"See what I mean, huh?" Evaluating Visual Inspection of F0 Tracking in Nasal Grunts

Nicolas Ballier, Aurélie Chlébowski

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

"See what I mean, huh?" Evaluating Visual Inspection of F_0 Tracking in Nasal Grunts

Aurélie Chlébowski, Nicolas Ballier

Université de Paris, CLILLAC-ARP, F-75013 Paris, France
aurelie.chlebowski@hotmail.fr, nicolas.ballier@u-paris.fr

Abstract

This paper proposes to evaluate the method used in Chlébowski and Ballier [1] for the annotation of F_0 variations in nasal grunts. We discuss and test issues raised by this kind of approach exclusively based on visual inspection of the F_0 tracking in *Praat* [2]. Results tend to show that consistency in the annotation depends on acoustic features intrinsic to the grunts such as F_0 slope and duration that are sensitive to display settings. We nonetheless acknowledge the potential benefits of such a method for automation and implementation in IA and in this respect, we introduce *Prosogram* [3] as an alternative material-maker.

Index Terms: Nasal grunts, paralinguistics, pitch, F_0 tracking, auditory perception, visual perception, annotations, CID.

1. Introduction

Non-lexical conversational sounds (hereafter, N-LC sounds; [4]) have become a central issue in several research areas, including linguistics, psychology, cognition and human-computer interactions [5]–[21]. Although ubiquitous in everyday interactions, these sounds have so far escaped both clear denomination and accurate definition [4], [11], [22]. Their "*liminality*" [11], appears to conflict with traditional conceptions of language and the mechanisms behind their semiotics have hitherto remained unclear. Numerous studies proposed to account for the functions of N-LC sounds in speech interactions have provided interesting classifications in this respect, *e.g.*, *backchannels*, *fillers*, *disfluencies*... Nonetheless, reasons why such a sound as *mm* can have multiple roles in interaction cannot be explained within interpretative approaches alone: N-LC sounds are, first and foremost, *sounds* and greater emphasis should be given to their acoustics [4]. Ward [4] thus offered to shift perspectives for these sounds by emphasizing their acoustic aspect. He proposed to consider N-LC sounds as compositional entities made of several acoustic components that convey specific meanings. Such a sound as *mm*, for instance, at least comprises a segmental component /m/ uttered with a specific phonation mode (*e.g.*, modal, creaky, or breathy), pitch contour (*e.g.* levelled, rising, or falling), intensity and duration [4], [19], [23], [24]. The variety of components an N-LC sound consist of could explain the difficulty to ascribe a single meaning to it. In addition, within this "Compositional Model" [4], variation in any of these components induces variation in meaning. Reported variability in functions for N-LC sounds could therefore be accounted for in terms of their acoustic compositions. Within this framework, Chlébowski and Ballier [1] gave detailed guidelines for primary annotations of acoustic components in *nasal grunts* (hereafter, NG), *e.g. han*, *hein*, *hum*, *mmhm* in French[1]. They focused their annotation guidelines on visual inspections of the signal. Annotators had to follow the guidelines to interpret noticeable acoustic cues as to characterize and provide basic comments on features under scrutiny (*e.g.*, noise on the spectrogram is considered a cue for /h/ component). Their guidelines, so far applied with *Praat* [2] by a single annotator on NG in the French *Corpus of Interactional Data* (CID) [26] and parts of the *Santa Barbara Corpus of Spoken American English* (SBC) [27] and the *Phonological Variation and Change in Contemporary Spoken English* (PVC) project [28], were deemed satisfactory for the characterization of the following components: non-modal phonations (creakiness, breathiness, and ingressive phonations), glottal stops, duration and variations in F_0[2]. Although their method was replicated on several corpora, inter-annotator agreement has not been controlled yet.

In this paper, we propose to examine the robustness of this kind of guidelines for the annotation of F_0. The rest of the paper is organized as follows: section 2 discusses the pros and cons raised by the method proposed in Chlébowski and Ballier [1] for the annotation F_0. Section 3 details material and experiment for evaluating this method with *Praat* software [2]. Section 4 discusses our results. Section 5 introduces *Prosogram* [3] as an alternative solution for visual annotations for the F_0 of N-LC sounds and section 6 concludes.

2. Related Work

N-LC sounds used to be investigated for their functions in interaction and their acoustics were subsequently often overlooked. Yet, the issue has gained increasing attention over time[3]. Prosodic features such as duration and intonation arguably constitute the most investigated acoustic aspects in N-LC sounds. This section discusses methods for the analysis of variations in pitch direction in N-LC sounds. We first provide a quick review of pitch-based investigations. We then introduce an alternative method based on a visual inspection of variations in F_0 tracking and discuss the benefits and issues of such a method.

[1] What they defined after Chlébowski and Ballier [25] as "a sub-category of *non-lexical conversation*[*al*] *sounds* based on a distinctive acoustic feature: nasality" ([1], p. 6514).
[2] See for instance [1], [24], [29].
[3] *Research on language and Social Interaction* dedicated a special issue to the matter, available at: https://www.tandfonline.com/toc/hrls20/53/1?nav=tocList (last accessed: March 4th, 2021).

http://dx.doi.org/10.21437/Interspeech.2021-129

2.1. Pitch-based investigations

As with most speech phenomena, progress in the analysis of variations in pitch direction in N-LC sounds was constrained by technological and theoretical advances[4]. Early studies being mostly concerned with providing functional categorizations of the sounds, variations in pitch were often acknowledged as an aside and seemingly based on perceptual categorizations. Improvement in the tracking of the fundamental frequency (F_0) then allowed for deeper acoustic measurements. Perceived pitch variations in N-LC sounds were enriched with measurements of mean, minimum, and maximum F_0 that provided valuable information to the understanding of these sounds. For instance, in a study that consisted of both perceptual and acoustic analyses, Duez ([30], [31]) showed, *inter alia*, that the French *filler euh* ([ø]) can display several intonational patterns, categorised under "*flat*", "*upward*", or "*downward*" labels. These can appear alone or paired and are neither influenced by duration nor by location. This kind of analyses based on perceptual categorizations of pitch variations together with acoustic measurements of the F_0 is slowly becoming a standard method for the study of pitch direction in N-LC sounds – when the quality of the recordings allows it (*cf.* [8], [17], [30]–[34]).

2.2. F_0-based investigations

N-LC sounds comprise a host of acoustic components that speakers can arrange to convey and vary meanings (for instance [4], [24], [29]). Considering N-LC sounds as compositional entities not only implies considering their acoustic components altogether but also monitoring their influence on the perception of a single component [35]. Perception thresholds for glissandi, for instance, were reported to be sensitive to other acoustic parameters such as intensity [36] and duration [3]. With the chief purpose to homogenize annotations and minimize perceptual biases, Chlébowski and Ballier [1] set up guidelines for the annotation of acoustic components in NG that consist of visual inspections of the signal. For the annotation of variations in pitch direction, they proposed to rely on variations in F_0 tracking with *Praat* [2]. Instructions are to set pitch in semitones (ST) and to annotate variations in F_0 ("rise" or "fall"), or lack thereof ("level"), for each segment in NG (*e.g.,* both /m/ in *mmhm*/m.m/; /œ/ and /m/ in *hum* /œm/)[5]. To that end, annotators are asked to zoom in a stimulus and zoom out of it only once. In case they needed to justify their choice, harmonics were displayed on a narrowband spectrogram.

2.3. Pros and cons of F_0-based investigations

The method has broader benefits but is not without drawbacks. The basic annotations can not only be used as primary material for deeper analysis as in Duez ([30], [31]), or Batliner *et al*. [8], but also to determine both the perceptual thresholds and the distinctive status of the components [1]. In addition, these could be submitted to automation insofar as they rely on acoustic cues [1], and even implemented in IA systems. On the other hand, F_0 tracking is ill-famed for pitch detection error. Moreover, components such as *non-modal phonation modes* are likely to impact the F_0 tracking [17], [30], [31], [37]. Finally, instructions in Chlébowski and Ballier [1] do not take account of the effect of duration on the display of F_0 tracking.

3. Evaluation Procedure

This section details material and method to evaluate guidelines proposed in Chlébowski and Ballier [1] for the annotation F_0.

3.1. Replication study

We focused our analysis on monosyllabic NG studied in Chlébowski and Ballier [1] that were produced by female speakers in the CID [26]. Reasons are three-fold. First, the CID [26] was recorded at the *Laboratoire Parole et Langage* (LPL) [6] and provides high-quality recordings of spontaneous conversations in French[7]. Given the goals of this paper, working with audio material where random noises are limited is an asset. Second, focusing on either female or male speakers allows keeping the same settings throughout the experiment – working with stimuli produced by both female and male speakers would have implied asking participants to adapt settings for pitch range every two stimuli[8]. Finally, we chose to restrict the evaluation to monosyllabic NG to keep instructions simple and avoid overloading the participants' task.

3.2. The stimuli

Our stimuli consist of 24 randomly selected monosyllabic NG[9]. Focus was restricted to NG with one segment only, which discards monosyllabic NG of *hum* (/œm/) type. Our set comprises 8 *hein* (/ɛ̃/), 4 *han* (/ɑ̃/) and 12 *mm* (/m/), either in modal or non-modal phonation, either high- or low-pitched, and with total lengths that range from 65ms to 893ms. Variables such as phonation modes, segment nature, F_0 label, pitch height[10], or duration were not controlled so as to evaluate the method independently of these features. Location of the NG in interaction does not weight much since the stimuli were extracted from context, see 3.1.

3.3. Experiment design and conditions

Stimuli were extracted from their environments, duplicated and randomly concatenated with around 100ms blank between each stimulus with *Praat* software [2] and stored in a .WAV file. The stimuli come with a TextGrid that consist of 2 tiers. The first tier recalls NG numbers as given in [1]. The second tier is dedicated to participants' labelling of F_0. Participants brought their own laptops. Settings for pitch range were set as follows: 100-500Hz in semitones re 1Hz. Instructions were the same as those in Chlébowski and Ballier [1] (see 2.2). Participants were exposed to dummy examples before performing the actual experiment.

[4] See Dingemanse [11] for a discussion about advances in theory and technology for the analysis of such sounds.
[5] Semi-tones were chosen instead of Hertz since they make available to the eye what is perceived with the ear.
[6] https://www.lpl-aix.fr/ (last accessed: March 4th, 2021).
[7] https://www.ortolang.fr/market/corpora/sldr000720 (last accessed: March 4th, 2021).
[8] Female and male speakers sharing different pitch range, *Praat* online manual recommends adapting pitch range settings accordingly. For additional information, please refer to: https://www.fon.hum.uva.nl/praat/manual/Intro_4_2__Configuring_the_pitch_contour.html (last accessed: March 4th, 2021).
[9] That is, around 4% of monosegmental NG in [1].
[10] Referred to as "*register*" in [1], [24], [25], and [29].

3.4. Participants

Three MA students (two females and a male) in linguistics at the University of Paris participated in the experiment. All were in their early twenties and native speakers of French – although one was born in India. As to the annotator in Chlébowski and Ballier [1], she was a female in her late twenties. A native speaker of French, she was completing a PhD in phonetics at the *University of Paris* as well.

4. Results

This section discusses results of the evaluation experiment. We first detail inter- and intra-rater agreement. We then investigate identification rates when the annotation in Chlébowski and Ballier [1] is considered a *gold standard*.

4.1. Inter- and intra-rater agreement

Inter-rater agreement was performed with R [38] using the kappam.fleiss() function in the {irr} package [39]. The percentage of agreement between participants in our experiment for the 48 stimuli is 70.8% with the Kappa coefficient showing moderate agreement ($\kappa = 0.684$, $p < .05$). When the annotator in Chlébowski and Ballier [1] is considered a fourth rater, the percentage of agreement decreases to 56.2% with the Kappa coefficient showing lower agreement ($\kappa = 0.601$, $p < .005$); which suggests some disagreement between the two sessions of annotation.

Participants' consistency against stimulus duplication was then measured with the kappa2 function from the {irr} package [39]. Participant 1 was consistent across stimuli 87.5% of the time, with the Kappa coefficient suggesting strong consistency ($\kappa = 0.804$, $p < .001$). Participants 2 and 3 were less consistent, with 79.2% and 75% consistency respectively, with the Kappa coefficient suggesting moderate consistencies ($\kappa = 0.687$; $\kappa = 0.603$, p-values $< .001$). Stimuli involved in rating inaccuracies vary across participants.

4.2. Identification of putative contours

Table 1 below presents the confusion matrix when the annotation in Chlébowski and Ballier [1] is used as gold standard. We calculated the F1 score to assess the classification of the contours by the participants. Falls are clearly easier to predict.

Table 1: *F1 score for each kind of contours as estimated by participants*

		Participant estimations (aggregated)			F1
		Fall	***Level***	***Rise***	
Gold standard	***Fall***	30	0	0	.908
	Level	3	40	11	.655
	Rise	3	28	29	.579

We investigated issues raised in 2.3 to determine their contribution to identification inaccuracies. Table 2 below recaps correct and incorrect identifications of F_0 variations according to phonation mode (modal vs. non-modal phonation), F_0 slope and duration of the stimuli. Slopes between 3 to 5 ST were considered moderate and lengths between 200 to 400ms were considered mid. During the experiment design, we noticed another phenomenon that could contribute to incorrect identifications of F_0 variations: micro-prosodic variations at the beginning and/or end of the F_0 tracking. These variations affect more than half of the stimuli and do not seem to be correlated with non-modal phonation but may reflect glottal aperture and/or closure. Figure 1 below illustrates stimuli both presumed rising, uttered in modal phonation, and which display (top) or not (bottom) micro-prosodic variations. While the F_0 curve is distinctly rising on NG#751, it is not clear whether it is rising, levelled, or even falling on NG#417.

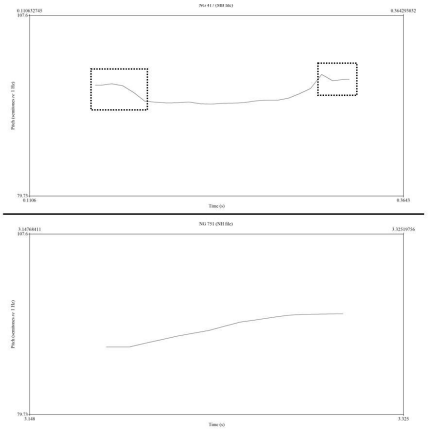

Figure 1: *F_0 curves of NG with micro-prosodic variations (NG#417- MB file; top) and without micro-prosodic variations (NG#751 - NH file; bottom) – Praat drawings*

We performed Chi Square tests to assess the independence of the results with the phonetic features. It is only for the category modal vs. non-modal that H_0 cannot be rejected ($\chi^2(1, N = 144) = .002$, $p > .001$). Nonetheless, most incorrect ratings in this case were related to creaky voice (> 80%). Overall, incorrect ratings are more frequent in cases of micro-prosodic variations, mid to short duration and low F_0 slope.

Table 2: *Frequencies of correct and incorrect identification according to several acoustic features*

		Correct	Incorrect	p-value
Modal phonation	***Yes***	53	25	> .001
	No	46	20	
Micro-prosody	***Yes***	61	41	< .001
	No	38	4	
Duration (ms)	***Short***	30	0	< . 001
	Mid	38	34	
	Long	31	11	
F_0 slope (ST)	***Low***	29	1	< .001
	Moderate	23	7	
	High	47	37	

5. Discussion

Results discussed in section 4 were to be expected. NG, alike any other N-LC sounds, consist of a certain number of acoustic components likely to interfere with each other and to disrupt not only auditory perception but also the acoustic signal. The idea to annotate acoustic components in NG from visual inspections of acoustic cues nonetheless remains interesting insofar as it could eventually enable IA systems to *read* what a speaker

means when producing an N-LC sound. As regards the feature at stake in this paper, there are other pieces of software and scripts that allow researchers to circumvent the issues raised by the compositional nature of NG. For instance, *Prosogram* [3] is a *Praat* script [2] which provides stylized representations of the F_0 curve that reflect human perception of glissandi, or lack thereof. In so doing, *Prosogram* not only provides accurate representations of variations in pitch directions (rise and fall *vs.* level) but also addresses the problems raised by duration of the sounds and micro-prosody – as well as that of disruptions in F_0 tracking induced by non-modal phonation.

Figure 2 below contrasts the F_0 curves of two NG as drawn in *Praat* [2] (left) and *Prosogram* [3] (right). Both NG were labelled as *levelled* in Chlébowski and Ballier [1] from the tracking provided in *Praat* and identified as so by our three annotators– although the legibility of the curves is in both cases affected by features intrinsic to the NG. NG#14 (top) was uttered in creaky voice and NG#477 (bottom) is of short length and has a low F_0 slope (if any). We used the *Prosogram* in an attempt to improve the legibility of F_0 tracking. The *Prosogram* was set to detect the smallest perceptible glissandi ($G = 0.16/T^2$) [3], [40] with pitch range from 0 ("autorange") to 500Hz. The script was run through the recordings for each of the participants in the CID [26] for more accurate results[11]. We chose to generate wide and rich prosograms with targets in semitones. Additional information displayed are *tokens* as given in TextGrids that come with the CID ([26], [41]), *NG numbers*, *NG transcription*, and *F_0* as labelled in Chlébowski and Ballier [1]. The prosograms show that NG#14 may indeed be *levelled* (around 93 ST) while NG#477 is likely *rising* (from 91.4 to 92.2 ST).

Prosogram has many other functions of interest here, such as the calculation and display of speaker's pitch range; a function that would allow for visual estimation of pitch height (just above the median for NG#14 on Figure 3). The only drawback we noted was that half of the stimuli used in this paper escaped the pitch detection in *Prosogram*. For instance, NG#417 and #751 on Figure 1 were not detected in *Prosogram*. We believe that the issue may be related to the relative intensity of the NG. The feature was not investigated in Chlébowski and Ballier [1] but nonetheless considered a significant component. Like other acoustic components in NG, intensity is likely to vary across NG. Since the calculation of F_0 in *Prosograms* is based on vowel nuclei [3] both vocalic and sonorant NG can be recognized by the program but low intensity NG are likely to be ignored in any case.

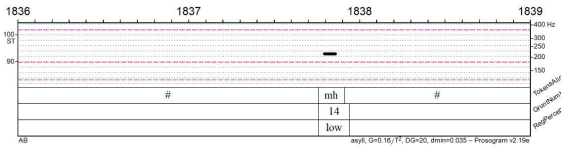

Figure 3: *Prosogram (settings: wide, light, with pitch range) for NG#14 (AB file)*

6. Conclusion

In this paper, we have tried to propose more robust methods for the analysis of detailed acoustic correlates as required for the investigation of paralinguistic items such as nasal grunts. Possibly paving the way for ulterior image detection of pitch contours in IA systems, we have advocated visual inspection of features observed in nasal grunts. We investigated the robustness of annotations of contours based on a visual inspection of the F_0 tracking in *Praat* [2]. It is likely that pitch tracking in *Praat* is sensitive to micro-prosodic phenomena (whether triggered by non-modal phonation or not). Human subjects, on the other hand, seemed overinfluenced by differences in visual displays (zoom distance, in particular). We introduced the *Prosogram* [3] as an alternative source of material. Stylized representation of F_0 tracking can help circumvent a few issues. As evidenced in Figure 2 (NG#14), the *Prosogram* seems to be less sensitive to micro-prosody in the case of creaky voice. Such a program, however, may fail to capture grunts (50% of the time with the *Prosogram* in our experiment). For future machine learning-based investigations of nasal grunts, annotations of features as provided by the *Prosogram* sound promising, with the proviso that up to half of the grunts may not be labelled.

7. Acknowledgements

We would like to thank Roxane Bertrand for giving us access to the CID. We would also like to thank the reviewers for their comments on an earlier version of this paper.

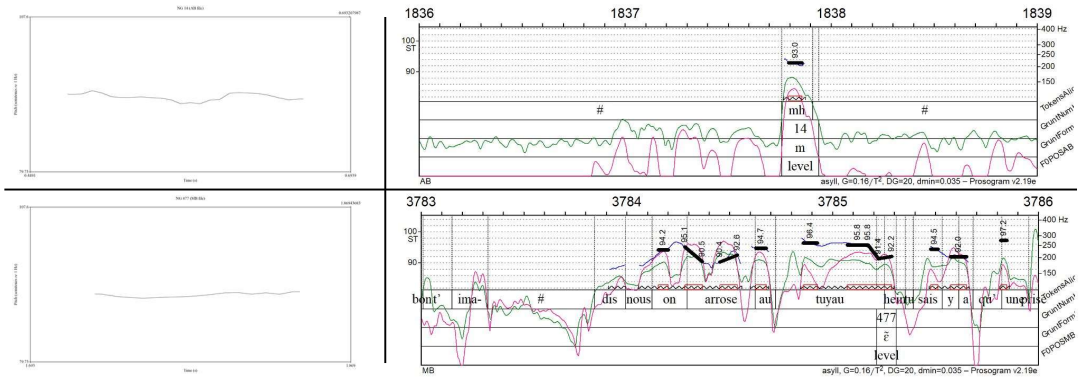

Figure 2: *F_0 curves as drawn in Praat (left; in ST) and in Prosogram (right; wide, rich with targets in ST) for the same grunts; with NG#14 (AB file; top) and NG#477 (MB file; bottom)*

[11] See additional recommendations on the *Prosogram* online user guide: https://sites.google.com/site/prosogram/home (last accessed: March 4th, 2021).

8. References

[1] A. Chlébowski and N. Ballier, 'A Manually Annotated Resource for the Investigation of Nasal Grunts', in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 6514–6522.

[2] P. Boersma and D. Weenink, 'Praat: Doing phonetics by computer [Computer program] (Version 6.1.38)', 2021.

[3] P. Mertens, 'The prosogram: Semi-automatic transcription of prosody based on a tonal perception model', in *Proceedings of Speech Prosody*, 2004, pp. 549–552.

[4] N. G. Ward, 'Non-lexical conversational sounds in American English', *Pragmat. Cogn.*, vol. 14, no. 1, pp. 129–182, 2006.

[5] K. Audhkhasi, K. Kandhway, O. D. Deshmukh, and A. Verma, 'Formant-based technique for automatic filled-pause detection in spontaneous spoken English', in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4857–4860.

[6] H. Buschmeier, Z. Malisz, M. Włodarczak, S. Kopp, and P. Wagner, ''Are You Sure You're Paying Attention?'- "Uh-Huh" Communicating Understanding as a Marker of Attentiveness', in *Proceedings of Interspeech 2011*, 2011, pp. 2057–2060.

[7] H. H. Clark and J. E. F. Tree, 'Using *uh* and *um* in spontaneous speaking', *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[8] A. Batliner, A. Kießling, S. Burger, and E. Nöth, 'Filled pauses in spontaneous speech', in *Proceedings of the XIIIth ICPhS*, 1995, vol. 3, pp. 472–475.

[9] M. Corley and O. W. Stewart, 'Hesitation Disfluencies in Spontaneous Speech: The Meaning of *um*', *Lang. Linguist. Compass*, vol. 2, no. 4, pp. 589–602, 2008.

[10] M. Corley and R. J. Hartsuiker, 'Why *Um* Helps Auditory Word Recognition: The Temporal Delay Hypothesis', *PloS One*, vol. 6, no. 5, p. e19792, 2011.

[11] M. Dingemanse, 'Between sound and speech: Liminal signs in interaction', *Res. Lang. Soc. Interact.*, vol. 53, no. 1, pp. 188–196, 2020.

[12] D. Duez, 'Signification des hésitations dans la parole spontanée', *Rev. Parole*, pp. 113–138, 2001.

[13] J. E. Fox Tree, 'Interpreting pauses and ums at turn exchanges', *Discourse Process.*, vol. 34, no. 1, pp. 37–55, 2002.

[14] R. Gardner, *When listeners talk: response tokens and listener stance*. John Benjamins Publishing, 2001.

[15] L. Keevallik and R. Ogden, 'Sounds on the Margins of Language at the Heart of Interaction', *Res. Lang. Soc. Interact.*, vol. 53, no. 1, pp. 1–18, 2020.

[16] S. Pammi and M. Schroder, 'Annotating meaning of listener vocalizations for speech synthesis', in *Proceedings of the Third International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–6.

[17] E. Shriberg, 'To `errrr' is human: ecology and acoustics of speech disfluencies', *J. Int. Phon. Assoc.*, vol. 31, no. 1, pp. 153–169, 2001.

[18] N. G. Ward, 'Responsiveness in dialog and priorities for language research', *Cybern. Syst.*, vol. 28, no. 6, pp. 521–533, 1997.

[19] N. G. Ward, 'The challenge of non-lexical speech sounds', in *Proceedings of the Sixth International Conference on Spoken Language Processing*, 2000, pp. 571–574.

[20] C.-H. Wu and G.-L. Yan, 'Acoustic Feature Analysis and Discriminative Modeling of Filled Pauses for Spontaneous Speech Recognition', in *Real World Speech Processing*, JF. Wang, S. Furui, and BH. Juang, Eds. Springer, 2004, pp. 17–30.

[21] T. Yamaguchi, K. Inoue, K. Yoshino, K. Takanashi, N. G. Ward, and T. Kawahara, 'Analysis and prediction of morphological patterns of backchannels for attentive listening agents', in *Proceedings of the 7th International Workshop on Spoken Dialogue Systems*, 2016, pp. 1–12.

[22] J. Trouvain and K. P. Truong, 'Comparing non-verbal vocalisations in conversational speech corpora', in *Proceedings of the 4th international workshop on corpora for research on emotion sentiment and social signals (es3 2012). Paris, France: European Language Resources Association (ELRA)*, 2012, pp. 36–39.

[23] N. G. Ward, 'Pragmatic functions of prosodic features in non-lexical utterances', in *Proceedings of the 7th International Conference on Speech Prosody*, 2004, pp. 325–328.

[24] A. Chlébowski and N. Ballier, 'C'est "mm-hm, oui" ou "mm-hm, non"? Propositions pour une grammaire des composantes acoustiques des interactions nasalisées', in *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e 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. 100–108.

[25] A. Chlébowski and N. Ballier, 'Nasal grunts" in the NECTE corpus, Meaningful interactional sounds', in *Proceedings of EPIP4-4th International Conference on English Pronunciation: Issues & Practices*, 2015, pp. 54–8.

[26] R. Bertrand *et al.*, 'Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle', *Trait. Autom. Lang.*, vol. 49, no. 3, pp. 105–134, 2008.

[27] J. W. Du Bois, W. L. Chafe, C. Meyer, S. A. Thompson, R. Englebretson, and N. Martey, *Santa Barbara corpus of spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium, 2000-2005.

[28] L. Milroy, J. Milroy, and G. J. Docherty, 'Phonological variation and change in contemporary spoken British English', Final report to the United Kingdom Economic and Social Research Council., 1997.

[29] A. Chlébowski, 'A Semasiological Approach to Non-Lexical Conversational Sounds: Issues, Benefits and Impact', *Proceedings of Laughter and Other Non-Verbal Vocalisations Workshop*, pp. 11–14, 2020.

[30] D. Duez, 'Caractéristiques acoustiques et phonétiques des pauses remplies dans la conversation en français', *Trav. Interdiscip. Lab. Parole Lang. Aix-En-Provence TIPA*, vol. 20, pp. 31–48, 2001.

[31] D. Duez, 'Acoustico-phonetic characteristics of filled pauses in spontaneous French speech: preliminary results', in *Proceedings of DISS'01*, 2001, pp. 41–44.

[32] M. Dingemanse, F. Torreira, and N. J. Enfield, 'Is "Huh?" a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items', *PloS ONE*, vol. 8, no. 11, p. e78273, 2013.

[33] E. Hofstetter, 'Nonlexical "Moans": Response Cries in Board Game Interactions', *Res. Lang. Soc. Interact.*, vol. 53, no. 1, pp. 42–65, 2020.

[34] S. Pehkonen, 'Response Cries Inviting an Alignment: Finnish *huh huh*', *Res. Lang. Soc. Interact.*, vol. 53, no. 1, pp. 19–41, 2020.

[35] A. Chlébowski, 'The Meaning of "Nasal Grunts" in the Necte Corpus. A Preliminary Perceptual Investigation', *Res. Lang.*, vol. 14, no. 1, pp. 43–59, 2016.

[36] M. Rossi, 'Interactions of intensity glides and frequency glissandos', *Lang. Speech*, vol. 21, no. 4, pp. 384–396, 1978.

[37] P. Keating, M. Garellek, and J. Kreiman, 'Acoustic properties of different kinds of creaky voice', in *Proceedings of ICPhS*, 2015.

[38] R. C. Team, 'R: A language and environment for statistical computing', 2021.

[39] M. Gamer, J. Lemon, M. M. Gamer, A. Robinson, and W. Kendall's, 'Package "irr"', *Var. Coeff. Interrater Reliab. Agreem.*, 2012.

[40] J. 't Hart, 'Psychoacoustic backgrounds of pitch contour stylisation', *IPO Annu. Prog. Rep.*, vol. 11, pp. 11–19, 1976.

[41] B. Bigi, 'SPPAS: a tool for the phonetic segmentations of Speech', in *Proceedings of the eighth international conference on Language Resources and Evaluation*, 2012, pp. 1748–1755.