An Analysis of Visual Speech Features for Recognition of Non-articulatory Sounds using Machine Learning

Francisco Carlos M. Souza
Federal University of Technology – Paraná
Dois Vizinhos, PR, Brazil

Alinne C. Correa Souza
Federal University of Technology – Paraná
Dois Vizinhos, PR, Brazil

Carolina Y. V. Watanabe
Federal University of Rondônia,
Porto Velho, RO, Brazil

Patricia Pupin Mandrá
University of São Paulo
Ribeirão Preto, SP, Brazil

Alessandra Alaniz Macedo
University of São Paulo
Ribeirão Preto, SP, Brazil

ABSTRACT
People with articulation and phonological disorders need exercise to execute sounds of speech. Essentially, exercise starts with production of non-articulatory sounds in clinics or homes where a huge variety of the environment sounds exist; i.e., in noisy locations. Speech recognition systems considers environment sounds as background noises, which can lead to unsatisfactory speech recognition. This study aims to assess a system that supports aggregation of visual features to audio features during recognition of non-articulatory sounds in noisy environments. Thehe methods Mel-Frequency Cepstrum Coefficients and Laplace transform were used to extract audio features, Convolutional Neural Network to extract video features, and Support Vector Machine to recognize audio and Long Short-Term Memory networks for video recognition. Report experimental results regarding the accuracy, recall and precision of the system on a set of 585 sounds was achieved. Overall, the results indicate that video information can complement audio recognition and assist non-articulatory sound recognition.

General Terms
Speech Recognition, down syndrome.

Keywords
Assistive technology; health information; speech recognition; machine learning; down syndrome.

1. INTRODUCTION
Human communication is mostly supported by speech, which is the expression of or the ability to express thoughts and feelings by articulate sounds. Speech Recognition (SR) is the process of automatically recognizing sounds emitted by a person or to determine a specific speaker on the basis of information about speech signal [1].

Speech impairments refers to a speaker's inability to produce speech sounds correctly. Disorders may be due to multiple reasons such as neurological, myofunctional, and/or congenital linguistic alterations and may have different levels of severity. Individuals with speech disorder face social difficulties and have problems integrating with the society. For example, children with speech disorders are at particularly higher risk of being bullied by peers [2]. Assistive technologies can help to overcome speech impairments, thereby avoiding bullying and social isolation. Actually, these technologies can be incorporated in the treatment of individuals with speech disorder.

Treatment of speech disorders requires speech therapy and substantial effort. People with such disorders need rigorous training to plan and to execute motor acts of speech. Speech training starts with facial motor praxia activities and oral myofunctional exercises that involve production of non-articulatory sounds, such as production of blow, tongue snap, and kisses (lip protrusion), which can be considered precursors in the production of phonemes and words (articulatory sounds). Essentially, a Speech Therapist supervises therapy exercises performed in therapeutic clinics. Depending on patient's condition, the use of multimedia devices and mobile technology can make it easier for individuals to achieve their goals through speech training exercises that they can carry out in a clinic or at home.

For several years, great effort in the field of multimedia processing has been devoted to addressing recognition of different types of sounds (speech and others) and to filtering noises [3-5]. In SR systems, therapy exercises for speech disorders start with non-articulatory sounds, which can be misclassified as noise. Besides this recognition issue, speech exercises conducted in noisy locations (clinics or homes) are recorded together with various environment sounds like music, bird song, rain noise, street traffic noise, TV sounds as well as in the presence of people speaking, baby cries, dog barking, etc. SR systems consider environment sounds as background noises, which can lead to false recognition or low performance. Some methods have been proposed to filter noises [6,7].

In this paper is presented an Audio-VIdeo Speech (AVIS) recognition system that supports aggregation of visual features to audio features during sound recognition tasks when the audio features are not sufficient to promote effective recognition[1] e created the AVIS system to analyze the hypothesis that video information can complement audio recognition and assist non-articulatory sound recognition in a real home environment. The proposed system employs a very popular method called Mel-Frequency Cepstrum Coefficients (MFCC), a Laplace Transform for audio recognition and a Viola-Jones method and facial landmarks for visual recognition. The use of well-known techniques aids

1 This proposal is part of a large project, called SofiaFala, in development at USP. SofiaFala (Sistema Inteligente de Apoio a Fala - Intelligent Speech Training Software). SofiaFala has funding from CNPq - Assistive Technology.

verification of the hypothesis. Overall, the contributions of the present work include: *(i)* an audio-video recognition system for non-articulatory sounds; *(ii)* an analysis of complementarity in an audio-video recognition system; and *(iii)* an experiment to demonstrate that video features effectively aid recognition of non-articulatory sounds.

The remainder sections are organized as follows: Section 2 depicts the background. Section 3 details the automated approach. Section 4 reports the experimental study conducted herein. Section 5 presents the results and discusses the benefits and relevance of the proposal. Section 6 shows related work to the present study. Finally, section 7 shows final remarks and future works.

2. BACKGROUND

Audio Visual speech recognition techniques have been widely investigated over the past years. Most studies have concentrated on articulatory speech sounds (words and phonemes). Some techniques like MFCCs, Laplace transform and Support Vector Machine (SVM) have been used to recognize these types of sounds.

2.1 Audio Visual Speech Recognition

Speech is an audiovisual signal consisting of audio vocalization and the corresponding mouth configuration. Although audio signal carries most information, visual signal also carries complementary and redundant information. Acoustic noise does not affect visual information, which can significantly improve speech recognition performance in noisy environments [8]. This improvement occurs because visual speech provides cues to both timing of the incoming acoustic signal (the amplitude envelope, which influences attention and perceptual sensitivity) and its content (place and manner of articulation, which constrains lexical selection) [9].

In this sense, a computer-aided Audio-Visual Speech Recognition (AVSR) system is a promising technique for reliable speech recognition, especially when noise corrupts audio [10]. AVSR uses visual information from the speakers' lip motion to complement corrupted audio speech input. Several studies have proven that visual information plays a key role in automatic speech recognition when background noise corrupts audio, for example, or even when the audio is inaccessible [11,12]. In the research context, this area is now known by different names including lip reading, speech reading and visual speech recognition [13].

Traditionally, visual speech recognition systems consist of two stages: feature extraction from the mouth region of interest (ROI) and classification. The feature extraction step is the process through which useful information is derived from an original signal; this information is relevant for the task and has a more compact representation, which is suitable for use in a classifier. This step simply involves selection, during which elements of the original data vector are kept, or a transform, which projects original data in a different, lower-dimensional space.

Image processing to extract visual information from the lips typically comprises three stages: face detection, ROI location, and lip segmentation [12]. Lip segmentation poses challenges to image processing. The first inherent challenge of lip segmentation is variability in the speaker's profile including skin color, lip color, lip shape, facial hair, and makeup. Second, the ROI contents are not static, and visibility of the teeth, tongue, and oral cavity changes as the lips move to form facial expressions and speech sounds. Finally, non-ideal environmental conditions such as lighting, speaker orientation, and background create a third layer of complexity.

For the classification step, audio and video information can be integrated by feature fusion or by decision fusion [14]. The feature fusion technique combines information at feature level and submits a single combined feature vector to a single classifier. This is generally simple to implement and allows correlation between audio and video to be modeled. The simplest feature fusion method corresponds to concatenation of the audio and video feature vectors. Unfortunately, this technique cannot explicitly model the relative reliability of each feature stream. Feature stream may vary significantly even within the duration of an utterance due to constant or instantaneous background noise or channel degradations.

In contrast, decision fusion systems assume independence between the two streams and combine the results of separate classifiers for audio and video, offering a mechanism that can model the reliabilities of each feature stream. These systems usually combine parallel classifier architecture. Capturing the reliabilities of the audio and video feature streams is possible through application of weights during the fusion process. Weights may be globally set to fixed values calculated by testing the system to find which weights produce optimal speech recognition [15, 16].

Because the system aims to present video information as complement to audio recognition, was proposed a decision fusion system to assist non-articulatory sound recognition in a real home environment.

2.2 Mel-Frequency Cepstrum Coefficients

Sound waves are mechanical waves that propagate through continuous media, including air and water. Sound waves can interact with thin surfaces and membranes, such as membranes in general purpose microphones, to produce local oscillations in a material. Oscillation amplitudes along time are converted to numbers, whereas the time-ordered sequence these amplitudes form generates the audio signal $\psi(t)$. For human speech, the corresponding audio signal may lack uniform or simple oscillatory patterns, which demands additional audio analysis techniques. Speech recognition shares deep ties with spectral analysis of sound waves in that sound waves are decomposed into several simpler waves with characteristic lengths and oscillation frequency. The most common way to decompose a given audio signal $\psi(t)$ is to employ Fourier transform [17]:

$$\psi(t) = \frac{1}{\sqrt{2\pi}} \int dv\, \psi(v) e^{2\pi i v t}, \quad (1)$$

where $\psi(v)$ are complex valued coefficients that correspond to monochromatic planar waves with frequency v. The Fourier transform is an invertible linear transform, which means that the original signal can be transformed and recovered. The audio signal spectrogram $P(v)$ complements the Fourier decomposition by informing each frequency contribution to the signal. For the Discrete Fourier Transform,

$$P(v) = |\psi(v)|^2 \Big/ \sum_{\varepsilon} |\psi(\varepsilon)|^2, \quad (2)$$

which plays the role of a likelihood estimator for the frequency v. For real valued signals, contributions from positive and negative frequencies mirror each other, so negative frequencies are usually discarded. Figure 1 exemplifies two spectrograms: one derived from harmonic

signal and another derived from one non-articulatory sound (kiss). For simpler signals, the spectrogram provides enough information. As sound complexity increases, as in the case of sounds produced by human speech, the number of modes available in a given signal also increases. Class intervals, or bins, circumvent this issue by breaking the spectrogram down into fewer frequency groups. The selection of class intervals depends on which frequency domain region or behavior one intends to highlight. The human auditory system perceives relative changes better than absolute changes, which suggests a logarithmic scale transformation for frequencies. The Mel scale satisfies this requirement νm = 2595 log10(1 +ν/700). Accordingly, uniform Mel-frequency domain division provides the desired class intervals. Finally, the Mel-frequency Cepstrum (MFC) summarizes P(νm) within a given class interval, and its numerical value follows from centroid or another average evaluation.

The MFC coefficients (MFCC) are the main features for speech recognition and classification [18]. They are evaluated by taking the log of MFCs, followed by the Cosine-DFT.

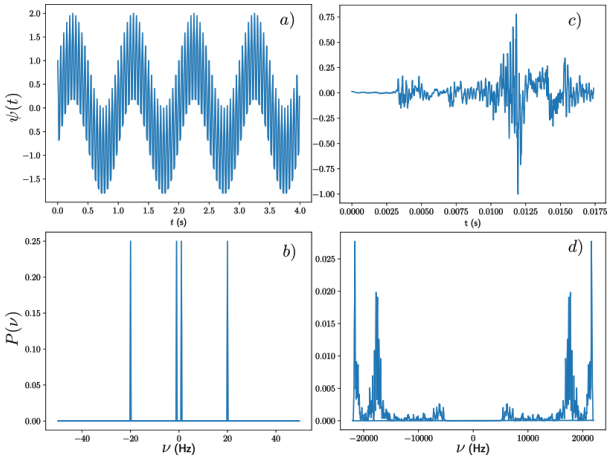

Figure 1. Signals and Spectrograms. (a) The signal ψ(t) = sin(2πt) + cos(40πt), which emerges from the combination of two frequencies ν = 1 and 20 Hz, respectively, and (b) the corresponding spectrogram (frequency centered). (c) A non-articulatory sound (kiss) audio signal and (d) the corresponding spectrogram.

2.3 Mel-Frequency Cepstrum Coefficients

For speech recognition, two techniques have been widely discussed in the literature: *(i)* Support Vector Machine and *(ii)* Neural Networks. Support Vector Machines (SVMs) represent a group of theoretically superior machine learning algorithms [19]. SVMs have proven to be much more effective than other conventional nonparametric classifiers (e.g., RBF neural networks, nearest neighbor (NN), nearest center (NC), and the NN classifier) in terms of classification accuracy, computational time, and stability to parameter settings [20]. Other works have shown that SVMs also are more effective than the traditional pattern recognition approaches based on the combination of a feature selection procedure with a conventional classifier.

A Neural Network (NN) is an information processing paradigm that is inspired by the way biological nervous systems, such as a brain and its information processing. The novelty of this paradigm is the structure of the information processing system, that is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. A NN is configured for a particular application, like pattern recognition or data classification, through a learning process. The following NN can be assigned these type of applications: *(i)* Convolutional Neural Network; and *(ii)* Recurrent Neural Network. This proposed work exploited these both types of NN.

In NN, Convolutional Neural Networks (CNNs) is one of the main categories to do images recognition and images classifications. Technically, deep learning CNN models are used to train and to test a series of convolving kernels filters each input image. CNNs utilize layers with convolving filters that are applied to local features [21]. In addition to classification and patterns recognition, a CNN can be used as a features extractor, since it is a particular skill in this kind of NN. A Recurrent Neural Networks (RNNs) is a NN model that involves directed cycles in memory for modeling time series such as speech, text, financial data, audio, video, etc. RNNs maps temporal dynamics through mapping input vectors to hidden states and hidden states to outputs that allow connections between hidden units associated with a time delay. This mechanism enables RNNs that can retain information of the past time, making it discover temporal correlations in events that are far away from each other in the data [22].

Although NN has been applied in different contexts successfully, it is not very efficient in problems with temporal information such as speech recognition and video classification. This is because traditional neural networks have a one-way flow of information and cannot store long-term information. Unlike traditional NNs, RNNs have cycles between their units. In other words, units can have connections to units of previous layers or from the same layer. This allows it to demonstrate temporal dynamic behavior for a time sequence.

It's important to highlight a type of RNN called Long Short-Term Memory networks (LSTM). LSTM was proposed by Hochreiter and Schmidhuber (1997) and they are able of learning long-term events. In general, RNNs have a structure in form of a chain of repeating modules of a neural network. The hidden state of LSTM units works with nonlinear mechanisms enabling the state to propagate without any modification, be updated, or be reset, using simple learned gating functions [23]. LSTMs have recently been demonstrated to be capable of large-scale learning of speech recognition [24] and language translation models [25, 26]. The proposed approach also used LSTM.

3. AUDIO VIDEO SPEECH RECOGNITION SYSTEM

This paper presents an Audio-VIdeo Speech recognition system (AVIS) based on machine learning and neural networks for non-articulatory sound recognition. AVIS will be part of a mobile application that intends to assist children with Down Syndrome during their training for the production of speech. In the SR area, recognizing speech distorted by noises or unwanted sounds coming from *(i)* the environment, *(ii)* a speaker with speech disturbance such as articulation problems, or *(iii)* non-articulatory sounds recognized as noises are admittedly difficult. The visual features can provide information that can be aggregated to corrupted or distorted speech.

The most common approaches to audio-visual recognition comprise a single system that performs audio and video processing simultaneously. Here, the AVIS system was poposed, which separates processing into an audio module and a visual module. The idea is to make speech recognition

more effective for noisy environments and even more feasible for low-performance hardware (such as mobile devices, Arduino, Raspberry, etc.) when just the audio module can be activated.

AVIS consists of three steps: (i) Signal Separation; (ii) Feature Extraction; and (iii) Speech Recognition. Figure 2 provides an overview of this system. The first step, "Signal Separation", separates audio and video signals from video files. Features are individually extracted for each type of signal (see "Audio Signal" and "Visual Signal" in Figure 2).

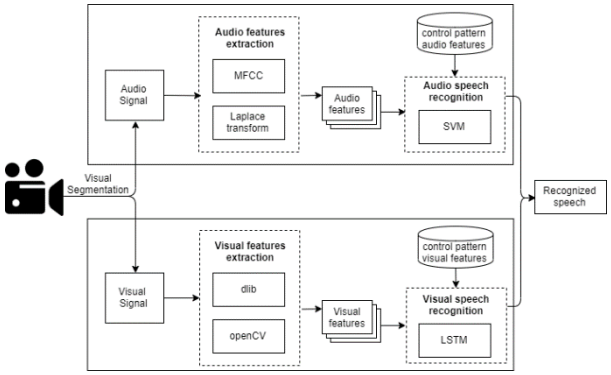

Figure 2. Audio and VIdeo recognition System (AVIS) for Recognition of nonarticulatory Sounds

The first step, "Signal Separation", separates audio and video signals from video files. Features are individually extracted for each type of signal (see "Audio Signal" and "Visual Signal" in Figure 2). For "Audio Feature Extraction", AVIS applies the methods MFCC and Laplace Transform. The "Visual Feature Extraction" is recognized by using CNN and the features sequence is sent to a separate LSTM.

In the third step, "Audio Speech Recognition" and "Visual Speech Recognition" verify whether SVM and LSTM from a previously trained audio and video datasets can correctly recognized a non-articulatory sound (e.g kiss, blowing, etc). Finally, the recognition rate of a sound is obtained for each module (audio and visual). The "Audio Feature Extraction" and the "Visual Feature Extraction" steps are detailed in the following sections.

3.1 Audio Feature Extraction

The audio signal extracted from the video file is the input data for the feature extraction process. For this process, the methods MFCC and Laplace transform were applied. The method MFCCs exhibits high degree of linear separability for non-articulatory sounds. This observation adds up to well-known evidence suggesting that MFCCs is a viable method to feature audio signals. Therefore, for a given audio signal ψ(t), the proposed approach extracts MFCCs corresponding to n = 13 in the Mel-frequency spectra, processes it, and assigns it to a feature factor with N = 14 entries. The last entry of the feature vector is set by Laplace transform of the signal and includes a feature that improves the separation of non-articulatory sounds.

Section 2 details the guidelines for the general procedure that was used to extract MFCCs. However, during evaluation of spectrogram P(ν), was considered the frequency resolution Δν in accordance to the uncertainty principle $\Delta t \Delta \nu \propto (4\pi) - 2$. In short, the principle asserts that achieving a fixed resolution Δν for arbitrary signal duration Δt is impossible. Instead, was subdivided the signal ψ(t) into smaller frames with fixed duration Δt (see Figure 3(a)), and to conduct Fourier analysis with fixed resolution Δν = 1/(44100 × 512) Hz.

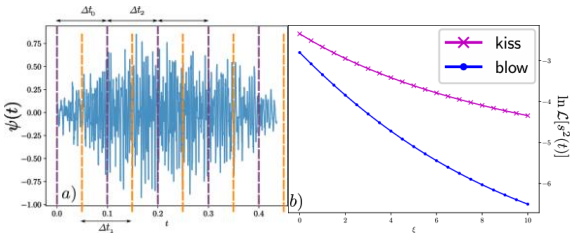

Figure 3. Signal division. (a) Frames and (b) Laplace transform.

After signal partitioning into shorter frames, each frame produced a sequence of MFCCs. Let $\emptyset_\ell(\tau)$ be the ℓ-th MFC coefficient corresponding to the τ-th frame. Because signals with different time duration would also produce collections of MFCCs with different sizes, they had to be processed further. AVIS intends to recognize nonarticulatory sounds, which are naturally short-lived, so was considered time-averaged MFCCs:

$$\langle \phi_\ell \rangle = \frac{1}{M} \sum_{\tau=0}^{M-1} \phi_\ell(\tau), \qquad (3)$$

where M > 0 is the total number of frames, and $\ell = 0, 1, \ldots, n - 1$. MFCCs were extracted by using python-speech-features package with the following default settings: frame duration length Δt = 20ms; time step between two frames = 10 ms; and 13 MFCCs extraction frames.

Signal processing employs several linear transformations to deal with the various signals. Fourier analysis features among the most popular transformation because it is useful during oscillating signal analysis. Laplace transform is another popular linear transformation that exploits the behavior of signal amplitude growth and decay. ANA applies Laplace transform to the signal squared amplitudes (always positive).

$$\mathcal{L}[\psi^2(t)] = \int_0^\infty dt e^{-\mathcal{E}t} \psi^2(t), \qquad (4)$$

to classify signals depending on how they evolve along time. We took the logarithm of Laplace transform to produce more pronounced signal separation. Figure 3(b) depicts the log of Laplace transform for various parameters $\mathcal{E} > 0$ for two non-articulatory sounds, in which the class separation increases for increasing $\mathcal{E}$ values. To avoid introduction of spurious correlations, we included $\mathcal{L}[\psi^2(t)]$ with parameter $\mathcal{E} = 10$ as the 14-th entry of the feature vector.

3.2 Visual Feature Extraction

The features extraction was performed by CNN and passing the features sequence to a separate LSTM. The goal of this stage is to extract features from image, for this, is necessary to perform a three step process: *(i)* conversion of videos into image; *(ii)* execution of images in a CNN; *(iii)* creation of a feature vectors.

In the first and second steps, every video was be subsampled down to 30 frames and, we run each frame from every video through Inception[2], saving the output in the final pool layer of the CNN.

In the last step, we created feature vectors. The sampled thirty

[2] https://tfhub.dev/google/imagenet/inceptionv3/

frames of each video in a single 2,048-d vector was defined, and saved it to disk. Next, the vector was ready to train different RNN models without needing to pass images through the CNN every time continuously. Therefore, we read the same sample or train a new network architecture.

3.3 Audio-Visual Speech Recognition

In the audio module, we performed the automatic speech recognition using SVM. This module comprises the training and prediction phases. In both phases, AVIS uses the well-documented python package *scikit-learn* [27].

In the training phase, SVM fit with linear kernel uses datasets consisting of feature vectors from segmented audio and video signals recorded by speech for participants and control patternf. Linear kernels appropriateness depends mostly on the feature linear separability. After the fit, the dataset is partitioned into distinct classes.

During the prediction phase, the SVM predicts the classes of feature vectors derived from target audio and video signals. The correct classification of each target audio and video signal are known *a priori*, which allows the evaluation of the accuracy corresponding to the SVM predictions.

For the visual speech recognition, the LSTMs to train and classify the samples was applied. Furthermore, for the second and third steps (video feature extraction and visual speech recognition) the tensorflow library[3] was used. In this step, a single, 4096-wide LSTM layer was used, followed by a 1024 dense layer, with some dropout in between. Dropout is a method for addressing over-fitting problems. The idea is to randomly drop units together with their connections from the neural network during training. This can avoid units from co-adapting too much.

In the training phase, LSTM create the model into three distinct classes (blow, tongue popping and kiss) and the output of prediction is a probability of a new sample belong to a specific class.

4. EXPERIMENTAL STUDY

The experiments intended to verify whether video can effectively be used as complementary information (signal or feature) to audio signals to recognize non-articulatory sounds. The methodology used to conduct this experiment is based on Wohlin [28].

4.1 Research Design

To reach the goal of this experiment, the following Research Question (RQ) was formulated:

RQ: How effectively does the video module complement information provided by the audio module to support recognition of non-articulatory sounds with the aid of AVIS?

To answer this RQ, was investigated whether the AVIS visual recognition module can separately improve speech production from the audio recognition module. The Goal/Question/Metric (GQM) model adapted from Wohlin [28] was used. The GQM model presents the objectives of the experiment divided into five parts:

- **Object of study:** the AVIS system is the object of study.
- **Purpose:** the experiment aims to verify how effectively the video module complements information provided by the audio module for recognition of nonarticulatory sounds in real scenarios.
- **Perspective:** this experiment is carried out from the researchers' standpoint.
- **Quality focus:** effectiveness of the AVIS system, measured by the number of correctly recognized sounds (blow, tongue popping, and kiss), is the main effect under investigation.
- **Context:** this experiment involves 11 people in one controlled and two simulated noisy scenarios.

4.2 Experiment Design

This investigative study comprises two different experiments ($E = e_1; e_2$). For both experiments, the audio (e_1) and visual (e_2) speech signal were recorded for three non-articulatory sounds (*blow*, *tongue popping* and *kiss*) in three different scenarios: *(i)* a controlled environment i.e. without any noises; and *(ii)* two simulated noisy scenarios evolving rain and TV sounds.

These experiments were conducted with 12 audio-visual datasets. Each dataset was related to one non-articulatory sound for the three scenarios. Nine datasets were recorded by 11 participants ($P = p_1; p_2\ ; p_3; \ldots; p_{11}$), eight male and three female, aged from 20 to 45 years. The three other datasets were recorded by one person representing our *control pattern* (cp_1).

The video data obtained for the participants and control pattern were recorded with the smartphone back camera. Video recording parameters were the following: Audio sample frequency = 8kHz (one sample 16bit.), Video frame rate = 30fps, and size of a single video image = 10920 x 1080 pixels.

The effectiveness of the audio and video recognition using precision (P), Recall (R), and Accuracy (A) metrics were evaluated. Precision is the ratio of correctly predicted positive observations to the total predicted positive observation. Thus, Precision (P) is defined as the number of true positives (TP) over the number of true positives plus the number of false positives (FP), as can be seen in equation 5.

$$P = \frac{TP}{TP + FP} \tag{5}$$

Recall (R) is the ratio of correctly predicted positive observations to the all observations in actual class. Thus, Recall (R) is defined as the number of true positives (TP) over the number of true positives plus the number of false negatives (FN), as can be seen in equation 6.

$$R = \frac{TP}{TP + FN} \tag{6}$$

Accuracy (A) measures how often the classifier takes the correct decision, determined as the ratio between the number of correctly classified non-articulatory sounds and the total number of non-articulatory sounds (see equation 7).

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Table 1 lists the activities we used to perform the experiments and the descriptions of such activities.

5. RESULTS AND DISCUSSION

To address RQ, 12 datasets of segmented non-articulatory signals were used: three for controlled scenario, three for tv sound scenario, three for rain sound scenario and three for control pattern that was used as the baseline. All these datasets correspond to the classes *tongue popping*, *blow* and

[3] https://www.tensorflow.org

kiss. The next sections describe the experimental results.

5.1 Audio Recognition

In the first experiment (e_1), only the audio signal was analyze. These datasets allowed us to execute the SVM training phase and to evaluate the effectiveness of audio recognition. Figure 4 depicts the classification accuracy of audio signals for each scenario when compared to the *control pattern*.

First, ninety percent of our control pattern dataset was used for training; the remaining 10\% was employed as test set in a 10-fold cross validation setup. The recognition accuracy for this dataset was about 97\%, represented by the first bar in Figure 4. In this case, accuracy was expected to be high because the aim of cross-validation is to evaluate the generalization capacity of the model.

Table 1. Activities description performed in experiments conduction

Activities	Description
Video recording	Video and audio were recorded together. For both experiments (e_1 *and* e_2), the non-articulatory sound (blow, tongue popping and kiss) was recorded by by p_1 to p_{11},five times in each scenario (controlled, tv sound and rain sound) totaling 495 videos and 30 times by one person (cp_1) totaling 90 videos.
Audio and video segmentation	Each video was segmented because *(i)* it was necessary to capture a unique temporal dynamics within speech and *(ii)* the feature extraction step was based on each framed speech segment.
Signals Separation	The audio (e_1) and video (e_2) signals were analyzed separately and the signals produced by participants were recorded with the signals of the control pattern dataset(cp_1).
Feature extraction	For the audio module, the methods MFCC and Laplace transform were applied to extract 14 sound features. For the video module, the tensorflow library and the CNN were used with 2,048-d vector of features, which passed the features sequence to a separate LSTM.
Audio and video recognition	SVM and LSTM were employed to recognize sounds and video signals, respectively. It was possible to assess precision, recall, accuracy of both modules. In addition, the accuracy of each setting was evaluated for experiment (e_2), and the frames setting with the highest accuracy were selected.
Analysis	An analysis was conducted in order to verify whether visual information could improve audio recognition of non-articulatory sounds in noisy environments.

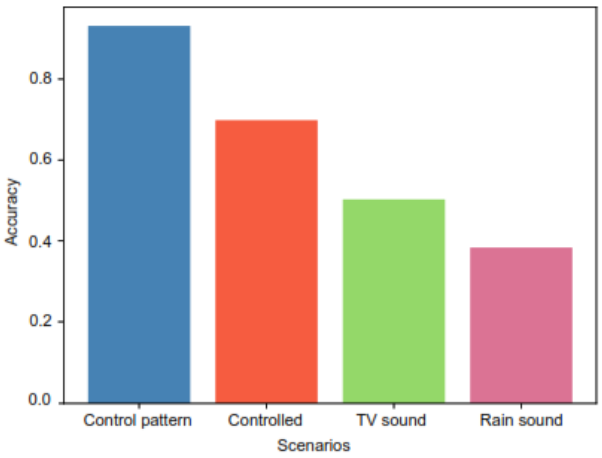

Figure 4. The accuracy of audio signals reference to non-articulatory sounds produced by participants in each scenario compared to the control pattern.

The accuracy of audio recognition obtained by using our proposed model was 75%, 55%, and 38% for the controlled scenario, tv sound scenario, and rain sound scenario, respectively, as compared to the *control pattern*. The accuracy of each scenario was the mean of all classes (*tongue popping, blow and kiss*).

The data in Figure 4 revealed that the rain sound was the worst scenario for audio recognition by our system because similar sounds in terms of spectral density were being overlapped. In addition, certain types of sounds; e.g., kisses, blowing, water, rain, alarm, etc., have harmonic structures, which can also hamper recognition of the target sound.

Figure 5 presents the precision and recall we achieved when we applied AVIS for audio recognition in each scenario. The results achieved were less than 60\% precision in the noisy scenarios (Figure 5.b and Figure 5.c). For the blow sound in Figure 5.c achieved 41\% recall at 38\% precision. In the case of the kiss in Figure 5.a achieved 65% recall at 83% precision.

It was also noticed that, even though achieved reasonable recognition rate for the controlled scenario, low variability in the baseline could lead the system to false classification. For instance, if we consider the blowing sound produced by different people, signal may be extremely distinct, i.e, they can be shorter, faster, higher, lower, more intense, etc., so that the target sound can be interpreted as a different sound.

However, for non-articulatory sounds (especially for sounds with in noisy environments), a baseline with only different speakers only may not be enough to achieve high recognition - features derived from other sources of information, such as facial movements may also be necessary.

5.2 Video Recognition

In the second experiment (e_2), only the video signal was analyzed. These datasets allowed us to execute the training and classifying phases by using LSTM and to evaluate the video recognition effectiveness.

First, thirty frames of each video were used for training, and the whole dataset (50 videos) was employed as test set in a 10-fold cross validation setup. After cross-validation, tests

were performed with different samples included in the training base. Each test dataset video of the 11 participants was classified using the trained model.

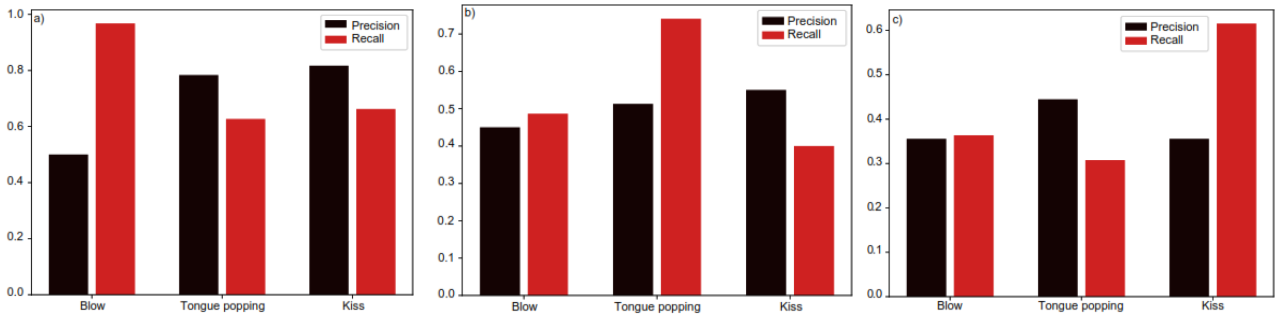

Figure 5. Precision and recall for non-articulatory sounds according to each scenario using SVM. (a) Data regarding the controlled scenario and (b) Data regarding the tv sound scenario. (c) Data regarding the rain sound scenario.

Figure 6 presents the accuracy regarding the correctness of non-articulatory movements of video signals in 20 epochs for cross-validation. Figure 6 shows the accuracy of LSTM hits in 20 epochs referring to the three non-articulatory movements for cross-validation. For each epoch, a gradual increase in accuracy was observed, reaching 100\% in epoch 20. It is natural to expect that: in the cross-validation, it achieves a high score, since the test is performed using a fold of the training dataset.

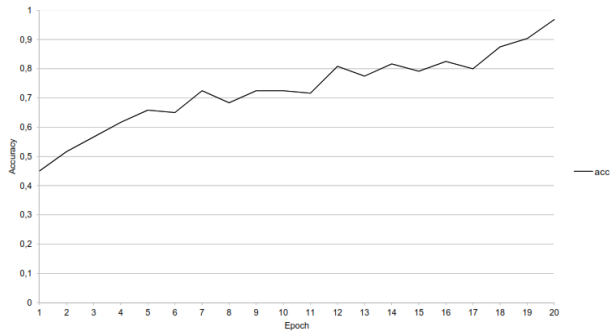

Figure 6. Accuracy of LSTM regarding the correctness of non-articulatory movements of video signals in 20 epoch

Figure 7 illustrates the prediction error of video signals in 20 epochs for cross-validation. Figure 7 is a complement of Figure 6, because the smaller the error, the better the generalization capacity of the model achieved.

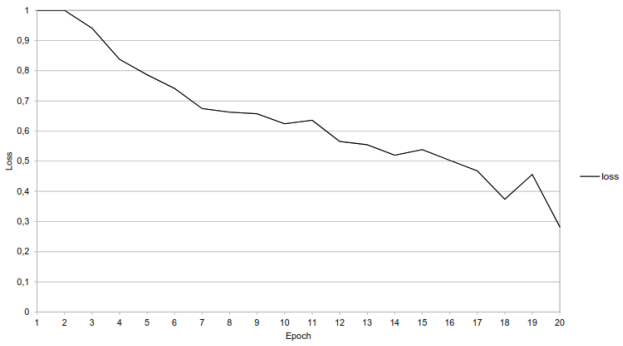

Figure 7. The prediction error of LSTM regarding video signals in 20 epochs for crossvalidation.

In Figure 8 each bar represents the average of hits of non-articulatory sounds produced by each participant of the trained model.

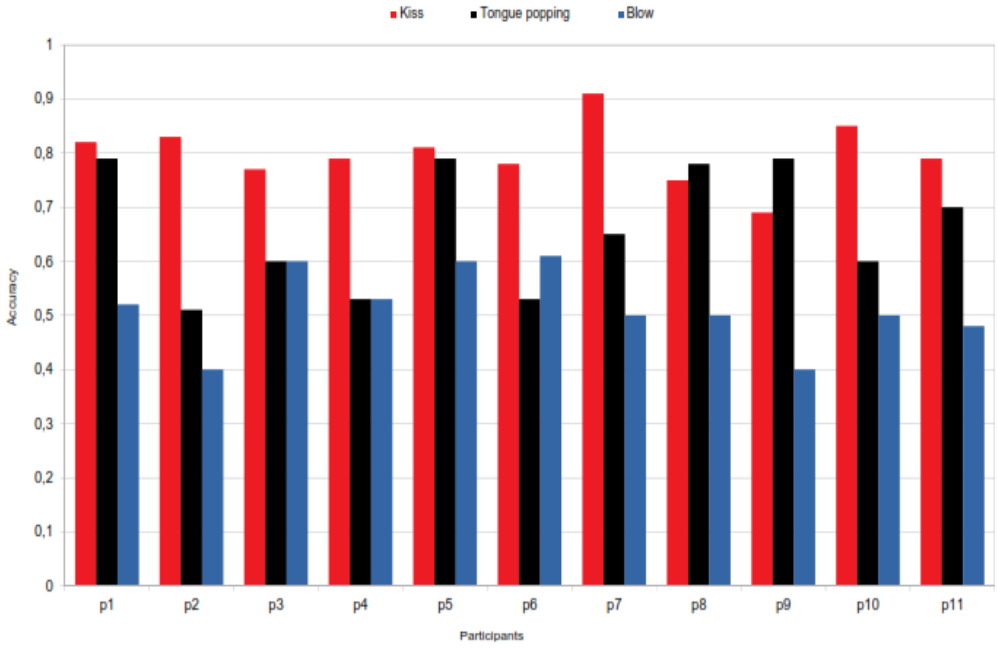

Figure 8. The accuracy of video signals reference to each non-articulatory sounds produced by participants.

Figure 8 reveals that the blowing video recognition was the worse, because this movement is similar to the kiss movement in terms of spectral density. For the blow sound, the median of accuracy achieved 51\%, tongue popping 66\% and kiss 81\%. Therefore, the recognition of the kiss video is more efficient than blow and tongue popping video, since its accuracy increase in 29\% and 14\%, respectively.

Based on outcomes, the hypothesis was confirmed: *video information can complement audio recognition and assist non-articulatory sound recognition in a real home environment*.

6. RELATED WORK

The speech produced by individuals with DS is limited due to their anatomical and motor specificities. Although some of these limitations of DS cannot be overcome, early speech training can improve the quality with which individuals with DS communicate. In this context, some studies have examined speech production by focusing on articulatory sounds made by children with DS [29-31].

In [29] is analyzed a system to assess and practice sounds aiming to improve the language ability of children with DS. This system uses the language ability of children with DS as input to generate graphs and to provide each child with suitable training. Moreover, the system acts as a DS information provider and a child data manager for parents and trainers.

In [30] was analyzed whether the addition of vision to audition can improve the intelligibility of speech produced by individuals with DS. Felix et al. [31] presented a computer-assisted learning tool for children with DS that uses mobile computing, multimedia design, and computer speech recognition to improve reading and writing abilities in Spanish through speech and drawing activities like letter identification, reading, spelling and handwriting.

Some efforts have been directed toward improving the speech produced by children with DS. However, the most of these efforts focus on systems for articulatory sounds that address phoneme and word production. These sounds are easier to recognize than non-articulatory sounds. Therefore, systems to deal with non-articulatory sound recognition are fundamental because they must serve as preparatory activities for speech production by children with speech disorder, justifying the importance of our study.

7. FINAL REMARKS

An Audio-VIdeo Speech recognition system (AVIS) based on machine learning and neural networks for non-articulatory sound recognition was created and experimented. AVIS intends to be part of a mobile application that assists children with Down Syndrome during their training for the production of speech.

Two experiments were conducted to assess the complementarity in audio and video recognition systems. We exploited MFCC and Laplace transform for audio recognition and, CNN and LSTM for video recognition. As expected, the controlled scenario was better than *TV sound* and *Rain sound*. Furthermore, the kiss precision was better than the others with 83\%. Moreover, the results indicated that video features effectively aid recognition of non-articulatory sounds

the. So, video information can complement audio recognition of non-articulatory sounds.

In the video analysis, we employed a LSTM for classification and a CNN for features extraction which is considered state-of-the-art and can be used with other classification techniques as well. Our visual feature extraction mechanism based on CNN and RNN effectively predicted the kisses and it was reasonably for blow sound and tongue popping samples. As a result, we can observe that a method for visual speech recognition totally depends on a large dataset for training, mainly if it works with no conventional sounds and face movements.

As future work, will be planned: *(i)* a new experiment with a protocol to collect the sounds and a larger number of participants with DS; *(ii)* an investigation of new non-articulatory sounds to be recognized in the same classification space; and *(iii)* the development of a mobile speech recognition prototype for children with DS.

8. ACKNOWLEDGEMENTS

Authors are grateful to CNPq (442533/2016-0) and FAPESP (2016/13206-4) for the funding.

9. REFERENCES

[1] Yu D, Deng L. Automatic speech recognition: A deep learning approach. Springer Publishing Company; 2014.

[2] Hughes S. Bullying: what speech-language pathologists should know; 2014.

[3] Sakoe H, Chiba S. Readings in speech recognition. Chapter Dynamic Programming Algorithm Optimization for Spoken Word Recognition; 1990. p. 159–165.

[4] Han W, fat Chan C, sing Choy OC, et al. An efficient mfcc extraction method in speech recognition. In: IEEE International Symposium on Circuits and Systems (ISCAS). IEEE; 2006.

[5] Alatwi A, So S, Paliwal KK. Perceptually motivated linear prediction cepstral features for network speech recognition. In: 10th International Conference on Signal Processing and Communication Systems, ICSPCS 2016, Surfers Paradise, Gold Coast, Australia, December 19-21, 2016; 2016. p. 1–5.

[6] Wang JC, Lee YS, Lin CH, et al. Robust environmental sound recognition with fast noise suppression for home automation. IEEE Transactions on Automation Science and Engineering. 2015 Oct;12(4):1235–1242.

[7] Yan X, Li Y. Anti-noise power normalized cepstral coefficients for robust environmental sounds recognition in real noisy conditions. In: 2012 Fourth International Conference on Computational Intelligence and Communication Networks; Nov; 2012. p. 263–267.

[8] Petridis S, Li Z, Pantic M. End-to-end visual speech recognition with lstms. arXiv preprint arXiv:170105847. 2017; 1–5.

[9] Peelle JE, Sommers MS. Prediction and constraint in audiovisual speech perception. Cortex. 2015; 68 (Supplement C):169–181.

[10] Noda K, Yamaguchi Y, Nakadai K, et al. Audio-visual speech recognition using deep learning. Applied Intelligence. 2015;42 (4): 722–737.

[11] Zhou Z, Zhao G, Hong X, et al. A review of recent advances in visual speech decoding. Image and Vision Computing. 2014; 32(9): 590–605.

[12] Gritzman AD, Aharonson V, Rubin DM, et al. Automatic

computation of histogram threshold for lip segmentation using feedback of shape information. Signal, Image and Video Processing. 2016; 10(5): 869–876.

[13] Heidenreich T, Spratling MW. A three-dimensional approach to visual speech recognition using discrete cosine transforms. arXiv preprint arXiv:160901932. 2016;1–27.

[14] Stewart D, Seymour R, Pass A, et al. Robust audio-visual speech recognition under noisy audio-video conditions. IEEE transactions on cybernetics. 2014; 44(2):175–184.

[15] Wu P, Liu H, Li X, et al. A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion. IEEE Transactions on Multimedia. 2016; 18(3): 326–338.

[16] Heckmann M. Audio-visual word prominence detection from clean and noisy speech. Computer Speech & Language. 2018; 48(Supplement C): 15–30.

[17] Courant R, Hilbert D. Methods of mathematical physics. Vol. 1. Interscience; 1953.

[18] Jothilakshmi S, Ramalingam V, Palanivel S. Unsupervised speaker segmentation with residual phase and mfcc features. Expert Systems with Applications. 2009; 36(6): 9799 – 9804.

[19] Mather P, Tso B. Classification methods for remotely sensed data. CRC press; 2016.

[20] Chien-Chang L, Shi-Huang C, Trieu-Kien T, et al. Audio classification and categorization based on wavelets and support vector machine. IEEE Transactions on Speech and Audio Processing. 2005; 13(5): 644–651.

[21] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. In: Proceedings of the IEEE; 1998. p. 2278–2324.

[22] Furlaneto DC. An analysis of ensemble empirical mode decomposition applied to trend prediction on financial time serie Mestrado em ciência da computação. Curitiba, PR, Brasil: Universidade Federal do Paraná; 2017.

[23] Donahue J, Hendricks LA, Rohrbach M, et al. Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Anal Mach Intell. 2017; 39(4): 677–69.

[24] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks. In: Proceedings of the 31st International Conference on Machine Learning; 2014. ICML'14.

[25] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2; 2014. p. 3104–3112; NIPS'14.

[26] Cho K, Merrienboer BV, Bahdanau D, et al. Towards end-to-end speech recognition with recurrent neural networks. In: Proceedings of the 31st International Conference on Machine Learning; 2014. ICML'14.

[27] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011; 12: 2825–2830.

[28] Wohlin C, Runeson P, H¨ost M, et al. Experimentation in software engineering: An introduction. 1st ed. Springer-Verlag Berlin Heidelberg; 2012.

[29] Kuan TM, Jiar YK, Supriyanto E. Language assessment and training support system (latss) for down syndrome children under 6 years old. WSEAS Transactions on Information Science and Applications. 2010;7(8):1058-1067.

[30] Hennequin A, Rochet-Capellan A, Dohen M. Auditory-visual perception of vcvs produced by people with down syndrome: Preliminary results. In: 17th Annual Conference of the International Speech Communication Association (Interspeech 2016); 2016.

[31] Felix VG, Mena LJ, Ostos R, et al. A pilot study of the use of emerging computer technologies to improve the e_ectiveness of reading and writing therapies in children with down syndrome. British Journal of Educational Technology. 2017;48(2):611-624.