The author's accepted version. Cite as: Anikin, A. (in press). A moan of pleasure should be breathy: the effect of voice quality on the meaning of human nonverbal vocalizations. *Phonetica*.

A moan of pleasure should be breathy: the effect of voice quality on the meaning of human nonverbal vocalizations

Andrey Anikin

Lund University

Andrey Anikin, Division of Cognitive Science, Department of Philosophy, Lund University,

Box 192, SE-221 00 Lund, Sweden. Tel.: (+46) (0)46-222 02 84.

Email: andrey.anikin@lucs.lu.se

Orcid: 0000-0002-1250-8261

1. Abstract

Prosodic features, such as intonation and voice intensity, have a well-documented role in communicating emotion, but less is known about the role of laryngeal voice quality in speech and particularly in nonverbal vocalizations such as laughs and moans. Potentially, however, variations in voice quality between tense and breathy may convey rich information about the speaker's physiological and affective state. In this study breathiness was manipulated in synthetic human nonverbal vocalizations by adjusting the relative strength of upper harmonics and aspiration noise. In Experiment 1 (28 prototypes x 3 manipulations = 84 sounds), otherwise identical vocalizations with tense vs. breathy voice quality were associated with higher arousal (general alertness), higher dominance, and lower valence (unpleasant states). Ratings on discrete emotions in Experiment 2 (56 x 3 = 168 sounds) confirmed that breathiness was reliably associated with positive emotions, particularly in ambiguous vocalizations (gasps and moans). Spectral centroid did not fully account for the effect of manipulation, confirming that the perceived change in voice quality was more specific than a general shift in timbral brightness. Breathiness is thus involved in communicating emotion with nonverbal vocalizations, possibly due to changes in low-level auditory salience and perceived vocal effort.

2. Introduction

Nonverbal vocalizations, such as moans or laughs, are ubiquitous in everyday interaction, express a wide range of easily recognizable emotions and attitudes (Anikin & Persson, 2017; Lima, Castro, & Scott, 2013; Sauter, Eisner, Calder, & Scott, 2010; Wood, Martin, & Niedenthal, 2017), and display significant cross-cultural similarities (Cordaro, Keltner, Tshering, Wangchuk, & Flynn, 2016). Their relatively simple acoustic structure, intuitiveness, and flexible meaning make nonverbal vocalizations attractive options for enriching speech synthesis (Campbell, 2006) and human-machine interaction (Haddad, Çakmak, Sulír, Dupont, & Dutoit, 2016). Furthermore, the acoustic structure and production context of some vocalizations, such as laughs (Ross, Owren, & Zimmermann, 2009) and screams (Högstedt, 1983), show marked similarities across species, suggesting that these sounds predate language and have deep biological roots. Insights from ethology can therefore prove instrumental for research on human nonverbal communication; in turn, learning more about human nonverbal repertoire can shed new light on acoustic communication in non-human animals.

In order to elucidate how humans communicate with nonverbal vocal cues, a crucial task is to understand the underlying acoustic code. This is an active area of research; several studies have reported detailed acoustic analyses of human nonverbal vocalizations, looking for acoustic correlates of emotion either

across all types of vocalizations (Anikin & Persson, 2017; Lima et al., 2013; Sauter et al., 2010) or in particular vocalizations or call types such as laughter (Szameitat et al., 2009; Wood et al., 2017). In many ways this research parallels the more extensive search for acoustic markers of emotion in speech (e.g., Banse & Scherer, 1996; Murray & Arnott, 1993; Pell, Paulmann, Dara, Alasseri, & Kotz, 2009). In both cases, the main focus is on easily measured prosodic characteristics such as intonation, intensity, and temporal features. More subtle acoustic features, such as the spectrum of laryngeal vocal source, the presence and spectrum of turbulent noise, the variability in the period (jitter) and amplitude (shimmer) of glottal pulses, or nonlinear vocal phenomena (e.g., pitch jumps and subharmonics), are more challenging to measure accurately and consequently less well-understood (Gobl & Ní Chasaide, 2003). At the same time, these aspects of vocal production, which give the voice a particular coloring or "voice quality", provide important information about the speaker's age, sex, and emotion in speech (Airas & Alku, 2006; Cummings & Clements, 1995; Grichkovtsova, Morel, & Lacheret, 2012; He, Lech, & Allen, 2010; Johnstone & Scherer, 1999; Laukkanen, Vilkman, Alku, & Oksanen, 1996; Murray & Arnott, 1993; Patel, Scherer, Björkner, & Sundberg, 2011; Waaramaa, Laukkanen, Airas, & Alku, 2010) and probably also in nonverbal vocalizations (Lima et al., 2013; Mittal & Yegnanarayana, 2014; Wood et al., 2017). In fact, non-speech sounds may be particularly suitable for studying the intrinsic link between voice quality and emotion because they are free both from semantic contents (Lavan, Scott, & McGettigan, 2016) and from the constraints imposed by language-specific phonemic structure or socio-cultural rules (Patel et al., 2011). For example, adding different types of nonlinear phenomena to human nonverbal vocalizations revealed that these acoustic features are perceptually salient and associated with distinct affective states: abrupt pitch jumps in screams enhance the impression of fear, episodes of unstable phonation with subharmonics or chaos in gasps and moans make the speaker sound hurt rather than pleased, and so on (Anikin, 2019b). The focus in the present paper is on the perceptual consequences of manipulating another aspect of voice quality, namely breathiness.

2.1. What makes a voice tense or breathy?

Vibrating vocal folds produce changes in air pressure that are periodic, but not perfectly sinusoidal. As a result, the glottal source of excitation contains the lowest frequency component determined by the rate at which the vocal folds oscillate (known as the fundamental frequency or f_0) and a number of other frequency components that are multiples of f_0 (harmonics). The energy of harmonics above f_0 dissipates, or rolls off - hence, "rolloff" - approximately exponentially at a rate of about 6-12 dB per octave, depending on the mode of phonation, alveolar pressure, f_0, and other factors (Stevens, 2000, Ch. 2). The buzz-like glottal pulses, as well as other excitation sources such as aspiration noise, are then modified by the resonances of the vocal tract before being perceived by the listeners, as described by the source-filter model (Fant, 1960). For a particular speaker, the exact shape of glottal pulses, and therefore the spectrum of glottal source, is primarily determined by the tension of the vocal folds and subglottal pressure, which are controlled by laryngeal and respiratory muscles (Gobl & Ní Chasaide, 2010). In addition, source spectrum can be significantly affected by nonlinear interactions between glottal source and filter, which are particularly relevant when f_0 is high enough to cross formant frequencies, as in many nonverbal vocalizations (Titze, 2008).

The spectrum of glottal pulses prior to their filtering by the vocal tract (source spectrum) has a major impact on the perceived voice quality (Gobl & Ní Chasaide, 2010; Kreiman, Gerratt, Garellek, Samlan, & Zhang, 2014). In fact, used in the narrow sense, the term "voice quality" may refer specifically to laryngeal source (Gobl & Ní Chasaide, 2003), although other aspects of vocal production are often also included, and the exact terminology varies across disciplines. One of the most widely recognized and perceptually important dimensions is breathiness, which describes variations in voice quality from tense to breathy, with modal phonation as the neutral type. A tense, or pressed, voice is characterized by complete and abrupt closure of the vocal folds, strong harmonics, and little or no aspiration noise. In

contrast, a breathy voice is characterized by loosely closed glottis, a strong f_0, weak harmonics, and audible aspiration noise caused by air leaking through the partially closed glottis (Gobl & Ní Chasaide, 2003, 2010; Stevens, 2000). The strength of upper harmonics and the amount of aspiration noise are thus the main acoustic correlates of voice quality changes along the tense-breathy continuum, which are referred to as "breathiness" in the remainder of the text.

2.2. Evidence linking breathiness and emotion

Because the prominence of upper harmonics depends on subglottal pressure and the activity of laryngeal muscles (Gobl & Ní Chasaide, 2010), breathiness has the potential to convey rich information about the speaker's physiological and affective state. To test whether listeners do utilize this information, it is necessary to manipulate, or at least to measure accurately, source spectrum. The most reliable way to estimate source spectrum is to monitor the oscillations of glottal folds directly using electroglottography (Laukkanen et al., 1996), but a more common indirect approach involves inverse filtering, which removes the contribution of the vocal tract by deconvolution of the signal with an estimated vocal tract transfer function (Drugman, Alku, Alwan, & Yegnanarayana, 2014). Several studies performed inverse filtering to demonstrate that source spectrum varied depending on the speaker's emotion in vowel sounds extracted from speech (Cummings & Clements, 1995; He et al., 2010; Johnstone & Scherer, 1999; Laukkanen et al., 1996; Patel et al., 2011), and there have been attempts to apply inverse filtering to laughs (Mittal & Yegnanarayana, 2014). Despite many methodological challenges and inconsistent definitions of voice quality in different studies (Gobl & Ní Chasaide, 2010), one of the most robust findings appears to be the association of pressed phonation with anger or other intense emotions, and of breathy phonation with more relaxed or subdued affective states. There are also reports that variations along the pressed-breathy continuum in synthesized speech are associated with perceived speaker's arousal or general alertness (Brady, 2005; Gobl & Ní Chasaide, 2003).

Researchers who do not have access to highly specialized recording facilities, or who wish to analyze large collections of recordings, are usually unable to perform electroglottography or inverse filtering and are confined to describing the observed spectrum (Gobl & Ní Chasaide, 2010). Conventional measures of the general shape of spectral envelope that have been reported in relation to emotion in speech or nonverbal vocalizations include spectral center of gravity or centroid (Lavan et al., 2016; Lima et al., 2013; Sauter et al., 2010), peak frequency with the highest amplitude within the spectrum (Scheiner, Hammerschmidt, Jürgens, & Zwirner, 2002), spectral slope or tilt (Goudbeek & Scherer, 2010; Schröder, Cowie, Douglas-Cowie, Westerdijk, & Gielen, 2001), ratios of energy above and below a certain frequency (Patel et al., 2011; Wood et al., 2017), dominant frequency bands and quantiles of spectral energy distribution (Fichtel, Hammerschmidt, & Jürgens, 2001; Hammerschmidt and Jürgens, 2007), or principal components combining several of these features. All these measures say something about the balance of low- and high-frequency energy in the spectrum, but their informativeness about source spectrum is limited by two factors. First, they do not distinguish between the contribution of source and filter. Increasing the frequency of one or more formants has the effect of raising the spectral centroid and making the voice "brighter" regardless of the glottal source (Fastl & Zwicker, 2006; Stevens, 2000). As a result, a vowel like [a] (high F1, average F2) will have noticeably stronger harmonics and sound brighter than [u] (low F1 and F2), even when both are pronounced or synthesized with the same glottal source. Second, glottal pulses are not the only source of excitation: the presence of turbulent noise can have a major effect on the shape of the resulting spectrum. For example, the spectrum flattens and its center of gravity rises as harmonics increase in strength in a tonal sound (Fig. 1, left panel), but this effect is much less noticeable in the presence of aspiration noise (right panel). In other words, harmonics in a breathy voice are weaker than might be expected based on the overall spectral slope (Gobl & Ní Chasaide, 2010). As a result, summary measures of spectral envelope may fail to capture variation in voice quality that is potentially informative to listeners.

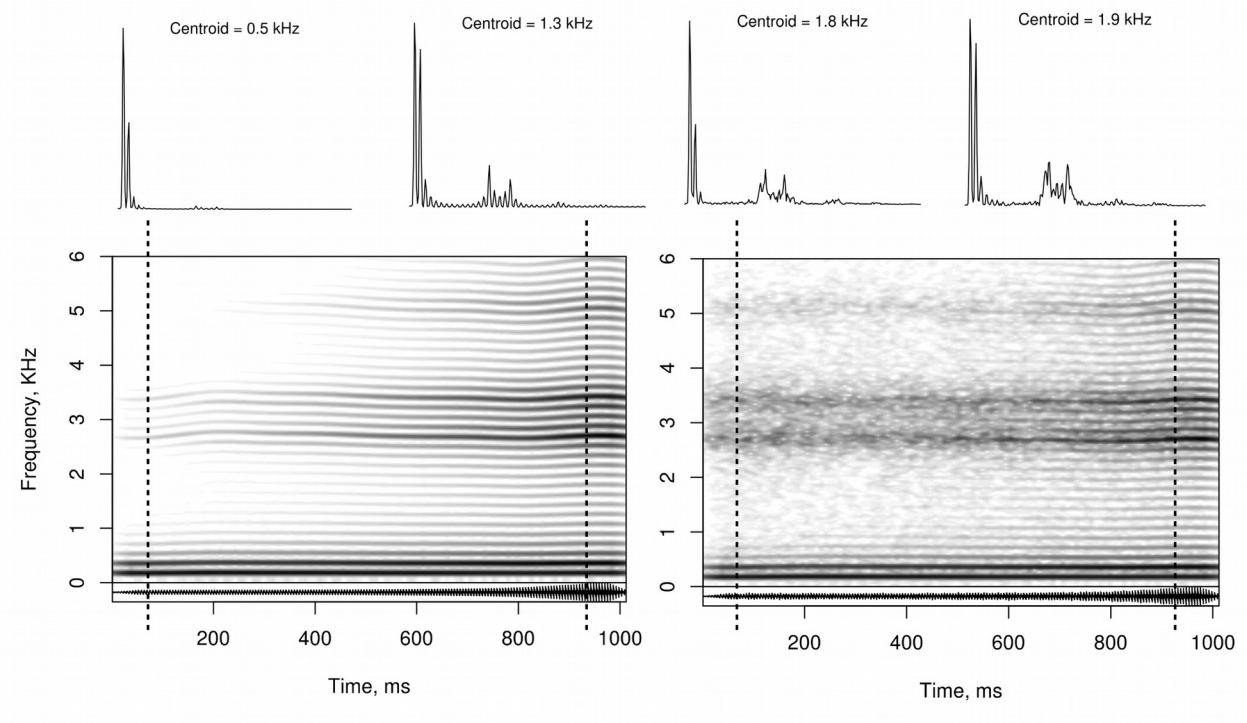

Fig. 1. Manipulation of rolloff from -16 to -6 dB/octave in a synthetic vowel that is purely harmonic (left panel) or contains turbulent noise at a constant level of -14 dB relative to the harmonic component (right panel). Spectrograms and spectral slices at ~50 ms and 950 ms. Observe that the spectral centroid is less dependent on the strength of harmonics in the presence of turbulent noise. AUDIO #1 in Supplements.

Keeping in mind these limitations of non-specific spectral measures, such as peak frequency or spectral centroid, there are several reports that listeners interpret high-frequency spectral energy as a sign of high arousal in speech (Johnstone & Scherer, 1999; Schröder et al., 2001) and in nonverbal vocalizations (Lavan et al., 2016; Lima et al., 2013). Raine, Pisanski, Simner, and Reby (2018) also report that listeners associate breathy voices with low pain intensity. In addition, there is extensive evidence from ethological literature that the spectrum contains more high-frequency energy when the animal is highly aroused (Briefer, 2012; Fichtel & Hammerschmidt, 2002) or distressed (Lingle, Wyman, Kotrba, Teichroeb, & Romanow, 2012). These studies support the generalization that stronger harmonics in source spectrum may be associated with higher general alertness (arousal) or emotion intensity not only in speech, but also in human nonverbal vocalizations and animal calls.

Besides the clearly motivated connection of a tense voice quality with arousal, various measures of high-frequency energy have been associated with unpleasant emotional experiences (negative valence) in human nonverbal vocalizations (Sauter et al., 2010; Scheiner et al., 2002) and speech (Goudbeek & Scherer, 2010; Hammerschmidt & Jurgens, 2007). On the other hand, spectral centroid was positively correlated with reward ratings of laughs in the study by Wood et al. (2017). The evidence is also mixed for animal vocalizations (Briefer, 2012), and the issue is further complicated by the fact that strongly negative affective states are usually associated with high arousal, making it difficult to know whether it is the intensity of emotion or its unpleasantness that is responsible for the observed acoustic characteristics. In one of the most methodologically rigorous studies, peak frequency was the best predictor of negative valence in squirrel monkeys, but only for relatively ambiguous vocalizations (Fichtel et al., 2001). Earlier maximum peak frequency has also been associated with positive valence, suggesting that a downward

trajectory of peak frequency may mark positive valence (Briefer, 2012; Hammerschmidt & Jurgens, 2007). The relationship between spectral envelope and perceived aversive or hedonistic nature of vocalizations may thus depend on temporal dynamics, the type of vocalization, and/or it may be mediated by arousal. Finally, several reports have linked high-frequency spectral energy to speaker's dominance or aggression in speech (Banse & Scherer, 1996; Gobl & Ní Chasaide, 2010; Hammerschmidt & Jurgens, 2007; McAleer, Todorov, & Belin, 2014) and in human nonverbal vocalizations (Sauter et al., 2010; Wood et al., 2017). However, many of these effects may be mediated by arousal or the intensity of affect in general, since high-frequency spectral energy has also been reported to correlate with the perceived intensity of several emotions (e.g., Banse & Scherer, 1996).

To summarize, the available evidence suggests that listeners interpret tense voices with strong harmonics as indicators of high alertness (arousal) or intense emotional states. Shifts from tense to breathy phonation may also be interpreted as a sign of unpleasant affective states (negative valence) or an assertive attitude (high dominance), but the evidence in this respect appears to be more mixed. In addition, there seems to be no experimental data showing that listeners attend to the strength of harmonics independently of the overall distribution of energy in the spectrum. Crucially, the most acoustically informative evidence of the role of laryngeal source – obtained with inverse filtering or direct manipulations of voice quality in synthetic stimuli – comes from studies of isolated vowels or short verbal utterances. The role of laryngeal voice quality in naturalistic nonverbal vocalizations or animal calls remains largely uncharted.

2.3. The present study

Affective speech synthesis has been slow in coming because of many technical challenges (Schröder, 2009), but it is an attractive complementary approach to inverse filtering that offers an ability to modify the laryngeal source according to stringent definitions and without acoustic confounds common in correlational studies (Gobl & Ní Chasaide, 2003). The aim of the present study was to capitalize on this underutilized methodological opportunity and to shed new light on the role of tense-breathy voice quality in emotional nonverbal vocalizations. *Soundgen*, an open-source formant synthesizer developed and validated specifically for parametric synthesis of nonverbal vocalizations (Anikin, 2019a), was used to synthesize a number of laughs, screams, and other non-speech vocalizations; each sound was created in three versions that differed only in voice quality and had identical duration, intonation and other acoustic characteristics. Most stimuli included both a harmonic component and some turbulent noise, and the manipulation had the effect of simultaneously modifying (1) the strength of higher harmonics relative to f_0 and (2) the strength of the harmonic component as a whole relative to the noise component. Qualitatively, this approximately corresponds to changing the perceived voice quality along the tense-breathy continuum, although this terminology was developed for speech and may not adequately describe acoustically “extreme” sounds like high-pitched screams.

Importantly, because the rolloff of harmonics was not the sole determinant of the observed spectrum in the presence of turbulent noise, it was possible to tease apart the effects of excitation source and overall spectral balance of energy. Statistically, this was achieved by analyzing the effect of manipulation after controlling for spectral centroid – the most common summary measure of spectral envelope and an excellent predictor of perceived timbral brightness of human voice (Fastl & Zwicker, 2006) and musical tones (Schubert, Wolfe, & Tarnopolsky, 2004). Based on the evidence reviewed above, it was hypothesized that shifts from breathy to tense phonation would produce an impression of intense or unpleasant emotional states.

To test this hypothesis, listeners rated nonverbal vocalizations with manipulated breathiness on valence, arousal, and dominance scales (Experiment 1) and then on discrete emotions (Experiment 2). Valence and arousal are among the most commonly used dimensions of emotional experience in both human (Belin,

Fillion-Bilodeau, & Gosselin, 2008; Lima et al., 2013) and animal (Briefer, 2012) research. Dominance is less well established as an emotional dimension, and the literature mentions a variety of conceptually related measures such as control, power, or potency (Goudbeek & Scherer, 2010). Categorization into discrete emotions is sometimes obtained alongside ratings on continuous dimensions in perceptual studies (e.g., Lima et al., 2013); in this study this was a natural choice given that (a) sounds in the original corpus were obtained from contexts related to several well-defined affective states, and (b) two previous studies established which emotions are most commonly perceived by people who hear these sounds (Anikin Bååth, & Persson, 2018; Anikin, 2019a).

3. Experiment 1

3.1. Methods

Stimuli. The stimuli were synthetic versions of human nonverbal vocalizations from the corpus collected by Anikin & Persson (2017). The original vocalizations were mostly non-staged, spontaneous emotional bursts that had been captured on video in real-life situations and then uploaded to social media. These sounds included little or no phonemic structure and were associated with a powerful emotional experience such as incurring a physical injury (pain), being the victim of a scare prank (fear), witnessing a funny accident (amusement), and so on for a total of nine emotions, whose recognition was tested in a validation study (Anikin & Persson, 2017). The sounds have also been sorted into call types based on linguistic labeling by English-, Swedish-, and Russian-speaking participants as well as acoustic analysis (Anikin et al., 2018).

For the present study, 28 prototypes were chosen from the larger corpus to represent the following seven call types: cry, gasp, grunt, laugh, moan, roar, and scream. These particular sounds were chosen primarily based on the relatively high authenticity ratings of their synthetic versions in the *soundgen* validation study (Anikin, 2019a), and together they represent a broad range of vocalizations from the human nonverbal repertoire. The 28 prototype vocalizations were all from different individuals (17 women and 11 men) and varied between 0.22 and 2.7 s in duration (Table 1).

Each of these 28 sounds was manually analyzed, and then a similar synthetic version was created with original / strengthened / weakened harmonics in source spectrum using *soundgen* 1.1.2 (Anikin, 2019a), for a total of 28 x 3 = 84 stimuli. *Soundgen* is an implementation of the source-filter model written in the *R* language that creates a separate sine wave for each harmonic, adds a noise component, and filters this excitation source with a simulated transfer function. All control parameters are set manually by specifying a few anchors that are interpolated to produce smooth curves for an entire syllable or bout; for example, the intonation contour is given by a few pitch anchors at different time points. The quality of this parametric synthesis had previously been validated, in the sense that synthetic vocalizations were shown to be similar to the original recordings in terms of their ratings on valence and arousal as well as perceived emotion, and in most cases the authenticity of synthetic stimuli was on a par with the originals (Anikin, 2019a). At the same time, the synthesized sounds in this study were not exact replicas of the 28 prototypes, and their fully synthetic nature made it possible to manipulate any desired acoustic characteristic, without any limitations associated with audio editing or resynthesis.

The harmonic structure of spectral source is controlled by several parameters in *soundgen*, but its overall slope can be manipulated with a single parameter called “rolloff”, which corresponds to the rate of exponential decay of the energy in harmonics above f_0, in dB/octave (Fig. 2). Depending on the sound, rolloff was changed by ±4-10 dB/octave, aiming to make the magnitude of manipulation comparable across all stimuli in terms of the perceptual salience of voice quality changes. All sounds were normalized for peak amplitude, so changes in rolloff did not entail major changes in the overall sound pressure level, although subjective loudness may have changed. Increasing or decreasing the strength of harmonics

changed the spectral centroid on average by +316 Hz and -183 Hz, respectively. However, 20 out of 28 prototype sounds included some turbulent noise, whose spectrum was not affected by the rolloff setting. Because the amplitude of the noise component was tied to the amplitude of the first harmonic, changing the amplitude of harmonics relative to f_0 also affected the relative amplitudes of the voiced and unvoiced (noise) components. In particular, noise partly replaced higher harmonics in manipulated sounds with steeper (more negative) rolloff, making the voice sound breathy, while in manipulated sounds with shallower (less negative) rolloff harmonics tended to replace the noise, creating the impression of pressed phonation (Fig. 2). High-pitched screams and roars were synthesized without a noise component because in real life these vocalizations are normally too loud for aspiration noise to be audible. In these sounds, the manipulation affected only the strength of harmonics relative to f_0.

Table 1. Acoustic characteristics of the synthetic stimuli in Experiment 1.

Call type	Number of stimuli		Acoustic characteristics: mean [range]					
	Total (with noise)	Female / male	Rolloff manipulation (dB/oct)	Duration (s)	Median voiced syllable (s)	Median f_0 (Hz)	Median HNR* (dB)	Median spectral centroid (kHz)
Cry	4 (2)	3/1	±4 [4, 4]	2.2 [1.6, 2.6]	0.3 [0.1, 0.5]	468 [255, 790]	15.5 [13.1, 19]	1.5 [0.8, 2.1]
Gasp	4 (4)	3/1	±7.5 [6, 8]	1.3 [1.2, 1.5]	1.0 [0.6, 1.2]	362 [255, 498]	11.2 [5.2, 16.6]	1.9 [1.0, 2.8]
Grunt	4 (4)	2/2	±5.5 [4, 10]	0.4 [0.2, 0.5]	0.3 [0.1, 0.4]	289 [190, 378]	4.1 [0.7, 9.5]	1.3 [0.8, 1.6]
Laugh	4 (4)	1/3	±5 [4, 8]	1.8 [1.4, 2.6]	0.1 [0.1, 0.2]	459 [330, 617]	7.9 [3.2, 12.7]	2.1 [1.4, 2.6]
Moan	4 (4)	2/2	±5.5 [4, 10]	1.3 [0.7, 1.8]	0.5 [0.4, 0.6]	293 [161, 490]	8.6 [1.7, 14.8]	1.5 [1.0, 2.2]
Roar	4 (2)	2/2	±4 [4, 4]	0.8 [0.5, 1.1]	0.8 [0.5, 1.0]	454 [288, 701]	1.2 [-3.0, 5.2]	1.8 [0.9, 2.9]
Scream	4 (0)	4/0	±7 [4, 10]	0.9 [0.6, 1.2]	0.8 [0.4, 1.2]	1800 [1420, 2102]	9.1 [3.4, 16.7]	2.6 [1.9, 4]

*HNR = harmonics-to-noise ratio, measured based on autocorrelation. HNR and spectral centroid were measured in the synthesized audio files using *soundgen*, and the rest of acoustic characteristics were derived directly from the settings used for synthesizing the stimuli.

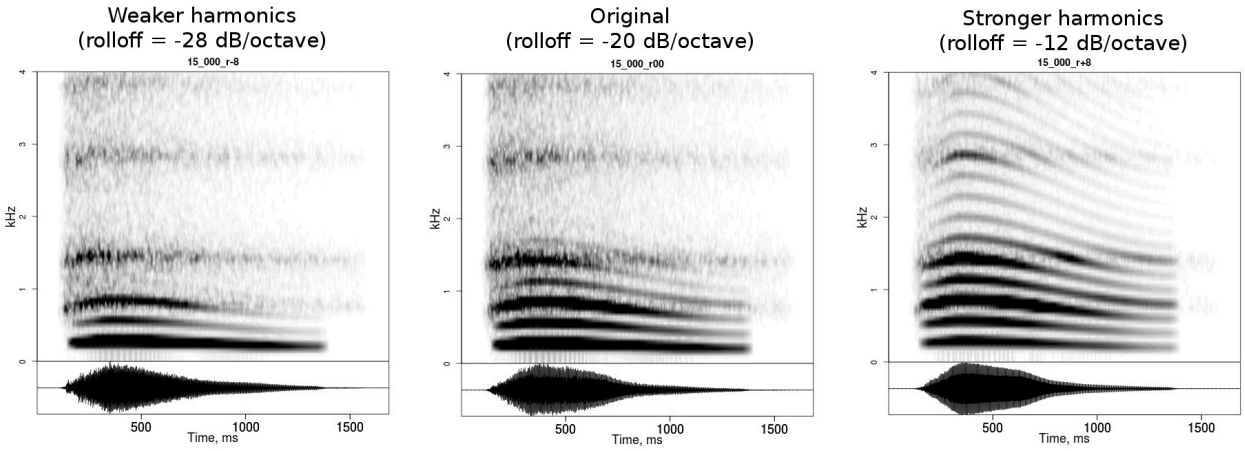

Fig. 2. Spectrograms illustrating the manipulation of rolloff in synthetic gasp #15 corresponding to changes in voice quality along the breathy-tense continuum. AUDIO #2 in Supplements.

Procedure. The rating experiment was performed online. To avoid presenting very similar sounds repeatedly to the same participants and to optimize the effectiveness of data collection, the rolloff manipulation was tested partly as a separate study, and partly in conjunction with another experiment on nonlinear vocal phenomena, which employed different manipulations, but the same prototype sounds and

design (Anikin, 2019b). The stimuli were divided in such a manner that each batch contained at most two manipulated versions of the same prototype stimulus, and each batch was rated by a different sample of participants. As a result, each participant heard only a subsample of experimental stimuli and rated each stimulus on three scales (valence, arousal, and dominance), which were explained to participants as follows:

> ***Valence*** *is high if the experience is pleasant, so the speaker is happy, pleased, relieved, etc. Valence is low if the experience is unpleasant, so the speaker is sad, afraid, in pain, etc.*
>
> ***Arousal*** *is high if the person is very energetic, alert, wide-awake. Arousal is low if the person is sleepy, relaxed, calm.*
>
> ***Dominance*** *is high if the speaker sounds assertive, self-confident, superior, perhaps aggressively so. Dominance is low if the person sounds submissive, uncertain, perhaps as someone who seeks reassurance or needs a hug.*

The rating was performed on a continuous horizontal visual analog scale. To minimize the correlation between valence, arousal, and dominance ratings, the experiment was divided into three blocks, in random order. For example, one participant might first rate the stimuli on valence, then on arousal, and finally on dominance; another participant would begin with dominance, etc. The order of sounds within each block was also randomized for each participant. Prior to each block the upcoming scale was illustrated with two contrasting examples: non-synthetic recordings of a person crying (low valence) or laughing (high valence), sighing (low arousal) or screaming (high arousal), and whimpering (low dominance) or roaring (high dominance; adapted from Puts, Gaulin, & Verdolini, 2006).

Participants. Data quality was ensured by carefully checking all submissions and reimbursing only participants with minimum 40 trials and responses that were not obviously faked (extremely fast and stereotypical). Out of 136 submissions that passed this minimal quality control, the responses of four participants were clear outliers in terms of their low correlation with the global median ratings across all three scales ($r < 0.2$), presumably indicating that they had not attended to the task. The responses of these four participants (2.2% of all data) were therefore excluded from the analysis. In addition, individual trials with a response time under 500 ms presumably represented accidental clicks and were removed from the dataset (<0.5% of all data). The final sample consisted of 132 participants, of whom 18 were unpaid volunteers contacted via online advertisements and 114 were recruited from https://www.prolific.ac. Each sound was rated on average 43 times (range 40 to 47) on each scale. No demographic characteristics were collected; according to the statistics on https://www.prolific.ac/demographics, over 80% of participants on this platform are native English speakers, and about 75% are between 20 and 40 years of age.

Statistical analysis. The response variable was the rating of a sound on a continuous scale (valence, arousal, or dominance) provided in a single trial by a particular participant. These ratings were modeled using the beta distribution in Bayesian mixed models with random intercepts per participant, per stimulus, and per prototype (shared by all sounds that were modifications of the same original vocalization). Rolloff manipulations were treated as a continuous variable with three values: 0 for more negative rolloff (breathy voice), 0.5 for original, and 1 for less negative (tense voice). To compare the relative contributions of the relative strength of harmonics and the overall spectral shape, rolloff manipulation and the logarithm of spectral centroid were entered simultaneously in multiple regression models. Mixed models were fit in Stan computational framework (http://mc-stan.org/) accessed with R package *brms* (Bürkner, 2017). To improve convergence and guard against overfitting, regularizing priors were used for all regression coefficients. The effects were summarized as the median of posterior distribution and 95%

credible interval. The stimuli, R code for their generation, experimental datasets, and scripts for statistical analysis can be downloaded from http://cogsci.se/publications.html.

3.2. Results

Inter-rater reliability was moderate for valence (ICC = .48, 95% CI [.41, .57]) and arousal (ICC = .53 [.46, .61]). These levels agreement of agreement among raters were similar to those in previous studies of both real (Anikin, 2019a) and synthetic (Anikin, 2019a, 2019b) versions of similar sounds. In contrast, dominance was rated less consistently by different participants: ICC = .22 [.17, .29]. The results for the dominance scale should therefore be treated with caution.

Increasing the rolloff parameter in *soundgen* (making it less negative) made the voice tenser (strong harmonics, less aspiration noise), while decreasing the rolloff made it more breathy (weak harmonics, more aspiration noise). The valence ratings were 3.4% (95% CI [0.4, 6.3]) lower in otherwise identical sounds with tense vs. breathy phonation. This small negative effect remained essentially unchanged after controlling for spectral centroid (-3.0% [-6.6, 0.5]), while the independent effect of spectral centroid after controlling for rolloff was highly uncertain: -2.6% [-17.8, 11.7] over the observed range of 0.8 to 4 kHz.

As predicted, making the voice tenser increased arousal ratings by 6.9% [3.6, 10.1]. The effect of rolloff on arousal became slightly weaker after controlling for spectral centroid (5.3% [1.0, 9.3]), whereas the effect of spectral centroid itself on perceived arousal was statistically uncertain after controlling for rolloff: 11.3% [-6.0, 29.0]. In other words, the effect of manipulations on arousal may be partly mediated by the general shape of the spectral envelope, but the strength of harmonics as such clearly makes an independent contribution. Dominance ratings were slightly higher for tense vs. breathy voice quality (2.8% [0.4, 5.1]). It is not clear to what extent this change may be mediated by spectral centroid: the effect of rolloff manipulation controlling for spectral centroid is predicted to be 2.1% [-0.8, 5.1], and that of spectral centroid 4.9% [-6.6, 16.2], which is too uncertain to draw any firm conclusions.

Although the three resynthesized versions of each prototype vocalization – with original / increased / decreased rolloff – differed only in the synthetic equivalent of laryngeal voice quality, there was a large variation among the 28 prototype sounds in terms of their duration, temporal structure, average pitch, and other prosodic characteristics. To test whether the effect of rolloff manipulation depended on some of these acoustic characteristics, likelihood ratio tests in non-Bayesian mixed models were employed to test the significance of interaction terms between rolloff manipulation and the following acoustic characteristics: duration, median f_0, median and maximum length of voiced syllables (taken directly from the control parameters supplied to the synthesizer), and the original speaker's sex. None of these interaction terms were significant after Bonferroni correction in models predicting valence, arousal, or dominance ratings, suggesting that the effects of rolloff manipulations were broadly consistent across acoustically diverse stimuli. On the other hand, the number of stimuli and effect sizes were not large enough to reveal relatively subtle interactions; the resolution of this analysis could be improved by creating and testing a larger number of stimuli.

3.3. Discussion

The aim of Experiment 1 was to perform an initial exploration of the perceptual consequences of manipulating laryngeal voice quality along the tense-breathy continuum in several types of human nonverbal vocalizations. The tested manipulations affected the ratings of synthetic vocalizations in a manner broadly consistent with theoretical predictions. Making the voice tenser enhanced the perceived level of speaker's general alertness or arousal; it also and made the vocalizations slightly more aversive and enhanced the speaker's perceived dominance, although the latter effect was uncertain. A commonly reported measure of timbral brightness or overall balance of low- and high-frequency energy in the spectrum – spectral centroid – appeared to contribute to the observed effects of voice quality, but did not

fully account for them. From a listener's perspective, the experimental manipulations of voice quality thus appeared to be both salient and more specific than a general shift of energy towards higher frequencies.

4. Experiment 2

Some perceptual effects of source spectrum may be specific to particular types of vocalizations, but these differences could not be estimated in Experiment 1 with only four prototype sounds per acoustic class. Accordingly, in the follow-up study the number of sounds of the same type, such as laugh or scream, was increased. To keep the overall number of stimuli manageable, only four acoustic classes were investigated: one that was predominantly positive in valence (laughs), one negative (screams or high-pitched roars), and two more ambivalent (gasps and moans). Laughs, screams, and moans are among the four most universally recognized human nonverbal vocalizations (Anikin et al., 2018). Gasps have not been extensively studied, but their ingressive nature makes them very distinct acoustically.

Only a few studies have looked at voice quality in specific vocalization types. Lavan et al. (2016) showed that ratings of breathiness provided by trained phoneticians were associated with higher valence ratings in all laughs and with higher arousal ratings in spontaneous, but not in volitional laughs. In contrast, laughs with a higher spectral centroid were rated as slightly higher on both dominance and reward scales in the study by Wood et al. (2017). Apart from laughs, some measures of voice quality have been reported in human screams. Mostly these relate to the presence of nonlinear phenomena (Arnal, Flinker, Kleinschmidt, Giraud, & Poeppel, 2015; Raine et al., 2018), but Hansen et al. (2017) also report more high-frequency energy and flatter spectral slopes in screams compared to neutral speech. This evidence is not sufficiently detailed to make specific predictions regarding the perceptual effects of breathiness in different call types. The most parsimonious assumption is that the same general acoustic code applies to both speech and all nonverbal vocalizations. On the other hand, there are significant differences between such vocalizations as gasps and screams in their production mechanism (e.g., ingressive or egressive) and general acoustic characteristics (the degree of voicing, pitch, syllable structure), which may affect the perceptual consequences of changes in voice quality.

Instead of the dimensions of valence, arousal, and dominance, participants in Experiment 2 rated the intensity of discrete emotions, aiming to provide a complementary outcome measure that was arguably more intuitive for participants. Three emotional labels were provided for each call type: *Pleased* / *Hurt* / *Surprised* for gasps, *Amused* / *Evil (jeering)* / *Polite* for laughs (adapted from Szameitat et al., 2009 and Wood et al., 2017), *Pleased* / *Hurt* / *Effortful* for moans, and *Pleased* / *Afraid* / *Aggressive* for screams. These emotions correspond to the most common classifications of acoustically similar vocalizations in earlier studies (Anikin et al., 2018; Anikin, 2019a). They may not cover every possible interpretation, but the objective in Experiment 2 was primarily to contrast responses to the same stimulus after a particular acoustic manipulation. The measure of interest was thus the difference in the weights of particular emotions caused by strengthening or weakening harmonics in the source spectrum of the same prototype sound.

4.1. Methods

Stimuli. The experimental stimuli were 168 modifications of 56 prototype vocalizations, selected from the same source as in Experiment 1 (Anikin & Persson, 2017) and re-synthesized with original, weakened or strengthened harmonics with *soundgen* 1.2.0 (Anikin, 2019a). The prototypes included 10 gasps, 14 laughs, 15 moans or grunts, and 17 screams or high-pitched roars with duration ranging from 0.4 to 3.4 s (Table 2). Out of 56 prototypes, 24 were by men, 29 by women, and 3 by preadolescence children; 15 were also used in Experiment 1. The range of rolloff manipulation was ±4-15 dB/octave.

Procedure. The experiment was performed in a web browser and began with training, in which participants rated eight human nonverbal vocalizations in order to become familiar with the rating tool

and the nature of stimuli. Following training, participants rated 81 or 84 synthetic vocalizations each – a subsample of a larger corpus prepared for this study and a companion study on nonlinear phenomena (Anikin, 2019b). The rating tool was a triadic scale designed for rating the proportional weight of three response categories with a single click. Three labels were placed in the corners of an equilateral triangle (Fig. 3), and the weights of these three categories were related to the position of the marker within the triangle via a nonlinear transformation under the constraint that the three weights should sum to 100%. Participants were instructed to set these weights, which were displayed as bars under the triangle.

Table 2. Acoustic characteristics of the synthetic stimuli in Experiment 2.

Call type	Number of stimuli		Acoustic characteristics: mean [range]					
	Total (with noise)	Female / male / child	Rolloff manipulation (dB/oct)	Duration (s)	Median voiced syllable (s)	Median f_0 (Hz)	Median HNR* (dB)	Median spectral centroid (kHz)
Gasp	10 (10)	6/4/0	±7.5 [4, 15]	0.8 [0.5, 1.4]	0.6 [0.2, 1.2]	345 [111, 634]	8.4 [0.4, 16.8]	1.8 [1.0, 2.7]
Laugh	14 (14)	6/6/2	±4.4 [4, 6]	1.7 [1.1, 3.1]	0.2 [0.1, 0.7]	517 [167, 846]	10.2 [1.5, 19.1]	1.8 [1.2, 2.9]
Moan	15 (15)	7/8/0	±5.9 [4, 8]	0.8 [0.4, 2.0]	0.6 [0.2, 1.8]	288 [135, 423]	13.3 [3.7, 20.1]	1.3 [0.5, 2.8]
Scream / roar	17 (4)	10/6/1	±6.1 [4, 8]	0.9 [0.3, 1.9]	0.7 [0.2, 1.6]	1205 [295, 3063]	18.4 [11.3, 22.4]	2.4 [0.9, 4.3]

*HNR = harmonics-to-noise ratio, measured based on autocorrelation. HNR and spectral centroid were measured in the synthesized audio files using *soundgen*, and the rest of acoustic characteristics were derived directly from the settings used for synthesizing the stimuli.

Participants. Participants were recruited via https://www.prolific.ac. As in Experiment 1, only submissions that were complete and not obviously faked were accepted and reimbursed. Beyond this initial quality check, no participants were excluded from the analysis because there were no clear outliers among participants in terms of how well their responses agreed with the typical response pattern (see Results). Individual trials with an unusually rapid response time suggestive of a technical problem or accidental clicking were excluded; they represented ~1.4% of data. The response time threshold was higher than in Experiment 1 (2000 vs. 500 ms) because of a different design involving at least two clicks per trial in Experiment 2. The responses of 151 participants provided on average 49 (range 46 to 51) ratings per sound.

Statistical analysis. The triadic rating scale returns a vector of three weights that sum to one – a so-called simplex – which was modeled with a redundantly-parameterized normal distribution that forces the means for the three categories to sum to one with a softmax transform (Gelman, Bois, & Jiang, 1996):

$$\mu_i = exp(\varphi_i) / (exp(\varphi_1) + exp(\varphi_2) + exp(\varphi_3))$$

for i in $\{1, 2, 3\}$, where μ_i is the mean of the normal distribution for the weight of each category and φ_i is normally distributed with a mean of zero. The φ parameters are not uniquely identifiable, but they do provide valid inference on μ. A corresponding Bayesian model was defined in Stan and extended to include main effects (condition, spectral centroid) and two random intercepts (per participant and per prototype sound) with regularizing priors. A separate model was fit for each of four call types.

4.2. Results

The emotion category with the greatest weight was identified for each of 151 participants and for the entire sample. The proportion of sounds for which a given participant assigned the greatest weight to the same category as the majority was then used as a measure of inter-rater agreement. This proportion was approximately normally distributed with no clear outliers and ranged from .26 to .71 (mean = .45),

suggesting that most participants understood the experimental procedure and were reasonably consistent in their responses.

The effect of voice quality manipulations can be visualized as the change in the average coordinates within the triangle per prototype sound, as shown in Figure 3. Since participants were instructed to set the relative weights of three emotions, the main analysis focused on how acoustic manipulations affected these weights, not coordinates as such. Similarly to Experiment 1, the outcome was modeled before and after controlling for spectral centroid (Table 3).

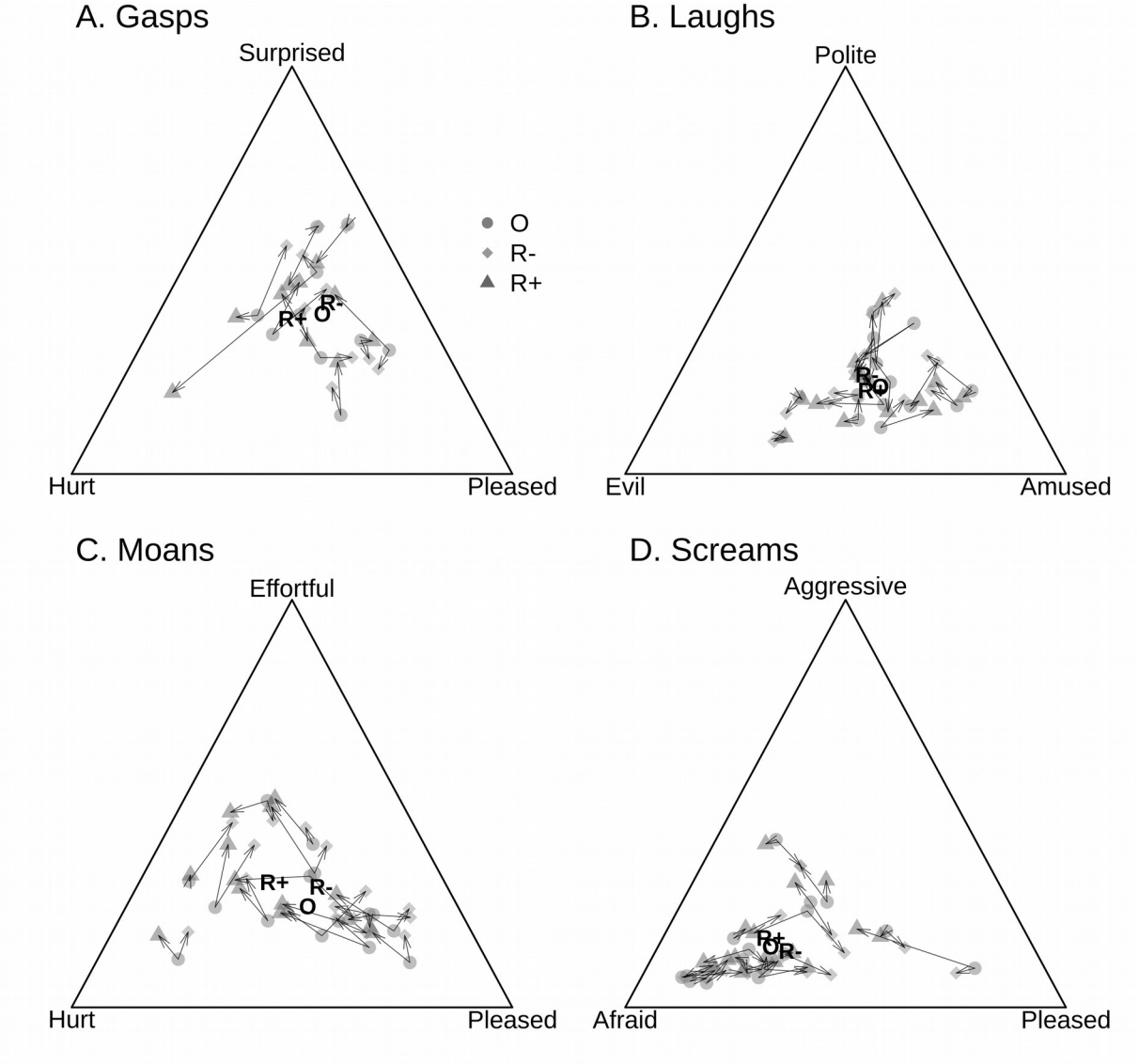

Fig. 3. Mean coordinates representing the perceived emotion of different call types and acoustic manipulations. Labels in bold show the average for all sounds, while individual symbols and arrows show the effect of manipulations for each prototype sound. O = original, R- = less energy in harmonics (breathy), R+ = more energy in harmonics (tense).

Table 3. Contrasts between the effect of different acoustic manipulations on the weight of emotion categories for each call type: median of posterior distribution (%) and 95% CI.

Model	Contrast	Gasps			Laughs		
		Hurt	**Surprised**	**Pleased**	**Evil**	**Polite**	**Amused**
Without SC*	R+ vs. R-	**13.4 [9.1, 17.7]**	-3.5 [-9.7, 2.7]	**-9.9 [-14.4, -5.5]**	0.7 [-4.0, 5.6]	**-5.1 [-8.5, -1.7]**	**4.4 [1.1, 7.9]**
With SC	R+ vs. R-	**13.7 [9.4, 17.9]**	-3.7 [-9.9, 2.4]	**-9.9 [-14.3, -5.5]**	**-8.1 [-14.1, -1.2]**	3.4 [-2.0, 8.5]	4.6 [-0.8, 9.8]
	SC***	**13.9 [1.4, 25.5]**	-11.6 [-29.6, 6.3]	-2.0 [-15.5, 11.0]	**26.7 [10.3, 39.0]**	**-25.4 [-37.5, -12.8]**	-1.0 [-13.0, 12.0]

Model	Contrast	Moans			Screams		
		Hurt	**Effortful**	**Pleased**	**Afraid**	**Aggressive**	**Pleased**
Without SC	R+ vs. R-	**12.8 [9.5, 16.1]**	-0.3 [-3.6, 3.0]	**-12.5 [-17.2, -7.8]**	**5.1 [2.4, 7.9]**	2.4 [-0.3, 5.1]	**-7.6 [-11.4, -3.7]**
With SC	R+ vs. R-	**12.7 [8.3, 17.3]**	0.4 [-4.0, 4.8]	**-13.1 [-19.2, -7.0]**	-3.6 [-10.0, 2.9]	1.9 [-3.1, 7.3]	1.7 [-6.8, 9.9]
	SC	0.6 [-13.8, 14.9]	-3.5 [-17.8, 10.8]	2.8 [-17.1, 22.4]	**4.6 [1.5, 7.6]**	0.3 [-2.2, 2.5]	**-4.9 [-8.6, -1.0]**

* SC = spectral centroid, R- = weaker harmonics, R+ = stronger harmonics.
** Cells in **bold** contain 95% CIs that exclude or nearly exclude zero. This is merely a visualization aid, not significance testing.
*** Effect over the observed range of 530 to 4260 Hz, controlling for rolloff manipulation.

Making the voice quality tenser in gasps (less negative rolloff, condition “R+” vs. “R-”) made the speaker sound more *Hurt* (+13.4%, 95% CI [9.1, 17.7]) and less *Pleased* (-9.9% [-14.4, -5.5]). For moans, a tense voice was likewise associated with being *Hurt* (+12.8% [9.5, 16.1]) rather than *Pleased* (-12.5% [-17.2, -7.8]). Interestingly, voice quality in moans had no effect on perceived effort (-0.3% [-3.6, 3.0]). Accounting for the spectral centroid did not substantively alter the observed effects on the perceived emotion in either gasps or moans, and spectral centroid had no independent effect on the weight of any emotion after controlling for rolloff, except for making the speaker sound 13.9% [1.4, 25.5] more *Hurt* in gasps (Table 3).

Tenser voice quality in laughs enhanced the impression that the speaker was genuinely *Amused* (+4.4% [1.1, 7.9]) rather than merely *Polite* (-5.1% [-8.5, -1.7]). However, this effect was relatively small and disappeared after controlling for spectral centroid (Table 3). Interestingly, a higher spectral centroid was strongly associated with sounding *Evil* (+26.7% [10.3, 39.0]) rather than *Polite* (-25.4% [-37.5, -12.8]), presumably because strongly aspirated giggles with little voicing are considered impolite or malicious.

Screams with stronger harmonics were associated with being *Afraid* (+5.1% [2.4, 7.9]) rather than *Pleased* (-7.6% [-11.4, -3.7]). Since screams were synthesized with little or no noise component, rolloff of harmonics was the sole determinant of the spectral shape, and therefore the effects of rolloff manipulation and spectral centroid in screams could not be separated in multiple regression, leading to uncertain results when both were entered simultaneously (Table 3).

4.3. Discussion

Experiment 2 aimed to verify and nuance the findings from Experiment 1 by increasing the number of stimuli and using discrete emotions instead of the dimensions of valence, arousal, and dominance. The results confirmed that a shift in voice quality from breathy to tense was associated with more aversive emotions, but with interesting differences between call types.

The clearest picture emerged for the most ambiguous of the tested call types – gasps and moans. Tense voices with strong harmonics and little aspiration noise were perceived as considerably more aversive than breathy voices with weak harmonics. In effect, manipulating only the voice quality, without changing the intonation or any other acoustic characteristic, was enough to turn a gasp or moan of pleasure into pain. Since this effect of voice quality persisted after controlling for spectral centroid,

listeners appeared to attend specifically to breathiness, and not simply to the amount of high-frequency energy in the spectrum.

In screams, the manipulation did not create a breathy voice as such, since no aspiration noise was added. Furthermore, screams were predominantly interpreted as an expression of fear, and this limited variation in responses partly masked the effect of manipulations. Nevertheless, strengthening upper harmonics relative to the fundamental frequency noticeably shifted the interpretation of screams from pleasure to fear, as predicted.

As for laughs, making the voice quality tense within one particular sound - without changing any prosodic characteristics - made the speaker sound slightly more amused rather than merely polite. When comparing different laughs, on the other hand, a higher spectral centroid (a measure of timbral brightness) was strongly associated with sounding malicious. Laughs come in a great variety of acoustic forms and contain a large amount of aspiration noise, complicating the relationship between summary measures of spectral shape, such as spectral centroid, and glottal source spectrum. This is possibly the reason for seemingly contradictory reports in previous correlational studies, which found that breathy laughs scored higher on both arousal and valence (Lavan et al., 2016), but also that laughs with a higher spectral centroid were rated as more rewarding (Wood et al., 2017). Taking the present results at face value, tensing the voice quality in laughs does not make them more negative - unlike other analyzed vocalizations - but it appears to enhance the perception of genuine amusement. Because high-intensity emotional expressions are often perceived as more authentic (Anikin & Lima, 2018; Lavan et al., 2016), this is in line with the association of tense voice quality with higher perceived arousal in Experiment 1. Considering that the manipulation effect for laughs was relatively weak and uncertain, however, it should be treated with caution. In addition, laughs present a formidable challenge to manual parametric synthesis because of their complex and dynamic spectrotemporal characteristics, and the perceived authenticity of synthetic laughs was lower than for other vocalizations in a previous validation study (Anikin, 2019a). It is therefore possible that the synthesis of laughs was not successful enough to analyze the effects of voice quality. The relatively short and noisy syllables that laughs consist of also make changes in laryngeal source harder to detect because breathiness is more salient in longer sustained syllables.

5. General Discussion

Two experiments were carried out to test the perceptual consequences of modifying the laryngeal voice quality in otherwise identical synthetic nonverbal vocalizations. The manipulation consisted in changing the rolloff of harmonics in source spectrum, which had two effects: it made upper harmonics more or less pronounced relative to the fundamental frequency and simultaneously changed the amplitude of aspiration noise relative to the voiced component. Perceptually, this manipulation approximately corresponds to shifting the voice quality along the tense-breathy continuum. As predicted, breathiness was associated with less intense and more pleasant emotions, particularly for those vocalizations that can be either positive or negative in valence. Some implications of these findings are discussed below.

In line with previous reports based on conventional measures of spectral shape (Briefer, 2012; Lima et al., 2013; Lingle et al., 2012; Schröder et al., 2001), stronger harmonics were associated with higher arousal ratings. This is not surprising, since physiologically this change in voice quality is caused by greater pharyngeal constriction (Briefer, 2012) and higher subglottal pressure (Stevens, 2000), both of which are associated with an active, aroused state. In addition, increasing the strength of harmonics had a negative effect on valence ratings in Experiment 1. This finding was strongly confirmed in Experiment 2, particularly for intrinsically ambiguous vocalizations such as gasps and moans. For example, a breathy gasp or moan with a strong f_0 and weak harmonics (breathy voice) was likely to be interpreted as a sign of pleasant surprise, whereas an otherwise identical sound with stronger harmonics and less aspiration noise

(tense voice) was interpreted as an expression of pain. Voice quality had a smaller effect on vocalizations that were interpreted as predominantly hedonistic (laughs) or predominantly aversive (screams). This tallies with the earlier observation that the distribution of energy in the spectrum correlates with valence only in the more ambiguous primate vocalizations (Fichtel et al., 2001). More generally, it calls for caution when generalizing acoustic observations across call types, since perceptual effects of acoustic features may be vocalization-specific (Linhart et al., 2015).

In addition to testing the relevance of laryngeal voice quality to the communication of emotion specifically in nonverbal vocalizations, an important novelty of the chosen experimental approach lies in the ability to distinguish between the contribution of "glottal source" (or rather, its synthetic counterpart) and the general balance of low- and high-frequency energy in the spectrum. A non-specific measure of timbral brightness (spectral centroid) appeared to contribute to some of the observed effects, but did not fully account for them. In line with speech research (Garellek et al., 2016), this suggests that listeners are sensitive to the relative strength of individual harmonics in nonverbal vocalizations when they decide what emotion the speaker is experiencing. As a result, relatively subtle modifications of voice quality – changes that would be nearly invisible to a conventional acoustic analysis – are sufficient to cause a major effect on perceived emotion, sometimes "flipping" the valence of a vocalization.

The reported manipulation of voice quality was somewhat simplified: in reality breathiness also includes other acoustic characteristics that were not modeled in this study, such as increased formant bandwidth and additional zero-pole pairs in the source spectrum due to coupling with supralaryngeal resonators (Gobl & Ní Chasaide, 2010). In addition, the manipulation affected all harmonics in source spectrum, whereas speech research suggests that humans are sensitive to the difference between specific harmonics, the overall spectral slope, and high-frequency noise excitation (Garellek, Samlan, Gerratt, & Kreiman, 2016; Kreiman, Gerratt, & Antoñanzas-Barroso, 2007; Kreiman et al., 2014). On the other hand, the narrow range of f_0 and moderate vocal effort typical of speech are different from the conditions of voice production in nonverbal vocalizations such as screams or roars, making it problematic to apply linguistic terminology or speech-specific measures of source spectrum. Furthermore, the role of laryngeal voice quality in nonverbal vocalizations in largely terra incognita, and the results reported here are only a preliminary investigation that will need to be confirmed and elaborated in future studies.

An interesting question for follow-up research is whether a tense voice with strong harmonics is intrinsically associated with intense and unpleasant affective states, or whether this effect is due to changes in perceived loudness and pitch. Although pitch is usually considered to be the perceptual equivalent of the fundamental frequency, it may in fact depend on other spectral characteristics (McPherson & McDermott, 2018), including voice quality. In particular, a tense voice literally sounds higher than a breathy voice (Kuang, Guo, & Liberman, 2016). Likewise, sounds with more high-frequency energy are subjectively experienced as louder because human hearing is more sensitive to high frequencies (Fastl & Zwicker, 2006), and loudness can enhance the impression of high activation states (Yanushevskaya, Gobl, & Ní Chasaide, 2013). It is therefore possible that strengthening the harmonics rather mechanistically makes the stimulus appear louder, brighter, and more high-pitched, thus enhancing its low-level perceptual salience to the auditory system.

In addition, listeners may interpret vocalizations with strong harmonics as originally produced with high vocal effort - loudly and close to the upper limit of the speaker's pitch range (Kuang, Guo, & Liberman, 2016; but see Bishop & Keeting, 2012). Because louder vocalizations tend to have stronger harmonics (Traunmüller & Eriksson, 2000) and therefore a higher peak frequency (Stout, 1938; Gustison & Townsend, 2015), some information about the loudness of the original utterance is still available to listeners even from recordings with normalized amplitude. This estimate of the original speaker's vocal effort may then be taken into account when interpreting the vocalization. For example, listeners may

expect that a moan of pleasure will be produced in a quieter and breathier voice than a moan of pain, that a person in pain will scream with greater vocal effort and thus a tenser voice than a person who is delighted, and so on. Indeed, hedonistic vocalizations, such as moans of pleasure, are more likely to occur in intimate, close-range contexts, whereas aversive vocalizations, such as screams of fear, are meant to broadcast the signal to distant observers, call for help, or warn others about the presence of predators, necessitating high vocal effort (Gustison & Townsend, 2015). Similarly, listeners in perceptual experiments may assume, often mistakenly, that intense emotional expressions are more likely to be aversive rather than hedonistic (Anikin & Persson, 2017). On the other hand, the association of tense voice quality with feeling more genuinely amused in laughs, although relatively uncertain, could indicate that higher perceived vocal effort is simply interpreted as a sign of greater emotion intensity, making purely hedonistic vocalizations such as laughs more positive, and aversive or ambiguous vocalizations more negative. It will be a productive avenue for future research to look more closely at the differences between vocalization types when investigating the link between voice acoustics and emotion, not least because it could elucidate the cognitive mechanisms involved.

To end on a methodological note, the present results further underline the importance of estimating glottal source when analyzing field recordings of non-speech vocalizations. In speech research it is common to compare the amplitudes of the first few harmonics with each other and with the harmonic nearest to the first formant or a specific frequency as an indirect measure of spectral source (Garellek et al., 2016; Gobl & Ní Chasaide, 2010; Kreiman et al., 2007; Kreiman et al., 2014), but this may be inappropriate for nonverbal vocalizations and animal calls with an extreme range of f_0 that routinely crosses formant frequencies. An interesting option is to estimate how high detectable harmonics reach in the spectrum (cf. "frequency range" in Fichtel et al., 2001; Fichtel & Hammerschmidt, 2002; Hammerschmidt & Jurgens, 2007), although high levels of jitter and noise will affect this measure. Above all, it is advisable to extract multiple measures of spectral shape instead of a single descriptive, such as mean or peak frequency, and to provide access to the original recordings.

6. Statements

6.1. Statement of ethics

Participants in the online experiments were informed of the nature and purpose of the study and, by clicking the corresponding box, confirmed that they were willing to take part in it and for their data to be published anonymously.

6.2. Disclosure statement

The author has no conflicts of interest to declare.

7. References

Airas, M., & Alku, P. (2006). Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient. *Phonetica*, *63*(1), 26-46.

Anikin, A. (2019a). Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior Research Methods*, *51*(2), 778-792.

Anikin, A. (2019b). The perceptual effects of manipulating nonlinear phenomena in synthetic nonverbal vocalizations. *Bioacoustics*, 1-22. doi: 10.1080/09524622.2019.1581839

Anikin, A., & Lima, C. (2018). Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *Quarterly Journal of Experimental Psychology, 71*(3), 622-641.

Anikin, A., & Persson, T. (2017). Nonlinguistic vocalizations from online amateur videos for emotion research: A validated corpus. *Behavior Research Methods*, *49*(2), 758-771.

Anikin, A., Bååth, R., & Persson, T. (2018). Human non-linguistic vocal repertoire: Call types and their meaning. *Journal of Nonverbal Behavior*, *42*(1), 53-80.

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, *25*(15), 2051-2056.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614-636.

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal Affective Voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*, *40*(2), 531-539.

Bishop, J., & Keating, P. (2012). Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. *The Journal of the Acoustical Society of America*, *132*(2), 1100-1112.

Brady, M. C. (2005). Synthesizing affect with an analog vocal tract: glottal source. In *Toward Social Mechanisms of Android Science: A CogSci 2005 Workshop* (pp. 25-26).

Briefer, E. F. (2012). Vocal expression of emotions in mammals: mechanisms of production and evidence. *Journal of Zoology*, *288*(1), 1-20.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Campbell, N. (2006). Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(4), 1171-1178.

Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. (2016). The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion*, *16*(1), 117-128.

Cummings, K. E., & Clements, M. A. (1995). Analysis of the glottal excitation of emotionally styled and stressed speech. *The Journal of the Acoustical Society of America*, *98*(1), 88-98.

Drugman, T., Alku, P., Alwan, A., & Yegnanarayana, B. (2014). Glottal source processing: From analysis to applications. *Computer Speech & Language*, *28*(5), 1117-1138.

Fant, G. (1960). *Acoustic theory of speech perception*. Mouton, The Hague.

Fastl, H., and Zwicker, E. (2006). *Psychoacoustics: facts and models, 2nd ed.* (Vol. 22). Springer Science & Business Media.

Fichtel, C., & Hammerschmidt, K. (2002). Responses of redfronted lemurs to experimentally modified alarm calls: Evidence for urgency based changes in call structure. *Ethology*, *108*(9), 763-778.

Fichtel, C., Hammerschmidt, K., & Jürgens, U. (2001). On the vocal expression of emotion. A multi-parametric analysis of different states of aversion in the squirrel monkey. *Behaviour*, *138*(1), 97-116.

Garellek, M., Samlan, R., Gerratt, B. R., & Kreiman, J. (2016). Modeling the voice source in terms of spectral slopes. *The Journal of the Acoustical Society of America*, *139*(3), 1404-1410.

Gelman, A., Bois, F., & Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, *91*(436), 1400-1412.

Gobl, C., & Ní Chasaide, A. (2010). "Voice source variation and its communicative functions". In Hardcastle, W. J., Laver, J., & Gibbon, F. E. (Eds.). *The handbook of phonetic sciences* (2nd ed.) (pp. 378-423). Singapore: Wiley-Blackwell.

Gobl, C., & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*(1-2), 189-212.

Goudbeek, M., & Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, *128*(3), 1322-1336.

Grichkovtsova, I., Morel, M., & Lacheret, A. (2012). The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, *54*(3), 414-429.

Gustison, M. L., & Townsend, S. W. (2015). A survey of the context and structure of high-and low-amplitude calls in mammals. *Animal Behaviour*, *105*, 281-288.

El Haddad, K., Çakmak, H., Sulír, M., Dupont, S., & Dutoit, T. (2016). Audio affect burst synthesis: A multilevel synthesis system for emotional expressions. In *2016 24th European Signal Processing Conference (EUSIPCO)* (pp. 1158-1162).

Hammerschmidt, K., & Jürgens, U. (2007). Acoustical correlates of affective prosody. *Journal of Voice*, *21*(5), 531-540.

Hansen, J. H., Nandwana, M. K., & Shokouhi, N. (2017). Analysis of human scream and its impact on text-independent speaker verification. *The Journal of the Acoustical Society of America*, *141*(4), 2957-2967.

He, L., Lech, M., & Allen, N. (2010). On the importance of glottal flow spectral energy for the recognition of emotions in speech. In *Eleventh Annual Conference of the International Speech Communication Association* (pp. 2346-2349).

Hogstedt, G. (1983). Adaptation unto death: function of fear screams. *The American Naturalist*, *121*(4), 562-570.

Johnstone, T., & Scherer, K. R. (1999). The effects of emotions on voice quality. In *Proceedings of the XIVth international congress of phonetic sciences* (pp. 2029-2032). San Francisco: University of California, Berkeley.

Kreiman, J., Gerratt, B. R., & Antoñanzas-Barroso, N. (2007). Measures of the glottal source spectrum. *Journal of Speech, Language, and Hearing Research*, *50*, 595-610.

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens*, *1*(1). doi:10.3989/loquens.2014.009.

Kuang, J., Guo, Y., & Liberman, M. (2016). Voice quality as a pitch-range indicator. In *Proceeding of Speech Prosody* (pp. 1061-1065).

Laukkanen, A. M., Vilkman, E., Alku, P., & Oksanen, H. (1996). Physical variations related to stress and emotional state: a preliminary study. *Journal of Phonetics*, *24*(3), 313-335.

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Laugh like you mean it: Authenticity modulates acoustic, physiological and perceptual properties of laughter. *Journal of Nonverbal Behavior*, *40*(2), 133-149.

Lima, C. F., Castro, S. L., & Scott, S. K. (2013). When voices get emotional: a corpus of nonverbal vocalizations for research on emotion processing. *Behavior Research Methods*, *45*(4), 1234-1245.

Lingle, S., Wyman, M. T., Kotrba, R., Teichroeb, L. J., & Romanow, C. A. (2012). What makes a cry a cry? A review of infant distress vocalizations. *Current Zoology*, *58*(5), 698-726.

Linhart, P., Ratcliffe, V. F., Reby, D., & Špinka, M. (2015). Expression of emotional arousal in two different piglet call types. *PloS one*, *10*(8), e0135414.

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say ‘Hello’? Personality impressions from brief novel voices. *PloS one*, *9*(3), e90779.

McPherson, M. J., & McDermott, J. H. (2018). Diversity in pitch perception revealed by task dependence. *Nature Human Behaviour*, *2*(1), 52-66.

Mittal, V. K., & Yegnanarayana, B. (2014). Study of changes in glottal vibration characteristics during laughter. In *Fifteenth Annual Conference of the International Speech Communication Association* (pp. 1777-1781).

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, *93*(2), 1097-1108.

Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological Psychology*, *87*(1), 93-98.

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, *37*(4), 417-435.

Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior*, *27*(4), 283-296.

Raine, J., Pisanski, K., Simner, J., & Reby, D. (2018). Vocal communication of simulated pain. *Bioacoustics*, 1-23. doi: 10.1080/09524622.2018.1463295

Ross, M. D., Owren, M. J., & Zimmermann, E. (2009). Reconstructing the evolution of laughter in great apes and humans. *Current Biology*, *19*(13), 1106-1111.

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *The Quarterly Journal of Experimental Psychology*, *63*(11), 2251-2272.

Scheiner, E., Hammerschmidt, K., Jürgens, U., & Zwirner, P. (2002). Acoustic analyses of developmental changes and emotional expression in the preverbal vocalizations of infants. *Journal of Voice*, *16*(4), 509-529.

Schröder, M. (2009). Expressive speech synthesis: Past, present, and possible futures. In J. Tao and T. Tan (Eds.) *Affective information processing* (pp. 111-126). London: Springer.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In *Seventh European Conference on Speech Communication and Technology* (pp. 1-4). Sep 3-7; Aalborg, Denmark.

Schubert, E., Wolfe, J., & Tarnopolsky, A. (2004). Spectral centroid and timbre in complex, multiple instrumental textures. In *Proceedings of the international conference on music perception and cognition, North Western University, Illinois* (pp. 112-116).

Stevens, K. N. (2000). *Acoustic phonetics* (Vol. 30). MIT press.

Stout, B. (1938). The harmonic structure of vowels in singing in relation to pitch and intensity. *The Journal of the Acoustical Society of America*, *10*(2), 137-146.

Szameitat, D. P., Alter, K., Szameitat, A. J., Darwin, C. J., Wildgruber, D., Dietrich, S., & Sterr, A. (2009). Differentiation of emotions in laughter at the behavioral level. *Emotion*, *9*(3), 397-405.

Titze, I. R. (2008). Nonlinear source–filter coupling in phonation: Theory. *The Journal of the Acoustical Society of America*, *123*(4), 1902-1915.

Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, *107*(6), 3438-3451.

Waaramaa, T., Laukkanen, A. M., Airas, M., & Alku, P. (2010). Perception of emotional valences and activity levels from vowel segments of continuous speech. *Journal of Voice*, *24*(1), 30-38.

Wood, A., Martin, J., & Niedenthal, P. (2017). Towards a social functional account of laughter: Acoustic features convey reward, affiliation, and dominance. *PloS one 12*(8), e0183811.

Yanushevskaya, I., Gobl, C., & Ní Chasaide, A. (2013). Voice quality in affect cueing: does loudness matter? *Frontiers in psychology*, *4*, 335, 1-14.