information

Article

The BioVisualSpeech Corpus of Words with Sibilants for Speech Therapy Games Development

Sofia Cavaco [1,*], Isabel Guimarães [2,3], Mariana Ascensão [2], Alberto Abad [4,5], Ivo Anjos [1], Francisco Oliveira [4], Sofia Martins [1], Nuno Marques [1], Maxine Eskenazi [6], João Magalhães [1] and Margarida Grilo [2,*]

[1] NOVA LINCS, Department of Computer Science, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, 2829-516 Caparica, Portugal; i.anjos@campus.fct.unl.pt (I.A.); str.martins@campus.fct.unl.pt (S.M.); nmm@fct.unl.pt (N.M.); jmag@fct.unl.pt (J.M.)
[2] Escola Superior de Saúde do Alcoitão, Rua Conde Barão, Alcoitão, 2649-506 Alcabideche, Portugal; isabel.guimaraes@essa.scml.pt (I.G.); mariana.ascensao@essa.scml.pt (M.A.)
[3] Clinical Pharmacological Unit, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Av. Prof. Egas Moniz, 1649-028 Lisboa, Portugal
[4] INESC-ID, Rua Alves Redol 9, 1000-029 Lisboa, Portugal; alberto.abad@inesc-id.pt (A.A.); francisco.campos@ist.utl.pt (F.O.)
[5] Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1 1049-001 Lisboa, Portugal
[6] Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA; max@cs.cmu.edu
* Correspondence: scavaco@fct.unl.pt (S.C.); margarida.grilo@essa.scml.pt (M.G.)

Received: 4 August 2020; Accepted: 23 September 2020; Published: 2 October 2020

Abstract: In order to develop computer tools for speech therapy that reliably classify speech productions, there is a need for speech production corpora that characterize the target population in terms of age, gender, and native language. Apart from including correct speech productions, in order to characterize the target population, the corpora should also include samples from people with speech sound disorders. In addition, the annotation of the data should include information on the correctness of the speech productions. Following these criteria, we collected a corpus that can be used to develop computer tools for speech and language therapy of Portuguese children with sigmatism. The proposed corpus contains European Portuguese children's word productions in which the words have sibilant consonants. The corpus has productions from 356 children from 5 to 9 years of age. Some important characteristics of this corpus, that are relevant to speech and language therapy and computer science research, are that (1) the corpus includes data from children with speech sound disorders; and (2) the productions were annotated according to the criteria of speech and language pathologists, and have information about the speech production errors. These are relevant features for the development and assessment of speech processing tools for speech therapy of Portuguese children. In addition, as an illustration on how to use the corpus, we present three speech therapy games that use a convolutional neural network sibilants classifier trained with data from this corpus and a word recognition module trained on additional children data and calibrated and evaluated with the collected corpus.

Keywords: sibilant consonants; children's speech corpus; speech sound disorders; serious games for speech and language therapy

1. Introduction

While most children learn how to speak in their native language and learn how to correctly produce the native language phonemes by the expected ages, for some children, the language acquisition process may be challenging [1]. As reported by Guimarães et al. for data on European Portuguese (EP), 8.8% of preschool-aged children suffer from some type of speech sound disorders (SSD) [2]. Many children can surpass their language acquisition difficulties as they grow older and their speech organs develop, but for some children, the speech distortions are not surpassed naturally. These children may need professional help to correct their SSD. Besides, it is important to address these difficulties as early as possible since SSD can affect the child's quality of life and literacy aquisition [3,4].

Sigmatism is a SSD that consists of pronouncing the sibilant consonants incorrectly. The sibilants, which include sounds like [s] in serpent and [z] in zipper, are consonants that are generated by letting the air flow through a very narrow channel towards the teeth [5]. Sigmatism is a very common SSD among children with different native languages [6,7], including EP [8,9].

Speech and language pathologists (SLPs) help children with sigmatism correct the production of sibilants with speech exercises that start with the isolated sibilants and then progress to the production of the sounds within syllables and words. While the repetition of speech exercises is important to practice and master the correct production of speech sounds, it may lead to the child's weariness and lack of interest on proceeding with the speech exercises. In order to keep children motivated and collaborative during the therapy sessions, SLPs need to adapt the speech and language exercises into fun and appealing activities.

This work is part of the BioVisualSpeech research project, in which we explore multimodal human computer interaction mechanisms for providing bio-feedback in speech and language therapy through the use of serious digital games. As a contribution to help SLPs motivating children to repeat the tasks that may lead to the correction of their speech disorder, we have been developing serious games for training the production of EP sibilants, which are controlled by the child's voice (more details in Sections 5.1–5.3). To make this possible, we have been developing machine learning approaches that are integrated into the games to detect the incorrect production of sibilant sounds [10] or to validate the correctness of produced words to help SLPs on assessing if the child has sigmatism (Section 5.4).

In order to develop the automatic speech processing modules for our serious games for sigmatism, we built a corpus of children's speech that was previously proposed in Reference [11]. This corpus contains isolated sibilants productions and productions of words with sibilants. Here, we focus on the data set of words with sibilants. The data set contains children's productions of 70 different EP words with sibilant consonants. The sibilant phoneme in these words occurs either at the start, middle or final position. The word productions were recorded in three schools, and 365 children from 5 to 9 years of age participated in the data collection task. One of the novelties of this work is that the data annotations include information on the quality of the sound productions according to SLPs criteria. Another novelty is that the set of chosen words focuses on the EP sibilant consonants.

In addition to proposing this EP corpus of words with sibilants, we illustrate how to use the corpus in the development of speech and language therapy games that address sigmatism. The first game can be used with the isolated sibilants therapy exercise, in which the child must produce isolated sibilants. In addition, the second game can be used to train the production of words that start with a sibilant consonant and to identify words that start with the same sibilant, while the third game allows for training the production of words containing one or more sibilant in a varying number of configurable positions and difficulties. The games are controlled by the child's speech and give visual feedback on the child's production. In this way, the games motivate the child on performing these speech exercises and also help the child understand when his/her sibilant productions are not correct. As an option, one of the games gives visual feedback on the point of articulation used for the speech production and on the use of the vocal folds. This visual feedback helps the child understand what he/she must do to correct the sibilant production.

The games use an EP sibilant classifier and word recognition module for children speech. The sibilant classifier is a convolutional neural network (CNN) classifier that uses two models: one that is able to distinguish sounds made with different points of articulation, and another that distinguishes between voiced and voiceless sounds. By combining the output of both models, the classifier can distinguish the EP sibilant consonants. The word recognition module is an automatic speech recognition system adapted to children speech and configured to operate in keyword spotting mode [12].

After discussing related work on the collection of children's speech productions of the EP sibilants in Section 2, and giving a short introduction to the EP sibilants in Section 3, this paper discusses the proposed corpus of words with sibilants and the protocol used to record it in Section 4. Section 5 presents the three game examples and the automatic speech processing modules developed for these games.

2. Related Work

The availability of speech resources with characteristics similar to the intended application—more specifically, corpora containing manually annotated speech—is of critical importance for the development of robust speech analysis tools. In particular, modern approaches based on deep neural networks depend on large amounts of training data. To the best of our knowledge, there has not been any previous large scale effort to collect EP speech from children in sibilant production tasks as the one targeted in BioVisualSpeech. Nevertheless, there has been some previous remarkable efforts to collect EP speech recordings of children in a variety of reading tasks. This is the case of the LetsRead database [13] and the CNG database [14]. While none of these corpora contain detailed phonetic annotations in spontaneous naming tasks, these two resources are still extremely valuable for the development of baseline speech and language technological modules tailored for the child population, which can eventually be adapted to specific tasks and/or to atypical children speech characteristics in a later stage.

The LetsRead database contains reading aloud recordings of 284 children, whose mother tongue is EP: 147 girls and 137 boys, distributed from 1st to 4th grade. Data of 104 participants, 58 girls and 46 boys, are manually annotated, equally distributed in the 4 grades (26 per grade) and correspond to approximately 5 h and 30 min of speech. The remaining data were automatically aligned with the prompts. The recording sessions were carried out in a classroom with low reverberation and low noise acoustic characteristics. Prompts of 20 sentences and 10 pseudo-words were presented to the children for reading, through a specially developed interface. The sentences used in these recordings were extracted from children fairy tales and grade-specific scholar books. The difficulty degree of the sentences was evaluated accordingly to their phonetic complexity and variety.

The CNG corpus contains reading aloud recordings from 510 children, whose maternal language is EP: 285 girls and 225 boys, distributed from ages 3 to 6 (153 participants) and from ages 7 to 10 (357 participants). The recording sessions were carried out at a room with low reverberation and low noise. The prompts were presented to the children for reading, through a specially made interface called "Your Speech". In total, 30 prompts were used for the 3 to 6 age group and 50 prompts for the 7 to 10 age group. The prompts were chosen from a set of four types of prompts: 292 phonetically rich sentences (CETEMPúblico corpus, https://www.linguateca.pt/CETEMPublico/); musical notes; isolated cardinals; and sequential cardinals.

3. Sibilant Consonants and Speech Sound Disorders

Fricative sounds are produced by letting a small amount of air pass through a narrow channel in the vocal tract. Different fricative sounds are made by using specific parts of the vocal tract and specific tongue shapes to configure the narrow channel. The vocal folds may be used, resulting in a voiced consonant. When the vocal folds are not used, the result is a voiceless consonant.

Sibilant sounds are a subset of the fricatives. There are two types of EP sibilant consonants: the alveolar sibilants, which are produced with the tongue nearly touching the alveolar region of the mouth, and the palato-alveolar sibilants, which are produced by positioning the tongue towards the palatal region of the mouth (Figure 1). The EP sibilant consonants are: [z] as in zebra, [s] as in snake, [ʃ] as the *sh* sound in sheep, and [ʒ] as the *s* sound in Asia [15] (international phonetic alphabet, IPA, symbols [16]). [z] and [s] are both alveolar sibilants, while [ʃ] and [ʒ] are palato-alveolar sibilants. Both [z] and [ʒ] are voiced sibilants, and [s] and [ʃ] are voiceless sibilants.

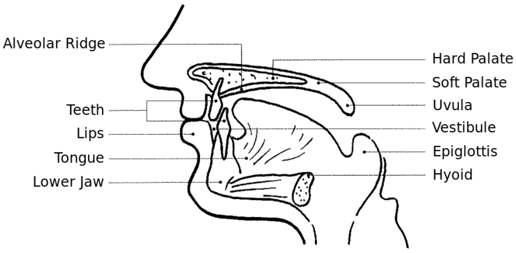


Figure 1. Main places of articulation in the vocal tract, adapted from Reference [5].

Other (non-sibilant) fricative sounds include the labiodental fricatives [f] and [v], which are produced with the lower lip nearly touching the upper front teeth [5]. [f] is voiceless, and [v] is voiced (Table 1).

Table 1. Classification of European Portuguese fricatives by place of articulation and use of vocal folds.

Use of	Place of Articulation		
Vocal Folds	**Labiodental**	**Alveolar**	**Palato-Alveolar**
Voiceless	[f]	[s]	[ʃ]
Voiced	[v]	[z]	[ʒ]

There are a few variations of the rhotic consonant *r* in EP, some of which are fricative sounds [17,18]. The rhotic consonant is the sound of *r* at the word initial position, like in *rato* (mouse), or double *r* in a medial position, such as in *carro* (car). This consonant is commonly pronounced as a voiced uvular fricative [ʁ] [15]. Other less common variations include the voiceless uvular fricative [χ] and the voiceless velar fricative [x]. This consonant can also be pronounced as a (non-fricative) trill sound [5]: it can be an alveolar trill, [r], which is made with vibrations of the tip of the tong against the upper alveolar ridge for longer than two or three periods, and an uvular trill, [ʀ], which is done by a vibration of the palatine uvula (Due to its different nature, this consonant is not included in Table 1).

SSD in sibilant consonants can occur due to oral structural problems, poor phonological awareness, or developmental disorders and may occur in different types of errors and phonological representations [3,4]. Difficulty in learning to produce and/or use sibilant sounds correctly can be manifested in a variety of types, and these can be classified as distortions, typical syllable structure errors (e.g., final consonant deletion), typical segmental errors (e.g., /s/ Y [t]), and atypical syllable structure errors (e.g., initial consonant deletion) (2). As distortion errors typically reflect an alteration in the production of a sound (e.g., a slight problem with tongue shape or placement, such as dentalized or lateralized [s]) are prevalent in SSD [3,19,20], this study aimed to develop clinical tools for SLP.

4. The Corpus of Words with Sibilants

The BioVisualSpeech EP sibilants corpus was built as part of a speech and language screening activity that took place in three schools. The screening activity had two purposes: (1) to assess children's speech in order to detect cases of SSD and (2) to collect data to build corpora of children's EP speech, which can be used for speech and language therapy and computer science research purposes, and to develop computer tools to assist speech and language therapy. In fact, while the data described here focuses on the sibilants, the screening activity assessed all EP consonant sounds and included several speech and orofacial exercises that helped the SLPs to detect not only sigmatism but also other SSD cases. In addition, as seen in Section 5.4.2, our sibilants classifier is trained with samples from the EP sibilants but also samples from the fricatives [f] and [v]. Section 4.1 discusses details about the screening activity.

Some of the exercises performed during the screening activity were recorded. In particular, we recorded an exercise in which the children produced isolated sibilants and another in which the children were prompted to say words with sibilant occurrences. These words were used to build the BioVisualSpeech EP corpus of words with sibilants. Section 4.2 describes in more detail the set of words and the protocol used to record the sibilant word productions. While here we focus on the words with sibilants data set, for more details on the isolated sibilants see Reference [11].

The recorded word productions were annotated according to SLPs criteria. The annotation task is discussed in Section 4.3.

4.1. The Screening Activity

The screening activity, which took six months, was done in three schools in the district of Lisbon (Portugal). The participating children were all either at pre-school or primary school. We obtained an informed consent from all parents or legal guardians, and the ethics approval was provided by the ethics committee of Escola Superior de Saúde do Alcoitão, Santa Casa da Misericórdia de Lisboa (process number 001/2017).

Several SLP graduate students participated in the screening task, under the supervision of a senior SLP from the BioVisualSpeech team. This means that different children could have interacted with different SLPs during the data collection and screening. During the screening task, the SLPs participating in the study filled in an individual report for each child to inform the parents or legal guardians about the results of the screening.

Our aimed age group was from 5- to 9-year-old children. There were 356 children from this age group participating in the study. Whilst data on age-appropriated speech sound production for EP speaking children between 3 and 6 and 11 years old is available, there have been limited studies including older EP speaking children [2,21]. Nonetheless, evidence for other languages supports the cutoff age between 8 and 9 years old for typical speech sound acquisition to be completed [22].

Children were assessed individually in a quiet room at their school setting by an SLP or an SLP graduated student. One or two other adults (SLP graduate students or researchers from the BioVisualSpeech team) could be present at the room, but only the SLP in charge gave the necessary instructions and interacted with the children. Each child had two different screening moments, on different days, to avoid that the children got tired and also to make shorter interruptions of their normal school day.

4.2. Data Collection of Words with Sibilant Consonants

The equipment used to collect the sibilants data consisted of a dedicated unidirectional condenser microphone, a portable battery powered digital audio tape (DAT) recorder (Sony TCD-D8) and acoustic foam to attenuate background noise (Figure 2). Due to the children's age group, we did not use head mounted microphones. The recordings were made in a reasonably quiet room at the schools, but, in many cases, it was still possible to hear the noise coming from the playground and corridors. While the recording conditions were not perfect, having background noise in the data samples is appropriate for our goal since we aim to develop automatic recognition models that are robust enough to be used in SLP's offices or at schools. The data was recorded with a 44,100 Hz sampling rate.

The data was recorded continuously, that is, the recorded speech signals include the SLP and the child's speech. Thus, after the data collection task was finished, we had to segment all the recorded speech signals, not only to extract the children relevant speech portions but also to discard all the speech data from the SLPs.

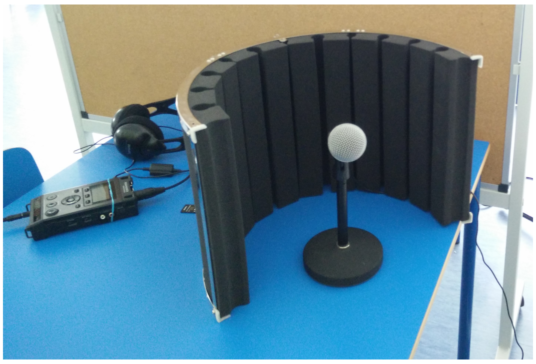


Figure 2. Equipment used for the recordings: a digital audio tape (DAT), a microphone, and acoustic foam.

We used a total of 70 words with sibilant consonants. The chosen words start with one of the four sibilant consonants (e.g., *sino*), have the sibilant in a middle position (e.g., *pijama*), or finish with the sibilant [ʃ] (e.g., *livros* or *peixe*). Ten of these words contain more than one sibilant phoneme (like *cereja*, in which the initial phoneme is [s] and there is a middle phoneme with [ʒ]). The 70 chosen words have 81 sibilant-phoneme occurrences (Table 2). The number of occurrences is not equal for all sibilants because these words were chosen taking into consideration their frequency of appearance in EP, and their semantic predictability [23,24]. In addition, [ʃ] is the only sibilant that can occur in word final positions in EP.

Table 2. Words with sibilants.

Phoneme	Initial Position	Middle Position	Final Position	Total
ʃ	6	20	13	39
ʒ	5	5		10
s	13	11		24
z	2	6		8

In order to have the children produce these words, the stimuli consisted of age-appropriate color images representing the words. We used plain images in a white background and printed in A5 paper (one image per paper sheet) to direct the attention to the aimed word as much as possible. As an example, Figure 3 shows the image used for one of the words (*mochila*, which is the EP word for backpack and which contains the [ʃ] phoneme in a middle position). In order to have the child pronouncing the word in his/her usual way and not having the child mimicking the SLP pronunciation, the SLP did not say the aimed word. Each picture was shown to the child, who was asked to name it. If the child did not answer, the examiner could give standardized semantic clues.

In total, we collected 22,830 word samples from these 70 words from the productions of 356 children (Tables 3 and 4). There are 20,198 correct word productions and 2632 word samples incorrectly produced. From these incorrectly word productions, there are 1138 word samples with incorrectly produced sibilants (fourth column of Table 3). In each word with sibilant phoneme x (where x is [ʃ], [ʒ], [s], or [z]), we considered that the sibilant is incorrectly produced when the word contains the sibilant phoneme x, but the child speech production does not contain x, substitutes or has a distortion of x. While this is a good approximation of the real number of word samples with incorrect sibilant productions, it may fail in some cases in which the word production contains the x phoneme but in a different word position/syllable. In addition, note that the total number of words for all sibilants presented in the table (22,830) does not result from adding the numbers in the fifth column since some words can contain more than one sibilant sound.

Figure 3. Stimulus used to suggest the word *mochila*.

Table 3. Word samples.

Phoneme	Incorrect Phoneme Occurrences	Total Phoneme Productions	Word Samples with Incorrect Sibilants	Total Number of Word Samples
ʃ	434	11,202	422	11,455
ʒ	240	2880	240	3333
s	354	6627	322	7024
z	154	2260	154	2616
Total	1182	22,969	1138	22,830

Table 4. Number of children.

Age	Girl	Boy	Total
5	20	19	39
6	35	35	70
7	51	33	84
8	39	50	89
9	37	37	74
Total	182	174	356

The second column of Table 3 shows the number of incorrect sibilant productions. The third column of this table shows the total number of occurrences (correct and incorrect) of each sibilant phoneme. Note that one word can have more than one sibilant production; thus, the total presented in this column (22,969) is higher than the total number of word samples (22,830).

4.3. The Annotation Task

The data was annotated according to SLPs criteria. The first annotation phase was done by the SLP in charge of the recordings and took place during the data collection task. The second annotation phase took place after the data collection task was finished and was done by an SLP and a software engineer. All annotations done by the software engineer were then verified by the SLP.

The annotation during the data collection consisted on marking each target word as either being produced, produced after repetition, or not produced by the recorded child. In the cases where the word was produced, the uttered word was annotated as a correct pronunciation or as an incorrect one. Moreover, each consonant group was labeled as correctly or incorrectly pronounced. Incorrect production was considered (i) for no response or (ii) for detected consonant omission, substitution or distortion. This process was useful for the individual screening reports sent to the children's parents or legal guardians and for speech-language pathology research purposes. All annotations in this phase were verified by a senior SLP from the team.

In order to make the recorded data and annotations useful for computerized methods, it was necessary to follow a second annotation stage. In this annotation stage, each recording was manually time aligned to identify the start and end boundaries of each word. Segments containing speech from the SLPs or other not interesting events were marked to be later rejected. Moreover, for each of the wrongly uttered words identified in the first stage, the speech assessment methods phonetic alphabet (SAMPA) transcriptions were produced.

The following example illustrates how the word *mochila* was annotated for a wrongly uttered occurrence of this word. The annotation includes the start and end times of the word within the speech sound file, the indication that the word is wrongly uttered (with the symbol * at the end of the word transcription, *mochila**) and the SAMPA transcription *m@sil6* (a correct occurence of the word would be annotated as *muSil6*, which means that this particular example has two pronunciation errors, one for the sibilant [ʃ] and another for the vowel sound [u]):

```
<Turn speaker="spk1" startTime="18.170" endTime="19.129">
<Sync time="18.170"/>
mochila*
<Event desc="m@sil6" type="pronounce" extent="previous"/>
</Turn>
```

5. Games for Sigmatism

In order to correct the sibilant distortion errors, SLPs use different speech exercises during the speech and language therapy sessions. They usually start assessing the child's capacity of distinguishing and producing the isolated sibilant consonants, and then proceed to have the child practice the production of the sibilants as isolated sounds. This consists of the *isolated sibilants exercise*, in which the child produces each sibilant with short and/or long duration. The main goal of this exercise is to teach the child to distinguish and correctly produce the different sibilant consonants.

At the next stage, SLPs use the *isolated sibilants exercise* for multiple alternate productions of the different sibilants, in which the SLP asks for sounds that alternate the point of articulation and the use of the vocal folds. Once the child can say the isolated sibilants correctly, the therapy activities can proceed to more complex exercises that use the sibilant consonants within words.

In the BioVisualSpeech project, we have developed different speech therapy games, some of which focus on helping children master the production of the EP sibilants (Figure 4). In particular the *VisualSpeech isolated sibilants game*, the *VisualSpeech pairs game* and the *BioVisualSpeech word naming game* give visual feedback on the child's sibilant production performance. The BioVisualSpeech isolated sibilants game uses the isolated sibilants exercise and can be used in the initial speech and language therapy stages (Section 5.1). On the other hand, the BioVisualSpeech pairs game is a cards game that can be used to practice the production of words with sibilants, and therefore, is aimed at subsequent stages of speech and language therapy (Section 5.2). The BioVisualSpeech word naming game—also designed for the subsequent stages of therapy—consists of a series of word naming exercises of varying difficulty containing sibilant sounds (Section 5.3).

An important characteristic of these games is that they are controlled by the child's voice. The games process the child's speech productions in real time and the sequence of game actions are determined by the quality of these productions. Thus, unlike with other speech and language therapy computer games that are manually controlled by the SLP, in these games the main character movement or the sequence of actions are controlled by the child's voice. In this way, the games give real time visual feedback about the sound production, which is an intuitive way of pointing out to the child whether his/her sound productions are correct.

In order to react to child's speech productions, the games use an isolated EP sibilants classifier trained with data extracted from the BioVisualSpeech EP sibilants corpus and a word recognition module for EP children trained with additional data. These modules are described in more detail in Section 5.4.

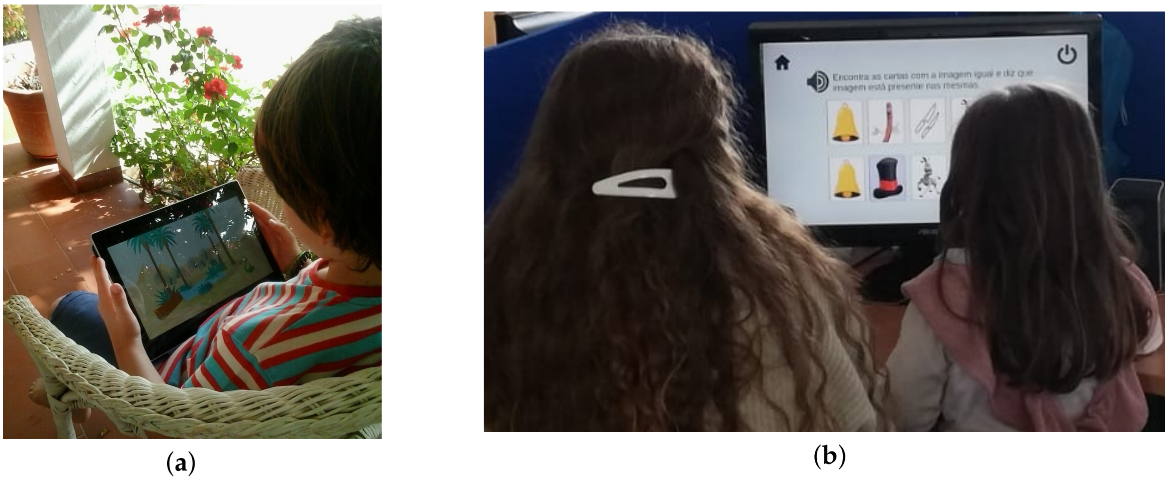

Encontra as cartas com a imagem igual e diz que imagem está presente nas mesmas.
(a)
(b)

Figure 4. The BioVisualSpeech games for sigmatism. (**a**) Child playing the isolated sibilants game. (**b**) Child playing the pairs game with the help of an adult.

5.1. The BioVisualSpeech Isolated Sibilants Game

The BioVisualSpeech isolated sibilants game was designed to help children train the production of the four EP sibilant consonants [25,26]. This game implements the isolated sibilant therapy exercise. In order to play the game, children have to do this speech therapy exercise.

The game has a different scenario and character for each of the four EP sibilants (Figure 5a). The game goal is to lead the character from its initial position to a target. The characters or goal were chosen to give visual cues on the sibilant sounds. More specifically, each scenario uses a character or goal whose EP name starts with the sibilant that must be used in that scenario: a bumblebee (*zangão* in EP) for the [z] sibilant, a serpent (*serpente* in EP) for the [s] sibilant, a ladybug (*joaninha* in EP) for the [ʃ] sibilant, and a boy running away from the rain (*chuva* in EP) for the [ʒ] sibilant.

In order to make the main character reach its goal, the child has to correctly produce the sibilant sound for that scenario (e.g., the [s] sound for the serpent scenario). The character only moves towards the target when the child correctly produces the sibilant. In this way, the movement of the character gives visual feedback about the child's speech performance and helps the child understand if he/she is producing the sibilant correctly.

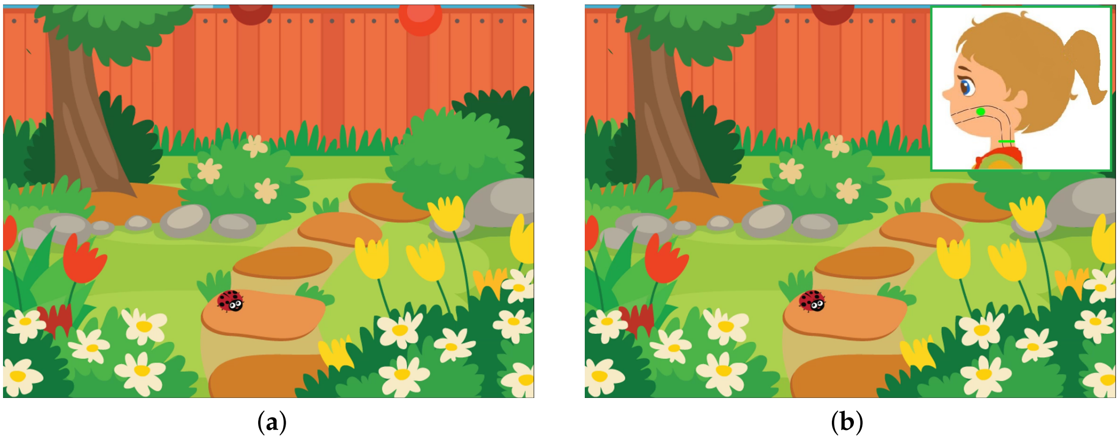


Figure 5. The BioVisualSpeech isolated sibilants game. The scenario for the [ʒ] sibilant in (**a**) plain mode and in (**b**) vocal tract feedback mode.

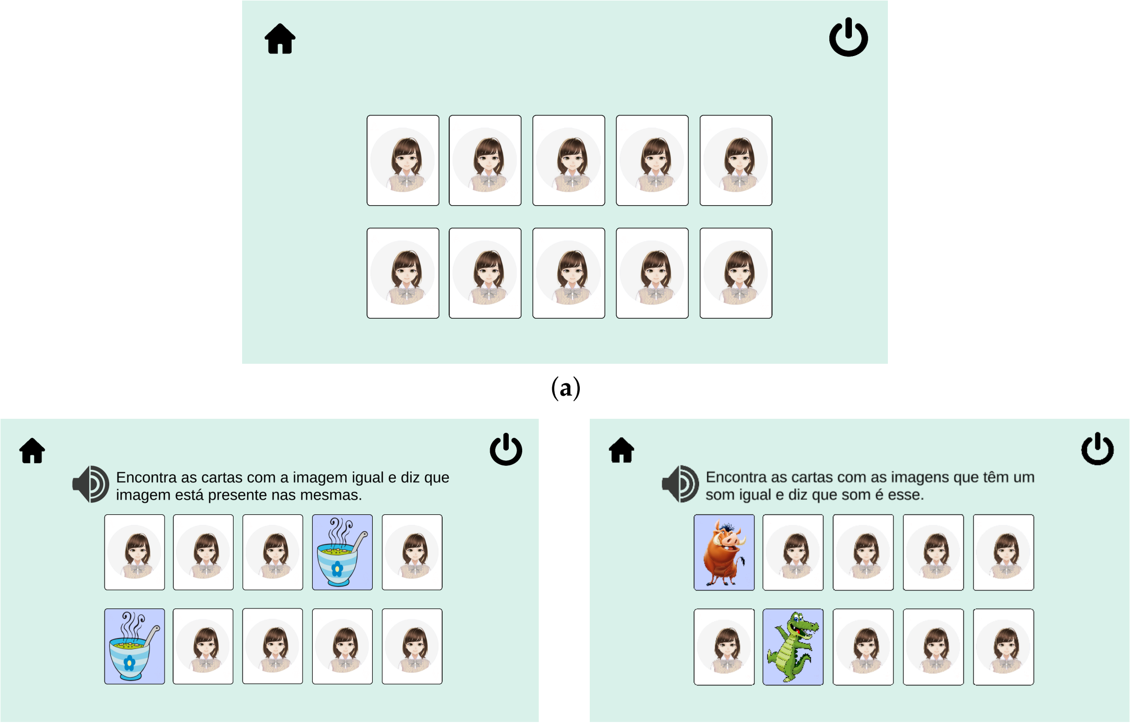
In addition to the visual feedback given by the character's movement, this game's current version includes the option of giving additional visual feedback about the use of the vocal tract. The user has the option of watching the face of a child character that gives feedback on the point of articulation and use of vocal folds (Figure 5b). The child character has a dot on the place of articulation that the child used. In addition, a line is drawn in the throat indicating if the vocal folds are used. This is a sinusoidal line when the vocal folds are used and a straight line otherwise. The dot and line are green if the place of articulation and use of vocal folds is correct, and red otherwise. Figure 5b shows an example for the correct production of the [ʒ] sibilant.

Both the plain mode and the vocal tract feedback mode use the classifier described in Section 5.4.2 to decide if the speech production is correct. In addition, the vocal tract feedback mode uses the output from that classifier to determine the point of articulation used by the child and if the child used the vocal folds.

5.2. The BioVisualSpeech Pairs Game

The BioVisualSpeech pairs game is based on the well known cards game in which the player has to find pairs of equal cards. Here, the cards game is adapted for speech and language therapy.

The BioVisualSpeech pairs game has different levels that can be used to help children with sigmatism. In particular, it motivates children on producing words that start with EP sibilants (level for *matching cards*) and helps them on identifying words that start with the same EP sibilant (level for *matching sounds*). Both levels start with all the cards facing down (Figure 6a), and the child can choose pairs of cards by clicking on them. The images used in these two levels start with sibilant consonants. (Other levels are being developed that contain words that start with other phonemes and can be used in other exercises).



(**a**)

Encontra as cartas com a imagem igual e diz que imagem está presente nas mesmas.

Encontra as cartas com as imagens que têm um som igual e diz que som é esse.

(**b**) (**c**)

Figure 6. The BioVisualSpeech pairs game. (**a**) The game starts with all cards facing down. The girl image is a character that the child can choose. (**b**) The level for *matching cards*. (The instructions in the figure are: *Find the cards with the same image and name that image.*) (**c**) The level for *matching sounds*. (The instructions in the figure are: *Find the cards with the same sound and say that sound.*) .

When the child is playing the *matching cards* level, he/she has to find pairs of cards with the same image (Figure 6b). Once the child finds two matching cards, she/he is asked to say the name of the object in the cards. The child wins that pair of cards when he/she says the correct word for the image in the cards and pronounces the initial sibilant slowly and correctly. For example, when the child finds the two soup bowls in Figure 6b, the child has to say *sopa*, which is soup in Portuguese, and must pronounce the [s] slowly and correctly.

For the *matching sounds* level, the child has to find pairs of cards with images that start with the same sibilant sound. When the child finds such a pair of cards, he/she has to say the initial sibilant for those images. The child wins that pair of cards when the sibilant production is correct. Figure 6c illustrates this level with a pair of cards that start with the EP sibilant [ʒ]: the Portuguese words for wild boar and alligator are *javali* and *jacaré*, respectively. If the child finds the pair of cards in Figure 6c, the child has correctly produce the sibilant [ʒ] to win this pair.

Just like the isolated sibilants game, the pairs game uses the classifier described in Section 5.4.2 to decide if the child's sibilant productions are correct. The *matching sounds* level gives the child's sibilant production as input to the classifier and uses its output to determine if the production is correct.

The *matching cards* level has to extract the initial phoneme from the child's speech production before sending it to the sibilant classifier. In addition, this level also sends the whole word production to the word recognition module described in Section 5.4.3 to determine if the whole word production is correct.

5.3. The BioVisualSpeech Word Naming Game

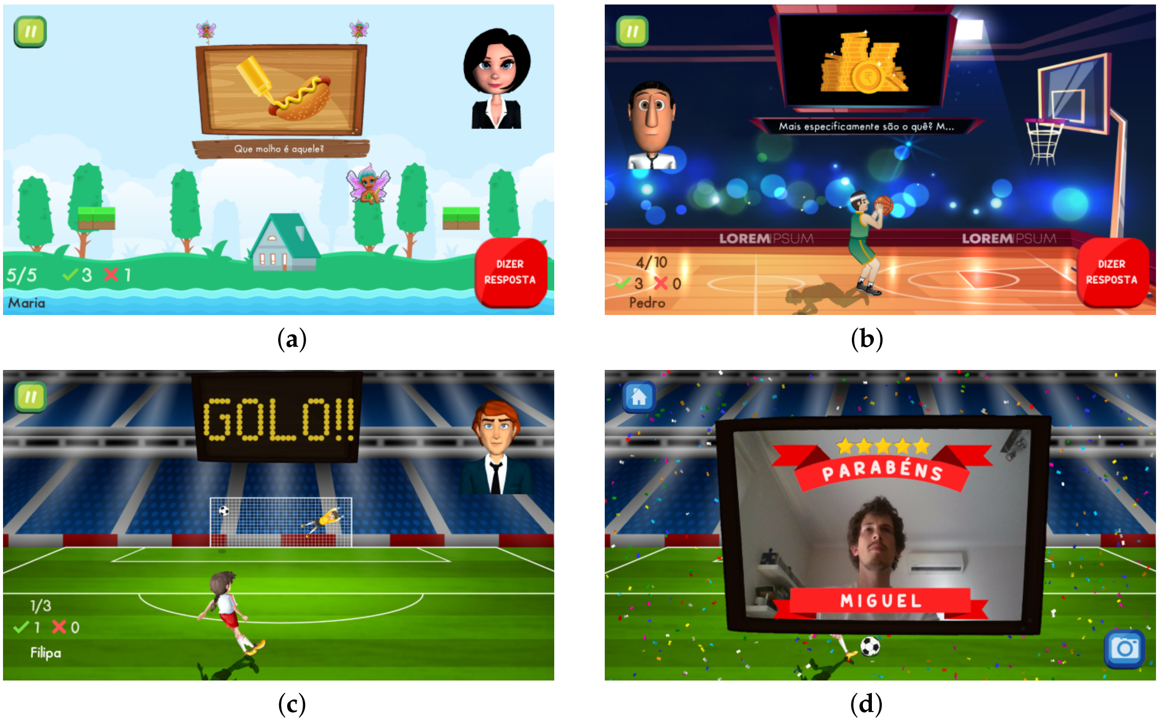
The BioVisualSpeech word naming game is a vocabulary game with words that include at least one of the four EP sibilants, designed to help children train the production of these sibilants. The basic mechanics of the game consist of showing a visual representation of a sibilant word, and the child having to say the word out loud. The objective of the game is to get the highest possible number of correct utterances out of the prompted words.

The game has three different backdrops the player can choose from; a floating fairy (Figure 7a), a basketball shootout (Figure 7b), and a penalty shootout (Figure 7c). Within each map, the image the child must name is presented in a framed rectangle in a big canvas (the mustard in Figure 7a and the coins in Figure 7b). If necessary, a text-based hint is presented alongside the image as seen in Figure 7a (the hint in this figure for the word mustard, *mustarda* in EP, says *What is this dressing?*). After the child names the object, different visual feedbacks are given depending on it being a correct or incorrect answer. In the case that a child said a synonymous word or one which was related to the image but not the desired word, an additional hint can be given which always includes the first letter of the word as seen in Figure 7b (the hint in this figure for the word coins, *moedas* in EP, says *What are these more specifically? C...*). After finishing the game, the child is given a star rating out of five and has the chance to take a picture (Figure 7d) as a form of positive feedback and to encourage engagement with the exercise.

Throughout all the stages of the game there is a 2D virtual character which acts as a "gamehost" as seen in Figure 7a–c. This character provides audio feedback to the player, complementing the visual feedback. This can be in the form of instructions on how to play the game, reading out hints, or saying what the correct word was in case the child answered incorrectly.

The objects that are presented to the child can be chosen before the game and are categorized in a way that allows the SLP to cater the session to specific characteristics that need to be trained by the specific child. These categories include word difficulty, which of the four EP sibilants is contained in the word, and the position of the sibilant within the word, among many others.

The game uses the word recognizer described in Section 5.4.3 to determine whether the word said by the child is the correct answer. In future versions of the game, we plan to incorporate a more fine-grained analysis of the production to provide specific feedback about the correctness of the sibilant contained in the word.



(**a**) (**b**)

(**c**) (**d**)

Figure 7. The BioVisualSpeech word naming game. (**a**) Fairy scenario; the host is providing a textual hint. (**b**) Basketball shootout scenario; the host is providing a follow-up hint after synonym answer. (**c**) Penalty shootout scenario; the answer is right. (**d**) Reward screen.

5.4. The EP Speech Processing Modules for Children

The BioVisualSpeech games integrate two European Portuguese automatic speech processing modules specifically developed for the particularities of children speech: a sibilant classifier and a word recognition module. The following sections describe the original sibilants classifier used in previous versions of the games (Section 5.4.1), the current sibilants classifier, which is an extension of the original sibilants classifier and that is able to determine the point of articulation and use of vocal folds (Section 5.4.2), and the word recognition module (Section 5.4.3).

5.4.1. The EP Four Class Sibilants Classifier

We compared multiple classification models for EP sibilants for our games. These included a model based on support vector machines (SVM) that used Mel frequency cepstral coefficients (MFCC) vectors as input [25]. The SVM based model was able to reach average accuracy test scores of 90.72%. The main drawback of this first model is the considerably high false negative rate of 8.73%, which means that our model classifies a considerable amount of correct sounds as false productions. This type of misclassifications can be very prejudicial to children, since they are producing a correct sound, and the game, by considering the sound incorrect, can induce them into error.

Other models included different neural network models, from simple hidden-layers artificial neural networks (ANN) to more complex 1D and 2D CNNs that used either MFCC or log Mel filterbanks as input [26]. We used the Adam optimizer as our loss function in all networks and we also used the stochastic gradient descent (SGD) in our CNNs [27].

The input to our initial 1D CNNs consists of 13 × 50 sub-matrices of the 13 × t original MFCC matrix. Our 1D convolutional model has three convolutional layers, each followed by the corresponding pooling layer. In the end, we flatten the output from the convolutional layers and use a fully connect layer with dropout, followed by another fully connect layer for the output (Figure 8). We used the rectified linear units (ReLU) as the activation functions of the convolutional and hidden layers. For the loss function, we performed multiple tests with both the Adam optimizer, and the

SGD. The biggest advantage of using the Adam optimizer, was that it allowed the model to reach local minima considerably faster than SGD. However, with the Adam optimizer the models had a tendency to overfit. So, in the end, we used SGD, since it helped to prevent overfitting, and we were able to reach similar results as with the Adam optimizer.

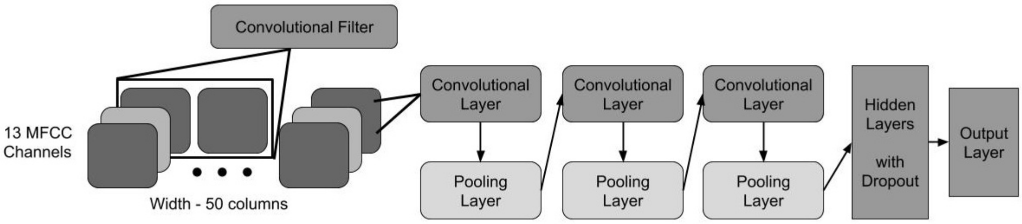


Figure 8. Representation of the 1D convolutional neural network (CNN) four class sibilants classifier.

Since log Mel filterbanks, from which the MFCCs are derived, are highly correlated both in time and frequency, they should benefit from more localized convolutions, i.e. they allow to extract more localized features from the input matrix. Thus, we used them in a 2D convolutional model instead of a 1D model, for trying to model joint correlations between time and frequency. In this model, spectro-temporal convolutions are applied across the whole 40 × 50 input matrix. The main architecture of this network is the same as above for our 1D CNN, but using 2D convolutional layers, instead of 1D layers. We use the 50 columns as our width, and 40 filters as our height, with just one channel. In this network, we use four fully connected layers with 1000, 500, 100, and 10 neurons, respectively.

Following the 1D model results with loss function, we started with the SGD optimizer for the 2D model. Again, SGD optimizer easily allowed us to prevent overfitting during training and validation for the 2D model. However, the model appeared to be converging to a local minimum. On the other hand, while it introduced some overfitting, the Adam optimizer reduced the number of epochs needed to reach the same results, and in some cases improved the results. In our testing, the overfitting was more severe with the 2D convolutional models and also more difficult to prevent. The validation loss easily converged to a minimum with the 1D convolutional models, and it did not overfit.

Regarding model comparison, simple ANN models for classifying EP sibilants can achieve very satisfactory results but not as high as our previous SVM-based model. Our best simple ANN model had three hidden layers and gave us a score of 88.76%, which was lower than the score obtained with the SVM-based model. Our best results were obtained with 2D CNN models that use matrices of log Mel filterbanks as input. The 1D convolutional model was able to increase the score to 94.04%. The 2D convolutional model with log Mel filterbanks achieved a classification score of 95.48% [26]. This could be expected since the convolutional layers can extract localized information that can contribute to a better classification than that of simpler ANN models.

In addition to a higher classification score, our CNN model has provided us with a great improvement in reducing the number of false negatives. With the CNN model, we now have an average false negative rate of 4.35%, a reduction of over 4% from the false negative rate obtained by our previous model. Taking into account that the purpose of the proposed models is to classify child speech productions in serious games for speech and language therapy, their false negative rate is an important factor to ensure that patients do not lose motivation in playing due to the models' incorrect classifications. Since the models have high classification scores and low false negative rates, they are suitable for use in speech and language therapy games for sigmatism. For a detailed evaluation of the different four class sibilant classifier models, please check reference [26].

5.4.2. The EP Sibilants Extended Classifiers

While the games' previous versions used the classifiers that distinguished between the four EP sibilants described in Section 5.4.1, the current versions' classifier uses a combination of two models: a model that determines the point of articulation, M_{pa}, and another that determines if the vocal folds have been used, M_{vf} [28]. Both models are independent models that reuse the architecture of the most successful CNN four class sibilants classifier discussed in Section 5.4.1, but that used different training sets.

The models were trained with single-phoneme samples. Instead of only training the models with samples from the four EP sibilants, we also used samples for both [f] and [v] fricatives (also recorded during the data collection task). Thus, M_{pa} can determine three points of articulation: labiodental, alveolar and palato-alveolar (Table 1). When an input phoneme sample is given to these two models, it is possible to determine if the phoneme is one of the six fricatives ([s], [z], [ʃ], [ʒ], [f], [v]) by combining the output of the two models.

Phoneme samples for the six fricatives were extracted from whole word samples. In order identify the locations of the individual phonemes within the whole word samples, an acoustic model of the Kaldi ASR was used [29]. The correct word productions with the six fricatives were downsampled to 16,000 Hz and used to train this new acoustic model. Table 5 shows the number of phoneme productions extracted from the word samples.

Table 5. Number of fricative phoneme productions used to train the classification models.

	Phoneme Productions
f	2983
v	2413
s	6625
z	2251
ʃ	11,201
ʒ	2879

These phoneme samples were then converted into log Mel filterbanks. These are 80 × 9 matrices (with 80 bins and 9 frames), that were extracted with a 25 ms window size and 10 ms shift size. Our approach consists of applying two dimensional, spectral-temporal, convolutions to the whole input matrix used to train two CNNs: one CNN to learn M_{pa}, and another to learn M_{vf}.

The architecture of these two CNNs mostly reuses the 2D CNN architecture described in Section 5.4.1, but with two convolutional layers. The first and second convolutional layers use 50 and 25 kernel filters of size 10 × 2, with a stride of 2 × 1 and a stride of 1 respectively. Max pooling with a 2 × 2 window and a stride of 1 was used for both convolutional layers. The LeCun normal initializer with a max norm of 2, was used for the filters in both layers. The size of the output layer depends on the number of output classes, with one neuron per class. Thus, we used three and two output neurons to learn M_{pa} and M_{vf}, respectively.

An analysis of the CNN filters and namely of first convolution layer representation shows that the relevant features are aggregated while also maintaining a direct relation with log Mel filter banks. So, the discriminating power given by composition of some the first layer features provides an important pool of (automatically learned) different encodings [28].

Both CNN classifier models (M_{pa} and M_{vf}) were trained using the production of 70% of the children in the training set, 20% in the test set, and 10% in the validation set. The M_{pa} model achieved an overall average accuracy of 90.40%, with the best results for the palato-alveolar fricatives, which achieve an F1-score of 93.05%. The labiodental and alveolar fricatives achieve an F1-score of 87.99% and 87.60%, respectively. The overall average accuracy of M_{vf} was 90.93%, and we have an F1-score of 83.32% for the voiced and 93.77% for voiceless fricatives [28]. These high F1-score values warrant that most times the game will classify the children's fricative productions correctly, which is an important aspect in a tool for speech therapy and to make an impact in the child's learning process. All remaining details about the training process and 2D CNN validation are discussed in Reference [28].

5.4.3. The EP Word Recognition Module

The word recognition module for Portuguese children addresses the task of deciding whether a claimed word (the word corresponding to the picture presented to the child) is uttered in a given speech segment or not (the answer uttered by the child). In the past, we have shown that keyword spotting (KWS) is an adequate solution for similar speech therapy applications [12], since it permits dealing with unexpected speech effects, such as hesitations, doubts, repetitions, and other speech disturbing factors.

In this work, like in Reference [30], we use a context-independent hybrid ANN/hidden Markov model (HMM) speech recognizer extended to incorporate a competing background speech model that is estimated without the need for acoustic model re-training. When operating in KWS mode, the recognition task simply consists of a finite state grammar composed of a background or filler word in contrast to the claimed target word (with optional background at the beginning and end of the utterance). A background penalty term can be used to weight the importance of the background model and calibrate the operation of the module. For this work, we adapted an acoustic model initially trained for Broadcast News speech recognition to the particular characteristics of Portuguese children speech. The initial model was trained with more than 1000 h of adult speech [31] and it was used both for initial forced alignment of the children speech data and for network weight initialization in the children adaptation experiments. The LetsRead and CNG databases described in Section 2 were used for acoustic model adaptation through simple back-propagation updates with a varying number of data training epochs. We experimented with several training data partitions and combinations. The best recognition results in a held out set of the LetsRead database were obtained when selecting for adaptation the speech data from both the LetsRead and CNG corpora of the "older children" group. This corresponds to a total of approximately 11.5 h of children adaptation data.

The word recognition module was validated with the BioVisualSpeech corpus of words with sibilants described in this work. In order to do so, we selected the set of correctly uttered words and we performed two word verification tests: one against the correct word and one against a random word selected among the other words of the collection. That way, we simulated a word verification task with half positive (the uttered word matches the expected word) and half negative (the uttered work does not match the expected word) trials. A sub-set of 5 children utterances was used to set the background penalty term of the KWS module, which was then fixed to conduct the evaluation on the remaining data. An F1-score of up to 89.4% was obtained in the word verification task described. It is worth noting that this result is achieved with absolutely no data from the BioVisualSpeech corpus used for acoustic model training, and, consequently, some performance degradation may be due to mismatch in acoustic characteristics. Nevertheless, we consider these results as a promising first step towards the future development of a more fine-grained analysis of the quality of pronunciation of each individual phoneme within the validated word.

6. Conclusions

In this work, we discussed the screening, collection, and annotation stages for the creation of a corpus of children speech composed of words with the EP sibilant sounds [s], [z], [ʃ], and [ʒ]. The corpus is the result of joint work between computer scientists and speech and language pathologists. It was built with children's speech recorded in schools and contains speech data from 356 children from 5 to 9 years of age. One of the novelties of the collected corpus is that the phoneme and word labeling annotations indicate if the sound productions are correct or incorrect according to SLPs criteria. In addition, the corpus includes samples from children with SSD. Both these features are important to assist the development of automatic speech classifiers that can deal with speech from children with SSD.

In addition, as an illustration on how to use the corpus for the development of tools for speech therapy, we presented three therapy games designed for helping children with sigmatism. These games can be used for exercises with isolated sibilants, exercises with sibilants at different positions within a word, the identification of words that start with similar sibilant sounds, and exercises in which the child is aware of the point of articulation and the use of the vocal folds. An important characteristic of these games is that they are controlled by the children's voices. The sequence of actions in the games depends on the output of two speech processing modules: a convolutional neural network sibilant classifier that is trained with the data from the collected corpus and a word recognition module that is calibrated with the same data. In this way, the games evolution provides visual feedback to the children about their speech performance.

The sibilant classifier uses two models: one that is able to identify the place of articulation (this model distinguishes three different places of articulation: labiodental, alveolar, and palato-alveolar), and another that identifies if the sibilant production is voiced or voiceless. This information is useful to help children understand their mistakes when they mispronounce the sibilant sounds. By combining the output of both models, the classifier can distinguish not only the four EP sibilants but also the two labiodental fricatives ([f] and [v]). The CNN classifier achieves F1 scores between 87.60% and 93.05% for place of articulation, and between 83.32% and 93.77% for voicing. On the other hand, the word recognition module allows for detecting whether an arbitrary word is uttered in a speech utterance. It is based on hybrid ANN/HMM automatic speech recognition configured to operate in KWS mode. The acoustic models were trained using external EP children and the recognizer achieved an F1-score of up to 89.4% in a word verification evaluation task conducted in the newly collected corpus. The remarkable performance of these two speech processing modules is a strong indicator of the benefits that can bring tools that use these classifiers in the child's learning process.

Author Contributions: Conceptualization, S.C., I.G., M.G., N.M., A.A.; Methodology, S.C., I.G., A.A., N.M., M.E., J.M., M.G.; Software, I.A., F.O., S.M., N.M.; Validation, N.M., I.A., S.C., A.A., F.O.; Investigation, S.C., I.G., M.A., A.A., I.A., F.O., S.M., N.M., M.E., J.M., M.G.; Resources, I.G., M.G.; Data Curation, M.G., I.G., I.A., A.A., M.A.; Writing Original Draft Preparation, S.C, A.A., N.M.; Writing Review and Editing, S.C., A.A., F.O., N.M.; Visualization, N.M., I.A., S.C.; Supervision, S.C., I.G., A.A., N.M., M.E., M.G.; Project Administration, S.C., A.A., M.E., M.G.; Funding Acquisition, S.C., M.E., A.A, J.M., I.G., M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Portuguese Foundation for Science and Technology under projects BioVisualSpeech (CMUP-ERI/TIC/0033/2014), NOVA-LINCS (UIDB/04516/2020) and INESC-ID (UIDB/50021/2020).

Acknowledgments: We thank Cátia Pedrosa, Diogo Carrasco and Catarina Botelho for the segmentation and annotation of our corpus, Mário Sansana for training the initial children speech acoustic model, and all postgraduate SLP students from Escola Superior de Saúde do Alcoitão who collaborated in the data collection task. We also thank the schools from Agrupamento de Escolas de Almeida Garrett, and the children who participated in the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McLeod, S. *The International Guide to Speech Acquisition*; Thomson Delmar Learning: Florence, KY, USA, 2007.
2. Guimarães, I.; Birrento, C.; Figueiredo, C.; Flores, C. *Teste de Articulaçã Overbal*; Oficina Didáctica: Lisboa, Portugal, 2014.
3. Preston, J.; Edwards, M.L. Phonological awareness and types of sound errors in preschoolers with speech sound disorders. *J. Speech Lang. Hear. Res.* **2010**, *53*, 44–60. [CrossRef]
4. Nathan, L.; Stackhouse, J.; Goulandris, N.; Snowling, M.J. The development of early literacy skills among children with speech difficulties: A test of the critical age hypothesis. *J. Speech Lang. Hear. Res.* **2004**, *47*, 377–391. [CrossRef]
5. Guimarães, I. *Ciência e Arte da Voz Humana*; Escola Superior de Saúde de Alcoitão: Alcabideche, Portugal, 2007.
6. Honová, J.; Jindra, P.; Pešák, J. Analysis of articulation of fricative praealveolar sibilant "s" in control population. *Biomed. Pap.* **2003**, *147*, 239–242. [CrossRef]

7. Weinrich, M.; Zehner, H. *Phonetiche und Phonologische Störungen bein Kindern*; Springer: Berlin/Heidelberg, Germany, 2005.
8. Figueiredo, A.C. Análise Acústica dos Fonemas /s, z/ Produzidos por Crianças com Desempenho Articulatório Alterado. Master's Thesis, Escola Superior de Saúde do Alcoitão, Santa Casa da Misericórdia de Lisboa, Lisboa, Portugal, 2017.
9. Rua, M. Caraterização do Desempenho Articulatório e Oromotor de Crianças Com Alterações da Fala. Master's Thesis, Escola Superior de Saúde do Alcoitão, Santa Casa da Misericórdia de Lisboa, Lisboa, Portugal, 2015.
10. Anjos, I.; Grilo, M.; Ascensão, M.; Guimarães, I.; Magalhães, J.; Cavaco, S. A Model for Sibilant Distortion Detection in Children. In Proceedings of the 2018 International Conference on Digital Medicine and Image Processing (DMIP), Okinawa, Japan, 12–14 November 2018.
11. Grilo, M.; Guimarães, I.; Ascensão, M.; Abad, A.; Anjos, I.; Magalhães, J.; Cavaco, S. The BioVisualSpeech European Portuguese Sibilants Corpus. In *International Conference on Computational Processing of the Portuguese Language (PROPOR)*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 23–33.
12. Abad, A.; Pompili, A.; Costa, A.; Trancoso, I.; Fonseca, J.; Leal, G.; Farrajota, L.; Martins, I. Automatic word naming recognition for an on-line aphasia treatment system. *Comput. Speech Lang.* **2013**, *27*, 1235–1248. [CrossRef]
13. Proença, J.; Celorico, D.; Candeias, S.; Lopes, C.; Perdigão, F. The LetsRead Corpus of Portuguese Children Reading Aloud for Performance Evaluation. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016; pp. 781–785.
14. Hämäläinen, A.; Rodrigues, S.; Júdice, A.; Silva, S.; Calado, A.; Pinto, F.; Dias, M. The CNG Corpus of European Portuguese Children's Speech. In Proceedings of the International Conference on Text, Speech and Dialogue, Pilsen, Czech Republic, 1–5 September 2013; Volume 8082. [CrossRef]
15. Cruz-Ferreira, M. Portuguese (European). In *Handbook of the International Phonetic Association, A Guide to the Use of the International Phonetic Alphabet*; Cambridge University Press: Cambridge, UK, 1999.
16. *Handbook of the International Phonetic Association, A Guide to the Use of the International Phonetic Alphabet*; Cambridge University Press: Cambridge, UK, 1999.
17. Rennicke, I.; Martins, P. *As realizações Fonéticas de /R/ em Português Europeu: Análise de um Corpus Dialetal e Implicações no Sistema Fonológico*; Encontro Nacional da Associação Portuguesa de Linguística, Universidade do Algarve: Faro, Portugal, 2013; pp. 509–523.
18. Grossinho, A.; Guimarães, I.; Magalhães, J.; Cavaco, S. Robust phoneme recognition for a speech therapy environment. In Proceedings of the IEEE International Conference on Serious Games and Applications for Health (SeGAH), Orlando, FL, USA, 11–13 May 2016.
19. Smit, A.B.; Hand, L.; Freilinger, J.J.; Bernthal, J.E.; Bird, A. The Iowa Articulation Norms Project and its Nebraska replication. *J. Speech Hear. Disord.* **1990**, *55*, 779–797. [CrossRef] [PubMed]
20. Shriberg, L.D.; Kwiatkowski, J. Developmental phonological disorders: I. A clinical profile. *J. Speech Lang. Hear. Res.* **1994**, *37*, 1100–1126. [CrossRef] [PubMed]
21. Lousada, M.; Mendes, A.; Valente, R.; Hall, A. Standardization of a Phonetic-Phonological Test for European-Portuguese Children. *Folia Phoniatr. Logop. Off. Organ Int. Assoc. Logop. Phoniatr. (IALP)* **2012**, *64*, 151–156. [CrossRef] [PubMed]
22. Wren, I.; McLeod, S.; White, P.; Miller, L.; Roulstone, S. Speech characteristics of 8-year-old children: Findings from a prospective population study. *J. Commun. Disord.* **2013**, *46*, 53–69. [CrossRef] [PubMed]
23. Charles-Luce, J.; Dressler, K.M.; Ragonese, E. Effects of semantic predictability on children's preservation of a phonemic voice contrast. *J. Child Lang.* **1999**, *26*, 505–530. [CrossRef] [PubMed]
24. Mestre, I. Sibilantes e Motricidade orofacial em CrianççAs Portuguesas dos 5;00 aos 9;11 Anos de Idade: Estudo Preliminar. Master's Thesis, Escola Superior de Saúde do Alcoitão, Santa Casa da Misericórdia de Lisboa, Lisboa, Portugal, 2017.
25. Anjos, I.; Grilo, M.; Ascensão, M.; Guimarães, I.; Magalhães, J.; Cavaco, S. A Serious Mobile Game with Visual Feedback for Training Sibilant Consonants. In *Advances in Computer Entertainment Technology*; Cheok, A.D., Inami, M., Romão, T., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 430–450.
26. Anjos, I.; Marques, N.; Grilo, M.; Guimarães, I.; Magalhães, J.; Cavaco, S. Sibilant consonants classification comparison with multi and single-class neural networks. *Expert Syst.* **2020**. [CrossRef]

27. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
28. Anjos, I.; Eskenazi, M.; Marques, N.; Grilo, M.; Guimarães, I.; Magalhães, J.; Cavaco, S. Detection of voicing and place of articulation of fricatives with deep learning in a virtual speech and language therapy tutor. In Proceedings of the Interspeech, Shanghai, China, 28 October 2020.
29. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi Speech Recognition Toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011.
30. Abad, A.; Pompili, A.; Costa, A.; Trancoso, I. Automatic word naming recognition for treatment and assessment of aphasia. In Proceedings of the 13th Annual Conference of the International Speech Communication Association, Portland, OH, USA, 9–13 September 2012.
31. Meinedo, H.; Abad, A.; Pellegrini, T.; Neto, J.; Trancoso, I. The L2F Broadcast News Speech Recognition System. Available online: http://lorien.die.upm.es/~lapiz/rtth/JORNADAS/VI/pdfs/0018.pdf (accessed on 24 September 2020).