NASALITY DETECTION FROM ACOUSTIC DATA WITH A CONVOLUTIONAL NEURAL NETWORK AND COMPARISON WITH AERODYNAMIC DATA

Amélie Elmerich[1], Lila Kim[1], Cédric Gendrot[1], Angélique Amelot[1], Lise Crevier-Buchman[1,2], Shinji Maeda[1]

[1] Laboratoire de Phonétique et Phonologie (UMR7018, CNRS - Sorbonne Nouvelle)
[2] Hôpital Foch: Service de Laryngologie Phoniatrie, Université Paris Saclay, Suresnes, France
{amelie.elmerich, lila.kim, cedric.gendrot, angelique.amelot}@sorbonne-nouvelle.fr

ABSTRACT

From a new acoustically transparent pneumotachograph mask, we simultaneously recorded aerodynamic (oral and nasal airflow) and acoustic data for 6 French male speakers, involving 3 oral and 3 nasal vowels out of logatoms (*i.e.* non words). A Convolutional Neural Network (CNN) trained on other acoustic corpora in French was tested on the data collected from the mask for the nasal/oral vowel distinction, with a 88% correct classification on average. We compared these CNN results with the nasal airflow extracted from all vowels of the corpora. Aerodynamic results showed a higher quantity of nasal airflow for the nasal vowels. However, for some speakers, distinction between nasal and oral vowels in terms of nasal airflow was less prominent, especially for /a/ *vs* /ɑ̃/, the 2 vowels for which the CNN have the least correct identifications. Finally, we discuss the discrepancies observed between aerodynamic data and CNN probabilities, and inter-speaker variations that can be approached by the CNN.

Keywords: Convolutional Neural Network (CNN), aerodynamic, nasality, speaker.

1. INTRODUCTION

Nasality is a distinctive feature in approximately a third of the world's languages [1]. The basic knowledge implies that the soft palate must be sufficiently lowered for the velopharyngeal port to be open allowing air to pass through the nose. The lowering of the soft palate as well as the passage of air through the nose will have an implication in the speech spectrum with acoustic zeroes that usually hinder the acoustic analysis for phoneticians [2].

There are temporal and spatial variations in the realisation of the [nasal] feature. It varies depending on the speaker's gender and anatomy [3 ; 4], speaker strategy [5 ; 6 ; 7], language [8], speaking style [9], speaking rate [10], the speech sound type, the phonetic and prosodic context [11], *etc.*

More specifically, the opening of the velopharyngeal port differs from one speaker to another [3, 4] and the morphology of the nasal cavities is highly variable between individuals. Nasal vowels and consonants are relevant for speaker identification as they contain more acoustic information relative to the speakers than the other sounds [12, 13].

Deep neural networks have recently shown an important development in the field of speech. Studies have been conducted in the clinical field with artificial neural networks to diagnose language pathologies, including hyper- or hypo-nasalisation [14, 15, 16]. Indeed it has been shown that CNNs have the ability to specialise on phonetic features such as place of articulation or articulatory mode [16, 18, 19].

From a new pneumotachograph developed at the Phonetics and Phonology Laboratory, composed of a paper-fiber mask which provides no acoustic distortion, it is possible to analyse acoustic and aerodynamic data in the same recording [20], which is not possible with most aerodynamic measurement systems. With this device, it is therefore possible to account for the aerodynamics in parallel with the resulting acoustics.

The purpose of the present study is to assess the detection of nasality from acoustic data with a CNN by comparing it with aerodynamic data collected during the acoustic recording. We take the vowel phonemic label 'nasal' or 'oral' as the reference and check whether CNN classification is correct from the acoustics. In a second time the nasal airflow provided will validate the CNN classification or help us understand the misclassifications. We will finally investigate to what extent the level of nasality for each speaker can be estimated with CNN.

2. MATERIALS AND METHODS

2.1. Corpora and data acquisition

The test data for this study was taken from 6 French male native speakers (mean age: 36 years) recorded in a soundproof room. The speech samples consist of VCV sequences, where C=[p,b,t,d,v,s,z,m,n] and V=[i,a,y,u,o,e,ã,ɛ̃,ɔ̃]. The sequences were inserted in the frame sentence, for example: « Non, tu n'as pas dit apa quatre fois, tu as dit aba et ada quatre fois » ('No, you didn't say apa four times, you said aba et ada four times'). Finally, we have a total of 270 sequences with C= 270 et V= 540.

Aerodynamic and acoustic data were recorded simultaneously with a pneumotachograph mask. The advantages of this mask are that i) oral and nasal airflow can be recorded separately, ii) it is possible to adapt the size and position of the plate (Fig. 1a) to separate the nasal airflow (NAF) from the oral airflow (OAF) for each speaker iii) there is no acoustic distortion [20]. This mask is connected to 2 sensors and each one measures the differential pressure inside the mask with respect to the atmospheric pressure (see Fig. 1). There may be slight differences in flow measurements depending on the mask, the position of the sensor and the size and the position of the plate separating the nasal and oral airflow. Consequently, a calibration must be operated separately for each mask (individual mask for each speaker): one calibration for the oral compartment and another one for the nasal compartment. The calibration of the 2 pressure sensor modules allows to convert airflow values in the physical unit (liters/s, see Fig. 2). The mask provides a small resistance, as required to measure the airflow without affecting the sound propagation [21].

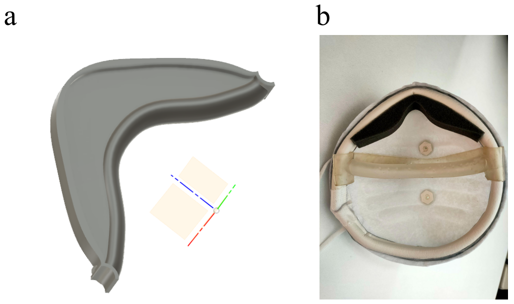

Figure 1: a. Flexible resin separation integrated into the mask to separate the nasal and oral airflow, b. Mask in fiber paper with plate and 2 adapters connected to the pressure sensors.

Acoustic data were captured with a microphone (AKG C520 L). All aerodynamic and acoustic sensors are linked to an acquisition card (DT9003). Audio and aerodynamic data were recorded at a sampling frequency of 20kHz.

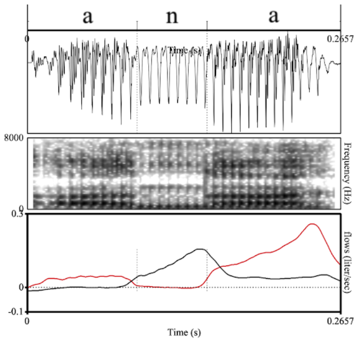

Figure 2: Example of acoustic and airflow recordings of [ana]. From top to bottom, (1) audio signal captured with a microphone, (2) spectrogram, (3) the nasal airflow (NAF in black) and oral airflow (OAF in red).

The data were manually segmented in Praat [22]. A Python script was used to automatically extract the mean of OAF and NAF for each vowel (l/sec). We used R [23] to perform graphics and statistical analysis.

2.2. Convolutional Neural Networks

For an acoustic nasal-non-nasal automatic classification task, we chose to work with Convolutional Neural Networks (CNN) and to focus only on 6 vowels /a,e,o,ã,ɛ̃,ɔ̃/, hence 3 vowel qualities /a,e,o/ and their nasal counterparts. We are aware that there is no exact articulatory match between our 3 oral and nasal vowels [24] and we decided to include all 6 vowels in the classification system (instead of comparing by pairs) so as to circumvent this asymmetry.

We opted for a CNN over other types of neural networks due to our intention of working with spectrograms of vowels, with the ultimate goal of applying a gradCam-like algorithm. This method is used to localise the specific parts of an image that have contributed to the final decision, in our case, the areas that contain information related to nasality. The training dataset is composed of the productions of these vowels extracted from 3 French corpora with various speech types: NCCFr [25], ESTER [26] and PTSVOX [27]. In all these corpora, automatic segmentations were provided at the phoneme level. Randomly chosen vowels were extracted at their boundaries (as determined by phonemic segmentation) in the form of a spectrogram, without any selection of the prosodic, lexical or phonemic context. For the first 2 corpora, the number of vowels of each type was checked. 10,887 productions of each type were taken from NCCFr and 9,186 from ESTER. For PTSVOX, we took all possible vowels respecting the natural frequency of phonemes as this brought better results.

The test dataset was derived from the acoustic data that was described in section 2.1. During this phase, we randomly selected vowel utterances and extracted their corresponding spectrograms. This set contains 66 productions of each vowel type (6 speakers * 11 occurrences), making 198 vowels for each category.

	Train & Val			Test
Source	NCCFR	ESTER	PTSVOX	Data recorded with the mask
nasal	32,661	27,558	65,669	198
non-nasal	32,661	27,558	135,119	198

Table 1: Number of vowels used for train and test sets according to the corpora.

All these spectrogram images with a frequency band from 0 to 8000 Hz were reduced in 48x48 pixel sizes and presented as input to our network. For the feature extraction part, the model is made of 2 pairs of convolution and pooling layers. The convolution layers were performed with a 5x5 kernel size and thus produced 32 and 64 filters respectively. After each convolution layer, a batch normalisation layer was inserted before applying an activation layer in order to allow the model to generalise [17] over different types of corpora and data. The max pooling layers were then used to reduce the size of the images with a 2x2 pool size. With the extracted features, 3 fully-connected layers performed the classification task with 1024 neurons. The ReLU activation function was applied after each batch normalisation layer and each fully-connected layer. Finally, a softmax activation function was used at the last fully-connected layer for nasal-oral classification. The labelling of "nasal" and "non-nasal" was based on the expected phonemic transcription in French. We performed a binary categorization to assign a probability to each class. The probability value for each class was also considered as a continuous variable for strength of nasality. During the model training, the model was improved by applying Adam as an optimisation technique and categorical cross entropy as a metric to measure model performance.

3. RESULTS

3.1 CNN results

For a given vowel, the classifier returns a value between 0 and 1, which we refer to as the probability of nasality obtained by the CNN model. When a vowel is identified as nasal by the classifier, the expected probability of nasality would be over 0.5 and conversely, a vowel classified as non-nasal would have a value close to 0 (or at least below 0.5).

Our image classifier could accurately identify 95% of non-nasal vowels and 82% of nasal vowels, reaching 88% overall accuracy and F1-score of 88% (k = 0.77) without fine-tuning between training and test.

The cross-speaker variability can be observed in Fig. 3. There are speakers for which the model makes many classification errors and others for which there are fewer errors from the model. For example, the model makes most misclassifications for speakers MT01 and MT04. Of the total misclassifications, 37% of the inaccurately classified vowels come from speaker MT04 (*i.e.* 17 out of 46 errors). Speaker MT01 gathers 13 incorrect occurrences (*i.e.* 28% of the total misclassifications) while speaker MT03 has only 1 error. In addition, errors on non-nasal vowels occur only for speakers MT01 and MT05. For the other speakers, the model accurately performed on non-nasal vowels, with errors occurring only for nasals.

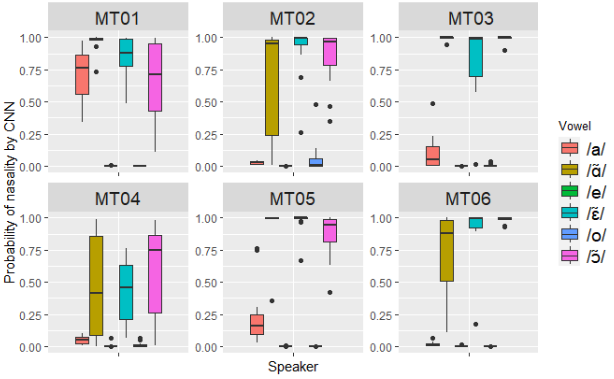

Figure 3: Nasality probabilities obtained by the CNN model for each speaker and each vowel type.

	Left	# Err	Right	# Err	Total
nasal	labial coronal	12 12	pause	9	36
non nasal	pause	5			10

Table 2: Contexts mostly found in misclassifications according to the nasal-oral categories with their number of occurrences in the errors.
('left' and 'right' for left and right context of vowels).

We observe that misclassifications mostly appear between /a/ and /ɑ̃/. Several phonetic contexts could be considered as factors in this confusion. On the one hand, as we can see in Table 2, when there is a pause in the left context, /a/ vowels tend to be classified as nasal by our model (5 out of 10 errors of /a/, i.e. 50%). On the other hand, the presence of a labial or coronal consonant before the vowels /ɑ̃/ can influence the decision of its class (respectively 5 and 6 out of 14 errors of /ɑ̃/). We notice the same influence when a pause is located after these vowels

(5 out of 14 misclassifications of /ɑ̃/). These 3 contexts for /ɑ̃/ vowels also appear in errors for other nasal vowels. Out of 36 incorrect nasal vowel classifications, 12 errors are caused by the labial and coronal consonant context preceding nasal vowels, and 9 errors occur with a pause in the right context.

3.2 Aerodynamic results

In Figure 4, we observe a higher quantity of nasal airflow for the 3 nasal vowels. Results of ANOVA were statistically significant with p<.001 for all speakers for the distinction of NAF between nasal and oral vowels.

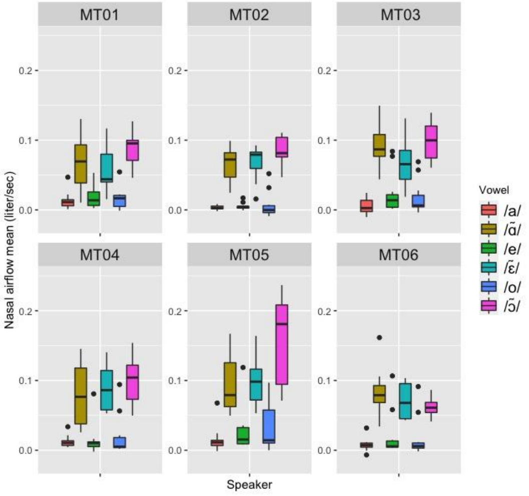

Figure 4: Mean of nasal airflow for each speaker and each vowel type.

	MT01	MT02	MT03	MT04	MT05	MT06	mean
/a/	0.010	0.003	0.005	0.012	0.015	0.008	0.010
s.d.	0,013	0.003	0.011	0.009	0.019	0.010	0.010
min	0.001	-0.002	-0.010	0.004	-0.002	-0.007	-0.002
max	0.047	0.008	0.025	0.033	0.068	0.032	0.035
/ɑ̃/	0.068	0.065	0.091	0.079	0.094	0.082	0.080
s.d.	0.039	0.026	0.030	0.044	0.041	0.036	0.040
min	0.010	0.025	0.044	0.026	0.050	0.034	0.031
max	0.130	0.01	0.149	0.145	0.167	0.161	0.142
/e/	0.020	0.005	0.023	0.015	0.027	0.019	0.020
s.d.	0.018	0.005	0.029	0.022	0.032	0.031	0.020
min	0.002	-0.0004	0.002	-0.002	0.008	0.003	0.002
max	0.053	0.017	0.084	0.081	0.119	0.106	0.076
/ɛ̃/	0.058	0.068	0.067	0.090	0.100	0.071	0.070
s.d.	0.031	0.025	0.033	0.031	0.036	0.029	0.030
min	0.015	0.016	0.019	0.053	0.053	0.043	0.033
max	0.117	0.092	0.131	0.140	0.164	0.103	0.124
/o/	0.016	0.007	0.018	0.020	0.034	0.017	0.020
s.d.	0.015	0.019	0.024	0.029	0.034	0.029	0.020
min	-0.001	-0.008	-0.003	0.00001	-0.002	0.054	-0.002
max	0.054	0.052	0.069	0.094	0.097	0.091	0.076
/ɔ̃/	0.087	0.086	0.099	0.100	0.154	0.061	0.100
s.d.	0.026	0.019	0.027	0.033	0.064	0.014	0.030
min	0.046	0.047	0.060	0.050	0.071	0.041	0.052
max	0.127	0.111	0.139	0.154	0.237	0.086	0.142

Table 3: Nasal airflow means (l/s) on vowels per speaker.

We can observe from Table 3 that each speaker has a minimum level of nasal airflow which differs between vowels. Also, for each pair of vowels (/a/ vs. /ɑ̃/, /e/ vs. /ɛ̃/, and /o/ vs. /ɔ̃/), the maximum nasal airflow for oral vowels can be larger than the minimum nasal airflow for nasal vowels. This obviously explains some of the misclassifications made by the CNN, in particular for speakers MT01, MT04 and MT05 who have a higher nasal airflow for oral vowels, or an overall lower nasal airflow for nasal vowels. For speaker MT02, the mean nasal airflow for /ɑ̃/ is 0.065 l/s, and all but 1 occurrence of /ɑ̃/ were misclassified below that threshold. As for speaker MT04, all but 2 occurrences of /ɛ̃/ were misclassified below a threshold of 0.059 l/s. Many examples of misclassifications were found according to these criteria. Overall, Pearson's correlation coefficient revealed that the prediction of nasality and non-nasality is correlated with the mean NAF measure (with r = 0.66).

4. DISCUSSION AND CONCLUSION

The main result of this work is the correct automatic classification of nasality for vowels up to 88% from a new acoustic corpus.

We showed that there is a significant connection between CNN probabilities and aerodynamic data. CNN misclassifications for speaker MT01 and MT04 or for the distinction between /a/ *vs.* /ɑ̃/ are found to be correlated with smaller differences in nasal airflow.

Our aim was also to correlate the CNN probabilities with the mean level of nasal airflow per speaker so as to evaluate the overall nasality per speaker and this point still has to be explored further. A first investigation indicated that the probability values given by the CNN were not related to the amount of nasal airflow, although misclassifications may give good hints for that matter.

It is unlikely that the values from the mask were erroneous due to the numerous calibration checks done during the recordings. However, we observed low values of nasal airflow that were unexpected (in particular preceding unvoiced obstruents) and the use of ratio [13] between nasal and oral airflow may provide useful answers.

In the near future, we will work on the activation functions so as to better relate the probability values with the level of nasal airflow. The analysis of spectral zones used by a CNN could be an important insight of our work since the modelling of the relationship between acoustics and articulation is still problematic for nasals. For example, it is known from articulatory studies that /a/ has a lowered velum compared to other vowels and this should have an impact on the rate of classifications [28]. Overall, the implications from these results should help phoneticians in their analysis of nasal vowels, and it is planned to share that model and aerodynamic mask with the community.

5. REFERENCES

[1] Maddieson, I. 1984. Patterns of sounds. *Phonology*, *2*(1), 343-353.

[2] Styler, W. 2017. "On the acoustical features of vowel nasality in English and French," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 2469–2482, doi: 10.1121/1.5008854.

[3] Clarke, W. M. 1975. "The measurements of the oral and nasal sound pressure levels of speech," *J. Phon.*, vol. 3, pp. 257–262.

[4] Amelot, A. 2004. Etude aérodynamique, fibroscopique, acoustique et perceptive des voyelles nasales du français. Université de la Sorbonne nouvelle - Paris III. Thèse de l'Université de la Sorbonne Nouvelle - Paris III.

[5] Croft, C. B., Shprintzen, R. J., & Rakoff, S. J. 1981. Patterns of velopharyngeal valving in normal and cleft palate subjects: A multiview videofluoroscopic and nasendoscopic study. *The Laryngoscope*, *91*(2), 265-271.

[6] Skolnick, M. L., McCall, G. N., & Barnes, M. 1973. The sphincteric mechanism of velopharyngeal closure. *The Cleft Palate Journal*, *10*(3), 286-305.

[7] Vaissière, J. 1988. "Prediction of velum movement from phonological specifications," *Phonetica*, vol. 54, pp. 122–139.

[8] Clumeck, H. 1976. "Patterns of soft palate movements in six languages," *J. Phon.*, vol. 4, no. 4, pp. 337–351, doi: 10.1016/S0095-4470(19)31260-4.

[9] Basset, P., A. Amelot, J. Vaissière, and B. Roubeau. 2001. Nasal airflow in French spontaneous speech. *J. Int. Phon. Assoc.*, vol. 31, no. 1, pp. 87–99, doi: 10.1017/S0025100301001074.

[10] Bell-Berti, F. and R. A. Krakow. 1991. Anticipatory velar lowering: A coproduction account. *J. Acoust. Soc. Am.*, vol. 90, pp. 112–123.

[11] Krakow, R. A. 1993. Nonsegmental influences on velum movement patterns: syllables, sentences, stress, and speaking rate. in *Phonetics and Phonology vol. 5: Nasals, Nasalization, and the Velum*, New York: Academic Press.: M.K. Huffman & R.A. Krakow, Eds., pp. 87–116.

[12] Ajili, M., Bonastre, J.-F., Rossetto, S., & Kahn, J. 2016. Inter-speaker variability in forensic voice comparison: a preliminary evaluation. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2114–2118.

[13] Kahn, J., Audibert, N., Bonastre, J.-F., & Rossato, S. 2011. Inter and Intra-speaker Variability in French: An Analysis of Oral Vowels and Its Implication for Automatic Speaker Verification. *ICPhS*, 1002–1005.

[14] Wang, X., Tang, M., Yang, S. *et al. 2019.* Automatic Hypernasality Detection in Cleft Palate Speech Using CNN. *Circuits Syst Signal Process* 38, 3521–3547.

[15] Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., Khanapi Abd Ghani, M., Maashi, M. S., Garcia-Zapirain, B., & Al-Dhief, F. T. 2020. Voice pathology detection and classification using convolutional neural network model. *Applied Sciences*, *10*(11), 3723.

[16] Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., & Woisard, V. 2022.. Validation of the Neuro-Concept Detector framework for the characterization of speech disorders: A comparative study including Dysarthria and Dysphonia. In *Interspeech 2022*.

[17] Ioffe, S., & Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning* (pp. 448-456). PMLR.

[18] Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., & Woisard, V. 2020. Towards interpreting deep learning models to understand loss of speech intelligibility in speech disorders - step 1: Cnn model-based phone classification. In *Interspeech 2020* (pp. 2522-2526). ISCA; ISCA.

[19] Pellegrini, T., & Mouysset, S. 2016. Inferring phonemic classes from CNN activation maps using clustering techniques. In *Annual conference Interspeech (INTERSPEECH 2016)* (pp. pp-1290).

[20] Elmerich, A., Amelot, A., Maeda, S., Laprie, Y., Papon, J. F., & Crevier-Buchman, L. 2020. F1 and F2 measurements for French oral vowel with a new pneumotachograph mask. In *ISSP 2020-12th International Seminar on Speech Production*.

[21] Honda, K., & Maeda, S. 2008. Glottal opening and airflow pattern during production of voiceless fricatives: a new noninvasive instrumentation. *The Jour. of the Acous. Soc. of Am.*, *123*(5), 3738-3738.

[22] Boersma, P. Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.10, retrieved 17 March 2022 from *http://www.praat.org/*.

[23] R Core Team. 2022. R: A language and environment for statistical computing [Computer software manual]. Retrieved from https://www.r-project.org/.

[24] Zerling, J. 1984. Phénomènes de nasalité et de nasalisation vocalique: étude cinéradiographique pour deux locuteurs. *Travaux de l'Institut de phonétique de Strasbourg* 16, 241-266.

[25] Torreira, F., Adda-Decker, M., & Ernestus, M. 2010. The Nijmegen Corpus of Casual French. *Speech Communication*, *52*(3), 201.

[26] G. Gravier, J.F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait and K. Choukri. 2004. The ESTER evaluation campaign of Rich Transcription of French Broadcast News, *Proc. Language Evaluation and Resources Conference*.

[27] Chanclu, A., Georgeton, L., Fredouille, C., &amp; Bonastre, J.-F. 2020. PTSVOX: une base de données pour la comparaison de voix dans le cadre judiciaire. *6e Conférence Conjointe Journées d'Études Sur La Parole (JEP, 33e Édition)*, 73–81.

[28] Durand, M. 1953. De la formation des voyelles nasales, *Stud. Linguist.*, vol. 7, no. 1–2, pp. 33–53, doi: 10.1111/J.1467-9582.1953.TB00489.X.