The BioVisualSpeech European Portuguese Sibilants Corpus

Chapter · February 2020

DOI: 10.1007/978-3-030-41505-1_3

7 authors, including:

Margarida Grilo
Escola Superior de Saude do Alcoitão
13 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Isabel Cristina Ramos Peixoto Guimarães
Escola Superior de Saude do Alcoitão
115 PUBLICATIONS **761** CITATIONS
SEE PROFILE

Mariana Ascensão
Alcoitão School of Healthy Sciences
9 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Alberto Abad
Technical University of Lisbon
113 PUBLICATIONS **1,028** CITATIONS
SEE PROFILE

Sibilant consonants classification with deep neural networks

Ivo Anjos[1 [0000−0002−9493−1564]], Nuno Marques[1 [0000−0002−3019−3304]], Margarida Grilo[2 [0000−0003−2187−8253]], Isabel Guimarães[2 [0000−0001−8524−8731]], João Magalhães[1 [0000−0001−6290−5719]], and Sofia Cavaco[1 [0000−0001−7315−4543]]

[1] NOVA LINCS, Department of Computer Science
Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa
2829-516 Caparica, Portugal
[2] Escola Superior de Saúde do Alcoitão
Rua Conde Barão, Alcoitão, 2649-506 Alcabideche, Portugal

Abstract. Many children suffering from speech sound disorders cannot pronounce the sibilant consonants correctly. We have developed a serious game that is controlled by the children's voices in real time and that allows children to practice the European Portuguese sibilant consonants. For this, the game uses a sibilant consonant classifier. Since the game does not require any type of adult supervision, children can practice the production of these sounds more often, which may lead to faster improvements of their speech.
Recently, the use of deep neural networks has given considerable improvements in classification for a variety of use cases, from image classification to speech and language processing. Here we propose to use deep convolutional neural networks to classify sibilant phonemes of European Portuguese in our serious game for speech and language therapy.
We compared the performance of several different artificial neural networks that used Mel frequency cepstral coefficients or log Mel filterbanks. Our best deep learning model achieves classification scores of 95.48% using a 2D convolutional model with log Mel filterbanks as input features.

Keywords: deep learning · sibilant consonants · speech and language therapy

1 Introduction

The sibilant sounds are a subgroup of the consonant sounds that are produced by letting the air flow through a very narrow channel in the direction of the teeth [12]. These include sounds like [s] in serpent and [z] in zipper. Sigmatism, the distortion of sibilant sounds production, is a very common speech sound disorder in European Portuguese (EP) children [10, 22]. Depending on the age of the child and the degree of the distortion, the speech problem can disappear naturally as the child grows up and his/her speech organ develop. Yet, in many cases these problems do not disappear naturally, and it is important that the child attends speech and language therapy to treat the disorder.

In order to help children to surpass sigmatism, speech and language pathologists (SLPs) start by teaching the child to produce these isolated sibilant sounds before progressing to the production of the sounds within words. SLPs use multiple and repeated tasks to allow the child to practice the sounds and learn how to produce them correctly. Yet, the repetition of the tasks can lead to the child's lack of interest and motivation on doing the exercises. In order to keep children motivated and collaborative during the therapy sessions, SLPs need to turn the speech and language monotonous tasks into fun and appealing activities.

As a contribution to overcome the children's lack of motivation to repeat productions of the sibilants, we have developed the *isolated sibilants game*, a serious game that allows children to practice the isolated sibilants exercise while playing a computer game [2]. An important characteristic of the game is that it is controlled by the child's voice. That is, instead of using a keyboard or other usual input device, the child must use his voice to play the game. To make this possible, at its core, the game uses a sibilant consonant classifier that processes the child's speech productions. This classifier was trained to distinguish the EP sibilant consonants, but can be adapted to other languages.

In addition to being appropriate for speech therapy sessions, the game can also be used for home training. Since the game has the ability to automatically classify the child's speech productions, it does not depend on the parents supervision, whose free time may be limited. Allowing children to practice the sounds often at home during their free time may contribute to faster speech improvements [3, 5, 9].

While traditional automatic speech recognition (ASR) systems typically use hidden Markov models (HMM) [11, 14], the classification of isolated phonemes and words can be successful with support vector machines (SVMs) [7, 26, 28]. As for the input features, statistical measures of Mel frequency cepstral coefficients (MFCCs) are commonly used as input to the ASR classification stages [8, 13, 19, 27]. MFCCs are a spectral representation of the sounds based on the Mel scale [11]. We have previously used a less common approach to classify sibilant consonants. We used the MFCCs as a raw feature [2], that is, without statistical measures over the MFCCs. This approach has also been successfully used by Carvalho *et al.* to classify the five EP vowels [6].

The classification of sibilant sounds is not a novelty. Benselama *et al.* focused on the sibilants that are mispronounced by people with Arabic occlusive sigmatism [4]. They used MFCCs in artificial neural networks for classifying Arabic sibilants, along with HMM with Gausian mixture models (GMM) for segmenting these phonemes. Valenini-Botinhao *et al.* use MFCCs along with other features to construct a GMM to classify German sibilants [30]. Miodońska *et al.* use MFCCs and other acoustic features on a SVM for the classification of Polish sibilants [18].

Currently, there has been a tendency to substitute the more traditional speech classifiers with deep learning models, which have proven to be quite robust [1, 24, 31]. The most common features for this type of models are log Mel filterbanks, which, like with MFCCs, are a spectral representation of the sounds

based on the Mel scale [23]. More recently, some studies also tested the use of convolutional neural networks (CNN) on the raw waveform [20, 25].

The sibilant classifier in the first version of our isolated sibilants game used a simple SVM model with the raw MFCCs. That simple classifier achieved an accuracy of 90.72% [2], with an average false negative rate of 8.73%. While the score of this sibilant classifier is already high, we were able to improve it by using deep learning techniques. We compared multiple artificial neural network (ANN) models, from simpler neural network models with one to three hidden layers, to convolutional models. We also compared using MFCCs and log Mel filterbanks as input features to these models. In the end we were able to develop models with CNN that surpassed the classification scores obtained by the game's previous SVM classifier.

The purpose of these classification models is to classify EP child sibilant consonants in a serious game for sigmatism. Therefore, the false negative rate of the models must be taken into account. A low false negative rate is important to ensure that patients receive suitable feedback when their speech productions are correct. It is important that the game does not misclassify correct speech productions so that it does not lead the child into error and ensures that the child does not loose motivation in playing due to the models' incorrect classifications. Our CNN models obtained high classification scores (95.48%) combined with low false negative rates (4.35%), which shows that the models are suitable to be used in the proposed game for training the EP sibilants.

In summary, the main novelty of this paper is the use deep learning models to classify EP sibilant sounds from child speech productions, to be used in tools for speech and language therapy.

2 Sibilant consonants and sigmatism

Different sibilant sounds can be produced by using different parts of the vocal tract. 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 vocal folds can either be used or not, resulting in a voiced or a voiceless sibilant, respectively.

There are four different sibilant consonant sounds in EP: [z] as in zebra, [s] as in snake, [ʃ] as the *sh* sound in sheep, and [ʒ] as the *s* sound in Asia. [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.

Most children with sigmatism cannot produce some sibilant sounds correctly either because they do not use the correct region of the vocal tract or because they exchange the voiced and voiceless sounds [21]. To correct these distortion errors, SLPs use different exercises and usually start with the *isolated sibilants exercise*, which consists of producing the sibilant sounds for a few seconds.

The continuous repetition and practice of the exercise is essential to master the isolated sibilant sounds. However this can be very demanding and wearing for

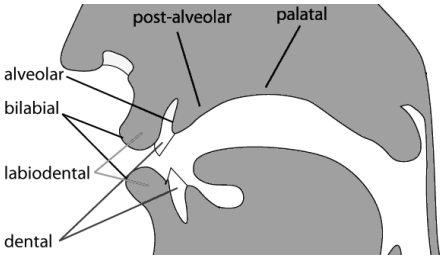

Fig. 1: Main places of articulation in the vocal tract [3].

(a) (b)

Fig. 2: Scenarios for (a) the [z] consonant, and for (b) the [s] consonant.

children, who can feel tired of repeating the exercise and thus stop collaborating with the SLP. In order to motivate the children on practicing the exercise as often as required, it is necessary to turn it into a fun activity. To this end, SLPs usually resort to rewards and make the exercise look like a game.

3 Serious game for training sibilant consonants

In order to motivate children on doing the isolated sibilants exercise and repeating it as often as necessary, and to help SLPs on turning this into a fun activity, our isolated sibilants game uses this exercise [2]. The game was designed for children from five to nine years old, because usually at these ages the regular phonological development exchange errors have already disappeared [17].

The game consists of leading the main character to a target. To make the main character move the child must use his voice. More specifically, the child has to produce one of the four EP sibilant consonants. An important characteristic of the game is that it processes the child's speech productions in real time. Thus, unlike with other speech and language therapy computer games that are manually controlled by the SLP, in this game the main character is controlled by

[3] Adapted from: The Mimic Method - Place of Articulation, https://www.mimicmethod.com/ft101/place-of-articulation (retrieved April/2019).

the child's voice. The character moves towards the target if the child produces the sibilant consonant correctly. In this way, the movement of the character gives 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.

The game includes a different scenario and main character for each of the four EP sibilant consonants (figure 2). These scenarios were created with the help of a visual artist and with images from *Freepik* [4]. In each scenario, the main character or game goal is related to the addressed sibilant sound, which is an additional visual cue that helps the child understand what sound she should produce: The main character for the [z] sibilant scenario is a bumblebee as the EP word for bumblebee, *zangão*, starts with the [z] sibilant (figure 2.a). The main characters for the [s] and [ʒ] sibilants are a snake (figure 2.b) and a ladybug, because the EP words for snake and ladybug, which are *serpente* and *joaninha*, start with the [s] and [ʒ] sibilants, respectively. In the [ʃ] sibilant scenario there is a boy who must run away from the rain to reach the end of the road. The EP word for rain, *chuva*, starts with the [ʃ] sibilant.

4 Data collection and data representation

In order to train our EP sibilant models we used sibilant sound productions from children. The sounds were collected in three schools in the greater Lisbon area. There was always a SLP present in the recording sessions. 145 children from 4 to 11 years old participated in the recordings, from which there were 83 girls and 62 boys (table 1). We recorded short and long versions of the four EP sibilants consonants (that is, a version that lasts less than a couple of seconds and another that last a few seconds), giving a total of over 1500 sound productions. These were manually labeled and the correct productions were used for training the models. Table 2 shows the number of correct sound productions for each sibilant.

In our previous work we used the raw MFCCs as input to the SVMs, that is, we used the columns of the MFCCs matrix (which were 13×1 MFCCs vectors) directly as feature vectors (and no statistical measure over the whole matrix) [2]. Given the good classification scores obtained using these input features, here we experimented to maintain the same type of features, that is, the raw MFCCs.

To extract the MFCCs we use a 25 ms window with a 10 ms shift. We extract the first 13 MFCCs. Therefore, each sound production is represented by a $13 \times t$ matrix, where t depends on the duration of the sound. Each column vector of this matrix consists of the 13 MFCCs of a 25 ms window. We also tried with log Mel filterbanks as input features. We use the first 40 filters to represent the data, and kept the same window and shift sizes of 25 ms and 10 ms, respectively.

5 Models for sibilant consonants

As explained above, here we propose to use ANN models to classify children's EP sibilant productions in a serious game for speech and language therapy for

[4] Freepik, http://www.freepik.com

Age	Girl	Boy	Total
4	0	1	1
5	8	3	11
6	8	8	16
7	19	9	28
8	21	20	41
9	23	19	42
10	3	2	5
11	1	0	1
Total	83	62	145

Table 1

Sibilant	# correct productions
ʃ	276
ʒ	257
s	278
z	264

Table 2

Table 1: Age and gender of children participating in the recordings. Table 2: Number of correct sound productions for each sibilant.

sigmatism. We compared multiple neural network models, from simple one hidden layer ANNs to more complex CNNs. As a base support for the comparison between all our models, we kept the input features as similar as possible. We used the Adam optimizer as our loss function in all networks and we also used the stochastic gradient descent (SGD) in our CNNs [15]. Below we discuss each of these models in detail.

We started by experimenting with simple feed-forward ANN consisting of multilayer perceptrons. The input layer of these networks receives the raw 13×1 column-vectors of the MFCCs matrix. We used the rectified linear unit (ReLU) [16] as the activation function for our neurons, and applied dropout to the hidden layers, with a drop rate of 40% [29].

The simplest model has just one hidden layer with 50 neurons and an output layer with 4 neurons. The goal of training this model was to test if such a simple network can achieve a good separation of the four sounds. We observed that while the model can distinguish the sounds, it shows worse results than those achieved by our previous SVM-based model used in the first version of our game (more details in section 6).

After experimenting increasing the number of hidden layers to two, and having no significant improvements, we further increased the number of hidden layers to four layers also with limited improvements. The first hidden layer has 100 neurons, and the second, third and fourth layers have 50, 25 and 10 neurons, respectively. In both models, we used 200 epochs for training.

As it will be discussed in section 6, while these models provide good results, their classification scores are lower than those achieved by our previous SVM-based model. Thus, we then proceeded to create more complex models with the goal of surpassing our previous SVM scores.

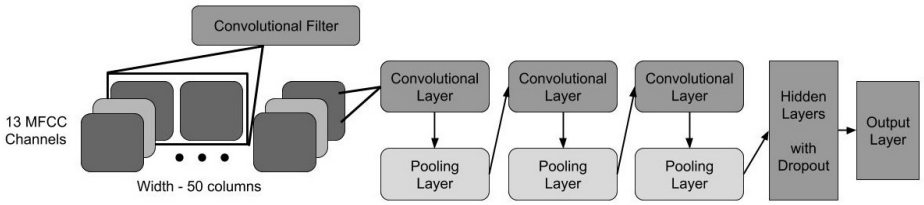

Fig. 3: Representation of our 1D CNN.

5.1 Convolutional models

Since convolutional layers have the potential to extract relevant local information, we explored using CNN to classify the EP child sibilant productions, to see if this type of knowledge helps improving our previous classification score. Here, instead of having 13×1 column input vectors, the input to our CNNs is given in the form of a matrix. We experimented using MFCCs matrices and log Mel filterbank matrices as input features to our CNNs.

1D CNNs with MFCCs

The input to our CNNs consists of sub-matrices of the $13 \times t$ original MFCC matrix. To build these matrices we fetch sets of 50 consecutive columns of the original MFCCs matrix with a step of 5 columns (and 45 columns overlap). In other words, we extract 50 consecutive columns (that is, 13×1 vectors), then skip 5 consecutive columns, and extract another 50 columns that have some overlap with the previous matrix. We then repeat the process until we reach the end of the original MFCCs matrix, obtaining matrices of size 13×50 for each sound production. Hereafter these matrices are considered as our data samples.

Our approach here consists of applying one dimensional convolutions to the data. We do this by considering the 50 columns as our width, and having our 13 MFCCs as channels (like with the RGB channels in images). In other words, we apply temporal convolutions to each of the 13 MFCCs receptive field (figure 3).

Our 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 3). We used the ReLU as the activation functions of the convolutional and hidden layers, and we trained the model for 100 epochs.

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.

2D CNNs with log Mel filterbanks
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, ie. 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, spacial convolutions are applied across the whole input matrix. We use the 50 columns as our width, and 40 filters as our height, with just one channel. The main architecture of this network is the same as above for our 1D CNN, but using 2D convolutional layers, instead of 1D layers.

Following the 1D model results, we started with the SGD optimizer for the 2D model. Again this optimizer easily allowed us to prevent overfitting. 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.

6 Results

The EP sibilant classifier of the game's first version used one SVM with a radial basis function kernel for each of the four EP sibilant sounds [2]. It used 13×1 MFCC vectors as input and was able to reach average accuracy test scores of 90.72% between all four SVM models. We also trained a SVM using the same 13×50 input matrix as in our convolutional models (section 5.1). Yet, with this input matrix, the SVM did not learn how to separate the four sibilant sounds. Indeed, unlike with CNN models, the MFCC input vectors high dimension and strong correlation makes this data representation less appropriated for SVM classifiers.

In order to train our neural networks, we divided the data set described in section 4 into training, validation and test sets. We randomly separated the data into these three sets, but to avoid any bias in the results, we were careful to put all productions of each child c either in the training, validation or the test set. Thus, while we have several feature vectors from each child, all the vectors from each child belong to the same data set. This way we avoid the insertion of bias in our results, since samples from the same sibilant of a particular child are likely to have some correlation. We used the productions of 20% of all children as the test set, 10% as the validation set and the remaining 70% productions form the training set. This type of split remained equal for all models.

This way, for the simple ANN models, we have around 250000 MFCC vectors for the training set, 30000 for the validation, and 70000 for the test set. We trained the simple ANN models for 200 epochs, using a batch size of 500 samples, and shuffling the dataset at every epoch.

While the data was split into training and test sets in the same manner for our previous SVM-based model, we did not use the whole MFCC matrices with these models due to the temporal learning complexity of the models. For these models, we used around 20000 MFCC vectors (more details in [2]).

Model	Test Accuracy	FNR
SVM	90.72%	8.73%
simple ANN	88.76%	11.15%
1D CNN	94.04%	5.56%
2D CNN	95.48%	4.35%

Table 3: Accuracy test scores of all our models.

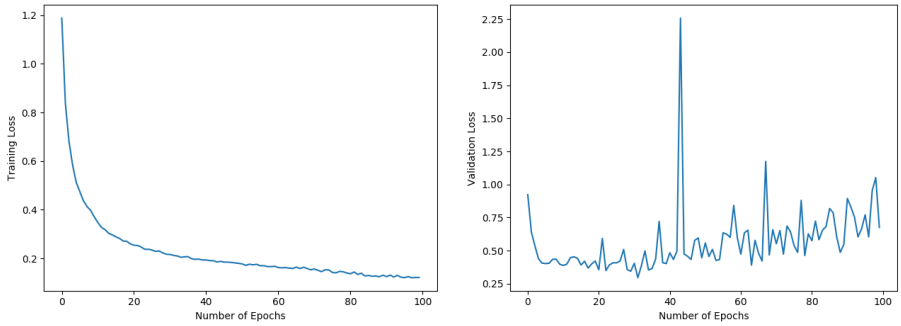

Fig. 4: 2D convolutional model. (Left) Training loss. (Right) Validation loss.

Using the matrices approach for the convolutional models reduces the total of data input samples. Nonetheless, we still have around 45000 samples for training, 5000 for validation, and 13000 samples for testing the convolutional models.

Table 3 shows the results achieved with the different models. It shows that simple ANN models for classifying EP sibilants can achieve very satisfactory results but not as high as our previous SVM-based model. A single hidden layer ANN model achieved an average classification score of 86.54%. Further increasing the number of hidden layers resulted in marginal improvements to the test score. Our best simple ANN model had three hidden layers and gave us a score of 88.76%, which was still lower than the score obtained with the SVM-based model.

The 1D convolutional model was able to increase the score to 94.04%, which is almost 6% higher than the score of our 4-hidden layer ANN model, and around 3% higher than the SVM score. This could be expected, since the convolutional layers have the ability to extract localized information that can contribute for a better classification than that of simpler ANN models. In particular, there may be relevant localized information in the 13×50 CNN input matrices that is not present in the 13×1 ANN input vectors. Using the concatenation of 50 MFCC columns as input to the CNN, can contribute for a better prediction of the sound.

The 2D convolutional model, which uses the highly correlated log Mel filterbanks, can be expected to achieve at least the same scores as our 1D model, or even to overcome them. Our experiments with the 2D CNN reached a test score of 95.48%, surpassing our 1D convolutional model. This shows that the convolutional layers were able to extract more significant information from the log Mel filterbanks than from the MFCCs.

Figure 4 shows the training and validation loss of the best 2D convolutional model. As we can see from the training loss graph, the model quickly converges. The validation graph shows that there is a lot of variation between epochs, and the model never reached a stable validation loss. For our final model, we chose the one that provided the lowest validation loss, which was at epoch 32, with a value of 0.29. After this epoch the model continued to adapt to our training set, but as can be seen by the validation loss, the model started to overfit. 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.

The main drawback of our previous model (SVM) was the considerably high false negative rate (FNR):

$$FNR = \frac{FN}{FN + TP},$$

where FN is the number of false negatives and TP is the number of true positives. This was also one of the reasons that led us to experiment with neural networks, and test if CNNs have an higher generalization ability. With the SVM model, we had an average 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. In addition, these misclassifications can be very frustrating for children. 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.

7 Conclusion and Future Work

Here we proposed deep CNN models to automatize the classification of child EP sibilant productions in a serious game for speech and language therapy for sigmatism. We compared the performances of different networks using either MFCCs or log Mel filterbanks. Our best model uses matrices of log Mel filterbanks as input and have three 2D convolutional layers. This model achieved a classification score of 95.48% and surpassed the classification scores obtained with our simple ANN model, and 1D model with MFFCs, and also our previous SVM-based model [2].

In addition to a higher classification score, the proposed model has a low average false negative rate, 4.35%. Taking into account that the purpose of the proposed models is to classify child speech productions in a serious game for speech and language therapy, their false negative rate is an important factor to ensure that patients do not loose 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.

As future work it will be interesting to experiment with other input features to our CNN, such as features that have been used in other state of the art models, like delta and delta-delta features, and the raw input sound data. A thorough study on the number of layers, and the choice of hyper parameters for all the layers, can help to further improve the classification scores.

8 Acknowledgements

This work was supported by the Portuguese Foundation for Science and Technology under projects BioVisualSpeech (CMUP-ERI/TIC/0033/2014) and NOVA-LINCS (PEest/UID/CEC/04516/2019). We thank Mariana Ascensão and the postgraduate SLP students from Escola Superior de Saúde do Alcoitão who collaborated in the data collection task. Finally, we thank Agrupamento de Escolas de Almeida Garrett, and the children who participated in the recordings.

References

1. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., et al.: Deep speech 2 : End-to-end speech recognition in english and mandarin. In: Proceedings of The 33rd International Conference on Machine Learning. vol. 48, pp. 173–182. PMLR (2016)
2. 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. pp. 430–450 (2018)
3. Barratt, J., Littlejohns, P., Thompson, J.: Trial of intensive compared with weekly speech therapy in preschool children. Archives of Disease in Childhood **67**(1), 106–108 (1992)
4. Benselama, Z., Guerti, M., Bencherif, M.: Arabic speech pathology therapy computer aided system. Journal of Computer Science **3**(9), 685–692 (2007)
5. Bhogal, S.K., Teasell, R., Speechley, M.: Intensity of aphasia therapy, impact on recovery. Stroke **34**(4), 987–993 (2003)
6. Carvalho, M.I.P., Ferreira, A.: Interactive Game for the Training of Portuguese Vowels. Master's thesis, Faculdade de Engenharia da Universidade do Porto (2008)
7. Clarkson, P., Moreno, P.J.: On the use of support vector machines for phonetic classification. In: Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. vol. 2, pp. 585–588 (1999)
8. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In: Readings in speech recognition, pp. 65–74. Elsevier (1990)
9. Denes, G., Perazzolo, C., Piani, A., Piccione, F.: Intensive versus regular speech therapy in global aphasia: A controlled study. Aphasiology **10**(4), 385–394 (1996)
10. Figueiredo, A.C.: Análise acústica dos fonemas /ʃ/ e /ʒ/ produzidos por crianças com desempenho articulatório alterado. Master's thesis, Escola Superior de Saúde de Alcoitão (2017)
11. Gold, B., Morgan, N., Ellis, D.: Speech and Audio Signal Processing: Processing and Perception of Speech and Music. Wiley-Interscience, 2nd edn. (2011)
12. Guimarães, I.: A Ciência e a Arte da Voz Humana. ESSA - Escola Superior de Saúde do Alcoitão (2007)

13. Hsu, C.W., Lee, L.S.: Higher order cepstral moment normalization for improved robust speech recognition. IEEE Transactions on audio, speech, and language processing **17**(2), 205–220 (2009)
14. Huang, X., Acero, A., Hon, H.W.: Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR, 1st edn. (2001)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Adv. in neural inf. proc. sys. pp. 1097–1105 (2012)
17. Mestre, I.: Sibilantes e motricidade orofacial em crianças portuguesas dos 5:00 aos 9:11 anos de idade. Master's thesis, Escola Superior de Saúde do Alcoitão (2018)
18. Miodońska, Z., Krecichwost, M., Szymańska, A.: Computer-aided evaluation of sibilants in preschool children sigmatism diagnosis. In: Information Technologies in Medicine. pp. 367–376 (2016)
19. Muda, L., Begam, M., Elamvazuthi, I.: Voice recognition algorithms using mel frequency cepstral coefficient and dynamic time warping techniques. Computing Research Repository (CoRR) **abs/1003.4083** (2010)
20. Palaz, D., Magimai.-Doss, M., Collobert, R.: Analysis of CNN-based speech recognition system using raw speech as input. In: Proc. of Interspeech. pp. 11–15 (2015)
21. Preston, J., Edwards, M.L.: Phonological awareness and types of sound errors in preschoolers with speech sound disorders. Journal of Speech, Language, and Hearing Research **53**(1), 44–60 (2010)
22. Rua, M.: Caraterização do desempenho articulatório e oromotor de crianças com alterações da fala. Master's thesis, Escola Superior de Saúde de Alcoitão (2015)
23. Sainath, T.N., Kingsbury, B., Mohamed, A.R., Saon, G., Ramabhadran, B.: Improvements to filterbank and delta learning within a deep neural network framework. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6839–6843 (2014)
24. Sainath, T.N., Kingsbury, B., Ramabhadran, B., Fousek, P., Novak, P., Mohamed, A.R.: Making deep belief networks effective for large vocabulary continuous speech recognition. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). pp. 30–35 (2011)
25. Sainath, T.N., Weiss, R.J., Senior, A., Wilson, K.W., Vinyals, O.: Learning the speech front-end with raw waveform CLDNNs. In: Proceedings of the Annual Conference of the International Speech Communication Association (2015)
26. Salomon, J., King, S., Salomon, J.: Framewise phone classification using support vector machines. In: Proc. of the Int. Conf. on Spoken Language Processing (2002)
27. Schuller, B., Steidl, S., Batliner, A.: The interspeech 2009 emotion challenge. In: Proc. of Interspeech (2009)
28. Solera-Ureña, R., Padrell, J., Iglesias, D., Gallardo-Antolín, A., Peláez-Moreno, C., Díaz-de María, F.: SVMs for automatic speech recognition: A survey. In: Progress in Nonlinear Speech Processing. pp. 190–216 (2007)
29. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1), 1929–1958 (2014)
30. Valentini-Botinhao, C., Degenkolb-Weyers, S., Maier, A., Nöth, E., Eysholdt, U., Bocklet, T.: Automatic detection of sigmatism in children. In: Proc. of the Workshop on Child, Computer Interaction (WOCCI) (2012)
31. Zhang, Y., Chan, W., Jaitly, N.: Very deep convolutional networks for end-to-end speech recognition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4845–4849 (2017)