TOWARDS INTERPRETING DEEP LEARNING MODELS TO UNDERSTAND LOSS OF SPEECH INTELLIGIBILITY IN SPEECH DISORDERS STEP 2 : CONTRIBUTION OF THE EMERGENCE OF PHONETIC TRAITS

Sondes Abderrazek[1], *Corinne Fredouille*[1], *Alain Ghio*[2], *Muriel Lalain*[2], *Christine Meunier*[2], *Virginie Woisard*[3]

[1]LIA, Avignon University, France
[2]Aix-Marseille Univ, LPL, CNRS, Aix-en-Provence, France
[3]UT2J, Octogone-Lordat, Toulouse University & Toulouse Hospital, France

ABSTRACT

Apart from the impressive performance it has achieved in several tasks, one of the most important factors remaining for the continuous progress of deep learning is the increased work related to interpretability, especially in a medical context. In a recent work, we presented competitive performance achieved with a CNN-based model trained on normal speech for the French phone classification and how it correlates well with different perceptual measures when exposed to disordered speech. This paper extends that work by focusing on interpretability. Here, the goal is to get insights into the way in which neural representations shape the final task of phone classification so that it can be used further to explain the loss of intelligibility in disordered speech. In this way, an original framework is proposed, relying firstly on the neural activity and a novel representation per neuron, here considering the phone classification, and, secondly, permitting to identify a set of neurons devoted to the detection of specific phonetic traits on normal speech. Faced to disordered speech, a degradation of that set of neurons is observed, demonstrating a loss of specific phonetic traits in some patients involved, and the potentiality of the proposed approaches to inform about speech alteration.

Index Terms— Deep learning, Interpretability, Phonetic traits, Intelligibility, Head and Neck Cancer, Speech disorders.

1. INTRODUCTION

There is a common agreement that deep learning has achieved impressive performance in several tasks. Their growing popularity has aroused interest in how these widely used tools really work and has given rise to a new line of research dedicated to their analysis. Interpretability becomes even more important when Deep Neural Networks (DNNs) are applied in a domain as medicine where transparency is crucial and errors cannot be tolerated. In such a case, understanding the nature of encoded representation performed by the model is a way to guarantee its reliability. In this context, several approaches have been proposed to get insights into the internal functioning of these models. The most straight-forward interpretation technique is based on the analysis of neuronal responses to a representative input sample. To this end, selectivity property of individual units has been of particular interest to researchers. Among works we can find those relying on *the whole activation* of neuron as the phone selectivity index attributed by [1] to units of an MLP trained for phone recognition task, while [2] and [3] characterized the selectivity of individual units through properties of samples *maximizing their activation*. Of relevance to the present work, several studies revealed that human-interpretable concepts can emerge without explicitly constraining the deep neural network to do it. Notably, [4] and [5] were able to discover automatically meaningful object detectors in their CNN trained for scene classification. As well, in [6], authors have shown that semantic part detectors emerge in object classifiers. While [7] introduced a more general method for interpretability, called Network Dissection, measuring the alignment between convolutional units and visual interpretable concepts. Most of the work reported above are interested in interpreting CNNs applied in computer vision domain. Even though CNNs have proven to be well suited for other tasks beyond computer vision such as speech signal processing, providing knowledge of how this input feature space is transformed into gradually higher levels of representation is still a challenging task due to the complex nature of speech signal, characterized within acoustic features.

The work presented here is part of a long-term project which aims to determine the linguistic units that contribute most to the maintenance or loss of the intelligibility in speech disorders. Different steps have been identified to reach this objective : (1) Modeling the characteristics of phonemic units of "normal" speech thanks to deep learning based system dedicated to a basic task of phone classification, (2) investigating the representational properties of the model in terms of phonetic contents, (3) the transfer of this deep learning modeling into a prediction task of intelligibility typically in the context of normal and disordered speech and the study of its capacity in yielding reasonable interpretation of the phonemic unit contribution in speech intelligibility and its variation (improvement or alteration). In a previous work [8], we proposed a CNN-based architecture, trained on healthy speech for French phone classification task, to respond partly to the first step of that project. In this context, the encoding capability of the CNN-based model for the targeted task was demonstrated through very high correlation rates between its phone classification rates and the perceptual measures available for both patients and control speakers present in the disordered speech corpus involved.

In this paper, the main contribution is to show that phonetic feature detectors emerge in the fully connected layers of the CNN-based architecture mentioned above. To this end, an original framework based on neuronal activity was proposed in order to identify the set of interpretable neurons giving rise to such information on normal speech. This framework involves a vector representation of neuronal activity of units belonging to classification layers, coupled with a scoring approach enabling to highlight the capacity of neurons to detect specific phonetic features characterizing French phones. Exposed to speech recordings produced by patients suffering from speech disorders, the aim of this framework is to reveal a degraded behavior of the identified set of neurons and the loss of some specific phonetic traits in those patients, in different manner according to their speech impairment and level of speech intelligibility.

2. CORPUS AND PHONETIC FEATURES

As mentioned in the introduction, the search of neurons giving rise to phonetic features through the CNN model was carried out on normal/healthy speech. In this purpose, the **BREF corpus** was involved. This corpus was also used to train the CNN architecture described in the next section, on which this work relies. The **BREF corpus** is composed of French read-speech records produced by 120 male and female speakers, while reading texts from newspapers [9], leading to about 115h of speech. All the speech productions were aligned automatically by using a forced-alignment system, commonly based on a Viterbi algorithm and three-state context-independent Hidden Markov Models (HMM) trained on separate French speech data. Thus, temporal boundaries of phones in speech records are available.

The ***C2SI-LEC*** **corpus** is a sub-part of the French speech corpus recorded within the C2SI project between 2015 and 2017 [10]. The overall corpus includes patients with Head and Neck Cancers (oral cavity or oropharynx) and control speakers. All patients underwent dedicated treatment consisting of surgery, and/or radiation therapy, and/or chemotherapy. During the recording protocol, all speakers were asked to record different speech production tasks (sustained /a/ vowels, isolated pseudo-words, text or sentences reading, image description and brief interviews to get spontaneous speech). Different perceptual evaluations were conducted by a jury of 5 to 6 experts (clinicians or speech therapists) including notably measures of speech severity (LEC-Sev) and intelligibility (LEC-Intel), on a 0-10 scale (0 - major speech disorder; 10 - no speech disorder) from the text reading task, and measures of phonemic alteration (DES-Phon) on a 0-3 scale (0 - no disorder; 3 - major disorders) from the image description task. Ratings given by the experts are averaged to provide unique values per speaker for the different perceptual evaluations.
In this study, the focus is made on the reading task only, considering 89 speech records produced by 82 patients (7 patients were recorded twice) and 25 records for 24 control speakers (a control speaker was recorded twice). Based on the reading text (systematically corrected in case of reading errors), all the speech productions were aligned automatically, in a similar way as described for the BREF corpus.

Phonetic Features of French Phones
French phones can be characterized within a set of phonetic features that distinguish them. Different categorizations exist to define this set. In this paper, we adopted the approach described in [11] to characterize the French phones in terms of phonetic traits. Here, the notion of phonetic traits imposes a binary status (i.e. equals 1 if phonetic trait is present in the phone or 0 if absent). In order to define phonetic traits in a way they are phonetically and acoustically pertinent, a distinction between vowels and consonants was done as in [12]. In this context, the following phonetic traits will be considered : *nasal, back, high, round, open* for vowels and *sonorant, continuant, nasal, voiced, compact, acute* for consonants. Following the same logic, we investigated later the capacity of the hidden neurons to detect phonetic traits in separate spaces for vowels and consonants. This will be detailed on the upcoming section.

3. FRAMEWORK DESCRIPTION

The fundamental goal of this work is to understand the nature of information encoding at various stages of our CNN detailed in 3.1 and the way in which its neural representations shape the final task of phone classification and, simultaneously, are able to emerge specific phonetic traits. As reported in the state of the art, the most straight forward interpretation approach is to investigate the neuron responses. As we go deeper in the layers of the neural network, observing the organization and structure of the activation space in response to specific inputs is a way to characterize how the learned representations evolve along the hierarchically structured layers.

3.1. Model Architecture

The CNN architecture used in this study was detailed in our previous work [8]. It is recalled that the CNN inputs consist of a sliding context window of 11 successive acoustic frames, each having 40 log Mel-filter bank energy features along with their first and second derivatives. These features were computed on a 20ms window with an overlap of 10ms between two adjacent frames and served as an input to the CNN. Inspired by [13], the CNN consists of two pairs of convolution and max-pooling layers followed by three fully connected layers of 1024 neurons, with a ReLU activation function. Finally, a softmax layer corresponds to the posterior probability of each class associated with 31 French phones and silence. It should be mentioned that the three fully connected layers, denoted FC1, FC2, and FC3 in the rest of the paper, will be the focus of the neuronal activity based framework proposed in next sections.

One key factor of a reliable interpretation is the selection of data used to ensure this task. Ideally, our CNN could be fully characterized if all possible input patterns could be presented, i.e. frames issued from the different French phones produced in all possible contexts, then neural responses measured for each of these inputs. It is obvious that in practice, the adopted **BREF corpus** does not cover all these possibilities, having to meet the need for training, validation and interpretation of the model. However, we tend to build the richest interpretation dataset possible covering the 31 French phones studied here, which the CNN was trained on for classification purpose. For this reason, we did not proceed to a random choice of frames for the interpretation set. Rather, we considered all the frames associated with speech segments (yielded by the automatic forced alignment) related to a complete phone production as well as involving different phone contexts and all the speakers available in the corpus. In addition to the diversity of the frames, the dataset was balanced in order to achieve a roughly equal distribution of frames over the 31 phones. Thus, the generated dataset used for interpretation purpose (different from datasets used for CNN training and validation), named *BREF-Int*, includes almost 85K of frames, and resulted in a classification accuracy of 80.8% once fed to the trained CNN. We already evoked in our previous work that our model is able to generalize to new data since it reached 74% accuracy when evaluated on healthy control utterances of the *C2SI-LEC* corpus described in section 2.

3.2. Representation Vectors of Neurons

In order to represent a neuron, the CNN is fed with the *BREF-Int* dataset and the activation $h_{n,i}$ of the neuron n is extracted, given the i^{th} input frame. A normalized activation $a_{n,i}$ is then calculated for each neuron by dividing the activation values of the neuron for different input frames of the dataset by the maximum of these values reached by the same neuron over all the samples; $a_{n,i} = \frac{h_{n,i}}{h_{max,n}}$ where $h_{max,n} = \max h_{n,j} \forall j$.
Afterwards, a process to generate representation vectors reflecting the neuronal activity was set up. An illustration of the steps we passed through is shown in fig.1. For a neuron n, a histogram is generated for each phone k in order to approximate the distribution of the neuron activations as response to all the frames associated with k.
The histogram displays the number of frames falling into each interval of normalized activation, also called bins, which have equal width and divide the entire range of normalized activation $[0; 1]$. In our case, we fixed the number of bins to 20. Subsequently, a vec-

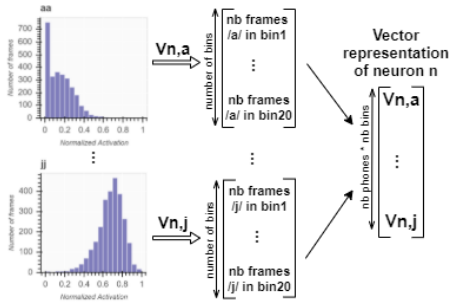

Fig. 1: Process of representation vector generation for a neuron

tor $V_{n,k} \in N^{20}$ containing the number of frames appearing in each bin is derived from each histogram. Finally, a concatenation of these vectors generated for each of the 31 phones for a given neuron results in a 620-dimensional representation vector, and is considered later as characterizing the neuron n for interpretation purpose. Interesting insights about the organization of neurons per layer and the emergence of interpretable ones, in regards with phonetic traits, will be revealed in next sections.

3.3. Capacity of Neurons to detect Phonetic Traits

In order to align each neuron with the phonetic traits mentioned in section 2 and to measure the hidden unit encoding degree for each of these features, a score has to be defined. The aim of this score is to quantify the degree to which a unit detects the presence of a phonetic trait based on the activation of phones associated with, and complementary those which do not present this phonetic trait. To this end, the analysis was still based on the values of normalized activations for the individual units as a response to the dataset, already presented in the previous section.

For each neuron n, let $A_{k,n}$ be the normalized activation values of n for all the frame samples belonging to the phone k. We note the median activation value of the neuron n for the phone k as $m_{A_{k,n}}$. The score S_{n,T_x}, quantifying the degree to which a unit detects the presence of a phonetic trait, is therefore calculated for each neuron n and phonetic trait T_x as follows:

$$S_{n,T_x} = \frac{1}{|T_x|} \sum_{k \in T_x} m_{A_{k,n}} - \frac{1}{|\bar{T}_x|} \sum_{k \in \bar{T}_x} m_{A_{k,n}} \qquad (1)$$

where $x \in [v, c]$ denotes the macro-class of either vowels or consonants, respectively v and c. Consequently, T_v denotes a vowel phonetic trait where: $T_v \in \{$*nasal, back, high, round, open*$\}$ and T_c denotes a consonant phonetic trait where: $T_c \in \{$*sonorant, continuant, nasal, voiced, compact, acute*$\}$. It has to be noted that since phonetic traits are binary concepts characterizing separately vowels and consonants, the score is calculated, consequently, taking into account either vowels or consonants. It is important to mention that the score $S_{n,T_x} \in [-1; 1]$. Therefore, a strong value close to 1 reflects the fact that the neuron is a strong detector for the phonetic trait in question since it distinguished by a high activation level the phones presenting the features. At the same time, a very low activation level represents the complementary set of phones not presenting this feature. Conversely, when a neuron has a very low score close to -1, it means that the neuron is a strong detector for the non phonetic trait, which is relevant as well.

Now once we have a score reflecting how well a neuron encodes a given phonetic trait, we consider that the neuron n is a detector of the phonetic trait T_x if S_{n,T_x} exceeds a threshold. And conversely, if S_{n,T_x} is below the threshold, then the neuron n is considered as detector of the absence of phonetic trait T_x. Clearly, different thresholds could lead to different numbers of neurons selected as phonetic trait detectors across layers. However, we observe that it does not result in a significant change in term of the distribution of this set of neurons over the different phonetic traits. Thus, we have empirically fixed the threshold to value ± 0.25.

Given that a neuron can be a detector for several phonetic traits (associated with relevant scores respecting the threshold), the top phonetic trait is chosen in this case. If the neuron is identified as detector for multiple phonetic features belonging to both vowel and consonant macro-classes, it will be considered as detector for the top phonetic trait for both vowels and consonants.

4. RESULTS

4.1. Emergence of phonetic features

As we previously detailed, a set of 620-dimensional vectors extracting the hidden representations of neurons was prepared. A projection of these representation vectors into a 2-dimensional space was performed by t-Distributed Stochastic Neighbour Embedding (t-SNE) [14] for each layer. Since t-SNE applies a non-linear dimensionality reduction technique where the focus is on keeping the very similar data points close together in lower-dimensional space, we expect to have a visualization where neurons with similar encoding properties appear in clusters. The first aim at this stage is to explore to what decomposition the hidden layers converged to ensure the final task of phone classification. The second aim is to determine if clusters of neurons associated with a specific phonetic trait (according to the score we defined) can be highlighted. Since our focus is basically to analyze neurons with phonetic feature encoding properties, only identified neurons respecting the fixed threshold will be displayed in the following t-SNE visualization plots. Thus, a specific color is attributed to each neuron according to the phonetic trait it detects based on its score.

Fig. 2 illustrates the neurons considered as detectors for the consonant phonetic traits. This visualization reveals impressive insights Firstly, the absence of FC1 visualization in the plot is due to the fact that none of its neurons were identified as a phonetic trait detector, neither for vowels nor for consonants. On the other hand, the number of phonetic trait detectors increases by a factor of close to two when we go deeper in layers towards the final layer performing phone classification. Indeed, although FC2 and FC3 have the same number of hidden neurons (1024), the total number of neurons detecting consonant phonetic traits has increased from 206 detectors in FC2 to 373 in FC3. This emergence of phonetic trait detectors when going deeper in layers suggests these features allow discrimination among phone classes. In fig. 2, we can observe the presence of dense neuron clusters automatically identified as encoding the same phonetic trait. To analyze in more details, fig. 2(b) and fig. 2(d) show the sorted count of neurons detecting consonant phonetic traits in FC2 and FC3 respectively. We observe that neurons detecting the nasality trait have the strongest presence in both layers. A similar study was performed on the vowel phonetic traits and generated roughly the same patterns. These results suggest that phonetic trait detection is an important part of the representation built by the CNN to obtain discriminative information for the final task of phone classification.

4.2. Application to disordered speech

As specified in the previous section, *BREF-Int* dataset issued from the *BREF* corpus was used to fix the set of interpretable neurons in each layer, that we considered as emerging phonetic features. Let $X_{BREF,L,x}$ be this set of neurons for the layer L and the macro-class

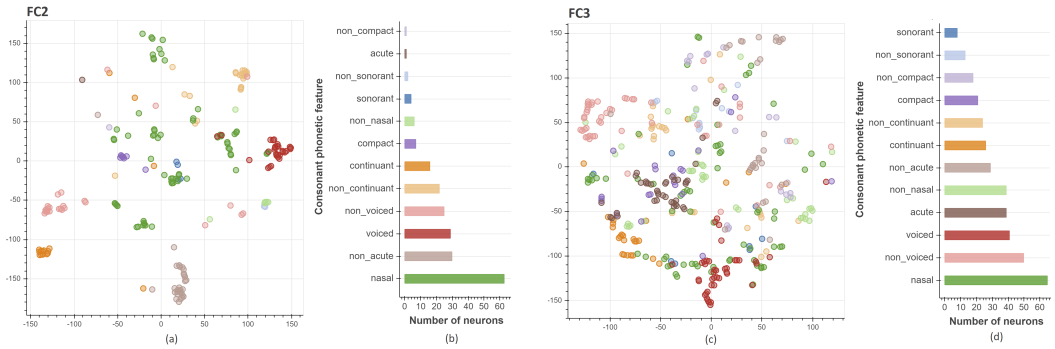

Fig. 2: t-SNE visualization highlighting neurons with phonetic trait encoding properties for consonants: (a) & (c) Plots of the embedded neurons of layers FC2 and FC3. (b) & (d) Sorted counts of neurons in FC2 and FC3 detecting each of the consonant phonetic traits.

x. In this section, we want to examine to which extent we can rely on this set of neurons to extract relevant interpretations about the speech alteration by injecting disordered speech.

To do so, each individual recording per speaker (involving controls and patients) available in *C2SI-LEC* corpus is injected in the CNN following the same process described in section 3. This permits to define the set of neurons having a score that exceeds the predefined threshold fixed on the *BREF-Int* dataset. It is important to note that even though we are no more dealing with balanced data in terms of number of samples per phone, we assume that the proposed score is relatively robust when exposed to highly imbalanced data (*C2SI-LEC* recordings) since it is based on the median calculus per phone distribution which is less sensitive to outliers than the mean. Once the set of neurons being selected for the corresponding speaker *spk*, noted $X_{spk,L,x}$, it will be compared to the corresponding set of interpretable neurons issued from the *BREF-Int* dataset. This comparison is based on the Sørensen–Dice index, noted *SDI*, used to measure the similarity between two sets of data. In our case, for a given layer L, macro-class x and *C2SI-LEC* speaker *spk*:

$$SDI_{spk,L,x} = \frac{2|X_{spk,L,x} \cap X_{BREF,L,x}|}{|X_{spk,L,x}| + |X_{BREF,L,x}|} \quad (2)$$

Ranging between 0 and 1, a strong value of SDI_{spk} close to 1 reflects that the two sets of neurons X_{spk} and X_{BREF} are almost equal. Precisely, the evoked neural responses of the set of interpretable neurons to the *C2SI-LEC* speaker were as strongly selective for phonetic features as they were for the same ones on the *BREF-Int* dataset. In other words, this means that the speech production of *C2SI-LEC* speaker in question presents acoustic characteristics close to *BREF* healthy speakers in terms of phonetic feature production (which is supposed to be the case for most of control speakers). While a low score close to 0 means that the two sets of neurons X_{spk} and X_{BREF} are almost disjoint. This implies that almost none of the phonetic feature detectors has expressly provided a selective response for a phonetic feature when exposed to the *C2SI-LEC* speaker *spk*. Consequently, we can assume that the speech production of the speaker in question does not exhibit typical acoustic characteristics. To confirm this assumption, Pearson correlation coefficients, noted *r*, were calculated between *SDI* and each of the perceptual measures for the overall set of C2SI speakers. Table 1 is a sum-up of the obtained correlations per layer and macro-class of either vowels or consonants. Firstly, it can be mentioned that these correlation rates are rather coherent with those reached in [8] and reported in Table 1 (CNN accuracy for the phone classification task). Secondly, *SDI* scores issued from FC3 correlates better with the different perceptual measures than those from FC2. This reflects that the more we get close to the phone-level decisions, perceptual measures are better "represented". Correspondingly, this is consistent with conclusions raised in section 4.1, when observing the increase in number of interpretable neurons detecting phonetic features when going deeper in layers. This finding comes to strengthen that phonetic features contribute to the discrimination of phones, and will be therefore very interesting for the characterization of speech disorders and related to intelligibility loss.

Table 1: Correlation between Sørensen–Dice index and perceptual measures considering vowels or consonants and layers FC2 and FC3.

		LEC-Sev	LEC-Intel	DES-Phon
CNN Accuracy		0.91	0.78	−0.88
FC2	Vowels	0.82	0.66	−0.76
	Consonants	0.85	0.75	−0.78
FC3	Vowels	0.84	0.73	−0.79
	Consonants	0.88	0.78	−0.82

5. CONCLUSION AND FURTHER WORK

In this paper, we argue that an interpretation involving phone-level decisions, as well as having access to more abstract level of representation as phonetic features, is of a great interest to go into the characteristics of disordered speech. Therefore, we propose a complete framework in order to analyze the role of dense layers and individual hidden units within a CNN model. The latter was trained for the task of French phone classification, but the proposed framework can be applied for other kinds of classification tasks or application domains. In our specific context, this framework reveals the presence of interpretable neurons detecting distinctive phonetic traits in the dense layers of the CNN while dealing with normal/healthy speech. Exposed to disordered speech (due to Head and Neck Cancers), we observed, based on strong correlations with perceptual evaluation of speech disorders by experts, that this set of phonetic trait detectors is less efficient while alterations in speech production increase. In further work, we will confront this loss of efficiency and related phonetic traits with the clinical information available per patient. Preliminary analyses show in particular that the distinction between some groups of patients according to the size of their tumors is possible within the proposed framework and, notably, the *SDI* scores. Finally, the sets of interpretable neurons, able to detect phonetic traits, will be at the core of the third step of our overall project, related to intelligibility.

6. ACKNOWLEDGMENT

This work has been carried out thanks to the French National Research Agency in RUGBI project untitled "Looking for Relevant linguistic Units to improve the intelliGiBIlity measurement of speech production disorders" (Grant n°ANR-18-CE45-0008-04).

7. REFERENCES

[1] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "Exploring how deep neural networks form phonemic categories," in *Proceedings of Interspeech'15*, Dresden, Germany, 2015.

[2] Ivet Rafegas, Maria Vanrell, Luís A. Alexandre, and Guillem Arias, "Understanding trained CNNs by indexing neuron selectivity," *Pattern Recognition Letters*, vol. 136, pp. 318–325, 2020.

[3] Matthew D. Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds., Cham, 2014, pp. 818–833, Springer International Publishing.

[4] B. Zhou, A. Khosla, Àgata Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," *CoRR*, vol. abs/1412.6856, 2015.

[5] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba, "Understanding the role of individual units in a deep neural network," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30071–30078, Sep 2020.

[6] Abel Gonzalez-Garcia, Davide Modolo, and Vittorio Ferrari, "Do semantic parts emerge in convolutional neural networks?," *International Journal of Computer Vision*, vol. 126, pp. 476–494, 2017.

[7] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2131–2145, 2019.

[8] S. Abderrazek, C. Fredouille, A. Ghio, M. Lalain, C. Meunier, and V. Woisard, "Towards interpreting deep learning models to understand loss of speech intelligibility in speech disorders, step 1 : CNN model-based phone classification," in *Proceedings of Interspeech'20*, Shanghai, China, October, 2020.

[9] L. F. Lamel, J. L. Gauvain, and M. Eskénazi, "BREF, a large vocabulary spoken corpus for french," in *Proceedings of European Conference on Speech Communication and Technology (Eurospeech'91)*, Genoa, Italy, 1991, pp. 505–508.

[10] Corine Astesano, Mathieu Balaguer, Jérôme Farinas, Corinne Fredouille, Pascal Gaillard, Alain Ghio, Laurence Giusti, Imed Laaridh, Muriel Lalain, Benoit Lepage, Julie Mauclair, Olivier Nocaudie, Julien Pinquier, Oriol Pont, Gilles Pouchoulin, Puech Michele, Danièle Robert, Etienne Sicard, and Virginie Woisard, "Carcinologic Speech Severity Index Project: A Database of Speech Disorders Productions to Assess Quality of Life Related to Speech After Cancer," in *Language Resources and Evaluation Conference (LREC), Miyazak, Japon*, http://www.elra.info, may 2018, European Language Resources Association (ELRA).

[11] Alain Ghio, Muriel Lalain, Laurence Giusti, Corinne Fredouille, and Virginie Woisard, "How to Compare Automatically Two Phonological Strings: Application to Intelligibility Measurement in the Case of Atypical Speech," in *12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May 2020, ELRA, pp. 1682–1687.

[12] Alain Ghio, *Achile : un dispositif de décodage acoustico-phonétique et d'identification lexicale indépendant du locuteur à partir de modules mixtes*, Phd thesis (in French), Université d'Aix Marseille, Nov. 1997.

[13] Thomas Pellegrini and Sandrine Mouysset, "Inferring phonemic classes from CNN activation maps using clustering techniques," in *Proceedings of Interspeech'16*, San Francisco, US, 2016.

[14] Laurens van der Maaten and Geoffrey Hinton, "Viualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.