# Towards Interpreting Deep Learning Models to Understand Loss of Speech Intelligibility in Speech Disorders - Step 1: CNN Model-Based Phone Classification

Sondes Abderrazek, Corinne Fredouille, Alain Ghio, Muriel Lalain, Christine Meunier, Virginie Woisard

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

# Towards Interpreting Deep Learning Models to Understand Loss of Speech Intelligibility in Speech Disorders
# Step 1 : CNN model-based phone classification

*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
(sondes.abderrazek,corinne.fredouille)@univ-avignon.fr

## Abstract

Perceptual measurement is still the most common method for assessing disordered speech in clinical practice. The subjectivity of such a measure, strongly due to human nature, but also to its lack of interpretation with regard to local alterations in speech units, strongly motivates a sophisticated tool for objective evaluation. Of interest is the increasing performance of Deep Neural Networks in speech applications, but more importantly the fact that they are no longer considered as black boxes. The work carried out here is the first step in a long-term research project, which aims to determine the linguistic units that contribute most to the maintenance or loss of the intelligibility in speech disorders. In this context, we study a CNN trained on normal speech for a classification task of phones and tested on pathological speech. The aim of this first study is to analyze the response of the CNN model to disordered speech in order to study later its effectiveness in providing relevant knowledge in terms of speech severity or loss of intelligibility. Compared to perceptual severity and intelligibility measures, the results revealed a very strong correlation between these metrics and our classifier performance scores, which is very promising for future work.

**Index Terms**: speech disorders, Head and Neck Cancer, perceptual evaluation, speech intelligibility, objective assessment, deep learning, phone classification.

## 1. Introduction

Speech intelligibility is defined in [1] by "the degree to which the speaker's intended message is recovered by the listener". When a patient suffers from speech disorders, whether due to dysarthria, a consequence of a neuro-degenerative disease such as Parkinson's disease, or due to a cancer of the head and neck, this intelligibility can be dramatically degraded. Different approaches are currently used in clinical practice to measure the intelligibility of a patient's speech and its progress in the event of rehabilitation or therapeutic treatment (pre / post surgical operation, radiotherapy, chemotherapy, ...), which are mainly based on perceptual evaluation [2][3]. However, it is well known in the literature that perceptual evaluation is subject to controversy because of its subjective nature, its intra- and inter-judgment variability and its lack of reproducibility [4]. In addition, the protocols available for this type of evaluation are not suitable for a precise analysis of the degree of intelligibility and its evolution, except from a global point of view. This could explain why there is no rehabilitation protocol specifically targeted at particular linguistic units to improve the intelligibility of patients. In this context, if clinicians need objective and reliable measures for the evaluation of speech intelligibility, the identification of the linguistic units that contribute most to the maintenance or loss of intelligibility is also a requirement to improve clinical protocols for patient management.

If the integration of deep learning-based approaches in the field of clinical phonetics can be considered as recent compared to the fields of speech or vision processing [5, 6], the literature now reports numerous studies linking deep learning and speech impairment. This reluctance towards deep learning came mainly from the lack of large corpora of pathological speech available, a major obstacle in the field of clinical phonetics, but a predominant factor in training systems based on deep neural networks. However, advances in neural architectures as well as transfer learning techniques applied to deep learning have largely enabled this opening up. Among these works, we can cite [7, 8, 9, 10].

The study presented in this paper is one of the main objectives of a long-term research project, which is the search for linguistic units playing a significant role in speech intelligibility, and therefore in its loss in the event of speech disorders. Inspired by [11] and [12] on the modeling of the characteristics of the different phonemic units of speech through deep learning, the long-term research work that we are carrying out is based on an original approach we propose, dedicated to the identification of these linguistic units from an acoustic point of view. This overall approach is based on three steps: (1) Modeling the characteristics of phonemic units of "normal" speech thanks to a system based on deep learning dedicated to a basic task of phone classification considered as the most relevant, (2) The transfer of this deep learning modeling into a prediction task of intelligibility typically in the context of normal and disordered speech, (3) Investigating the representational properties of the model and its capacity in yielding reasonable interpretation of the phonemic unit contribution in speech intelligibility and its variation (improvement or alteration). Concerned by the step 1, this paper presents the deep neural network architecture chosen for the modeling of the phonemic units on "normal" speech and examines its behaviour when exposed to disordered speech (here patients with Head and Neck Cancer). The rest of the paper is dedicated as follows. Section 2 describes corpora implied in this work. Section 3 provides a detailed description of step 1 of the overall proposed approach. Section 4 presents the experimental validation of this step 1. Finally, conclusions and perspectives on the rest of the work will be given in section 5.



http://dx.doi.org/10.21437/Interspeech.2020-2239

## 2. Corpus

Both corpora used in this paper are described below.

The **BREF corpus** is composed of French read-speech records produced by 120 speakers, recruited in the region of Paris, while reading texts from newspapers. Developed in 90's, this corpus was designed to provide continuous speech for the development and the evaluation of Automatic Speech Recognition systems and for phonological variation modeling [13]. In this paper, read-speech records from 65 female and 47 male speakers are used, representing about 115h of speech.
Based on the reading texts, 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 frontiers of all the 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 [14]. The overall corpus includes patients with Head and Neck Cancers (oral cavity or oropharynx) and control speakers, recruited in the southwest of France. All patients underwent dedicated treatment consisting of surgery, and/or radiation therapy, and/or chemotherapy. During the recording protocol designed specifically for the C2SI project, 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 measures of speech severity and intelligibility, on a 0-10 scale (0 - major speech disorder; 10 - no speech disorder), on both the text reading and image description tasks, and measures of the voice quality, the degree of alteration of resonance, prosody and phonemic production on a 0-3 scale (0 - no disorder; 3 - major disorders) on the image description task. Ratings given by the experts according to each kind of measurement are averaged to provide unique values for each speaker. It is important to point out that, even designated as speech intelligibility, the related perceptual assessment task has to be more considered as a comprehensibility measurement as reported in [15] since it integrates contextual information in addition to the acoustic-phonetic information, in the speech decoding process. Indeed, the text used in the protocol is relatively short and may be memorized during evaluation by experts (if it is not already known). This probably leads to an overestimation of speech intelligibility measures for the patients notably since experts can deduce the heard text despite speech production errors. Severity being assessed taking into account the overall speech signal degradation, it is less influenced by the undesirable effect of text memorizing. This difference can explain why statistics presented in figure 1 show a smaller variation of intelligibility ratings compared to those of severity (as well as to phonemic alteration even if the scale is not comparable).
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 during two different sessions) and 25 records for 24 control speakers (a control speaker was recorded twice in the same session). This sub-corpus is named C2SI-LEC in the rest of the paper. The perceptual assessment of the speech severity (LEC-Sev) and intelligibility (LEC-Intel) on the reading task as well as the phonemic alteration (DES-Phon) on the image description task is concerned here.

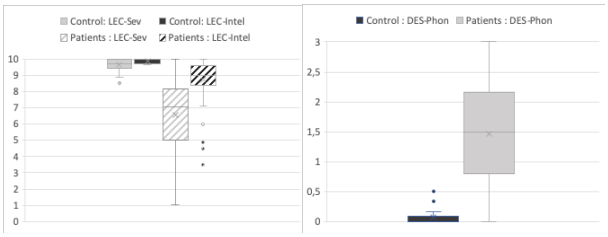



Figure 1: *Perceptual evaluation by an expert jury : box plots of the different measures of speech severity (LEC-Sev), speech intelligibility (LEC-Intel), and phonemic alteration (DES-Phon)*

Based on the reading text (systematically corrected in case of reading errors), all the speech productions were aligned automatically, by applying the same system as described for the BREF corpus.

## 3. Methodology

As mentioned above, this study is part of a long-term project, which aims to determine the linguistic units playing a significant role in the maintenance or loss of the speech intelligibility in speech disorders. As the first step, the modeling of the characteristics of a set of French phonetic units, based on deep learning and centered on a task of phone classification is investigated. We believe that these choices are the most relevant for the objective sought.
The deep neural network architecture, we describe below, is trained on a large corpus of normal speech to permit the modeling of the phonetic units - 31 French phones plus silence. It is then tested on disordered speech corpus, including patients and control speakers, in order to examine its behaviour in terms of classification performance (capacity of generalization on unseen data) but also in terms of correlation with different kinds of perceptual measures, and notably measures of speech intelligibility, given our long-term research objective.

### 3.1. CNN based modeling

In light of its competitive accuracy for the phone classification task in the literature, we choose the Convolutional Neural Network (CNN) architecture in our work. The input of our CNN is a context window of 11 acoustic frames, each having 40 log Mel-filter bank energy features along with their first and second derivatives. These features are computed on a 20ms window with an overlap of 10ms between two adjacent frames, and thus stable acoustic features for classes such as phones serving as an input to two pairs of convolution and pooling layers.
Following a recipe similar to [12], the convolution layers apply a set of 3x5 filters to extract the local characteristics concentrated along the frequency axis, then producing respectively 32 and 64 activation maps. The max-pooling layers apply respectively 1x3 and 1x2 filters providing a lower frequency resolution features that contain more useful information to be processed by higher layers of the neural network. The task of classification is then performed by three fully connected layers of 1024 neurons. We apply a ReLU activation function followed by a dropout of 0.4 to the output of each of the fully connected layers. Our main goal is to minimise the categorical crossentropy loss function using the stochastic gradient descent algorithm. Finally, an output softmax layer corresponds to the posterior probability of each class associated with the 31 French phones and silence.

### 3.2. CNN training and validation

To train our proposed model, we use the BREF corpus described in section 2. After extracting the features as reported above, we conduct an input data normalization by subtracting the mean and by dividing by the standard deviation, both statistics are computed at the speaker utterance level.
Since the phone distribution within the BREF corpus is highly imbalanced, we adopt a random undersampling technique to handle disproportional distribution of classes and thus prevent the classifier from being biased towards the majority class. Then, we partition our data into 3M samples for the training set and 300K for the validation set, an almost 90%-10% data partitioning. For the CNN training, an initial learning rate of 0.001 following an exponential decay schedule and an early stopping strategies is utilized. For a first evaluation of our model, a BREF test set is prepared aside containing a total of 1M samples, almost 2 hours and 45 minutes of speech spread over 1489 utterances. We fix such a complete set for further analysis. In a later phase, the evaluation is made on C2SI-LEC dataset (about 350K samples), described in section 2, seeking the goal of this work.

## 4. Results

Our main purpose from training a CNN for a task of phone classification is to evaluate its phonetic feature encoding capability in order to prepare the ground for the long-term objective : the extraction of relevant linguistic units related to speech intelligibility variation. To be able to explore this CNN capability, its performance is measured, at the frame level, using a balanced accuracy metric. In order to deal with the classification task involving phone imbalanced datasets, this metric consists in averaging the correct classification rate computed for each concerned phone. Therefore, we will be referring to this balanced accuracy whenever talking about CNN performance metric in the rest of the paper.

### 4.1. Confusion Matrix analysis

In this section we evaluate our classification model performance on two corpora, BREF and C2SI-LEC (control speakers only) test sets on the basis of balanced accuracy, and through an analysis of phone confusion matrices given in figures 2 and 3 respectively. For readability purposes and to highlight the most relevant confusions, we split each of the confusion matrices into two parts: the first one regroups oral and nasal vowels as well as nasal consonants while the second regroups the voiced and voiceless plosive and fricative consonants.

The model performance is impressive when evaluated on BREF test data reaching a balanced accuracy of 82% with coherent phone confusions. When evaluated on the C2SI-LEC dataset, our model gives 74% of balanced accuracy on healthy control utterances, which is still a significant accuracy. This can clearly confirm that our CNN is able to generalize to new data, considering the fact that BREF and C2SI-LEC do not share neither the same conditions (recording equipment and location), nor the same speakers on which the model was trained.

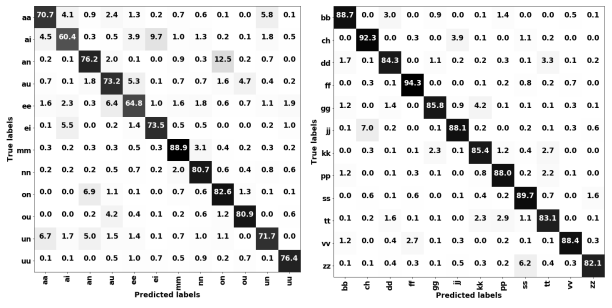



Figure 2: *Confusion matrices (CM) on BREF test - (left) Sub-CM grouping oral/nasal vowels, and nasal consonants - (right) Sub-CM grouping voiced and voiceless consonants*

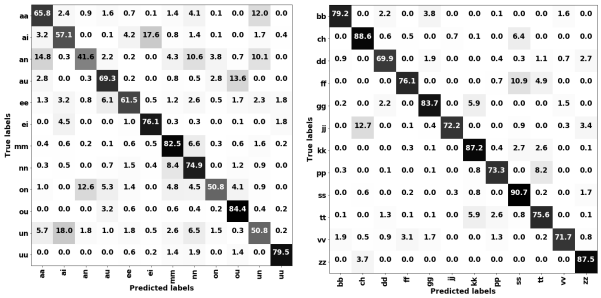
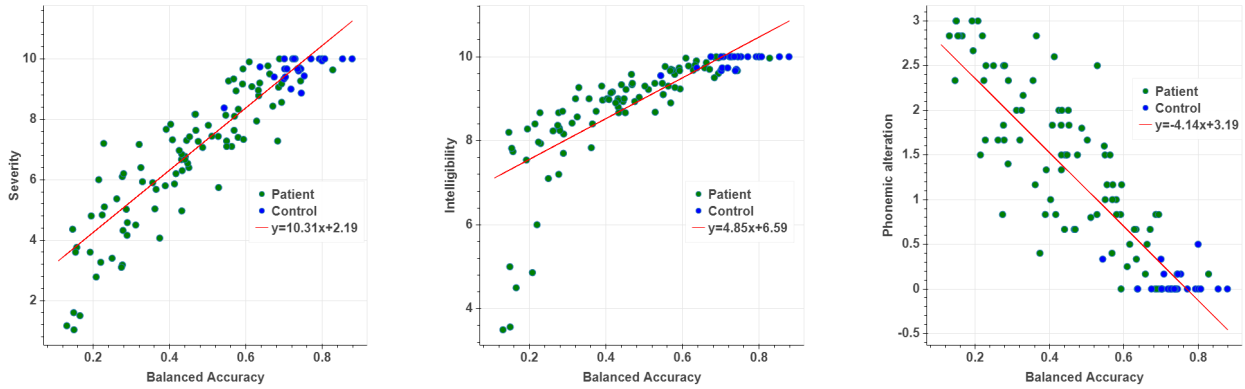



Figure 3: *Confusion matrices (CM) on C2SI-LEC control - (left) Sub-CM grouping oral/nasal vowels and nasal consonants - (right) Sub-CM grouping voiced and voiceless consonants*

While analyzing individually the confusion matrices for both BREF test and C2SI-LEC corpora, we can clearly observe that classification errors are somehow logical and have sense since confusions are generally made between phones sharing most of their phonemic features. Indeed, regarding fricative consonants - voiced (/v/, /z/, /ʒ/, noted as "vv", "zz", "jj" in CM) and voiceless (/f/, /s/, /ʃ/, noted as "ff", "ss", "ch" in CM) - or plosive consonants - voiced (/b/, /d/, /g/) and voiceless (/p/, /t/, /k/), illustrated in the right sub-CMs in figures 2 and 3, confusions are observed inside a phonetic class (for instance 3% confusion between /b/ and /d/ on BREF corpus) or, on the voicing distinctive feature inside a phonetic class (4.2% confusion between /g/ and /k/ belonging to the voiced and voiceless plosive class respectively on BREF corpus). By comparing both left sub-CMs, two major differences can be observed. Regarding the left sub-CM issued from the C2SI-LEC corpus in figure 3, the first one concerns the nasal vowels, which are subject to strong confusions with both oral vowels and nasal consonants (/n/ and /m/). This difference can be explained by the recruitment region of speakers for both corpora, exhibiting a major Parisian accent for the BREF corpus (the closest to standardized French), and a major southwestern accent for the C2SI-LEC corpus. Indeed, it is reported in the literature that nasal vowels can be produced with a less complete nasalization in speech exhibiting a southwestern accent, no more dealing with nasal vowels but rather with a combination of an oral vowel followed by a nasal consonant, typically the case of the LEC-C2SI corpus and the observed nasal vowel confusions. The second difference reflects the fact that speakers from the southwest region can have a more reduced phonological system of vowels compared to the Parisian speakers, since the mid vowels are not in a distinctive opposition (e.g. "épée" vs "épais"). This can explain the strong confusions observed notably between /ɛ/ and /e/ vowels, respectively noted as "ai" and "ei" in the CM.
One way to overcome this drop in CNN performance, in terms of accuracy, is to fine-tune our model using the C2SI-LEC control data. This alternative goes beyond all consideration because we deal with a very limited amount of data that we require for further analysis purposes. We can thus far justify the accuracy degradation observed on the C2SI-LEC corpus and consider that we achieve a low generalization error reflecting the classification performance and eliminating the possibility that the model is subject to an overfitting.




(a) *Scatter plot for severity rating (LEC-Sev) vs balanced accuracy:* $r$=0.91

(b) *Scatter plot for intelligibility rating (LEC-Intel) vs balanced accuracy:* $r$=0.78

(c) *Scatter plot for phonemic alteration rating (DES-Phon) vs balanced accuracy:* $r$=−0.88

Figure 4: *Scatter plots of different perceptual measures versus model balanced accuracy on the LEC-C2SI control and patient speakers*

### 4.2. Correlation between perceptual measures and model performance

Now that we have a relatively accurate model against dataset variation, we can assume that any significant degradation in the model performance, while typically testing on patient's utterances issued from C2SI-LEC corpus, is consistent with the degree of speech quality degradation and thus with the perceptual ratings of that patient. To highlight this idea, we evaluate our model per speaker (control speakers and patients), and we calculate the correlation between the CNN balanced accuracy and their corresponding perceptual measures notably intelligibility, severity and phonemic ratings. Figure 4 plots respectively the LEC-Sev, LEC-Intel, and DES-Phon perceptual measures against the balanced accuracy calculated on each speaker record. The corresponding Pearson correlation coefficients, noted $r$ and calculated between the balanced accuracy scores and perceptual measures for the overall set of speakers are also provided. These figures, whatever the perceptual measures observed, show a coherent behaviour by comparing the control speakers and patients, represented by blue and green dots respectively, in terms of balanced accuracy but also of perceptual ratings. Indeed, the blue dots are concentrated on the upper right (resp. down right for the phonemic alteration) where we have the highest severity and intelligibility scores (resp. the lowest phonemic alteration scores) as well as the highest balanced accuracies, reflecting a high quality of speech. Moreover, we can clearly see that the severity rating has the strongest correlation, with a 0.91 *r-value*, with our model accuracy. While we expect a value roughly the same as the LEC-Sev *r-value*, the LEC-Intel *r-value* deteriorates to 0.78. The first interpretation coming to mind is that our CNN model is not as efficient in encoding speech intelligibility characteristics as for severity. However, with regard to the remark differentiating intelligibility and comprehensibility underlined in section 2, this *r-value* decrease as well as the difference with the severity rating could be easily explained, the severity focusing more on speech sounds - which is closer to our CNN model objective - rather than on the spoken message as with intelligibility. Finally, correlation between the model accuracy and the phonemic alteration rating (DES-Phon), shown in figure 4c, is slightly lower than for severity score but still very strong. Although this perceptual rating was not assessed on the reading task, but on the image description, a very high *r-value* up to −0.88 is reached. Regarding the initial goal of modeling the characteristics of phonemic units, this *r-value* still confirms the phonetic modeling capabilities of the CNN-based architecture chosen.

## 5. Conclusions and perspectives

This paper investigates the encoding capability of a CNN-based deep learning architecture to finely model distinctive phonetic characteristics. This first study is part of a long-term research project dedicated to the characterization of speech intelligibility in disordered speech. Involved in a basic task of phone classification, even the most relevant for the final objective, the CNN-based model is trained on a large corpus of normal speech. Confronted to disordered speech, the encoding capability of this CNN-based model for the targeted task is demonstrated through a very high correlation between its phone classification rates and different perceptual measures available for both patients and control speakers present in the disordered speech corpus. Indeed, correlation coefficients of 0.91, −0.88, 0.78 with speech severity, degree of phonemic alteration and speech intelligibility ratings respectively are reached. The high performance of the CNN-based model observed for the task of phone classification, in terms of global accuracy rates, as well as these *r-values* make us confident in its involvement in the second step of the long-term project dedicated to the prediction of the speech intelligibility, still based on phonemic unit characteristic modeling.

In these future works, particular attention will be paid to the perceptual measurement of speech intelligibility available within the corpus of speech disorders involved, which is nevertheless necessary (for prediction purposes), but which we have pointed out in because of its moderate reliability (patients' overestimation of speech intelligibility). Indeed, a more reliable measure of speech intelligibility specifically developed for this purpose [15] will be investigated in this context.

## 6. Acknowledgements

This work has been carried out thanks to the French National Research Agency in 2018 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] R. D. Kent, *Intelligibility in speech disorders: theory, measurement and management*. John Benjamins Publishing, 1992, vol. 1.

[2] P. Enderby, "Frenchay dysarthric assessment," *Pro-Ed, Texas*, 1983.

[3] A. Lowit and R. D. Kent, *Assessment of motor speech disorders*. Plural publishing, 2010, vol. 1.

[4] S. Fex, "Perceptual evaluation," *Journal of voice*, vol. 6, no. 2, pp. 155–158, 1992.

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29(6), pp. 82–97, 2012.

[6] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Hindwai, Computational Intelligence and Neuroscience*, 2018.

[7] T. B. Ijitona, J. J. Soraghan, A. Lowit, G. Di-Caterina, and H. Yue, "Automatic detection of speech disorder in dysarthria using extended speech feature extraction and neural networks classification," in *IET 3rd International Conference on Intelligent Signal Processing (ISP 2017)*, 2017, pp. 1–6.

[8] B. Vachhani, C. Bhat, B. Das, and S. K. Kopparapu, "Deep autoencoder based speech features for improved dysarthric speech recognition," in *Proc. Interspeech 2017*, 2017, pp. 1854–1858. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1318

[9] L. Bin, M. C. Kelley, D. Aalto, and B. V. Tucker, "Automatic speech intelligibility scoring of head and neck cancer patients with deep neural networks," in *International Congress of Phonetic Sciences (ICPHs'19)*, Melbourne, Australia, 2019.

[10] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of parkinson's disease: A deep learning approach," *IEEE J. Biomed. Health Informatics*, vol. 23, no. 4, pp. 1618–1630, 2019. [Online]. Available: https://doi.org/10.1109/JBHI.2018.2866873

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

[12] T. Pellegrini and S. Mouysset, "Inferring phonemic classes from cnn activation maps using clustering techniques," in *Proceedings of Interspeech'16*, San Francisco, US, 2016.

[13] 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.

[14] C. Astesano, M. Balaguer, J. Farinas, C. Fredouille, P. Gaillard, A. Ghio, L. Giusti, I. Laaridh, M. Lalain, B. Lepage, J. Mauclair, O. Nocaudie, J. Pinquier, O. Pont, G. Pouchoulin, P. Michele, D. Robert, E. Sicard, and V. 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: European Language Resources Association (ELRA), may 2018.

[15] M. Lalain, A. Ghio, L. Giusti, D. Robert, C. Fredouille, and V. Woisard, "Design and development of a speech intelligibility test based on pseudo-words in French: why and how?" *Journal of Speech, Language, and Hearing Research*, p. in press, 2020.