INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Automatic classification of phonation types in spontaneous speech: towards a new workflow for the characterization of speakers' voice quality

Anaïs Chanclu[1], Imen Ben Amor[1], Cédric Gendrot[2], Emmanuel Ferragne[2], Jean-François Bonastre[1]

[1]Laboratoire Informatique d'Avignon, EA 4128, Avignon Université, France
[2]Laboratoire de Phonétique et Phonologie, UMR7018, CNRS - Sorbonne Nouvelle, France
anais.chanclu@univ-avignon.fr

Abstract

Voice quality is known to be an important factor for the characterization of a speaker's voice, both in terms of physiological features (mainly laryngeal and supralaryngeal) and of the speaker's habits (sociolinguistic factors). This paper is devoted to one of the main components of voice quality: phonation type. It proposes neural representations of speech followed by a cascade of two binary neural network-based classifiers, one dedicated to the detection of modal and nonmodal vowels, and one for the classification of nonmodal vowels into creaky and breathy types. This approach is evaluated on the spontaneous part of the PTSVOX database, following an expert manual labelling of the data by phonation type. The results of the proposed classifiers reaches on average 85 % accuracy at the frame-level and up to 95 % accuracy at the segment-level. Further research is planned to generalize the classifiers on more contexts and speakers, and thus pave the way for a new workflow aimed at characterizing phonation types.

Index Terms: voice quality, speaker characterization, phonation type classification, neural network, explainability

1. Introduction

Voice quality is considered in the literature to have great implications for the characterization of speakers [1]. It can be a permanent component of a speaker's voice due to physiological particularities (mostly laryngeal and supralaryngeal) or speaker habits (sociolinguistic factors), but it may also be subject to intra-speaker variability, notably speech style or emotion [2]. Some known examples of voice quality are nasality, tenseness/laxness, dark/clear voice, phonation type, *etc*.

In this study, we focus on phonation type, which is the mode of vibration of the vocal folds during voiced phonation. It is typically regarded as an articulatory continuum between closed and open glottal state, respectively for speech from creaky to modal and finally breathy voice [3, 4]. Phonation type can be a phonological feature in quite a few languages. For example, Jalapa Mazatec has creaky/modal/breathy /i/, /a/, /æ/, /o/, /u/, but other languages can be mentioned such as Hindi or Tsonga. Phonation type has also been demonstrated to be used for phonetic purposes: for example, creak can be an allophone of intervocalic unvoiced stops in British English [5], and breathiness can be used to signal prosodic finality in French [6], hence a potential role of phonation type as speaker stylistic marker. Obviously, phonation type can be used in the phonetic characterization of a speaker or a group of speakers: women are said to exhibit a more breathy voice than men in British English [7], high prestige speakers in Edinburgh have been reported to use more creaky voice [8], see also [9] for a review.

Voice quality, including phonation type, is often studied perceptually due to limitations in phoneticians' understanding of which acoustic parameters are most relevant, and how exactly they should be measured [10, 11]. We aim in this article at developing a neural-network system allowing for the automatic detection of these three phonation types (modal, creaky, breathy) as they are known to be useful for the characterization of speakers' voices. The system is designed to help phoneticians learn more about phonation types. This knowledge can be easily fed back into the system, to build a virtuous circle.

Section 2 describes the system. The corpus, including the manual annotations dedicated to the targeted task, is described in Section 3. Then, Section 4 presents the experimental protocol and Section 5 the corresponding experimental results. Finally, Section 6 proposes some takeaways and future work.

2. PASE-MLP phonation type detection system

Our goal is to provide an automatic system able to predict the type of phonation for a given speech frame or segment. The system must be bootstrapped on a small set of monitored data because manual annotation of this type of phenomenon takes time. Its ability to use large amounts of unlabelled data later is also of great importance to improve generalization.

At the feature extraction level, the system should allow great flexibility since discovering the relevant features is a part of the task. In addition, the knowledge gathered by the expert thanks to the system should be easily reinjected into the system. The problem-agnostic speech encoder (PASE) [12] looks promising for this purpose. PASE is a variant of autoencoder [13, 14], a kind of neural network trained without supervision to reproduce input to output. During the operating phase, a vector, the "embedding", is extracted from the weights of the central layer for each audio signal frame. PASE is trained for multitasking: the embedding extractor is optimized to achieve different objectives, materialized by a set of "workers", such as retrieving the waveform, cepstral coefficients, the phoneme or certain prosodic parameters. So, the embedding is a compressed information capable of satisfying each worker. In this work we use the PASE+ [15] encoder with a set of workers we consider most useful for the task (MFCC, waveform, prosody, LIM and GIM). By selecting other workers or by defining new workers based on new knowledge, it will be possible to modify the extraction characteristics to take new knowledge into account. PASE is trained with 6609 audio files from PTSVOX [16]. The outputted frames (the embeddings) have a size of 256 coefficients, with a frame rate of 1/100s.

For the classification part of the system, we built a cascade of two binary neural network-based classifiers. The first one aims to detect modal and nonmodal frames (nonmodal frames are frames labelled creaky or breathy) in a vowel segment.

http://dx.doi.org/10.21437/Interspeech.2021-1765

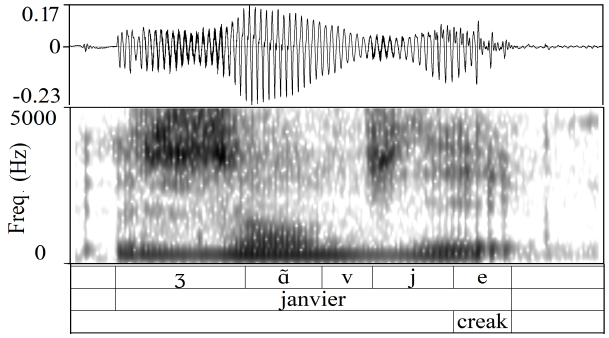

Figure 1: *Labelling example of a creaky vowel*

The second one classifies the nonmodal frames into creaky and breathy types. We selected for both classifiers the same solution, a classical multilayer perceptron (MLP) with one hidden layer. The hidden layer has 256 nodes with a ReLU activation function and a dropout of 0.15. The output layer is a softmax layer. The training is done at frame level in 150 epochs with batch sizes of 50 frames and a learning rate of 10^{-4}.

We selected a balanced training strategy, where the number of examples of each class is balanced for each binary classifier. In the majority class, we randomly selected the same number of frames as that available for the minority class. Details on the training and test sets are provided in Section 4.

3. Corpus

This work is using audio data extracted from the PTSVOX [16] corpus. A specific expert-based annotation process was applied to enrich the meta information of the corpus in terms of phonation type, for the prepausal vowels. The remainder of this section shows the details of the data and the annotation process.

3.1. PTSVOX

The PTSVOX database is composed of French audio recordings. It is specifically designed to study the factors of inter- and intra-speaker variability in forensic voice comparison. These factors include speaking style (reading or spontaneous speech), recording equipment (microphone or telephone), gender, and various information about the speaker (smoking, health issues, *etc*.). All recordings are re-sampled to 16000Hz.

PTSVOX contains 369 speakers split in two sets: *Intra* composed of 24 speakers recorded multiple times and *Inter*, with 345 speakers recorded only once. For this work, we selected the microphone recordings of spontaneous speech from the *Intra* set. This set is composed of 12 female and 12 male speakers recorded once a month, using a microphone and a telephone, over a 3-month period (approximately 8 to 10 minutes per speaker whose excerpts were transcribed and segmented manually).

3.2. Voice quality annotation on prepausal vowels

More specifically, we focused on prepausal vowels because they contain linguistic key points since they occur at prosodic prominences, and they are perceptually important in communication. Prepausal vowels were perceptually annotated by one of the authors in three categories: creaky, modal, and breathy. For this annotation process, we automatically isolated vowels immediately followed by a pause for each speaker with a threshold of 50 ms for minimum vowel and pause duration. Then, by visualizing spectrograms and listening to the vowels, labels were assigned to them (see Figure 1). A single vowel could have two labels if it contained a change in voice quality. The 'other' label was assigned to ambiguous or noisy cases, which were then discarded. As there can be more than one label within a vowel, we use the term *segment* to refer to the labels within the vowel. The resulting corpus contains 10,361 segments for 8,889 vowels.

4. Experimental protocol

We decided to evaluate the two binary classifiers separately: modality detection between modal and nonmodal labels and voice quality detection between creaky and breathy labels. In the rest of this section, we present the reference system we defined, the split of the data into training and testing sets and the decision rules.

4.1. MFCC-SVM reference system

In order to serve as a baseline, we built a classical MFCC+SVM recognizer. Acoustic features are composed of 30 MFCC parameters with a 0-8kHz bandwidth, extracted using torchaudio. The window length is 20 ms with an overlap of 10 ms between 2 consecutive windows, giving a frame rate of 100 frames per second. We used a support vector machine (SVM) with a radial basis function. Similarly to our MLP system, the models were trained on a balanced and randomized amount of data between classes.

4.2. Training and test subsets

For each gender, we randomly selected two speakers and used the corresponding audio data to form a fully separated test set. The selected speakers for the test set are `LG001` and `LG009` for female speakers and `LG012` and `LG024` for male speakers. 20 % of the remaining data are then randomly selected to form a validation set, and the rest is dedicated to the training set. The speakers of the test set are never seen in the training or in the validation sets. The train and validation sets contain recordings spoken by the same speakers.

Table 1: *Number of segments and duration (in 1/100s frames) for the train and test subsets*

	Train & Val		Test	
	Seg.	Frames	Seg.	Frames
modal	4,090	89,478	687	15,773
nonmodal	2,220	33,554	464	7,410
creaky	1,088	17,441	251	3,569
breathy	1,132	16,113	213	3,841

Table 1 presents the number of annotated prepausal vowel segments and the corresponding duration (in number of frames) for both sets and vowel label.

4.3. Scoring rules

The predictions are done frame by frame for both tasks (modal/nonmodal and creaky/breathy) and both recognizers, with a rate of 100 frames per second (recall that the parameterization steps of the two recognizers used a signal window longer than 1/100s to compute a given frame). Decisions were made using an simple rule: if the score was higher than 0.5, the rec-

ognizer answered yes, and no otherwise (there was no threshold setting at all). Performance at the segment-level was also computed. To predict voice quality classification at the segment-level, we first performed a prediction at the frame-level. Then we computed the mean prediction over all frames of the segment before applying the same decision rule as previously.

5. Results

We first present the results for the modal/nonmodal task before moving on to those of the creaky/breathy classification task. Performance is presented using confusion matrices for a clearer visualization of class imbalance in the test set. Performance on the test set only (i.e. using data coming from speakers never seen during the training) is shown.

5.1. Modal/nonmodal classification

Table 2 presents the performance of the modal/nonmodal classifiers at the frame-level. 87 % of the modal frames and 76 % of the nonmodal frames are well classified with the PASE-MLP system ($\kappa = 0.62$). The reference system shows lower performance with, respectively, 79 % and 72 % ($\kappa = 0.49$). Table 3 shows the performance for the same task but gathered at the segment-level. Overall, the same difference between our system PASE-MLP and the baseline appears. As expected, the general level of performance is significantly higher than for the frame-level, with up to 93 % of good classification for modal frames ($\kappa = 0.76$). We can notice the performance of the baseline system at the segment-level is close to the performance of our PASE-MLP system at the frame-level ($\kappa = 0.65$).

Table 2: *Confusion matrices between modal and nonmodal frames in the test set for both systems*

	PASE-MLP		
	modal	nonmodal	Total
modal	13,640 (87 %)	2,133 (13 %)	15,773
nonmodal	1,768 (24 %)	5,642 (76 %)	7,410
	MFCC-SVM		
modal	12,430 (79 %)	3,343 (21 %)	15,773
nonmodal	2,070 (28 %)	5,334 (72 %)	7,410

Table 3: *Confusion matrices between modal and nonmodal segments in the test set for both systems*

	PASE-MLP		
	modal	nonmodal	Total
modal	638 (93 %)	49 (7 %)	687
nonmodal	80 (17 %)	384 (83 %)	464
	MFCC-SVM		
modal	595 (87 %)	92 (13 %)	687
nonmodal	102 (22 %)	362 (78 %)	464

5.2. Creaky/breathy classification

Table 4 presents the performance of the creaky/breathy classifiers at the frame-level. The baseline system has a good recognition rate for breathy frames (88 %) but shows a clear weakness for creaky frames with 67 % of correct recognition ($\kappa = 0.51$). Our PASE-MLP approach outperforms the baseline with 85 % for creaky frames and 88 % for breathy frames ($\kappa = 0.73$). This favours the cascade architecture: it seems easier to separate breathy and creaky frames than modal and nonmodal frames, although additional experiments are required before a definitive conclusion can be reached. Table 5 shows the results of similar experiments to Table 4 but at the segment-level. The performance confirms our previous findings. The performance at the segment-level for this task with our PASE-MLP system appears very promising, with about 95 % of correct classification ($\kappa = 0.88$, and $\kappa = 0.67$ for MFCC-SVM).

Table 4: *Confusion matrices between creaky and breathy frames on the test set for both systems*

	PASE-MLP		
	creaky	breathy	Total
creaky	3,048 (85 %)	521 (15 %)	3,569
breathy	465 (12 %)	3,376 (88 %)	3,841
	MFCC-SVM		
creaky	2,376 (67 %)	1,193 (33 %)	3,569
breathy	586 (15 %)	3,255 (85 %)	3,841

Table 5: *Confusion matrices between creaky and breathy segments on the test set for both systems*

	PASE-MLP		
	creaky	breathy	Total
creaky	235 (94 %)	16 (6 %)	251
breathy	11 (5 %)	202 (95 %)	213
	MFCC-SVM		
creaky	193 (77 %)	58 (23 %)	251
breathy	19 (9 %)	194 (91 %)	213

6. Discussion and conclusion

This article aims to lay the foundations of a new workflow dedicated to voice quality, and more precisely to phonation type. The final objective is to propose an automatic tool to help phoneticians to improve their understanding of phonation phenomena.

We started here at a low level by defining a new neural network solution to automatically detect three phonation types (modal, creaky and breathy) on prepausal vowels. We implemented a cascade of two binary classifiers, based on a representation learning approach, PASE, for feature extraction and a classical MLP for classification. These choices were driven by two wishes: to be able to generalize the knowledge gathered from the -limited in size- manual annotation provided by phoneticians, and to embed easily the phoneticians understanding in the system, following a virtuous circle model. The proposed solution yielded very satisfactory performance, with on average ≈84 % of correctly labelled frames for both classifiers. The performance reached ≈91 % when the decision was taken at the segment level. Our neural network-based solution, PASE-MLP, clearly outperformed a classical MFCC-SVM approach (≈75 % on average at the frame level for the latter, to be compared to ≈85 % for the PASE-MLP system). More importantly, the performance of PASE-MLP system appeared more stable than the MFCC-SVM solution.

The results open the door for further investigation. First, the PASE-MLP system should be generalized to more phonetic contexts and speakers. Our next step is to use transfer learning [17] to pretrain both PASE and MLP modules (it will use the current system to propose initial labels as suggested in [18]). It will allow us to apply the system on all the PTSVOX data as well as on other, larger, databases. Second, the system will provide expert phoneticians with the best subset of examples for perceptual study and the possibility to incorporate their feedback. Finally, we wish to help expert phoneticians to study speaker phonotypes on a large scale, including intra-speaker variability, thanks to the automatic tools we will provide.

7. Acknowledgements

The research reported here was supported by the ANR-17-CE39-0016 VoxCrim project.

8. References

[1] E. Gold and P. French, "International practices in forensic speaker comparison," *International Journal of Speech, Language and the Law*, vol. 18, no. 2, pp. 293–307, 2011.

[2] F. Nolan, "Forensic speaker identification and the phonetic," *A figure of speech: A Festschrift for John Laver*, p. 385, 2005.

[3] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[4] R. Wright, C. Mansfield, and L. Panfili, "Voice quality types and uses in north american english," *Anglophonia. French Journal of English Linguistics*, no. 27, 2019.

[5] M. Garellek and S. Seyfarth, "Acoustic differences between english /t/ glottalization and phrasal creak," in *Interspeech 2016*, 2016, pp. 1054–1058. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1472

[6] C. Smith, "Marking the boundary: utterance-final prosody in french questions and statements," in *14th International Conference of Phonetic Sciences*, 1999, pp. 1181–1184.

[7] C. Henton and A. Bladon, *Language, speech, and mind: Studies in honour of Victoria A. Fromkin*, v fromkin, larry m hyman and c n li ed. London; New York: Routledge, 1988, ch. Creak as a Sociophonetic Marker, pp. 3–29.

[8] J. Esling, "The identification of features of voice quality in social groups," *Journal of the International Phonetic Association*, vol. 8, no. 1/2, pp. 18–23, 1978.

[9] E. San Segundo, P. Foulkes, P. French, P. Harrison, V. Hughes, and C. Kavanagh, "The use of the vocal profile analysis for speaker characterization: Methodological proposals," *Journal of the International Phonetic Association*, vol. 49, no. 3, pp. 353–380, 2019.

[10] P. Keating and M. Garellek, "Acoustic analysis of creaky voice," *Poster apresentado em sessão especial sobre voz crepitante no Encontro Anual da Linguistic Society of America em Portland (OR)*, 2015.

[11] P. F. Katharina Klug, Christin Kirchhübel and P. French, "Analysing breathy voice in forensic speaker comparison. using acoustics to confirm perception," in *19th International Conference of Phonetic Sciences*, 2019, pp. 795–799.

[12] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks," in *Proc. of the Conf. of the Int. Speech Communication Association (INTERSPEECH)*, 2019, pp. 161–165. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2605

[13] A. Ng *et al.*, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, no. 2011, pp. 1–19, 2011.

[14] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, vol. 2013, 2013, pp. 436–440.

[15] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for Robust Speech Recognition," *ArXiv:2001.09239*, 2020.

[16] A. Chanclu, L. Georgeton, C. Fredouille, and J.-F. Bonastre, "Ptsvox: une base de données pour la comparaison de voix dans le cadre judiciaire," in *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)*. ATALA; AFCP, 2020, pp. 73–81.

[17] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[18] A. Gresse, M. Quillot, R. Dufour, and J.-F. Bonastre, "Learning Voice Representation Using Knowledge Distillation For Automatic Voice Casting," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.