Interspeech 2024
1-5 September 2024, Kos, Greece

Using wav2vec 2.0 for phonetic classification tasks: methodological aspects

Lila Kim[1], *Cédric Gendrot*[1]

[1]Laboratoire de Phonétique et Phonologie (CNRS & U. Sorbonne Nouvelle), France
lila.kim@sorbonne-nouvelle.fr, cedric.gendrot@sorbonne-nouvelle.fr

Abstract

Self-supervised learning, particularly in the context of speech, has been shown to be effective in a variety of tasks such as speaker recognition and speaker verification. Our research question focuses on the effectiveness of vector representations extracted from shorter versus longer phoneme sequences in detecting nasality. Two distinct approaches were studied: extracting vectors over the duration of the phoneme and taking longer sequences with one second added on each side of the phoneme, then recovering the central part a posteriori. The results show that the models react differently depending on the phone and the speaker, with variability observed at both levels. The long sequence model outperformed the short sequence model by correlating more robustly with nasal airflow.
Index Terms: speech, wav2vec 2.0, nasality, physiology

1. Introduction

Since the recurrent use of self-supervised learning in automatic speech recognition tasks, a number of studies have applied these Transformer models to domains such as speaker verification, code-switching or emotion detection, and so on [1, 2, 3, 4]. Furthermore, [5] have shown in particular that information differs in vector representations according to Transformer layers [6, 5]. It should be noted that a Transformer model such as wav2vec 2.0 takes into account contextual information within a sequence, and generally works on sequences lasting several seconds [7]. This sets the stage for our research question: Do vector representations extracted from single phones deliver better nasality detection performance than those extracted from longer sequences?

Our work therefore consists in exploring the length of the sequence for vector extraction: the first consists in taking vectors over the duration of the phoneme, while the second consists in taking a longer sequence by adding one second on both sides of the phoneme and recovering the central part a posteriori.

Firstly, we will describe the resources used in training and testing, and the extraction methods used with the wav2vec 2.0 self-supervised model. Secondly, we will look at two different approaches to vector extraction and nasality detection using logistic regression. We will then evaluate the trained model on acoustic data, comparing the results with physiological data obtained simultaneously with acoustics, thus serving as a reference.

1.1. Acoustic modelling

Speech is a complex phenomenon influenced by various elements such as articulation, the speaker's geographical or social origin, emotional state and pragmatic aspects such as the listener. Approaches to automatic speech transcription and acoustic signal modeling have evolved from expert systems to neural approaches. The initial approaches focused on the linguistic aspect, while probabilistic methods, notably hidden Markov models, began to dominate from the 1990s onwards. [8, 9]. With the advent of connectionist models, phonetization approaches have evolved towards fully neural methods, combining spectral representations such as MFCCs with multilayer perceptrons. To overcome Deep Learning challenges such as the need for large amounts of data and the lack of manual annotations, lightly supervised or self-supervised learning approaches have been undertaken. These models are previously trained on large numbers of hours of unannotated audio, then tuned on smaller annotated datasets for specific tasks [7]. Using the probing method, [5] analyzes the information contained in the different layers of these models, seeking to better understand the nature of the data at these levels [5, 6].

1.2. Nasality

Nasality is a ubiquitous feature of the world's languages, occurring when the soft palate is lowered and creating distinct acoustic effects on nasal sounds [10]. It is essential in speech production to distinguish nasal sounds phonologically from oral sounds, whether in the case of vowels (such as /a/ and /ɑ̃/, for example) or consonants (such as /b/ and /m/, for example). The nasality of a sound can be propagated to its oral neighbor by articulatory realizations, such as premature lowering or late raising of the velum [11, 12]. This nasal coarticulation, influenced by phonemic context, can occur in languages where nasality is a distinctive phonological feature (such as French, where /a/ in "maman" is nasalized), but also in language systems where this distinction is not present (as in English, for example in "can't").

Voice quality has major implications for speaker characterization. It can be a permanent feature of a speaker's voice due to physiological factors, but it is also subject to intra-speaker variability, notably in speech style or emotion [13]. Nasal sounds provide a reliable indicator for speaker recognition [14] because although the morphology of the nasal cavity differs between individuals, it remains consistent and not very malleable for each person during speech production [15, 16]. However, the acoustic analysis of nasality is complex, as the coupling of two cavities causes acoustic modifications by generating poles and zeros on the acoustic spectrum. Although analysis methods have been undertaken for nasality [17, 18], they are heavily influenced by the articulatory characteristics specific to each sound, and to each speaker.

10.21437/Interspeech.2024-1155

2. Experimental protocol

2.1. Methodology

2.1.1. wav2vec 2.0

Our research is part of an exploration of how the wav2vec 2.0 model encodes nasality information in its vector representations. We focus in particular on the "wav2vec 2.0-FR-3K-large-LeBenchmark" model, pre-trained on 2,900 hours of various types of French speech (spontaneous, read and broadcast). This model has been specifically designed to optimize its performance in French-related tasks [19].

The wav2vec 2.0 model works by taking the raw signal as input data, processed by the convolutional encoder. Every 25 milliseconds of audio is converted into a sequence of vectors with 20ms of overlap between them. These sequences undergo normalization and a GELU activation function before being routed to the transformer. During the pre-training phase, the quantization module is used to discretize the encoder output values. The latent representations obtained from the encoder then undergo analysis and contextualization by the Transformer layers, which capture information about the entire sequence. The wide model, in particular, comprises 24 Transformer layers, each producing a vector of 1,024 dimensions in latent representations. The feed-forward dimensions are 4,096, with 16 attention mechanisms [7].

2.1.2. Vector representation generation

Our approach is based on two methods for obtaining vector representations of our data. The first, inspired by the phonetic approach, involves extracting vector representations directly from phonemes cut at their boundaries. We used the max pooling strategy to aggregate the different latent representations of a recording into a single vector, which represents the entire signal. The second approach is to use longer sequences, adding one second on the beginning and end of the phoneme. Once wav2vec 2.0 takes contextual features from the whole sequence, we recover the middle vector a posteriori. For example, if the vowel lasts 200 ms, we extracted a sequence of 2.2 seconds, then performed max pooling on the middle 200 ms when retrieving vector features. In this way, contextual information on the entire 2.2-second sequence can be captured by the transformation blocks and be present in the medium vector that would represent the phoneme in question. Middle recovery was performed by removing the added seconds, representing 50 start and end vectors, as the wav2vec 2.0 model returns a vector every 20 ms. The resulting vector representations were labeled using the phonological labels of the sounds [+ nasal] and [- nasal].

2.1.3. Feature probing

A logistic regression model was implemented to determine whether the pronounced phone is realized with nasality (=1) or without nasality (=0). For this purpose, the "scikit-learn" python machine learning library was used, with hyperparameters set by default. The procedure of our methodology is described in figure 1. The embeddings for all datasets were obtained during the feature extraction phase. In the feature probing phase, a logistic regression model was trained on the training and validation datasets and then applied to the test data to obtain probability values.

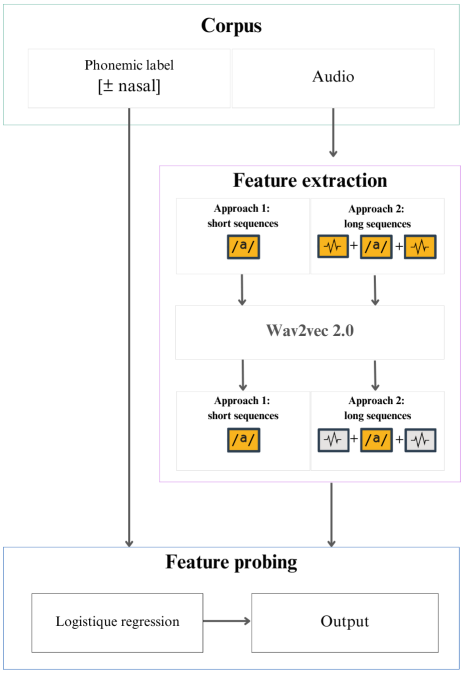

Figure 1: *Overview of experimental methodology including logistic regression model architecture*

2.2. Data for training

For training and validation, we extracted the different phone types from four separate corpora, each representing a specific speech type. The data corpora used in this study include:

1. NCCFr (The Nijmegen Corpus of Casual French): Comprising a total of 36 hours of friendly conversations, involving 46 French speakers. [20] ;
2. ESTER (Evaluation de Systèmes de Transcription enrichie d'Emissions Radiophoniques): Radio conversations in French, totaling 100 hours of prepared and read speech [21, 22]. Only a portion of the 30 hours was retained for this training.
3. PTSVOX: Corpus, created to evaluate intra- and inter-speaker variations, comprises speech recordings of around 90 hours, involving 369 French speakers. [23]. We have retained only a small part of this corpus with verified alignments, for the productions of only 24 speakers;
4. BREF: Developed for the development and evaluation of speech recognition systems, this project includes recordings of 120 French speakers in Paris reading articles from the newspaper "Le Monde", for a total of 100 hours of continuous speech. [24].Here again, as we were not provided with all the phoneme alignments, only half of the BREF corpus was used for training.

In this work, we decided to extract 8 vowels and 6 nasal and oral consonants combined. The vowels to be extracted are 3 pairs of vowels /a,ɛ,o,ɑ̃,ɛ̃,ɔ̃/ which can be distinguished by the nasal feature [± nasal]. We are aware of the articulatory distinction between an oral vowel and its nasal counterpart, however, in the context of this study, we have chosen to focus our attention on nasality in particular. Two vowels /e,ɔ/ have

been added, as the phonetization of mid vowels in French is not always consistent. As far as consonants are concerned, we have selected four oral consonants and three nasal consonants: /b,d,v,l,m,n,ɲ/. They are articulated in different ways and places (bilabial, labio-dental, dental, alveolar, and stop or fricative).

2.3. Acoustic and physiological data for test

The test data consisted of two parts: acoustic and physiological. They were collected simultaneously using an "Aeromask" mask developed at the Phonetics and Phonology Laboratory [25]. This mask records the voice as well as nasal and oral airflow without disturbing sound propagation, enabling acoustics to be used to evaluate the neural network and airflow to check for the presence of nasality in the phones being evaluated. Carrier sentence recordings were made with six male speakers, all native French speakers. Stimuli were inserted into words without literal meaning (i.e., logatomes) in the form of VCV or VNV, where C represents [p,b,t,d,v,s,z], N represents [m,n], and V represents [i,a,y,u,o,e,ɑ̃,ɛ̃,ɔ̃]. [25]. From these lists of phones, we selected the same phones as for training, with the exception of /l/ which is not present in the list, giving /a,e,ɛ,o,ɑ̃,ɛ̃,ɔ̃,b,d,v,m,n/. In summary, a total of 269 sounds from each class were extracted. The data used for training, validation and testing are summarized in table 1.

	Phone [+ nasal]	**Phone [- nasal]**
Training	60 000	60 000
Validation	15 000	15 000
Test	269	269

Table 1: *data used for training, validation and testing*

The aerodynamic data we will be using as a reference consists of three values: nasal airflow (NAF), buccal airflow (OAF) and proportional nasal airflow. The calculation of these values is explained in [26].

3. Results

Analyzing the results obtained with deep neural networks by physiological measurement helps to verify the nasality index in phone realization. In section 3.1, our aim is to establish whether nasality is detectable when a classifier is based on features extracted by the wav2vec 2.0 self-supervised model, and whether the errors produced by the networks can be explained by nasal and oral airflow rates. Finally, in section 3.2, We investigate whether the classifiers have learned phoneme identities rather than detecting nasality, using the same vector representations.

3.1. Nasality detection : system performance by extraction approach

The overall accuracy rates for the nasality feature [± nasal] across the different Transformer layers are shown in figure 2. In order to determine the optimal layer to exploit, we have examined the performance evolution of the different layers with regard to nasality, from the CNN encoder to the last Transformer layer. The figure shows the presence of nasal information in practically all layers when the audio extract is long. In short sequences, nasality is particularly marked in the output of the CNN encoder and in the first Transformer layers.

According to [5], the first layers of the wav2vec 2.0 model, including the CNN encoder, are associated with acoustic identity and spectrogram features [6, 5]. In the light of these observations, we decided to focus on the first Transformer layer in order to improve nasality identification using acoustic rather than phonemic features. Thus, when classifying nasality with features extracted from the first layer, performance was better for long sequences, with an overall accuracy of 94.05%, compared with 81.04% for short sequences.

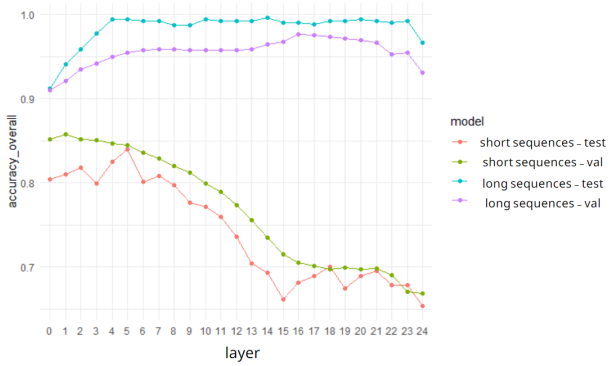

Figure 2: *Distribution of overall accuracy according to wav2vec 2.0 layers by audio extract length*

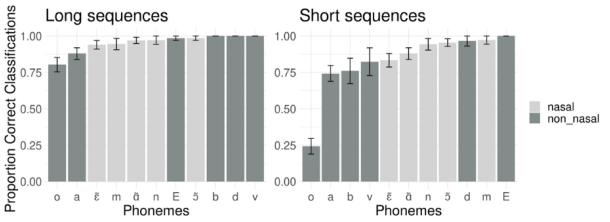

Figure 3: *Correct classification rate for each phoneme (long sequences on the left and short sequences on the right)*

In figure 3, the proportion of correct class assignments for each phoneme is shown. Note that in this visualization, [E] represents [e,ɛ]. Performance varies across phonemes and models. For instance, the model demonstrates strong classification accuracy for certain phonemes such as [ɔ̃,E,m,n,d], while struggling with the classification of oral vowels [o,a], yielding lower accuracy rates for both models. Moreover, nasal vowels show similar levels of difficulty: [ɛ̃] is considered the most difficult to detect in terms of nasality, while [ɔ̃] is identified as the easiest. For the long-sequence model, the most difficult consonant to predict is the nasal [n,m], while for the short-sequence model, it is the oral consonant [b,v].

3.2. Comparison of classifier results with physiological data

In this study, Pearson's correlation coefficient is used to examine the linear relationship between nasality probability and nasal airflow. Correlations were measured in three distinct ways: (i) we used nasal airflow as obtained from the aeromask (ii) normalization of the nasal airflow with respect to its oral counterpart, e.g. /a/-/ɑ̃/. (iii) normalization was performed for each nasal-oral pair and for each speaker. For the consonant /v/ without a nasal correspondent, the value was normalized to its whole.

Whether using raw or normalized airflow, the long-sequence model shows a stronger correlation than the short-

Nasal airflow	Speaker	MT03	MT04	MT05	MT06	MT07	MT08	ALL
mean	Long sequences	0,75	0,73	0,68	0,70	0,76	0,77	0,70
phonemes		0,66	0,76	0,68	0,68	0,72	0,72	0,68
phonemes+speakers		0,70	0,79	0,69	0,68	0,73	0,68	0,71
mean	Short sequences	0,61	0,65	0,59	0,59	0,46	0,69	0,55
phonemes		0,52	0,69	0,58	0,51	0,45	0,64	0,53
phonemes+speakers		0,55	0,70	0,61	0,52	0,48	0,60	0,57

Table 2: *Comparison of classification model results with nasal airflow (NAF) using Pearson's correlation coefficient*

sequence model. We note two observations in common for both models. Overall, nasality probabilities are most strongly correlated with normalized values per phoneme and per speaker. This shows that nasal airflow is phoneme- and speaker-specific. Secondly, the correlation is strongest for speaker MT04, and this observation is common to both models. However, the speaker with the lowest correlation differs according to audio extract length and nasal airflow measurements.

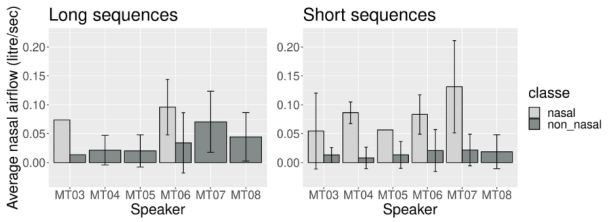

Figure 4: *Distribution of nasal airflow values for incorrectly identified phonemes*

Figure 4 shows the distribution of nasal airflow values for phonemes incorrectly identified by models trained with the first Transformer layer. It has been observed that misattributions behave differently depending on the length of the audio extract. With longer sequences, errors in oral realizations are more likely to be explained by nasal airflow, while with shorter sequences, misclassified nasal productions are more likely to be justified by nasal airflow. For example, in long sequences, incorrectly identified [+ nasal] phonemes display typical nasal airflow, which is included in the range of means with standard deviation. However, the oral productions of speakers MT07 and MT08 sometimes show a high nasal airflow, sometimes above the mean. For short sequences, the incorrectly classified productions of speakers MT03 and MT07 display nasal airflow outside the range of the mean including standard deviation. The nasal phoneme for which the short sequence model has assigned an oral label has the lowest nasal airflow. However, some incorrectly classified [o] vowels show nasal airflow above its maximum in the NAF for the [+ nasal] class.

4. Discussion and conclusion

Our aim was to investigate sequence length for vector extraction to facilitate a phonetic classification task, in particular that of nasality. Two lengths were examined: a sequence of one phoneme and a more extended sequence with 1 second added on each side of the phoneme. Both approaches performed satisfactorily in the proposed task. The longer sequences achieved an overall accuracy of 94.05%, while the shorter sequences scored 81.04%.

Both our models succeeded in specializing to nasality in speech, but with behavior varying according to phoneme and speaker. In the section 3.1, it was shown that the behavior of the models differed according to phoneme, a phenomenon that can be explained by the variation in articulator positions during phoneme realization and by the fact that the soft palate position itself differs according to vowel [27, 11]. For example, the mouth opening is minimal for the vowel /ɔ̃/, causing the velum to rest on the tongue, which prevents the soft palate from descending [27]. This minimal opening allows pressure to build in the oral canal, enabling air to pass through the velopharyngeal port into the nasal cavity [11].

Comparison of nasality probabilities with physiological data reveals a correlation between nasal airflow and the probabilities obtained with our models. This correlation varies according to phoneme and speaker. For example, the correlation is stronger when nasal airflow is normalized by phoneme and speaker than for raw NAF. Speaker MT04 shows a particularly strong correlation. This speaker can be considered to have a good distinction between oral and nasal voice production. In addition, we investigated whether nasal airflow could explain misattributions of labels, and observed two trends. Firstly, nasal airflow can be used to explain incorrectly identified [- nasal] phone realizations for long sequences, whereas it is used to justify misattributions of [+ nasal] phones for short sequences. Secondly, when the nasal airflow of a phoneme does not fall within the range of means (with standard deviation), both classifiers recognize it as not expected for the class in question. This pattern is associated with atypical values.

In conclusion, our study demonstrated the use of two sequence lengths to extract vector information for a specific nasality identification task. By comparing our classifiers with aerodynamic measurements, a significant correlation was observed between nasal airflow rates and nasality probabilities. The results reveal the differentiated behavior of the models across phonemes and speakers, with interlocutor variability observed. Performance remains constant for speakers who clearly distinguish between oral and nasal production, compared to those with atypical nasal airflow patterns in oral sounds or atypical oral airflow patterns in nasal sounds. Embedding analysis shows that our models detect nasality by acquiring phoneme-related knowledge. However, the short sequences enabled to capture phonetic or acoustic nasality globally, which partly explains the classification errors. Indeed, it is not uncommon for oral class phonemes to be nasalized by context (and vice versa), and these phenomena can be confirmed by aerodynamic measurements. On the other hand, long sequences better captured phonemic nasality and phonological contrasts between phonemes, with performance close to 100%.

5. Limits and future studies

In this study, we carried out a comparison between the probabilities assigned by the classifier and an aerodynamic measurement to confirm that our model is capable of detecting phonetic as well as phoneme nasality. However, in certain situations, nasality remained perceptible in the segment, even when the nasal airflow was as reduced as that of an oral phone. A typical example is seen with nasal second vowels in a logatome, such as [ɔ̃tɔ̃]. In this case, the second nasal vowel consistently had a lower nasal airflow than the first nasal vowel in the logatome. Yet, when listening to the second vowel excerpt, the nasality remained perceptible. While these specific variations can be articulatorily explained [28] without calling into question the results of our models, a perceptual study to validate the nasality identified by naïve listeners will be relevant not only for a better interpretation of the results, but also for a better characterization of the speaker's voice.

6. References

[1] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," Jan. 2021, arXiv:2012.06185 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2012.06185

[2] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings," Apr. 2021, arXiv:2104.03502 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2104.03502

[3] L.-H. Tseng, Y.-K. Fu, H.-J. Chang, and H.-y. Lee, "Mandarin-english code-switching speech recognition with self-supervised speech representation models," *arXiv preprint arXiv:2110.03504*, 2021.

[4] P. Cormac English, J. D. Kelleher, and J. Carson-Berndsen, "Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features," in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Seattle, Washington: Association for Computational Linguistics, 2022, pp. 83–91. [Online]. Available: https://aclanthology.org/2022.sigmorphon-1.9

[5] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," Mar. 2023, arXiv:2211.03929 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2211.03929

[6] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise Analysis of a Self-supervised Speech Representation Model," Dec. 2022, arXiv:2107.04734 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2107.04734

[7] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Oct. 2020, arXiv:2006.11477 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2006.11477

[8] B. H. Juang and L. R. Rabiner, "Hidden Markov Models for Speech Recognition," vol. 33, no. 3, 1991.

[9] I. Patel and Y. Srinivas Rao, "Speech Recognition Using HMM with MFCC-An Analysis Using Frequency Specral Decomposion Technique," *Signal & Image Processing : An International Journal*, vol. 1, no. 2, pp. 101–110, Dec. 2010. [Online]. Available: http://www.aircconline.com/sipij/V1N2/1210sipij09.pdf

[10] S. Maeda, "Acoustic cues for vowel nasalization: A simulation study," *The Journal of the Acoustical Society of America*, vol. 72, no. S1, pp. S102–S102, Nov. 1982. [Online]. Available: https://pubs.aip.org/jasa/article/72/S1/S102/733010/Acoustic-cues-for-vowel-nasalization-A-simulation

[11] A. Amelot, P. Basset, S. Maeda, K. Honda, and L. Crevier-Buchman, "Etude simultanée des mouvements du voile du palais et de l'ouverture du port vélopharyngé," *XXVIIè JEP*, pp. 65–68, 2008.

[12] A. Brkan, "Etude comparative des phénomènes de coarticulation nasale en anglais américain, bosnien, français, norvégien et ourdou," Ph.D. dissertation, Université Sorbonne Paris Cité, 2018.

[13] F. Nolan, "Forensic Speaker Identification and the Phonetic," *A Figure of Speech: A Festschrift for John Laver*, p. 385, 2014, publisher: Routledge.

[14] J. Kahn, "Parole de locuteur: performance et confiance en identification biométrique vocale," Ph.D. dissertation, Université d'Avignon, 2011.

[15] J. Dang, K. Honda, and H. Suzuki, "Morphological and acoustical analysis of the nasal and the paranasal cavities," *The Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2088–2100, 1994, publisher: Acoustical Society of America.

[16] A. Serrurier, "Modélisation tridimensionnelle des organes de la parole à partir d'images irm pour la production de nasales-caractérisation articulatori-acoustique des mouvements du voile du palais." Ph.D. dissertation, Institut National Polytechnique de Grenoble-INPG, 2006.

[17] M. Y. Chen, "Acoustic correlates of English and French nasalized vowels," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2360–2370, Oct. 1997. [Online]. Available: https://pubs.aip.org/jasa/article/102/4/2360/562446/Acoustic-correlates-of-English-and-French

[18] W. Styler, "On the acoustical features of vowel nasality in English and French," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2469–2482, Oct. 2017. [Online]. Available: https://pubs.aip.org/jasa/article/142/4/2469/853233/On-the-acoustical-features-of-vowel-nasality-in

[19] T. Parcollet, H. Nguyen, S. Evain, M. Z. Boito, A. Pupier, S. Mdhaffar, H. Le, S. Alisamir, N. Tomashenko, M. Dinarelli, S. Zhang, A. Allauzen, M. Coavoux, Y. Esteve, M. Rouvier, J. Goulian, B. Lecouteux, F. Portet, S. Rossato, F. Ringeval, D. Schwab, and L. Besacier, "LeBenchmark 2.0: a Standardized, Replicable and Enhanced Framework for Self-supervised Representations of French Speech," Sep. 2023, arXiv:2309.05472 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2309.05472

[20] F. Torreira, M. Adda-Decker, and M. Ernestus, "The Nijmegen Corpus of Casual French," *Speech Communication*, vol. 52, no. 3, pp. 201–212, Mar. 2010. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167639309001629

[21] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. McTait, and K. Choukri, "The ester evaluation campaign for the rich transcription of french broadcast news." in *LREC*, 2004.

[22] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news." in *LREC*, 2006, pp. 139–142.

[23] A. Chanclu, L. Georgeton, C. Fredouille, and J.-F. Bonastre, "Ptsvox: une base de données pour la comparaison de voix dans le cadre judiciaire (ptsvox: a speech database for forensic voice comparison)," in *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1: Journées d'Études sur la Parole*, 2020, pp. 73–81.

[24] L. F. Lamel, J.-L. Gauvain, and M. Eskénazi, "Bref, a large vocabulary spoken corpus for french," *training*, vol. 22, no. 28, p. 5.

[25] A. Elmerich, J. Gao, A. Amelot, L. Crevier-Buchman, and S. Maeda, "Combining acoustic and aerodynamic data collection: A perceptual evaluation of acoustic distortions," in *INTERSPEECH 2023*. ISCA, Aug. 2023, pp. 3078–3082. [Online]. Available: https://www.isca-archive.org/interspeech_2023/elmerich23_interspeech.html

[26] A. Elmerich, L. Kim, C. Gendrot, A. Amelot, L. Crevier-Buchman, and S. Maeda, "Nasality detection from acoustic data with a convolutional neural network and comparison with aerodynamic data."

[27] V. Delvaux, T. Metens, and A. Soquet, "Propriétés acoustiques et articulatoires des voyelles nasales du français," *XXIVèmes Journées d'étude sur la parole, Nancy*, vol. 1, pp. 348–352, 2002.

[28] J. J. Ohala *et al.*, "Phonetic explanations for nasal sound patterns," in *Nasálfest: Papers from a symposium on nasals and nasalization*. Stanford University Language Universals Project Palo Alto, CA, 1975, pp. 289–316.