INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Detection of voicing and place of articulation of fricatives with deep learning in a virtual speech and language therapy tutor

Ivo Anjos[1], *Maxine Eskenazi*[2], *Nuno Marques*[1], *Margarida Grilo*[3], *Isabel Guimarães*[3], *João Magalhães*[1], *Sofia Cavaco*[1]

[1]NOVA LINCS, Department of Computer Science
Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa
2829-516 Caparica, Portugal

[2]Language Technologies Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA

[3]Escola Superior de Saúde do Alcoitão
Rua Conde Barão, Alcoitão, 2649-506 Alcabideche, Portugal

i.anjos@campus.fct.unl.pt, max@cs.cmu.edu,
{margarida.grilo, isabel.guimaraes}@essa.scml.pt, {nmm, jmag, scavaco}@fct.unl.pt

Abstract

Children with fricative distortion errors have to learn how to correctly use the vocal folds, and which place of articulation to use in order to correctly produce the different fricatives. Here we propose a virtual tutor for fricatives distortion correction. This is a virtual tutor for speech and language therapy that helps children understand their fricative production errors and how to correctly use their speech organs. The virtual tutor uses log Mel filter banks and deep learning techniques with spectral-temporal convolutions of the data to classify the fricatives in children's speech by place of articulation and voicing. It achieves an accuracy of 90.40% for place of articulation and 90.93% for voicing with children's speech. Furthermore, this paper discusses a multidimensional advanced data analysis of the first layer convolutional kernel filters that validates the usefulness of performing the convolution on the log Mel filter bank.

Index Terms: fricatives, speech and language therapy, convolutional neural networks

1. Introduction

Speech sound disorders (SSD) can affect children's health, literacy acquisition, and development processes, but also their social interaction and quality of life [1, 2, 3]. While most children with language acquisition difficulties can surpass these speech distortions as they grow older and their speech organs develop, some children may have a difficult time to overcome their SSD [4], thus delaying their normal cognitive development. These children may need to have speech and language therapy.

Many children's speech production mistakes are distortion errors [2, 5]. These typically consist of a slight alteration in the production of a sound due to the use of an incorrect vocal tract region, an incorrect tongue shape or placement, or exchanging voiced and voiceless sounds. Speech and language pathologists (SLPs) help children overcome distortion errors with speech exercises and by explaining how to correctly use their speech organs. Yet, it may be difficult for children to understand why their speech production is incorrect and how to correct it.

In order to help children understand how to correctly use their speech organs, this paper describes a virtual tutor, Frica, that can be used in speech and language therapy sessions for fricatives production correction and training. Frica automatically detects the place of articulation of European Portuguese (EP) fricatives and whether the vocal folds have been used.

Most previous work on automatic classification of fricatives by place of articulation and voicing has centered on the features that lead to good accuracy results. Studies on the classification of American English fricatives into voicing and place of articulation have proposed features that include the duration of the voiceless portion, relative amplitude and spectral flatness, spectral shape and peak location [6], and 2-D feature matrices obtained by applying higher order singular value decomposition on modulation spectrograms [7]. Chang *et al.* discuss a spectrogram frame selection procedure to improve the accuracy of a multilayer perceptron neural network for place and manner of articulation detection [8]. Cepstral coefficients have been proposed for Greek fricatives [9]. These studies achieved accuracies between 11% and 91% for place of articulation, and between 79% and 93% for voicing detection.

Convolutional layers in deep neural architectures can capture essential data patterns at different levels of abstraction. This paper uses convolutional neural networks (CNN) with log Mel filter banks to learn models for classifying children's EP fricative production. One of the models classifies fricative voicing while the other is used to detect place of articulation. The models were trained on EP speech samples from 356 children, and achieve an average accuracy of 90.40% for place of articulation and 90.93% for voicing, with F1 scores between 87.60% and 93.05% for place of articulation, and between 83.32% and 93.77% for voicing. While the average accuracy values are within the range of previous studies, those studies addressed adult speech. On the contrary, since Frica is meant to be used in speech therapy for children, the proposed classification models are trained on children's speech, with their intrinsic characteristics which are quite different from those of adult voices.

In addition, we carried out a multidimensional advanced data analysis on the convolutional filters of the proposed CNN first convolution layer aimed at contributing to a better technical understanding of how relevant features are learned from the log Mel filter bank. This analysis shows that relevant features, which characterize the individual classes are learned as early as in the first convolutional layer.

http://dx.doi.org/10.21437/Interspeech.2020-2821

Use of	Place of articulation		
vocal folds	Labiodental	Alveolar	Palato-alveolar
Voiceless	[f]	[s]	[ʃ]
Voiced	[v]	[z]	[ʒ]

Table 1: *Classification of fricatives by place of articulation and use of vocal folds.*

	Word position			Occurrences	Correct
	Initial	Medial	Final	in words	productions
f	7	4		11	2 983
v	2	7		9	2 413
s	13	11		24	6 625
z	2	6		8	2 251
ʃ	6	20	13	39	11 201
ʒ	5	5		10	2 879

Table 2: *Number of fricative phoneme occurrences within words and number of fricative productions.*

The main contributions of this paper are: (1) the analysis that demonstrates that the convolutional filters are efficient feature extractors for the classification of fricatives by place of articulation and voicing, and (2) the combination of CNN models for the real time classification of children's fricative productions in a virtual tutor for fricative production training. Frica gives children instant visual feedback on their fricative performance. With the help of this tool, children easily visualize their mistakes, and understand how to correct them.

2. Data collection

In order to identify the place of articulation and voicing characteristics of fricative production, Frica uses two classification models: one to identify the place of articulation, which we call M_{pa}, and another to identify whether the vocal folds were used, which we call M_{vf}. The classification models were trained with fricative sounds extracted from recordings of children's speech.

Fricatives in EP include [f][1] as in fish, [v] as in vulture, and the sibilants [s] as in snake, [z] as in zebra, [ʃ] as the *sh* sound in sheep, and [ʒ] as the s sound in Asia [11]. Table 1 characterizes these fricatives by place of articulation and voicing. Other fricatives in EP include variations of s and z [12] (not addressed due to lack of samples) and variations of the rhotic consonant *r* [13, 14] (not addressed due to its different nature.)

The speech samples consist of single words that were collected in three schools in the greater Lisbon area. Word samples were collected at a 44 100 Hz sampling rate, and later downsampled to 16 000 Hz. 356 children (182 girls and 174 boys) from 5 to 9 years old participated in the recordings. The recording sessions were led by SLPs. For more details see [15].

The recording protocol contained 79 words with 101 fricative occurrences in different word positions (table 2). Children produced 31 291 words, from which 27 723 were correct productions. The rightmost column in table 2 shows the number of correct productions for each fricative, which add up to 35 327.

These word productions were manually labeled by an SLP and a software engineer. In order to recognize the isolated phonemes, a new acoustic model for the Kaldi ASR was trained using only the correct word productions [16]. This was used to obtain the locations of the individual phonemes, which were then segmented. The IPA transcription was used to select only the fricative phonemes.

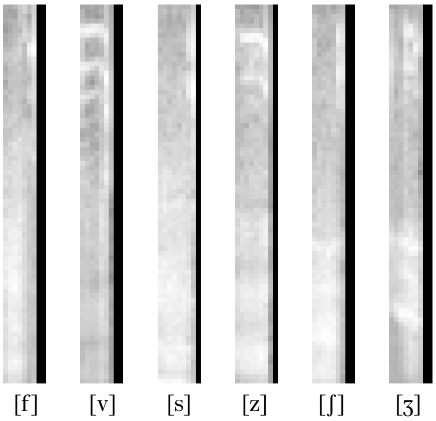


Figure 1: *Log Mel filter bank features for fricative productions.*

After being segmented, the fricatives were cut, or zero padded, so that the single waveforms all had the same length ($\approx 1/4$ s). This length was determined by observing the average length, and the maximum length for each phoneme class, in order to avoid padding or cutting the samples excessively.

3. Frica and the classification models

The same CNN architecture was used to train the classification models, M_{pa} and M_{vf} (section 3.1), used by the virtual tutor (section 3.4). Each individual model was obtained by using a different training set in the learning phase (section 3.2). After training the models, the codification of the convolution filters was validated (section 3.3).

3.1. The CNN architecture

After trying several neural network architectures for classification, the best results were obtained using a CNN and log Mel filter banks [17]. Building on these earlier results, this paper uses a CNN architecture adjusted to the current classification goals, that is, the classification of fricatives by place of articulation and voicing.

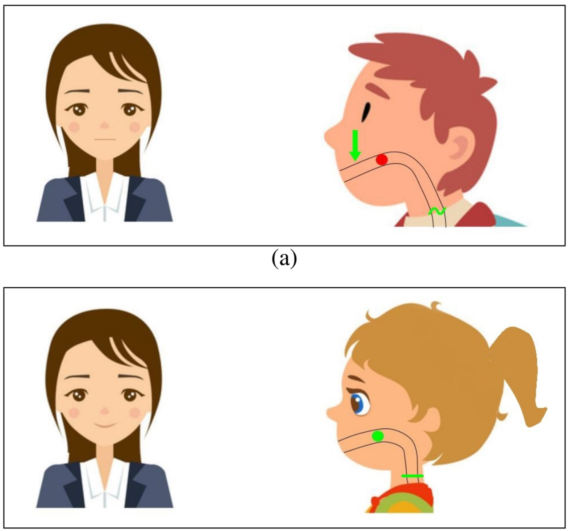
The input to the CNN consists of matrices of log Mel filter banks. These are 80×9 matrices (with 80 bins and 9 frames), that were extracted with a 25 ms window size and 10 ms shift size (figure 1 - the black regions in the figure are due to zero padding). Our approach consists of applying two dimensional, spectral-temporal, convolutions to the whole input matrix.

The CNN has two convolutional layers, each followed by the corresponding pooling layer. 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 [18], with a max norm of 2, was used for the filters in both layers.

The output from the last convolutional layer is flattened and then fed to a fully connected network with four hidden layers with 1 000, 500, 100 and 10 neurons, respectively. 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.

The convolutional and hidden layers used the rectified linear unit (ReLU) activation function, while the output layer used the softmax function. We applied dropout to the hidden layers,

[1] International phonetic alphabet (IPA) symbols [10].


(a)

(b)

Figure 2: *Frica gives visual feedback about sound production. The red and green symbols help the child visualize how to correctly produce the fricative sound. (a) Incorrect production of [z] - incorrect point of articulation, but correct use of vocal folds. (b) Correct production of [ʃ] - correct point of articulation and the vocal folds were not used.*

Class	# samples	Duration (mm:ss)
Labiodental	6 498	10:05
Alveolar	10 606	24:37
Palato-alveolar	18 223	45:18
Voiceless	9 164	16:28
Voiced	26 163	63:32

Table 3: *Fricative samples in each class.*

with a drop rate of 30% [19], and used the Adam optimizer as the loss function. The models were trained for 100 epochs, with a batch size of 10.

3.2. Training data

In order to learn M_{pa} and M_{vf}, the fricative sound samples were combined in different ways. We created a place of articulation training set (T_{pa}) and a voicing training set (T_{vf}), and the CNN was trained with each of these sets.

M_{pa} was trained by labeling the samples in table 2 by place of articulation. More specifically, T_{pa} has three classes: labiodental ([f] and [v] samples), alveolar ([s] and [z] samples), and palato-alveolar ([ʃ] and [ʒ]). On the other hand, T_{vf}, which was used to train M_{vf}, has two classes: unvoiced ([f], [s] and [ʃ]), and voiced ([v], [z] and [ʒ]). Table 3 shows the distribution of the data samples across the three place of articulation classes and the two voicing classes.

In order to learn each model, the data was randomly divided into training, validation and test sets. Since samples of the same fricative of a specific child are likely to have some correlation, to avoid any bias in the results, all productions of a given child c were placed either in the training, validation or test set. Thus, while there are several feature vectors from each child, all those vectors belong to the same part of the data set. The production of 70% of the children was used in the training set, 20% in the test set, and 10% in the validation set.

3.3. CNN feature representation

An additional analysis was carried out in order to partially validate the codification made by the proposed neural network's first convolution layer. The direct analysis of the fricatives' log Mel filter banks is not feasible, since their values are highly correlated and do not have a direct mapping to the intended classifications. The study focuses on the first level of convolution since the representation of input in this layer should allow access to a neural encoding that transforms and aggregates the input information based on the classification model. Indeed, the first layer is the one that is more easily explained regarding input and minimizes dependencies regarding the final classification. Recall that the first layer learns a convolution of size 10 × 2, i.e., ten log Mel bins at two consecutive time frames. This way, the maps learned by the first layer can be understood as relevant combinations of related log Mel features. Also, the second layer has a much more complex interpretation, since it combines several filters and is more influenced by target classification.

The convolution over the log Mel filter bank feature matrix (figure 1) with each filter in the first convolution layer results in 36 × 8 maps. These maps were analysed with help of `MultiSOM`, a software for exploring multidimensional data [20]. `MultiSOM` allows easy qualitative visual inspection of the input data as it is represented in the convolution map, e.g. [21]. The results are discussed in section 4.

3.4. Frica - a virtual tutor for fricative distortion correction

Frica can be used by SLPs in speech and language therapy sessions. The SLP can choose which fricative to practice, and Frica uses the output from M_{pa} and M_{vf} to offer visual feedback on voicing as well as the place of articulation used by the child.

To identify the place of articulation and voicing characteristics of the child's production, Frica looks into the values of the output layer for each CNN model. By using softmax on the fully connected output layer, the models have the scores of all classes summing to one, which can be used as probabilities. Fica's current version detects isolated fricatives by requiring a classification of at least 60%, to consider a given place of articulation or voicing characteristics for a sound production. This value already gives some confidence on the detected production characteristics, but higher values, like 90%, give even more confidence on the results. Adjusting these detection thresholds, Frica can be adapted to classify fricatives within words.

Visual feedback is given in real time on a simplified image of the vocal tract drawn on top of a child's face (figure 2). Using the output from M_{pa}, Frica draws a dot on the place of articulation that the child used. This is a green dot if the place of articulation is correct or red otherwise. When the point of articulation is not correct, Frica draws a green arrow indicating the expected point of articulation (figure 2.a). Similarly, Frica uses the output from M_{vf} to indicate if the child used the vocal folds. If the child's fricative production was voiced, a sinusoidal line is drawn on the vocal folds (figure 2.a). A straight line is used for voiceless productions (figure 2.b). The line is green if the vocal folds are correctly used, and red otherwise.

The proposed algorithm works in real time, allowing the tutor to respond to children's sound productions changes in an interactive manner. In this way, when children do not produce the desired sound correctly, they can immediately vary their speech production and see if they achieved the desired effect (by seeing if the dots and lines turn green).

Disabled filter	Average	Alveolar	Palato--alveolar	Labio-dental
–	90.40%	87.60%	93.05%	87.99%
23	**73,29%**	80,90%	73,41%	64,95%
37	**79,48%**	75,27%	86,44%	72,16%
1	86,42%	81,90%	90,06%	83,09%
38	88,73%	85,70%	92,12%	85,21%
50	88,98%	86,95%	91,70%	84,45%
44	89,01%	86,13%	92,72%	83,58%
35	89,22%	86,60%	92,43%	84,69%
18	89,38%	86,24%	92,55%	86,28%
29	89,49%	87,76%	92,01%	86,18%
43	89,60%	87,55%	92,53%	85,83%

Table 4: *Scores after disabling the 10 most relevant filters. The first column shows the filter that was disabled, followed by the average accuracy score, and the F1 scores for each class. The first row shows the values when no filter is disabled.*

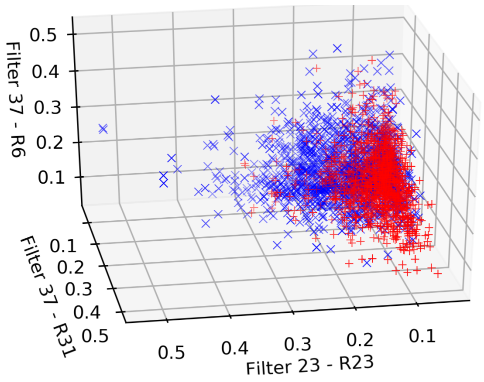


Figure 3: *$3D$ Scatter plot of target sounds [s] ($\times$) and [ʃ] ($+$) using the values of representative rows $R6$, $R31$ in filter 37, and row $R23$ in filter 23 in first convolutional map.*

4. Results

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.

These high F1-score values warrant that most times the virtual tutor 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.

As mentioned in section 3.3, we performed an analysis on the CNN filters' representation. The discussion here focuses on the first convolution layer but an analysis was also done on the original input and on the second convolution layer. The original input was found to have high feature correlation. Also, the difficult relation between the log Mel features and target classification has shown to lead to poor accuracy in previous studies [17]. The second layer convolution matrix presents more evolved features, but with a much stronger relationship with the target classes. Likewise, the direct relationship of these characteristics with the log Mel filter bank becomes too difficult.

The sounds with less complexity in the time domain are the [s] and [ʃ] fricatives (figure 1). Without loss of generality and for better illustrative purposes, both fricatives, and their representation in M_{pa} were selected for our discussion. This setting allows a better focus on the effects of the convolution transformation applied to log Mel filter banks. The analysis focused on time frames 2 to 4 (the ones with more pure sound representations) and discarded all representations smaller than 0.05. Such convolution patterns did not induce further activation in the neural classifier and so were not considered in this analysis.

There are 50 distinct convolution kernel filters generated by the deep training process for the first layer. We selected the most relevant ones in the M_{pa} model by a simple sensibility analyses. Each filter in the first convolution layer was turned off and the full model without that filter was applied to test data. Table 4 presents such information with each filter ordered by the score resulting from the impact of removing this filter on the model's accuracy. Filters 23 and 37 were selected for further study since these filters are ranked as having the highest impact on the average classification accuracy. The `multiSOM` advanced data exploration and data mining process helped us to identify the most representative features in each filter: features $R11$, $R23$ and $R30$ for filter 23 and features $R6$, $R15$, $R31$ for filter 37.

The 3D scatter plot of a conjunction of three of these representative features (features $R6$ and $R31$ from filter 37, and $R11$ from filter 23) shows relevant separation of both sounds (figure 3). The higher values for filter 37 $R31$ with lower enough values for filter 37 $R6$ and filter 23 $R23$ provide a distinctive cluster for [ʃ] sounds, while most values for the [s] sound values are surrounding the [ʃ] sounds. We also noticed that — as it could be expected on an internal neural encoding representation– the convolution filters presented very diverse and sometimes overlapping classes. Indeed the neural network classifier is trying to maximize the representative encoding power of each individual feature for all the sounds and data sets. Therefore, only by combining three characteristics it was possible to overcome this overlap of sounds.

This analysis shows that the relevant features, which characterize the individual classes, are learned as early as in the first convolutional layer. The features are more aggregated while also maintaining a direct relation with log Mel filter banks. The illustrative example also shows how the discriminating power given by composition of some features provides an important pool of (automatically learned) different encodings.

5. Conclusion

Frica, the proposed virtual tutor gives real time visual feedback in an easily understandable and appealing manner. In this way, it helps children understand how to achieve correct fricative productions. Frica uses CNN models to classify children's fricative productions. The models achieve F1 scores between 87.60% and 93.05% for place of articulation, and between 83.32% and 93.77% for voicing. Also, we showed that the convolutional layers are efficient feature extractors for the classification of fricatives by place of articulation and voicing.

6. 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 SLP students from Escola Superior de Saúde do Alcoitão who collaborated in the data collection, as well as Agrupamento de Escolas de Almeida Garrett, and the participating children. We also thank Cátia Pedrosa, Diogo Carrasco and Alberto Abad for the annotation of our corpus.

7. References

[1] L. Furlong, S. Erickson, and M. E. Morris, “Review: Computer-based speech therapy for childhood speech sound disorders.” *Journal of Communication Disorders*, vol. 68, pp. 50 – 69, 2017.

[2] J. Preston and M. L. Edwards, “Phonological awareness and types of sound errors in preschoolers with speech sound disorders,” *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 1, pp. 44–60, 2010.

[3] L. Nathan, J. Stackhouse, N. Goulandris, and M. J. Snowling, “The development of early literacy skills among children with speech difficulties: A test of the critical age hypothesis,” *Journal of Speech, Language, and Hearing Research*, vol. 47, pp. 377–391, 2004, american Speech-Language-Hearing Association.

[4] S. McLeod, *The international guide to speech acquisition*. Thomson Delmar Learning, 2007.

[5] I. Guimarães, C. Birrento, C. Figueiredo, and C. Flores, *Teste de articulação verbal*. Oficina Didáctica, Lisboa, Portugal, 2014.

[6] A. M. Abdelatty Ali, J. Van der Spiegel, and P. Mueller, “Acoustic-phonetic features for the automatic classification of fricatives,” *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2217–2235, 2001.

[7] K. D. Malde, A. Chittora, and H. A. Patil, “Classification of fricatives using novel modulation spectrogram based features,” *Pattern Recognition and Machine Intelligence*, pp. 134–139, 2013.

[8] S. Chang, S. Greenberg, and M. Wester, “An elitist approach to articulatory-acoustic feature classification.” *Speech Communication*, pp. 1725–1728, 2001.

[9] A. Athanasopoulou and I. Vogel, “The classification of greek fricatives with cepstral coefficients,” *Journal of The Acoustical Society of America*, vol. 129, 04 2011.

[10] *Handbook of the International Phonetic Association, A guide to the use of the international phonetic alphabet*. Cambridge University Press, 1999.

[11] M. Cruz-Ferreira, “Portuguese (european),” in *Handbook of the International Phonetic Association, A guide to the use of the international phonetic alphabet*. Cambridge, University Press, 1999.

[12] M. Mateus, “A mudança da língua no tempo e no espaço,” in *A Língua Portuguesa em Mudança*, M. Mateus and F. Bacelar do Nascimento, Eds. Editorial Caminho, Portugal, 2005.

[13] I. Rennicke and P. Martins, “As realizações fonéticas de /R/ em português europeu: análise de um corpus dialetal e implicações no sistema fonológico,” in *Encontro Nacional da Associação Portuguesa de Linguística*, 2013, pp. 509–523.

[14] A. Grossinho, I. Guimarães, J. Magalhães, and S. Cavaco, “Robust phoneme recognition for a speech therapy environment,” in *Proceedings of IEEE International Conference on Serious Games and Applications for Health (SeGAH)*, 2016.

[15] M. Grilo, I. Guimarães, M. Ascensão, A. Abad, I. Anjos, J. Magalhães, and S. Cavaco, “The BioVisualSpeech European Portuguese sibilants corpus,” in *International Conference on Computational Processing of the Portuguese Language (PROPOR)*. Springer, 2020, pp. 23–33.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[17] I. Anjos, N. Marques, M. Grilo, I. Guimarães, J. Magalhães, and S. Cavaco, “Sibilant consonants classification with deep neural networks,” in *Proceedings of 19th European Conference on Artificial Intelligence (EPIA)*, 2019.

[18] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, “Efficient backprop,” in *Neural Networks: Tricks of the Trade: Second Edition*, G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–48.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[20] N. C. Marques, B. Silva, and H. Santos, “An interactive interface for multi-dimensional data stream analysis,” in *Information Visualisation (IV), 2016 20th International Conference*. IEEE, 2016, pp. 223–229.

[21] N. C. Marques, M. Monteiro, and B. Silva, “Analysis of a token density metric for concern detection in matlab sources using UbiSOM,” *Expert Systems*, vol. 35, no. 4, 2018.