2015, she received the Ph.D. degree in Linguistics from the University of Antioquia, Medellín, Colombia. She is an Associate Professor and the coordinator of the Phonetics Laboratory of the Linguistics area at the Communications Faculty of the University of Antioquia. She is a member of the Sociolinguistic Studies Group, head of the Corpus Preseea-Medellín project and researcher in the field of phonetics and variationist sociolinguistics.

**Juan Rafael Orozco-Arroyave** received the B.S. degree in Electronics Engineering and the M.Sc. degree in Telecommunications Engineering from the University of Antioquia, Medellin, Colombia, in 2004 and 2011, respectively. He received the Ph.D. degree in Computer Science from the Friedrich-Alexander-Universität Erlangen-Nürmberg, Germany, in 2015. He is currently an Associate Professor and the head of the GITA Lab at the University of Antioquia and an adjunct researcher with the Pattern Recognition Lab at the Friedrich-Alexander-Universität Erlangen-Nürmberg. His main research interests include speech and language processing, signal processing, pattern recognition, machine learning, and their applications to different fields in medicine.