“See what I mean, huh?” Evaluating Visual Inspection of F0 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 F0 variations in nasal grunts. We discuss and test issues raised by this kind of approach exclusively based on visual inspection of the F0 tracking in Praat [2]. Results tend to show that consistency in the annotation depends on acoustic features intrinsic to the grunts such as F0 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, F0 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 paralinguistics, 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 French1. 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 Interational 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 F02. 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 F0. 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 F0. 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 F0 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 time3. 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 F0 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].

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. 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 ([œ]) 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 munhm/m.m/ and /m/ in hum /œm/). 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) b and provides high-quality recordings of spontaneous conversations in French. 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. 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. Focus was restricted to NG with one segment only, which discards monosyllabic NG of hum (/œm/) type. Our set comprises 8 hein (/œ/), 4 han (/a/) 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, 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.
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 (κ = 0.684, p < 0.05). 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 (κ = 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 (κ = 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 (κ = 0.687; κ = 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

<table>
<thead>
<tr>
<th>Participant estimations (aggregated)</th>
<th>Fall</th>
<th>Level</th>
<th>Rise</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold standard</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0.908</td>
</tr>
<tr>
<td>Level</td>
<td>3</td>
<td>40</td>
<td>11</td>
<td>.655</td>
</tr>
<tr>
<td>Rise</td>
<td>3</td>
<td>28</td>
<td>29</td>
<td>.579</td>
</tr>
</tbody>
</table>

We investigated issues raised in 2.3 to determine their contribution to identification inaccuracies. Table 2 below recaps correct and incorrect identifications of F0 variations according to phonation mode (modal vs. non-modal phonation), F0 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 F0 variations: micro-prosodic variations at the beginning and/or end of the F0 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 Fo curve is distinctly rising on NG#751, it is not clear whether it is rising, levelled, or even falling on NG#417.

![Figure 1](image)

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

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modality</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>53</td>
<td>25</td>
<td>&gt; .001</td>
</tr>
<tr>
<td>No</td>
<td>46</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td><strong>Micro-prosody</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>61</td>
<td>41</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>No</td>
<td>38</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td><strong>Duration (ms)</strong></td>
<td></td>
<td></td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Short</td>
<td>30</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Mid</td>
<td>38</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>Long</td>
<td>31</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td><strong>F0 slope</strong></td>
<td></td>
<td></td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Low</td>
<td>29</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>23</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>47</td>
<td>37</td>
<td></td>
</tr>
</tbody>
</table>

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 F0 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 F0 tracking induced by non-modal phonation.

Like other acoustic components in NG, intensity is likely to vary across NG. Since the calculation of F0 is a Praat script [2] which provides stylized representations of the F0 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 F0 tracking induced by non-modal phonation.

Both NG were labelled as levelled in Chłebowski 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 F0 slope (if any). We used the Prosogram in an attempt to improve the legibility of F0 tracking. The Prosogram was set to detect the smallest perceptible glissandi (G = 0.16/T0) [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 results11. 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 F0 as labelled in Chłebowski 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 Chłebowski 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 F0 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 2 below contrasts the F0 curves of two NG as drawn in Praat [2] (left) and Prosogram [3] (right). Both NG were levelled in Chłebowski 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 F0 slope (if any). We used the Prosogram in an attempt to improve the legibility of F0 tracking. The Prosogram was set to detect the smallest perceptible glissandi (G = 0.16/T0) [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 results11. 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 F0 as labelled in Chłebowski 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 Chłebowski 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 F0 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 F0 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 F0 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.

---

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


