How to analyze and manipulate nonlinear phenomena in voice recordings

Andrey Anikin[1,2*] and Christian T. Herbst[3,4]

1. Division of Cognitive Science, Lund University, 22100, Lund, Sweden
2. ENES Bioacoustics Research Lab / Lyon Neuroscience Research Centre (CRNL), University of Saint-Etienne, CNRS UMR5292, INSERM UMR_S 1028, Saint-Etienne, France
3. Bioacoustics Laboratory, Department of Behavioral and Cognitive Biology, University of Vienna, Djerassiplatz 1, Vienna 1030, Austria
4. Janette Ogg Voice Research Center, Shenandoah Conservatory, 1460 University Drive, Winchester, VA 22601, USA
ORCIDs: A. Anikin 0000-0002-1250-8261, C. T. Herbst 0000-0001-9095-3953

Keywords: voice, vocal communication, nonlinear phenomena, acoustic analysis, voice synthesis

Summary

We address two research applications in this methodological review: starting from an audio recording, the goal may be to characterize nonlinear phenomena (NLP) at the level of voice production or to test their perceptual effects on listeners. A crucial prerequisite for this work is the ability to detect NLP in acoustic signals, which can then be correlated with biologically relevant information about the caller and with listeners' reaction. NLP are often annotated manually, but this is labor-intensive and not very reliable, although we describe potentially helpful advanced visualization aids such as reassigned spectrograms and phasegrams. Objective acoustic features can also be useful, including general descriptives (harmonics-to-noise ratio, cepstral peak prominence, vocal roughness), statistics derived from nonlinear dynamics (correlation dimension), and NLP-specific measures (depth of modulation and subharmonics). On the perception side, playback studies can greatly benefit from tools for directly manipulating NLP in recordings. Adding frequency jumps, amplitude modulation, and subharmonics is relatively straightforward. Creating biphonation, imitating chaos, or removing NLP from a recording is more challenging, but feasible with parametric voice synthesis. We describe the most promising algorithms for analyzing and manipulating NLP and provide detailed examples with audio files and R code in supplementary materials (https://osf.io/gs8u3/).

Introduction

Nonlinear phenomena (NLP) is an umbrella term for certain oscillatory states of the vertebrate sound generation system, or for transitions (bifurcations) between these states. They encompass phenomena such as frequency jumps, subharmonics, deterministic chaos, and biphonation (see Herzel *et al*. and Fig. 1 in Dunn *et al*. in this volume [*update once DOI is available*] for an overview of NLP classifications). The various NLP can be considered on three levels: (1) as oscillatory features of the voice production apparatus; (2) as features of the radiated acoustic voice signal; and (3) as phenomena that evoke certain sensory impressions and behavioral effects on the perceiving end of vocal communication. The study of NLP is a vast, technically demanding research domain. To limit the scope of this review and to keep it relevant for applied research in bioacoustics and psychology, we focus on the acoustic domain and provide methodological guidelines and tools for answering two main research questions:

Question 1: What do NLP reveal about the caller? Here, the ultimate objective is to learn what biologically relevant information is available in the signal due to the presence of NLP. For

[*] Author for correspondence (andrey.anikin@lucs.lu.se).
†Present address: Division of Cognitive Science, Department of Philosophy, Lund University, Box 192, SE-221 00 Lund, Sweden.

example, we may be interested in whether NLP encode information about the age, health, mate quality, or motivational and affective state of the caller (e.g., [1,2]). Answering this question requires knowledge of the physiological reality of NLP production. Because this information is not directly observable, a variety of techniques are employed to infer it. Specifically, the task we focus on here is detecting NLP from a recorded acoustic signal (the blue path in Fig. 1). We also list the main methods for investigating voice production more directly, such as with high-speed imaging *in vivo* or via excised larynx experimentation (Table S1, see [3,4] for recent reviews), but do not cover them in detail because these methods are not widely available outside clinical voice science. Direct *in vivo* observation of voice production is particularly challenging in non-human animals, and empirical evidence is very sparse. Instead, we assume that researchers will normally only have access to audio recordings, although many of the discussed analytical techniques are also applicable to physiological signals such as electroglottographic (EGG) recordings [5]. Likewise, computational simulation methods – while very useful for creating biophysical models of phonation and understanding NLP at a fundamental level – are less relevant to NLP detection in audio recordings, and they are reviewed elsewhere [6,7].

Question 2: What effect do NLP have on listeners? It is one thing to establish what information is potentially available in a signal, and another to show that receivers actually attend to it. Listeners may also possess sensory biases, in which case the effect of a perceptual feature may be partly decoupled from the biological information it encodes. For example, humans and many nonhuman animals strongly associate low frequencies [8], high intensity [9], and acoustic roughness [10] with size and formidability, which callers can exploit to achieve acoustic size exaggeration. Ideally, this requires models linking objective acoustic properties with percepts, but human psychoacoustics is a long way from achieving comprehensive perceptual models, and even less is known about the perception of NLP in nonhuman animals. Fortunately, this is not an insuperable barrier to progress because a link between the presence of NLP in a perceived signal and the listeners' responses can be demonstrated empirically without fully understanding the underlying psychoacoustics, in two ways: with correlational designs or direct manipulation.

The traditional approach has been to compare the listeners’ reaction to otherwise similar calls with vs. without NLP in playback experiments (e.g., [11–13]; the orange arrow in Fig. 1). Apart from methodological simplicity, this approach has the further advantage that natural calls are used directly for playback, ensuring maximum ecological validity of the stimuli. The main drawback is that it is difficult to infer causality because the presence of NLP in natural vocalizations is strongly associated with other voice characteristics, especially with high intensity and f_o [10,14]. Even if the stimuli with and without NLP are carefully matched on other relevant acoustic characteristics (e.g., as in [13]), it is difficult to ascertain that listeners attend specifically to NLP. Thus, a more powerful solution for inferring causality is to manipulate NLP experimentally while preserving all other acoustic characteristics of a vocalization (the green arrow in Fig. 1). It is desirable to repeat the manipulation in a wide range of stimuli that vary in their duration, f_o range, caller characteristics such as sex and age, etc. to ensure that the results generalize to a broad range of vocalizations and to increase the statistical power, which depends both on the number of stimuli and the number of times each stimulus is evaluated [15]. Accordingly, perceptual studies of NLP require tools that can manipulate them in recordings with high precision and flexibility, and preferably in a user-friendly framework that will streamline the creation of many stimuli.

In this paper we aim to provide an up-to-date review of the analytical techniques and practical tools for working with NLP in the context of these two research questions. We begin with a discussion of the simplest and most common approach to NLP detection – manual NLP annotation – highlighting its pitfalls and offering possible solutions and complementary measures such as general acoustic descriptives. We then consider each NLP type in turn, discussing suitable methods for their analysis and experimental manipulation. Given the intended research application, we do not cover all possible methods of creating NLP (e.g., with biomechanical/computational models of phonation), but focus specifically on their *manipulation* in recordings, defined here as adding or removing NLP episodes or changing their type in recorded or synthesized vocalizations that are naturalistic enough for playback experiments. The key theoretical considerations and algorithms are covered in the main text at a conceptual level, but we also provide the datasets and complete code for all presented examples and simulations in supplementary materials (https://osf.io/gs8u3/).

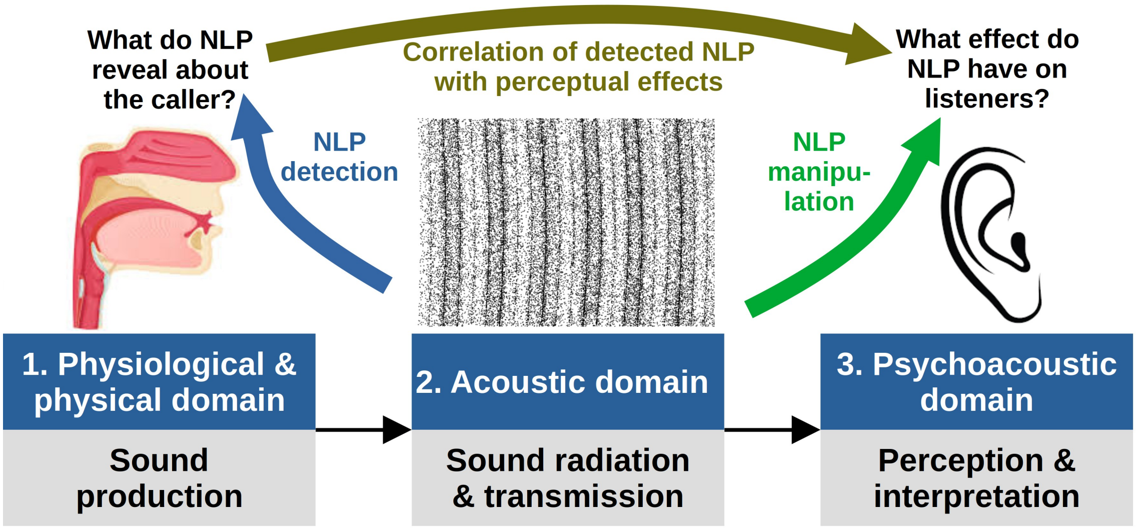
What do NLP reveal about the caller?
Correlation of detected NLP with perceptual effects
What effect do NLP have on listeners?
NLP detection
NLP manipu-lation
1. Physiological & physical domain
Sound production
2. Acoustic domain
Sound radiation & transmission
3. Psychoacoustic domain
Perception & interpretation

Fig. 1 The Brunswik's lens model of communication [16] applied to NLP research. Explanation in the text.

NLP Annotation and Quantification

NLP analysis in bioacoustics typically begins with researchers manually annotating NLP episodes by means of listening to each recording and either inspecting the raw (acoustic) signal in the time domain or scrutinizing some sort of signal feature visualization. The most common of these is the spectrogram, but a number of other visualization tools exist, including phase space embeddings, recurrence plots, etc. (Table S1). When it comes to classifying the different NLP, this approach is time-consuming and often highly subjective, potentially suffering from several issues:

1. Implicit or explicit variation of parameters for display generation has a major impact on NLP detection. For instance, variation of the spectrogram's dynamic range (typically not reported in scientific publications) can be critical for whether subharmonics are discernible in the generated spectrogram or not (Fig. 2D).
2. As soon as an experimenter listens to a sound to decide whether NLP are present, two sorts of biases are potentially introduced. The first is an individual bias, which depends on the experimenter's previous experiences and training. The second is an overall anthropocentric bias. In most cases, very little is known about NLP perception in the investigated species, and there is no guarantee that the features identified by human listeners are relevant to the studied animals. For example, humans are unusually good at pitch discrimination compared to many other mammals [17], potentially making minor frequency jumps in mammalian calls more salient to us than they would be to conspecifics. In contrast, high-frequency biphonation in dog whines may not even be audible to humans [18], not to mention NLP in ultrasonic calls like rodent vocalizations, which can only be detected visually or by transposing the recordings several octaves down in frequency.
3. Finally, both manual and automatic detection of NLP would require a consensus among the bioacoustic research community concerning NLP classification, as well as a standardized approach to their assessment and interpretation. Despite notable efforts, such as signal typing conventions proposed for the human voice [19,20] and the NLP workshop in St. Etienne (2023) that led to this special issue, so far there is no consensus regarding even the nature of NLP and their basic types, much less the best approach to their detection and analysis.

For all these reasons, manual NLP annotations do not necessarily correspond to the ground truth of vocal production, or even of NLP perception in nonhuman species. Nevertheless, given its feasibility, manual annotation remains the go-to approach in NLP research, so it would be important to ascertain how well it captures the reality of vocal production. In the absence of suitable datasets for doing so, we tested at least the inter-rater reliability with which several trained raters performed manual annotation of NLP episodes. Specifically, we asked the attendants of the NLP workshop in St. Etienne in June 2023 to note all NLP episodes in a randomly selected subset of 23 vocalizations from a published corpus, all of which were reported as containing some NLP in the original publication [21]. The recordings included 10 human nonverbal vocalizations (5F + 5M), 10 speech samples (5F + 5M), and three samples of a cappella singing (2F + 1M); the duration varied from 2 to 10 s. Ten raters independently annotated four NLP types (frequency jumps, sidebands, subharmonics, and chaos). We then calculated the agreement between all possible pairs of raters about the status of each 100 ms frame. The average agreement was 80% for the presence or absence of NLP per frame, 60% for NLP type (excluding frequency jumps, which have no duration), and 60% for the presence of a frequency jump within ±50 ms of one annotated by another rater (see vignette *manual_annotation* for the complete analysis of inter-rater agreement). Thus, highly trained and motivated raters are reasonably consistent at detecting NLP episodes in audio recordings of human voice (sidebands, subharmonics, or chaos), but the classification of NLP types appears to be less reliable. Crucially, manual annotations have better internal validity as measures of *perceived* nonlinearities or general harshness in the voice (especially when working with human voices), making them more suitable for research on NLP perception rather than NLP production.

To ensure that the results of manual annotation are as accurate as possible, it is important to avoid a few common pitfalls. Given the importance of the chosen visual representation (Fig. 2), it is helpful to compare several approaches and settings. In particular, the spectrogram can be juxtaposed with the raw waveform (oscillogram), which is an under-utilized, but very informative medium for detecting NLP such as slow amplitude modulation (Fig. 2A). The spectrogram itself may need to be adjusted depending on the acoustic characteristics of each analyzed fragment. Short windows are good for resolving rapid transitions that could otherwise be blurred (e.g., frequency jumps) or misclassified as NLP (e.g., rapid frequency sweeps in bird songs, which may appear as sidebands if the window is too long). On the other hand, extremely long windows upwards of 400-500 ms may be necessary for visualizing stable, but very low-frequency amplitude modulation in calls such as alligator bellows [22]. It may also be helpful to try less familiar visual representations such as time-frequency reassigned spectrograms or wavelet-based transforms [23,24], phasegrams [25], and modulation power spectra [26,27]. These approaches are demonstrated in vignette *visualization* in the supplements (https://osf.io/gs8u3/), and we provide ready-to-use R code for their implementation.

Quantitative analysis of acoustic signals is often used as a complement to manual annotation, or even as the only feasible approach when the analyzed corpus of recordings is very large. This saves time, dispenses with the need for trained raters, and the results are both objective and reproducible. The main drawback is that most measures are only indirectly affected by NLP ("General acoustic measures" in Table S1), and individual metrics are typically affected by a number of NLP in a complex fashion. For example, a drop in signal periodicity (e.g., as measured by the harmonics-to-noise ratio [HNR]) might be due to NLP or some other causes (background noise, breathy phonation, etc.), and this lack of specificity can have major implications for the substantive interpretation of obtained results. In addition, most software for voice analysis, such as the popular open-source toolbox Praat [28] and its algorithms, is designed for analyzing nearly-periodic signals, typically speech. Other measures, discussed in the sections on each NLP type below, are more theoretically grounded, being derived from nonlinear dynamics ("Nonlinear time series analysis" in Table S1), or designed to capture particular NLP types ("NLP-specific acoustic measures" in Table S1).

As an exemplary check of NLP specificity, we calculated a variety of acoustic features (generic, NLP-specific, and derived from nonlinear time series analysis), frame by frame, in 5000 fully synthetic vocalizations (with ground truth of NLP presence and type known *a priori*), as well as in 1518 audio recordings of human nonverbal vocalizations, singing, and speech from [21] with a total duration of two hours and nearly 300,000 overlapping STFT frames 50 ms each (with NLP annotated manually). We then compared the values of each acoustic feature in STFT frames depending on the presence and type of NLP (see vignette *analysis_any-NLP*). The main conclusion was that the presence of NLP explained vastly more variance of the analyzed acoustic measures in synthetic sounds compared to annotated recordings, which could indicate that manual NLP annotations are not entirely accurate (as suggested by the analysis of inter-rater reliability above), and/or that real-life recordings are too “messy” for this kind of acoustic analysis to pick up NLP-specific features. Two measures, the amount of amplitude modulation in the “roughness” frequency range and Cepstral Peak Prominence (CPP), appeared to be most robust to noise in real recordings. Notably, however, none of the tested features were really suitable for discriminating between vocalizations with and without NLP, and especially not for distinguishing deterministic chaos in voiced fragments from turbulent noise in unvoiced fragments.

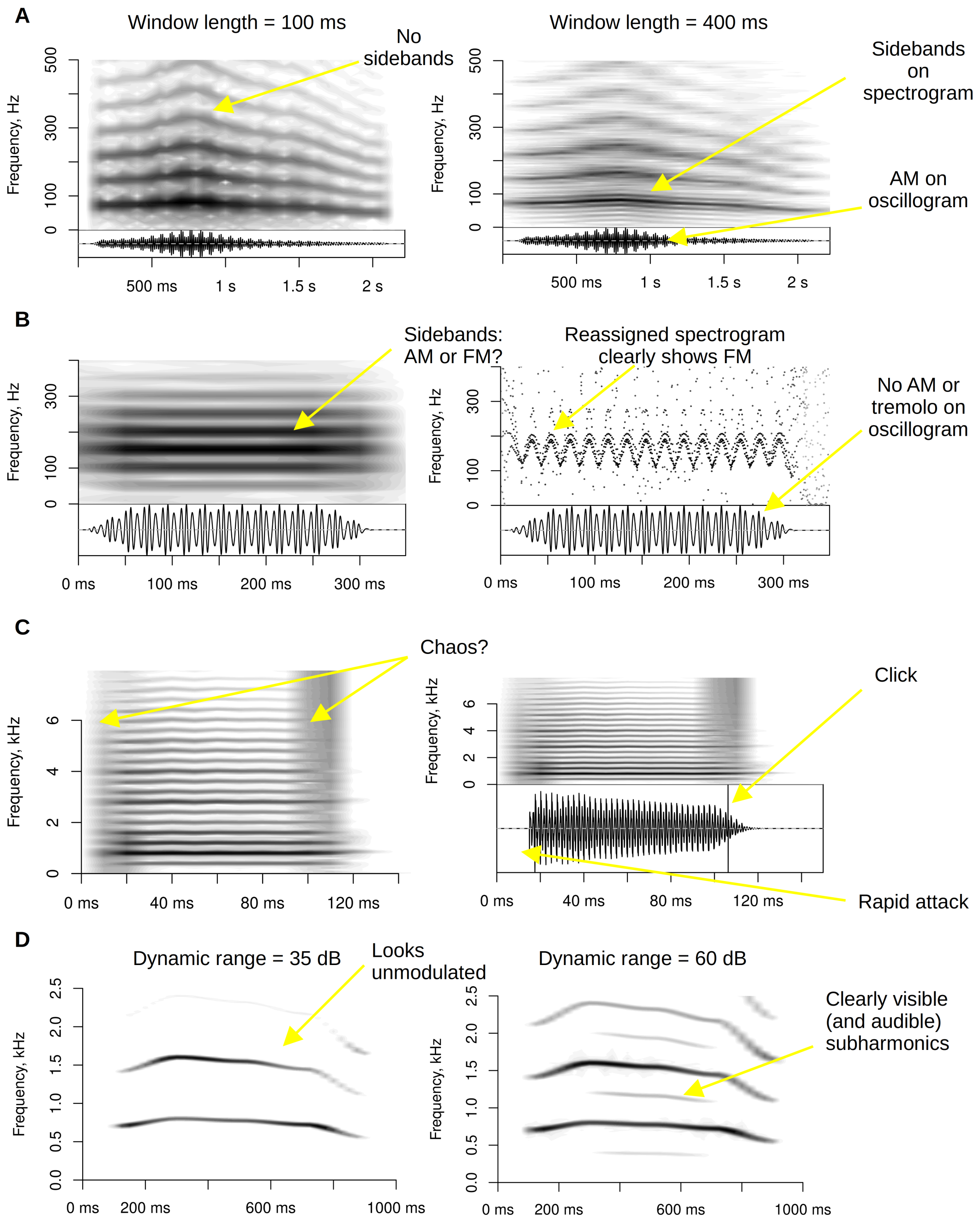

Fig. 2 Visualizing NLP: some common problems and solutions.

Having briefly considered the difficult challenge of annotating NLP manually or using proxy acoustic measures, we now turn to the algorithms that have been designed for analyzing and manipulating specific NLP types.

Frequency jumps

Sudden changes of f_o, known as frequency jumps or pitch jumps, have primarily been researched in the context of human singing [29–31], but they are also found in a variety of animal calls [13,32,33] and in human nonverbal vocalizations such as screams [21,34] and baby cries [35]. Their possible causes include both conditions intrinsic to the vocal folds [36,37] and source-filter interaction with the resonances of either the supralaryngeal vocal tract or the tracheal vocal tract [38–41] ; see Herbst & Elemans in this issue for more details). The best understood example is the transition between vocal registers in human singing, often between the modal or chest voice and the falsetto [31,42,43]. Such voice breaks can occur in both upward and downward directions [29,31,32,36,41,44,45]. Crucially, frequency jumps during register transitions are not merely rapid f_o glides: the larynx suddenly transitions into a different vibratory mode, often with brief episodes of subharmonics or other nonlinearities during the transition [31,37]. Thus, true frequency jumps constitute bifurcations and should be considered a type of NLP.

Analysis

Sophisticated algorithms have been tested in human voice science to detect frequency jumps (often in the context of register transitions) from the electroglottographic signal (EGG) based on statistics such as sample entropy [30,42,46]. In practice, frequency jumps are often annotated manually based on inspecting narrowband spectrograms, and perhaps also listening to the recordings, although the exact method is seldom specified [13,21,33]. As with all NLP, this introduces subjectivity in the analysis, particularly when human listeners annotate the vocalizations of other species because of potential differences both in voice production mechanisms and in the perception of pitch discontinuities. A further challenge is posed by rapid f_o variation caused by super-fast muscles in some animal species that might – without proper analysis tools – be mistaken for a frequency bifurcation [47].

To ensure reproducibility and make the analysis more objective, could frequency jumps be detected automatically? Despite the apparent simplicity, even basic f_o detection is not a trivial task in itself [48]. Furthermore, once f_o contours are extracted, and assuming that these are correct, it is not always obvious what constitutes a discontinuity. An algorithmic jump detector might look for f_o changes in continuously voiced fragments that exceed the average rate of f_o change before and after the focal frame by a certain threshold. We compared the results of such automatized analysis with manual annotation of frequency jumps and obtained a rather poor match (see vignette *analysis_freqJump*). As noted above, inter-rater agreement was also far from perfect for frequency jumps. An additional difficulty is that frequency jump detection requires good time resolution of f_o tracks, making tracking errors more likely and time-consuming to correct manually.

Manipulation

Frequency jumps are relatively straightforward to manipulate (add, remove, or modify) in recordings of human voice or animal vocalizations using high-fidelity pitch-shifting algorithms that can modify f_o in existing recordings, while preserving other spectral properties such as vocal tract resonances [49]. The two most common approaches are pitch-synchronous overlap-and-add (e.g., in *Praat* [28]) and phase vocoding (e.g., in the CLEESE toolbox [50]), which operate in the time domain and frequency domain, respectively. Manipulation of existing recordings is preferable when the required fundamental frequency shift is relatively small because this method preserves all other characteristics of the original recording (Fig. 3A-B), but it may not work very well when the jumps are numerous and/or large. Voice synthesis offers more control (Fig. 3C-D) and is less likely to introduce artifacts, but it may require more work unless the calls are acoustically very simple (e.g., short pure tones).

Whichever method is used, the manipulated vocalizations should sound natural, which is not merely a matter of avoiding artifacts of (re)synthesis [49], but also of ensuring a plausible spectro-temporal context for each jump (e.g., f_o would normally jump upward during an ascending frequency sweep, and vice versa), ideally with a naturally-sounding concomitant change in amplitude and voice quality. It may thus be safer to find a recording with a frequency jump and to remove it by smoothing out the f_o contour, rather than to introduce a new frequency jump where there was none originally (the same reasoning applies to other NLP as well). Finally, the average f_o of a recording should not be greatly affected by manipulating frequency jumps, which would otherwise constitute a confound. For instance, the elimination of a frequency jump in Fig. 3A-B has no effect on the mean f_o, whereas the manipulation of frequency jumps in Fig. 3C-D changes the mean f_o from 420 to 440 Hz, which should perhaps be compensated for by means of slightly adjusting f_o in the preceding fragment.

Despite being relatively simple to manipulate, frequency jumps are arguably the least understood NLP outside human singing, having been experimentally investigated in only a few bioacoustic studies [34,51–53]. More studies with direct manipulation of frequency jumps are needed to shed light on their communicative significance.

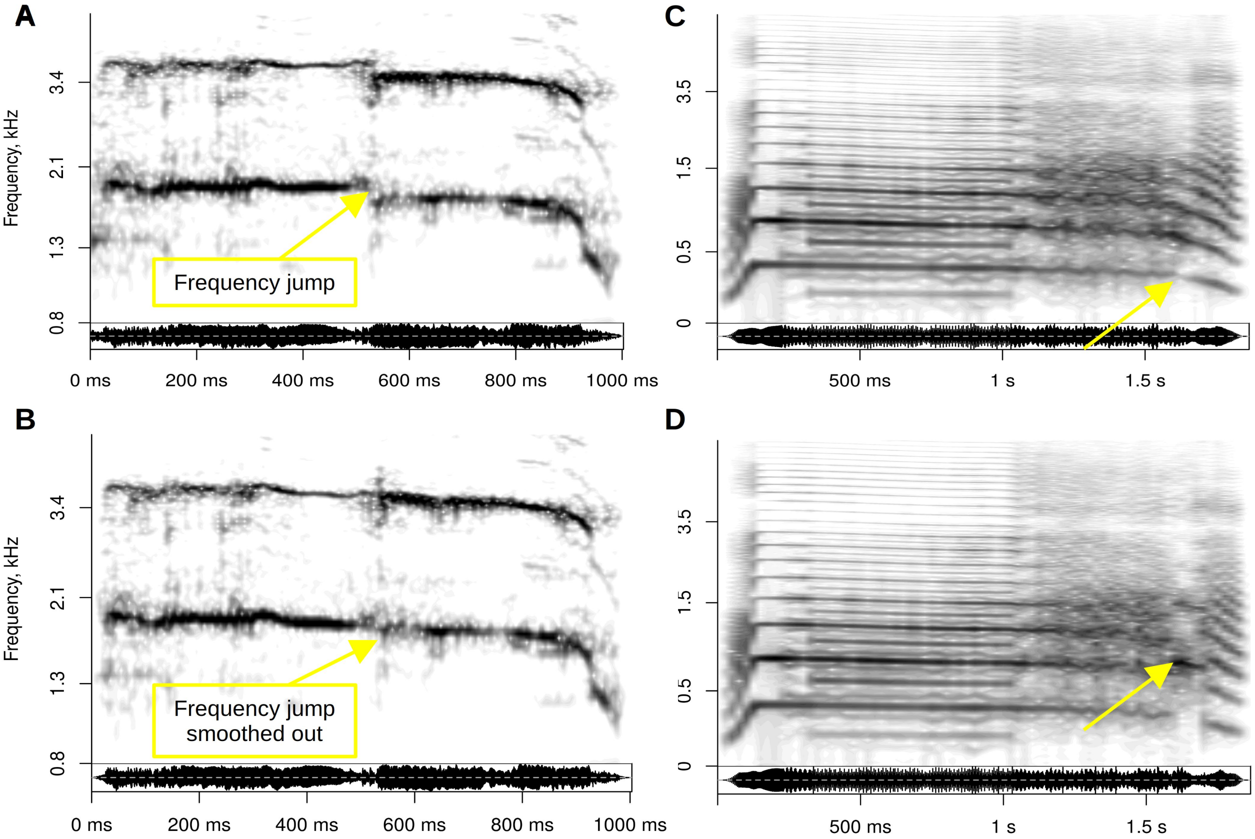

Fig. 3 Manipulating frequency jumps in recorded (A-B) and synthetic vocalizations (C-D). The original recording of a woman's scream shown in panel A was pitch-shifted with a phase vocoder to smooth out the frequency jump at ~520 ms from 1880 to 1750 Hz in panel B. The human roar shown in panels C-D is fully synthetic, making it straightforward to add or remove two rapid frequency jumps at ~1.5 s without affecting other acoustic characteristics.

Low-Frequency Amplitude and Frequency Modulation

Modulation can be of two basic types. Frequency modulation (FM) corresponds to cyclic changes of f_o itself; a familiar example is FM in the range of 4-8 Hz in classical Western singing, known as *vibrato* [54]. Amplitude modulation (AM, related to musical *tremolo*) corresponds to cyclic changes of the waveform amplitude envelope, as when a trilled /r/ modulates the airflow. At least in human singing, vibrato (FM) is typically accompanied by some amplitude modulation (AM), so in practice these two phenomena are often present simultaneously [55,56]. In terms of system dynamics, modulation turns a limit cycle in the phase space into a torus [57,58]. Looking at a spectrogram, both AM and FM can produce sidebands around each f_o harmonic if the modulation frequency is much lower than f_o (Fig. 4).

The modulation frequency can be independent of f_o or coupled with it in some rational fraction such as 3:2 or 2:1, in which case we speak of subharmonics instead of modulation [59]. To complicate matters further, irregular AM with variable frequency and amplitude does not create visible sidebands, but simply produces a noisy-looking spectrogram and an irregular, harsh-sounding voice quality that may look and sound similar to chaos (Fig. 4). For example, using high-speed imaging, [60] showed that rock singers can voluntarily make supralaryngeal mucosa vibrate, creating either subharmonics or irregular AM even when the vocal folds are vibrating in a quasi-periodic fashion. Furthermore, the same production mechanism – simultaneous vibration of aerodynamically coupled vocal folds and supralaryngeal structures such as the ventricular folds – can also create true deterministic chaos [61]. In sum, low-frequency modulation is a rather complex and vaguely defined, albeit very common, NLP category. Bioacousticians often label anything that produces visible sidebands simply as *sidebands*, remaining agnostic as to the underlying mechanism.

Analysis

FM is relatively straightforward to visualize on a spectrogram and to measure if it is slow relative to f_o. For a slow sinusoidal FM, the frequency can be estimated directly from a spectrogram as the reverse of a single period of f_o oscillation (5 or 10 consecutive periods can be averaged to improve the precision). When conceptually considering a nonsinusoidal FM, the spectrum of the f_o contour can be searched for spectral peaks that correspond to FM frequencies. For instance, [59] found two such peaks – two dominant frequencies – in the vibrato produced by Freddie Mercury. Two main metrics are typically computed: modulation rate (the vibrato frequency) and modulation extent (the vibrato amplitude). Depending on the method of visualization and FM frequency, FM can visually appear either as vibrato (slow FM, short analysis window) or as sidebands (fast FM, long analysis window). If FM becomes very rapid relative to f_o, the individual cycles of modulation can no longer be resolved with a standard spectrogram because the instantaneous frequency changes too much within an analysis window (as this window must be kept long enough to resolve the f_o itself). As a result, the spectrogram of a signal with rapid FM will show sidebands around f_o harmonics instead of vibrato, making the situation visually indistinguishable from AM (Fig. 4). For instance, a conventional spectrogram with a window length of 100 ms in Fig. 2B shows strong and stable sidebands, whereas a time-frequency reassigned spectrogram with a window length of 10 ms still captures the 50 Hz FM while preserving reasonable frequency resolution (a conventional spectrogram with a 10 ms window cannot even resolve the f_o itself at 150 Hz). Perceptually, rapid FM also no longer resembles vibrato; for instance, the tone in Fig. 2B sounds like a steady, unmodulated note with the same pitch as FM frequency (50 Hz).

The easiest method of estimating AM frequency is to measure the period of one or several adjacent modulation cycles on an oscillogram (Fig. 4). More formally, AM can be measured from the amplitude envelope or from the modulation spectrum. The first method is based on extracting a smoothed envelope (e.g., as the root mean square amplitude or the magnitude of the analytical signal obtained with the Hilbert transform), which is then bandpass-filtered to focus on the range of AM frequencies of interest to the researcher. Peaks in the spectrum of this amplitude envelope correspond to relatively stable AM frequencies, whose magnitude is directly proportionate to AM depth. The second method begins with generating a modulation spectrum, which is a two-dimensional Fourier transform of the spectrogram often used for calculating the perceptual roughness of a sound [26,27]. Again, peaks along the temporal modulation dimension of the modulation spectrum, averaging across the spectral modulation dimension, indicate the presence of relatively stable and pronounced AM, and the magnitude of these peaks is related to AM depth. Depending on which method is used, the range of detectable modulation frequencies depends on the amount of smoothing of the envelope or the window length and step used to produce a modulation spectrum.

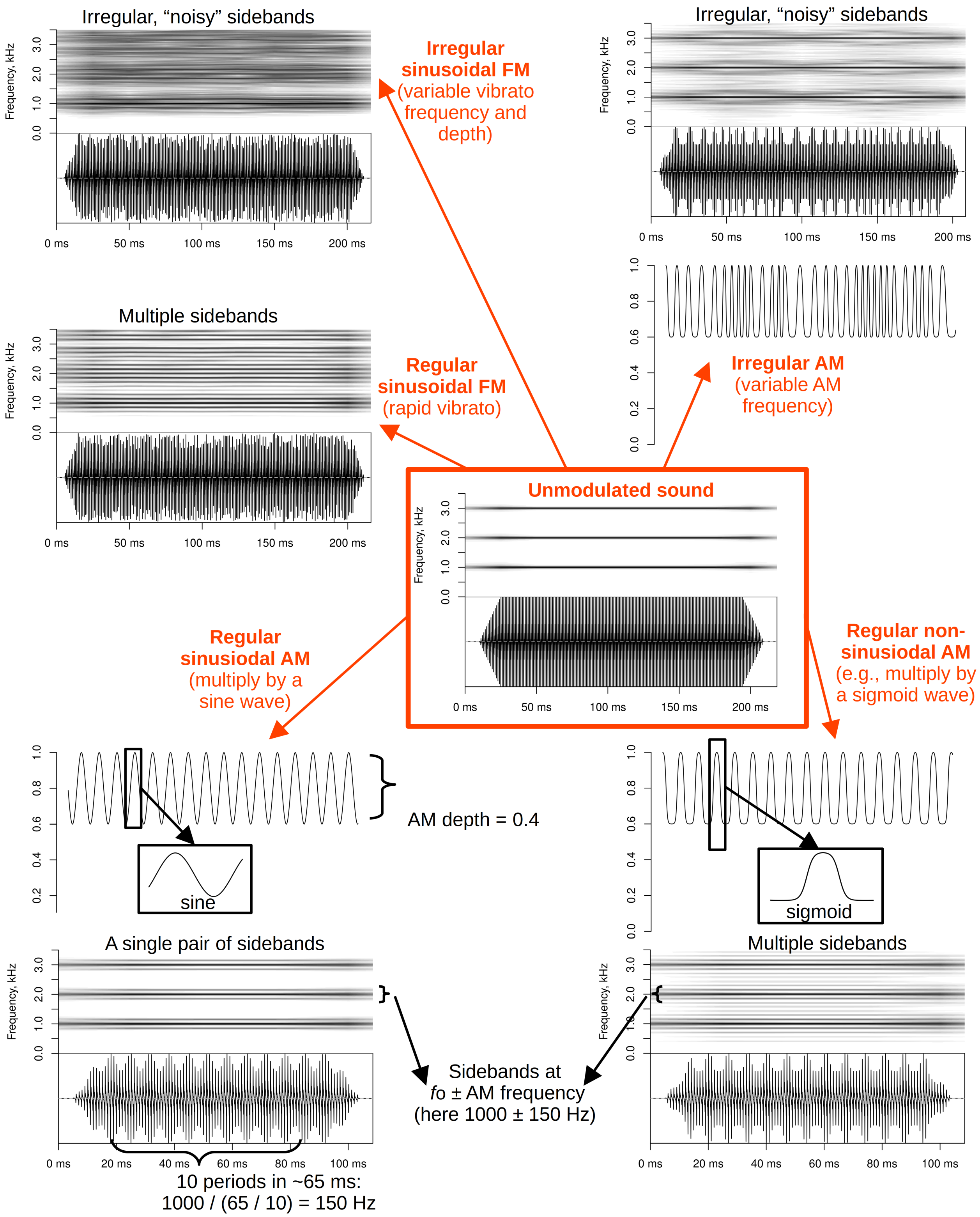

Fig. 4 Signal modulation produces a wide variety of sidebands about the harmonics of f_o. A steady harmonic sound with an f_o of 1 kHz is amplitude- or frequency-modulated; all spectrograms use a window length of 50 ms. See also vignette *synthesis_AM-FM*.

Both methods of measuring AM with the *soundgen* function *analyze()* were optimized by means of synthesizing and analyzing 10,000 vocalizations (see vignette *analysis_amDep*). In our simulations, AM frequency estimates derived from the envelope were found to be less reliable when the true modulation frequency dropped close to the lower end of the analyzed frequency range, whereas the modulation spectrum estimates erred when AM was too rapid relative to the window length used to create the modulation spectrum (here, 15 ms). The two methods can thus be combined to capture a wider range of AM frequencies. An important practical tip is to narrow down the range of considered AM frequencies as much as possible, based on what is known about the species' biology and vocal behaviors.

Manipulation

Slow, vibrato-like FM can be introduced in a recording using the same pitch-shifting techniques that were described above for frequency jumps. Singing voice synthesis is also routinely performed with a controlled amount of vibrato (e.g., [62–64]). Rapid FM is an uncommon manipulation to perform on an existing recording because pitch-shifting algorithms can introduce artifacts at high modulation rates, but it is achievable with parametric voice synthesis. AM is probably the easiest type of NLP to add to a recording: all that is needed is to multiply the sound by a modulating waveform ranging from *(1 – AM depth)* to 1. If the modulating waveform is a pure sine, which may be uncommon in animal vocalizations, it creates a single pair of new harmonics at ±*AM frequency* around each partial of f_o. More complex, nonsinusoidal AM creates multiple harmonics at ±*AM frequency x integer*, which form characteristic sidebands around each f_o partial (Fig. 4). This is a popular method for creating rough-sounding voices, which can even be applied in real time [50,65]. The manipulation of AM and FM is demonstrated together with subharmonics in vignette *synthesis_AM-FM*.

Subharmonics

Subharmonics are additional frequency components (f_{sub}) at a rational fraction of f_o, typically at $f_{sub} = f_o/2$ or $f_o/3$, but potentially more complex situations are possible such as f_o:f_{sub} = 3:2 [66,67]. They can be produced by partly desynchronized, but still strongly coupled vocal folds or parts thereof that vibrate at harmonically related frequencies, which can be caused by the entrainment of two vibratory modes of the vocal folds [67–69], asymmetric tension on the two vocal folds [70–72], or source-filter interactions with supraglottal [39] or subglottal [41] resonances. Another possible origin of subharmonics is simultaneous frequency-locked vibration of two oscillators such as the vocal folds and the ventricular folds [59,73–75] or aryepiglottic folds [76]. When subharmonics are caused by AM, the modulation depth can be defined as the difference in the amplitude of adjacent glottal cycles expressed as a proportion of the sum of the two amplitudes; in the case of FM, it is defined as the difference in the periods instead of amplitudes of adjacent cycles [77,78]. Amplitude and frequency modulation is thus a more general phenomenon, of which subharmonics are an important and common special case [77].

Analysis

Subharmonics can be detected and quantified in several ways. One approach is to literally compare, in the frequency domain, the amplitude of spectral peaks corresponding to f_o partials and subharmonics. The only available open-source algorithm of this kind [79] appears to produce valid estimates of the strength of subharmonics provided that it is properly tuned [77]. Our own simulations suggest that, with realistic noisy recordings, a more robust method is to work with the cepstrum, looking for peaks at a fraction of f_o. The ratio of the magnitude of two cepstral peaks, one corresponding to f_o and the other to a potential subharmonic, is an approximately linear function of the logarithm of subharmonic depth, which makes it straightforward to calibrate the subharmonic detector against a benchmark of synthetic sounds with known subharmonic depth. Once calibrated, this algorithm can approximately measure the depth of subharmonics for a wide range of f_o and signal-to-noise ratios, achieving an overall Pearson's correlation of $r = .64$ between true and measured subDep when f_o is detected correctly (vignette *analysis_subh*). Obviously, there exists a "chicken-and-egg" problem: f_o tracking must be accurate, otherwise estimates of subharmonic frequency and depth are meaningless, yet the very presence of subharmonics complicates f_o detection. One approach is to correct f_o contours manually prior to analyzing subharmonics; another is to label episodes of subharmonics first and to pass on this information to the f_o tracker. If neither of these options is feasible, it may be safer to exclude this particular signal from the analysis.

Manipulation

At least three distinct ways of manipulating subharmonics have been described in the literature. The first is to multiply the signal by a low-frequency modulating waveform, which can produce subharmonics if the modulation frequency is kept at $f_o/2$ (or any other integer rate) at each time point [65]. This method is straightforward to apply to any sound, including instrumental music [80]. The main limitation is that it is critically dependent on accurate f_o tracking, which can be problematic when working with aperiodic or relatively noisy signals. The second approach is to vary the amplitude and timing of adjacent glottal pulses during voice synthesis [81]. For instance, the *diplophonia* parameter in the well-known Klatt synthesizer controls the extent to which every second glottal cycle is delayed in time and attenuated in amplitude [82]. Finally, a brute-force approach is to literally synthesize new partials spaced by a fraction of f_o [83]. Like the diplophonia method, this requires parametric synthesis, but the connection between control parameters and the output is more transparent, and any f_o:f_{sub} ratio can be achieved. In particular, the depth of subharmonics is then specified in the frequency domain as the amplitude of subharmonic (f_{sub}) partials relative to f_o partials, which is more perceptually relevant compared to time-domain definitions [81].

Biphonation

Signals with two frequency components are identified as *quasi-periodic* in the nonlinear dynamics literature, and their phase trajectory forms a torus [84]. Biphonation is commonly distinguished from other tori by stipulating that both frequencies should evoke pitch sensations (unlike low-frequency modulation), change independently, and not form a rational ratio (unlike subharmonics) [57,58]. An additional requirement is sometimes proposed that both frequencies should be produced by a single sound source – for example, by the same half of the syrinx [85]. The term *diplophonia*, common in the voice literature, does not make such distinctions and includes any two-frequency phenomena [66].

Biphonation *senso stricto*, with two or even three independent audible frequencies, has been reported in human pathological voices [57] and singing [86], as well as in vocalizations of other mammals [18,87–89] and birds [85,90]. The two sources may be physically coupled: for example, the higher frequency (g_o) can be amplitude-modulated by the lower f_o, generating complex sidebands [88,89]. On the other hand, two fully independent frequencies may be produced without noticeable coupling, as in bird songs consisting of two independently controlled sound sources in the syrinx [90].

Analysis

The task of tracking two independent frequencies is challenging and rather exotic in bioacoustics. There are numerous proposed solutions for multi-pitch estimation in polyphonic music [91] and conversation analysis [92], but their robustness and applicability to biphonation remain largely unknown. A rare exception is the work by Aichinger, who developed and tested algorithms for simultaneous tracking of two frequencies in pathological human voices, essentially by means of testing several paths through f_o candidates [93,94]. Tools for AM analysis may also be applicable when one frequency is many times lower and amplitude-modulates the higher one, but otherwise manual annotation and description based on the spectrogram and the phase space are the only viable approaches at present.

Manipulation

Biphonation can be created by mixing separate recordings or (re)synthesized versions of two or more vocal sources. The key challenge is that these vocal sources may be coupled: for example, one may be modulating the other, or f_0 and g_0 may be momentarily locking to each other and/or to resonance frequencies. There are some workaround solutions such as adding a multiplicative term when mixing the two sounds in order to achieve modulation (see vignette *synthesis_biphonation*). Potentially tractable cases for synthesizing biphonation involve fully independent sound sources, such as in bird and whale songs, but there is only limited work in this area [95]. Overall, however, two-frequency calls like horse whinnies [87] or wapiti bugles [88] are some of the most challenging vocalizations to work with.

Chaos

An entire branch of mathematics, known as chaos theory, has been developed to model the behavior of deterministic systems that nevertheless display seemingly random or chaotic behavior. More formally, chaotic systems are deterministic, bounded, aperiodic, and sensitive to initial conditions [96]. Belying this apparent complexity, the attractors of chaotic systems can be relatively simple when appropriately reconstructed in the phase space. Applied to vocal production, the term *chaos* designates “nonrandom noise” [39], namely highly irregular vibration of the vocal folds and other coupled structures [97] that is deterministic, being the result of a mechanically simple system with relatively few degrees of freedom, yet seemingly random or noisy. Thus, aspiration noise as in /s/ would not be described as chaos [19] because the source of randomness in this case is turbulence in a very high-dimensional system, and its attractor in low-dimensional phase space is structureless.

Analysis

Like other NLP, chaos is often annotated manually in audio recordings. There is typically a residual trace of the original f_0 and perhaps even a few of the lower harmonics on the spectrogram, but with a varying amount of spectral smearing, making chaotic phonation look superficially similar to turbulent noise (e.g., whispered speech). Furthermore, given the limitations of spectrograms in terms of the tradeoff between time and frequency resolution, they are potentially inappropriate for distinguishing between chaos and irregular AM/FM. Given its similar appearance, chaos is also difficult to distinguish visually in the time domain from noise or signals containing FM. Therefore, inspection of the waveform is not very helpful, except that it sometimes reveals other causes of chaos-like spectral noise such as rapid sound onsets, clicks, or recording artifacts (Fig. 2C).

The most powerful visual tool for distinguishing between (high-dimensional) noise and (low-dimensional) chaos is the phase space (see vignette *phasegrams*). Formal mathematical methods of nonlinear time series analysis have also been applied to voice analysis, mostly in the context of quantifying voice pathology (dysphonia). A great variety of visualization aids and measures have been explored, including the correlation dimension D2, Lyapunov exponents, Poincare map, fractal dimension, Kolmogorov entropy, Shannon entropy, Renyi entropy, correlation entropy, Hurst exponent H, Lempel-Ziv complexity, etc. Some measures may be more noise-robust and tolerant of nonstationary signals: for instance, the correlation dimension D2 is claimed to tolerate noise levels of 8% or even up to 20%, and to perform robustly with relatively short voiced frames of only 20 ms in duration [98]. D2 was found useful for detecting voice pathology in several studies [98,99]. For instance, it was elevated in patients with vocal tremor caused by Parkinson’s disease or vocal polyps [100] and moderately correlated with subjectively rated dysphonia [101]. D2 and the largest Lyapunov's exponent also discriminated between normal and irregular phonation in excised larynges [102–104]. However, other work suggests that D2 is only meaningful in noise-free and relatively periodic signals [105], making it less suitable for detecting episodes of chaos in field recordings. In fact, D2 and shimmer were reported to be elevated in a cappella opera singing compared to other musical genres, possibly because of the vibrato [106]. Accordingly, even when dealing with high-quality recordings of steady vowels, some authors recommend classifying the voices manually into three or four types, from mostly periodic to fully aperiodic, and excluding the unsuitable voice types from the analyses of f_o perturbation and nonlinear measures [19,20], which brings us full circle back to manual classification and annotation.

Our own analysis and simulations suggest that D2 is indeed among the most robust measures derived from nonlinear dynamics, but still far from a reliable "litmus test" for chaos in relatively noisy real-life recordings (see vignette *analysis_any-NLP*). Computationally cheaper measures include summaries of Poincare sections such as their Shannon entropy or the Phasegram Complexity Estimate, which is calculated as the one-dimensional correlation dimension along each Poincare section [97]. An important direction for future work would be to better validate D2 and related nonlinear measures for detecting NLP in various species because most research to date has focused rather narrowly on diagnosing human voice pathology.

Tools for nonlinear time series analysis can be found in specialized software like TISEAN [107] and in various R libraries. An important proviso is that the effective use of these tools requires advanced mathematical expertise for diagnostics and customization. An additional – and often neglected – condition is that the input for nonlinear analysis is supposed to be a noise-free and stationary signal, where the mean and variance do not change over time. Real-life recordings like continuous speech or nonverbal vocalizations routinely violate the key assumptions of nonlinear analysis, being relatively short, noisy, and nonstationary – for example, f_o can noticeably change within an analysis window [96,101]. In particular, just like perturbation measures (jitter and shimmer) are not meaningful when f_o cannot be detected, nonlinear dynamic analysis is not applicable to high-dimensional noise of the kind found in breathy or whispery voices [19] and in recordings with a low signal-to-noise ratio.

Another challenge is to determine the optimal lag for reconstructing the phase space when the signal is no longer periodic. This lag is normally set to approximately 1/4 of the fundamental period to reduce the correlation of the original and time-delayed versions of the signal, and the optimal value can be estimated from the autocorrelation or mutual information function [96]. Alternatively, we can perform a Hilbert transform and plot its real versus imaginary components [25,46]. Because the phase of each frequency component is shifted by ±pi/2, this decorrelates the components and produces a suitable phase space without the need to determine f_o, which is nearly impossible during episodes of chaos (see vignette *phasegrams*).

Manipulation

Deterministic chaos in animal vocalizations is sometimes imitated simply by inserting a short episode of white noise into a tonal call [108,109]. There is some evidence that birds respond similarly to synthetic chaotic time series and white noise [110]. However, chaos found in natural calls retains residual periodicity and formant structure, making it different from both white noise and mathematically constructed chaos such as the output of a logistic map. We have therefore tested an alternative method of imitating chaos by means of stochastically perturbing the periodicity of a vocal source. Because of the stochastic implementation, this is not true low-dimensional deterministic chaos, but it has been shown to create a perceptually adequate imitation thereof both in human nonverbal vocalizations [34,111,112] and in puppy whines [52]. As implemented in *soundgen* [83], this is achieved by adding strong and very rapid Gaussian jitter to f_o. This works for any f_o and jitter depths because, in contrast to most algorithms for voice synthesis, jitter in *soundgen* corresponds to perturbing the instantaneous frequency repeatedly within glottal cycles rather than resetting it at cycle boundaries [113]. Alternatively, random and very rapid jumps of the instantaneous frequency (again, not synchronized with cycle boundaries) can be added between two f_o values: a latent f_o contour and another value such as a nearby formant. This is based on the observation that chaotic behavior is often brought about by an interaction between two or more oscillation modes [97], which can be activated simultaneously and superimposed [114]. Examples of both approaches are demonstrated in vignette *synthesis_chaos*.

Conclusions

There is mounting evidence that NLP encode a wealth of biologically important information about the caller, from individual identity to emotional state, and that listeners carefully attend to these acoustic features. In order to unlock the potential of NLP for a better understanding of vocal communication in the animal world and in human societies, it is crucial to have effective, accurate, and user-friendly tools for working with these acoustic phenomena. In this methodological review, we provide a necessarily brief, but relatively comprehensive description of modern techniques for NLP analysis, as well as for their manipulation and perceptual testing in playback experiments.

When it comes to NLP analysis, many challenges and exciting opportunities still lie ahead. To make claims about the presence of NLP, their episodes must be annotated, but both manual and automatic approaches have their particular drawbacks. Even when performed by trained researchers, the annotation task is far from trivial, and we describe several common pitfalls and suggest solutions for improving its accuracy. Several automatically extracted acoustic descriptives do capture some of the NLP-related variation in voice quality, but these are typically not specific enough to infer the presence of NLP in general or of their specific types. The mathematical tools of nonlinear time series analysis are potentially useful for detecting chaos, but the measures proposed so far are not sufficiently robust or user-friendly to be of much practical use, with the possible exception of the correlation dimension D2 and the use of the phase space as a visual aid. This calls for further work on applied NLP analysis and a better consensus on the best practices in the scientific community.

The domain of applied NLP manipulation and synthesis appears to be more mature as several effective experimental techniques have now been developed and validated. One major tool to exploit is pitch-shifting algorithms, which can create frequency jumps and frequency modulation with high precision, sometimes even in real time. Amplitude modulation and subharmonics are also relatively straightforward to add to a recording. Voice synthesis is the ultimate tool for the most demanding NLP manipulations, including chaos. A major advantage of parametric synthesis is the ability not only to add a specific NLP, but also to remove it. A promising approach is to find prototype recordings with various NLP and then resynthesize them for playback experiments, either removing the NLP completely or modifying their type (e.g., turning an episode of chaos into subharmonics). With this method, the manipulated NLP will appear in their natural spectro-temporal context, greatly improving the ecological validity of manipulated stimuli compared to simply inserting NLP in an arbitrarily chosen location in a recording that originally did not contain NLP at all.

In future, the work on NLP detection and manipulation can greatly benefit from better theoretical and methodological integration between clinical voice research and bioacoustic or psychological studies aiming to understand the communicative function of NLP. A particularly welcome development would be the release of freely available, extensive, and well-documented audio collections covering various NLP in a range of species. This will facilitate further methodological advances and provide suitable training datasets for machine-learning algorithms. NLP detection and annotation, in particular, is a natural task for neural networks – with the important proviso that the training data must be valid. Because expert annotations of specific NLP types in audio recordings may not be entirely reliable, the training data should ideally be more objective (e.g., EGG or high-speed imaging) if the goal is to understand vocal production. On the other hand, expert annotations can provide highly valid training data for algorithms whose aim is to capture the perceptual salience of NLP rather than the precise mechanism of their production. Psychoacoustic modeling of NLP perception is another major area for future research, particularly in nonhuman animals. We hope that this special issue will catalyze the multifaceted field of NLP research, promoting better integration and data sharing between different disciplines and research groups.

Acknowledgments

A.A was supported by grant 2023-00850 from the Swedish Research Council (Vetenskapsrådet). We are grateful to Paul Boersma and David Reby for their help with the analysis of harmonics-to-noise ratio. We also thank our colleagues who provided NLP annotations for testing the inter-rater reliability: Anna Terrade, Daria Valente, David Reby, Elodie Briefer, Floriane Fournier, Katarzyna Pisanski, Mathilde Massenet, and Romane Filippe.

References

1. Cazau D, Adam O, Aubin T, Laitman JT, Reidenberg JS. 2016 A study of vocal nonlinearities in humpback whale songs: from production mechanisms to acoustic analysis. *Scientific reports* **6**, 1–12.
2. Riede T, Arcadi AC, Owren MJ. 2007 Nonlinear acoustics in the pant hoots of common chimpanzees (Pan troglodytes): vocalizing at the edge. *The Journal of the Acoustical Society of America* **121**, 1758–1767.
3. Deliyski DD, Hillman RE. 2010 State of the art laryngeal imaging: research and clinical implications. *Current opinion in otolaryngology & head and neck surgery* **18**, 147.
4. Garcia M, Herbst CT. 2018 Excised larynx experimentation: history, current developments, and prospects for bioacoustic research. *Anthropological Science* **126**, 9–17.
5. Herbst CT. 2020 Electroglottography–an update. *Journal of Voice* **34**, 503–526.
6. Calvache C, Solaque L, Velasco A, Peñuela L. 2023 Biomechanical models to represent vocal physiology: a systematic review. *Journal of Voice* **37**, 465-e1.
7. Döllinger M, Zhang Z, Schoder S, Šidlof P, Tur B, Kniesburges S. 2023 Overview on state-of-the-art numerical modeling of the phonation process. *Acta Acustica* **7**, 25.
8. Aung T, Puts D. 2020 Voice pitch: a window into the communication of social power. *Current opinion in psychology* **33**, 154–161.
9. Anikin A, Valente D, Pisanski K, Cornec C, Bryant G, Reby D. 2023 The role of loudness in vocal intimidation. *Journal of Experimental Psychology: General* (doi:https://doi.org/10.1037/xge0001508)
10. Fitch WT, Neubauer J, Herzel H. 2002 Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal Behaviour* **63**, 407–418.
11. Karp D, Manser MB, Wiley EM, Townsend SW. 2014 Nonlinearities in meerkat alarm calls prevent receivers from habituating. *Ethology* **120**, 189–196.
12. Reby D, Charlton BD. 2012 Attention grabbing in red deer sexual calls. *Animal cognition* **15**, 265–270.
13. Wu Y, Luo X, Chen P, Zhang F. 2023 Frequency jumps and subharmonic components in calls of female Odorrana tormota differentially affect the vocal behaviors of male frogs. *Frontiers in Zoology* **20**, 39.
14. Herzel H, Berry D, Titze I, Steinecke I. 1995 Nonlinear dynamics of the voice: signal analysis and biomechanical modeling. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **5**, 30–34.
15. Westfall J, Kenny DA, Judd CM. 2014 Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General* **143**, 2020.
16. Brunswik E. 1956 *Perception and the representative design of psychological experiments*. Berkeley: University of California Press.
17. Shofner WP. 2005 Comparative aspects of pitch perception. In *Pitch: Neural Coding and Perception*, pp. 56–98. Springer.
18. Sibiryakova OV, Volodin IA, Volodina EV. 2020 Polyphony of domestic dog whines and vocal cues to body size. *Current Zoology*
19. Sprecher A, Olszewski A, Jiang JJ, Zhang Y. 2010 Updating signal typing in voice: addition of type 4 signals. *The Journal of the Acoustical Society of America* **127**, 3710–3716.
20. Titze IR. 1995 *Workshop on acoustic voice analysis: Summary statement*. Denver, Co: National Center for Voice and Speech.
21. Anikin A, Canessa-Pollard V, Pisanski K, Massenet M, Reby D. 2023 Beyond speech: exploring diversity in the human voice. *iScience* (doi:https://doi.org/10.1016/j.isci.2023.108204)
22. Jensen TR, Anikin A, Osvath M, Reber SA. 2024 Knowing a fellow by their bellow: acoustic individuality in the bellows of the American alligator. *Animal Behaviour* **207**, 157–167.
23. Fulop SA, Fitz K. 2006 Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *The Journal of the Acoustical Society of America* **119**, 360–371.

24. Sainburg T, Gentner TQ. 2021 Towards a computational neuroethology of vocal communication: from bioacoustics to neurophysiology, emerging tools and future direction. *Frontiers in Behavioral Neuroscience* , 330.

25. Herbst CT, Herzel H, Švec JG, Wyman MT, Fitch WT. 2013 Visualization of system dynamics using phasegrams. *Journal of the Royal Society Interface* **10**, 20130288.

26. Elliott TM, Theunissen FE. 2009 The modulation transfer function for speech intelligibility. *PLoS computational biology* **5**, e1000302.

27. Singh NC, Theunissen FE. 2003 Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America* **114**, 3394–3411.

28. Boersma P. 2006 Praat: doing phonetics by computer.

29. Bloothooft G, Wijck M van, Pabon P. 2001 Relations between vocal registers in voice breaks. In *Seventh European Conference on Speech Communication and Technology*,

30. Echternach M, Burk F, Köberlein M, Selamtzis A, Döllinger M, Burdumy M, Richter B, Herbst CT. 2017 Laryngeal evidence for the first and second passaggio in professionally trained sopranos. *PloS one* **12**, e0175865.

31. Švec JG, Schutte HK, Miller DG. 1999 On pitch jumps between chest and falsetto registers in voice: Data from living and excised human larynges. *The Journal of the Acoustical Society of America* **106**, 1523–1531.

32. Riede T, Owren MJ, Arcadi AC. 2004 Nonlinear acoustics in pant hoots of common chimpanzees (Pan troglodytes): frequency jumps, subharmonics, biphonation, and deterministic chaos. *American Journal of Primatology* **64**, 277–291.

33. Riede T, Kobrina A, Pasch B. 2024 Anatomy and mechanisms of vocal production in harvest mice. *Journal of Experimental Biology* , jeb-246553.

34. Anikin A. 2020 The perceptual effects of manipulating nonlinear phenomena in synthetic nonverbal vocalizations. *Bioacoustics* **29**, 226–247.

35. Mende W, Herzel H, Wermke K. 1990 Bifurcations and chaos in newborn infant cries. *Physics Letters A* **145**, 418–424.

36. Lehoux S, Herbst CT, Dobiáš M, Švec JG. 2023 Frequency jumps in excised larynges in anechoic conditions: A pilot study. *Journal of Sound and Vibration* **551**, 117607.

37. Uezu Y, Kaburagi T. 2016 A measurement study on voice instabilities during modal-falsetto register transition. *Acoustical Science and Technology* **37**, 267–276.

38. Maxfield L, Palaparthi A, Titze I. 2017 New evidence that nonlinear source-filter coupling affects harmonic intensity and fo stability during instances of harmonics crossing formants. *Journal of voice* **31**, 149–156.

39. Titze IR. 2008 Nonlinear source–filter coupling in phonation: Theory. *The Journal of the Acoustical Society of America* **123**, 1902–1915.

40. Tokuda IT, Zemke M, Kob M, Herzel H. 2010 Biomechanical modeling of register transitions and the role of vocal tract resonators. *The Journal of the Acoustical Society of America* **127**, 1528–1536.

41. Zhang Z, Neubauer J, Berry DA. 2006 The influence of subglottal acoustics on laboratory models of phonation. *The Journal of the Acoustical Society of America* **120**, 1558–1569.

42. Henrich DN. 2006 Mirroring the voice from Garcia to the present day: Some insights into singing voice registers. *Logopedics Phoniatrics Vocology* **31**, 3–14.

43. Herbst CT. 2020 Registers-The Snake Pit of Voice Pedagogy. Part 1: Proprioception, perception, and laryngeal mechanisms. *Journal of Singing* **77**, 175–190.

44. Miller DG, Švec JG, Schutte HK. 2002 Measurement of characteristic leap interval between chest and falsetto registers. *Journal of Voice* **16**, 8–19.

45. Van den Berg J. 1968 Register problems. *Annals of the New York Academy of Sciences* **155**, 129–134.

46. Selamtzis A, Ternström S. 2014 Analysis of vibratory states in phonation using spectral features of the electroglottographic signal. *The Journal of the Acoustical Society of America* **136**, 2773–2783.

47. Elemans CP, Laje R, Mindlin GB, Goller F. 2010 Smooth operator: avoidance of subharmonic bifurcations through mechanical mechanisms simplifies song motor control in adult zebra finches. *Journal of Neuroscience* **30**, 13246–13253.

48. Roark RM. 2006 Frequency and voice: perspectives in the time domain. *Journal of Voice* **20**, 325–354.

49. Arias P, Rachman L, Liuni M, Aucouturier J-J. 2021 Beyond correlation: acoustic transformation methods for the experimental study of emotional voice and speech. *Emotion Review* **13**, 12–24.

50. Burred JJ, Ponsot E, Goupil L, Liuni M, Aucouturier J-J. 2019 CLEESE: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition. *PloS one* **14**, e0205943.

51. Blesdoe EK, Blumstein DT. 2014 What is the sound of fear? Behavioral responses of white-crowned sparrows Zonotrichia leucophrys to synthesized nonlinear acoustic phenomena. *Current Zoology* **60**, 534–541.

52. Massenet M, Anikin A, Pisanski K, Reynaud K, Mathevon N, Reby D. 2022 Nonlinear vocal phenomena affect human perceptions of distress, size and dominance in puppy whines. *Proceedings of the Royal Society B* **289**, 20220429.

53. Slaughter EI, Berlin ER, Bower JT, Blumstein DT. 2013 A Test of the Nonlinearity Hypothesis in Great-tailed Grackles (Q uiscalus mexicanus). *Ethology* **119**, 309–315.

54. Dejonckere PH, Hirano M, Sundberg J. 1995 *Vibrato*. San Diego: Singular Publishing Group.

55. Horii Y. 1989 Acoustic analysis of vocal vibrato: A theoretical interpretation of data. *Journal of voice* **3**, 36–43.

56. Horii Y, Hata K. 1988 A note on phase relationships between frequency and amplitude modulations in vocal vibrato. *Folia Phoniatrica et Logopaedica* **40**, 303–311.

57. Herzel H, Reuter R. 1996 Biphonation in voice signals. In *AIP Conference Proceedings*, pp. 644–657. American Institute of Physics.

58. Wilden I, Herzel H, Peters G, Tembrock G. 1998 Subharmonics, biphonation, and deterministic chaos in mammal vocalization. *Bioacoustics* **9**, 171–196.

59. Herbst CT, Hertegard S, Zangger-Borch D, Lindestad P-Å. 2017 Freddie Mercury—acoustic analysis of speaking fundamental frequency, vibrato, and subharmonics. *Logopedics Phoniatrics Vocology* **42**, 29–38.

60. Borch DZ, Sundberg J, Lindestad P-Å, Thalen M. 2004 Vocal fold vibration and voice source aperiodicity in \ lqdist\rqtones: A study of a timbral ornament in rock singing. *Logopedics Phoniatrics Vocology* **29**, 147–153.

61. Inoue T, Shiozawa K, Matsumoto T, Kanaya M, Tokuda IT. 2024 Nonlinear dynamics and chaos in a vocal-ventricular fold system. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **34**.

62. Hono Y, Hashimoto K, Oura K, Nankaku Y, Tokuda K. 2021 Sinsy: A deep neural network-based singing voice synthesis system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **29**, 2803–2815.

63. Saitou T, Unoki M, Akagi M. 2005 Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis. *Speech communication* **46**, 405–417.

64. Sundberg J. 2006 The KTH synthesis of singing. *Advances in cognitive Psychology* **2**, 131–143.

65. Liuni M, Ardaillon L, Bonal L, Seropian L, Aucouturier J-J. 2020 ANGUS: Real-time manipulation of vocal roughness for emotional speech transformations. *arXiv preprint arXiv:2008.11241*

66. Aichinger P, Roesner I, Schneider-Stickler B, Leonhard M, Denk-Linnert D-M, Bigenzahn W, Fuchs AK, Hagmüller M, Kubin G. 2017 Towards objective voice assessment: the diplophonia diagram. *Journal of voice* **31**, 253-e17.

67. Švec JG, Schutte HK, Miller DG. 1996 A subharmonic vibratory pattern in normal vocal folds. *Journal of Speech, Language, and Hearing Research* **39**, 135–143.

68. Berry DA, Herzel H, Titze IR, Krischer K. 1994 Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions. *The Journal of the Acoustical Society of America* **95**, 3595–3604.

69. Berry DA, Zhang Z, Neubauer J. 2006 Mechanisms of irregular vibration in a physical model of the vocal folds. *The Journal of the Acoustical Society of America* **120**, EL36–EL42.

70. Giovanni A, Ouaknine M, Guelfucci B, Yu P, Zanaret M, Triglia J-M. 1999 Nonlinear behavior of vocal fold vibration: the role of coupling between the vocal folds. *Journal of Voice* **13**, 465–476.

71. Maunsell R, Ouaknine M, Giovanni A, Crespo A. 2006 Vibratory pattern of vocal folds under tension asymmetry. *Otolaryngology—Head and Neck Surgery* **135**, 438–444.

72. Steinecke I, Herzel H. 1995 Bifurcations in an asymmetric vocal-fold model. *The Journal of the Acoustical Society of America* **97**, 1874–1884.

73. Bailly L, Henrich N, Pelorson X. 2010 Vocal fold and ventricular fold vibration in period-doubling phonation: Physiological description and aerodynamic modeling. *The Journal of the Acoustical Society of America* **127**, 3212–3222.

74. Fuks L, Hammarberg B, Sundberg J. 1998 A self-sustained vocal-ventricular phonation mode: acoustical, aerodynamic and glottographic evidences. *KTH TMH-QPSR* **3**, 49–59.

75. Lindestad P-Å, Södersten M, Merker B, Granqvist S. 2001 Voice source characteristics in Mongolian “throat singing” studied with high-speed imaging technique, acoustic spectra, and inverse filtering. *Journal of Voice* **15**, 78–85.

76. Sakakibara K-I, Fuks L, Imagawa H, Tayama N, others. 2004 Growl voice in ethnic and pop styles. In *Proc. Int. Symp. on Musical Acoustics*,

77. Herbst CT. 2021 Performance evaluation of subharmonic-to-harmonic ratio (SHR) computation. *Journal of Voice* **35**, 365–375.

78. Titze IR. 2000 *Principles of voice production. Second printing*. Iowa City, IA: National Center for Voice and Speech.

79. Sun X. 2000 A pitch determination algorithm based on subharmonic-to-harmonic ratio. In *Sixth International Conference on Spoken Language Processing*, Citeseer.

80. Bedoya D, Arias P, Rachman L, Liuni M, Canonne C, Goupil L, Aucouturier J-J. 2021 Even violins can cry: specifically vocal emotional behaviours also drive the perception of emotions in non-vocal music. *Philosophical Transactions of the Royal Society B* **376**, 20200396.

81. Sun X, Xu Y. 2002 Perceived pitch of synthesized voice with alternate cycles. *Journal of Voice* **16**, 443–459.

82. Klatt DH, Klatt LC. 1990 Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America* **87**, 820–857.

83. Anikin A. 2019 Soundgen: an open-source tool for synthesizing nonverbal vocalizations. *Behavior research methods* **51**, 778–792.

84. Bergé P, Pomeau Y, Vidal C. 1987 *Order within chaos: Towards a deterministic approach to turbulence*. New York: John Wiley and sons.

85. Zollinger SA, Riede T, Suthers RA. 2008 Two-voice complexity from a single side of the syrinx in northern mockingbird Mimus polyglottos vocalizations. *Journal of Experimental Biology* **211**, 1978–1991.

86. Neubauer J, Edgerton M, Herzel H. 2004 Nonlinear phenomena in contemporary vocal music. *Journal of Voice* **18**, 1–12.

87. Briefer EF, Maigrot A-L, Mandel R, Freymond SB, Bachmann I, Hillmann E. 2015 Segregation of information about emotional arousal and valence in horse whinnies. *Scientific reports* **5**, 1–8.

88. Reby D, Wyman M, Frey R, Passilongo D, Gilbert J, Locatelli Y, Charlton B. 2016 Evidence of biphonation and source–filter interactions in the bugles of male North American wapiti (Cervus canadensis). *Journal of Experimental Biology* **219**, 1224–1236.

89. Volodina EV, Volodin IA, Filatova OA. 2006 *The Occurence of Nonlinear Vocal Phenomena in Frustration Whines of the Domestic Dog (Canis Familiaris)*. Slovenska akademija znanosti in umetnosti.

90. Suthers RA. 1990 Contributions to birdsong from the left and right sides of the intact syrinx. *Nature* **347**, 473–477.

91. Bhattarai B, Lee J. 2023 A Comprehensive Review on Music Transcription. *Applied Sciences* **13**, 11882.

92. Zhang J, Tang J, Dai L-R. 2016 RNN-BLSTM Based Multi-Pitch Estimation. In *Interspeech*, pp. 1785–1789.

93. Aichinger P, Pernkopf F, Schoentgen J. 2019 Detection of extra pulses in synthesized glottal area waveforms of dysphonic voices. *Biomedical signal processing and control* **50**, 158–167.

94. Aichinger P, Hagmüller M, Schneider-Stickler B, Schoentgen J, Pernkopf F. 2017 Tracking of multiple fundamental frequencies in diplophonic voices. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **26**, 330–341.

95. Zúñiga J, Reiss JD. 2019 Realistic procedural sound synthesis of bird song using particle swarm optimization. In *Audio Engineering Society Convention 147*, Audio Engineering Society.

96. Huffaker R, Huffaker RG, Bittelli M, Rosa R. 2017 *Nonlinear time series analysis with R*. Oxford University Press.

97. Herbst CT, Nishimura T, Garcia M, Migimatsu K, Tokuda IT. 2021 Effect of ventricular folds on vocalization fundamental frequency in domestic pigs (Sus scrofa domesticus). *Journal of Voice* **35**, 805-e1.

98. Zhang Y, Jiang JJ, Wallace SM, Zhou L. 2005 Comparison of nonlinear dynamic methods and perturbation methods for voice analysis. *The Journal of the Acoustical Society of America* **118**, 2551–2560.

99. Henríquez P, Alonso JB, Ferrer MA, Travieso CM, Godino-Llorente JI, Díaz-de-María F. 2009 Characterization of healthy and pathological voice through measures based on nonlinear dynamics. *IEEE transactions on audio, speech, and language processing* **17**, 1186–1195.

100. Shao J, MacCallum JK, Zhang Y, Sprecher A, Jiang JJ. 2010 Acoustic analysis of the tremulous voice: assessing the utility of the correlation dimension and perturbation parameters. *Journal of communication disorders* **43**, 35–44.

101. Awan SN, Roy N, Jiang JJ. 2010 Nonlinear dynamic analysis of disordered voice: the relationship between the correlation dimension (D2) and pre-/post-treatment change in perceived dysphonia severity. *Journal of Voice* **24**, 285–293.

102. Jiang JJ, Zhang Y, McGilligan C. 2006 Chaos in voice, from modeling to measurement. *Journal of Voice* **20**, 2–17.

103. Orozco JR, Vargas JF, Alonso JB, Ferrer MA, Travieso CM, Henríquez P. 2012 Voice pathology detection in continuous speech using nonlinear dynamics. In *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pp. 1030–1033. IEEE.

104. Vaziri G, Almasganj F, Behroozmand R. 2010 Pathological assessment of patients' speech signals using nonlinear dynamical analysis. *Computers in biology and medicine* **40**, 54–63.

105. Behrman A, Baken R. 1997 Correlation dimension of electroglottographic data from healthy and pathologic subjects. *The Journal of the Acoustical Society of America* **102**, 2371–2379.

106. Butte CJ, Zhang Y, Song H, Jiang JJ. 2009 Perturbation and nonlinear dynamic analysis of different singing styles. *Journal of Voice* **23**, 647–652.

107. Hegger R, Kantz H, Schreiber T. 1999 Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **9**, 413–435.

108. Blumstein DT, Recapet C. 2009 The sound of arousal: The addition of novel non-linearities increases responsiveness in marmot alarm calls. *Ethology* **115**, 1074–1081.

109. Ruiz-Monachesi MR, Labra A. 2020 Complex distress calls sound frightening: the case of the weeping lizard. *Animal Behaviour* **165**, 71–77.

110. Blumstein DT, Whitaker J, Kennen J, Bryant GA. 2017 Do birds differentiate between white noise and deterministic chaos? *Ethology* **123**, 966–973.

111. Anikin A, Pisanski K, Reby D. 2020 Do nonlinear vocal phenomena signal negative valence or high emotion intensity? *Royal Society open science* **7**, 201306.

112. Anikin A, Pisanski K, Massenet M, Reby D. 2021 Harsh is large: nonlinear vocal phenomena lower voice pitch and exaggerate body size. *Proceedings of the Royal Society B* **288**, 20210872.

113. Schoentgen J. 2001 Stochastic models of jitter. *The Journal of the Acoustical Society of America* **109**, 1631–1650.

114. Herbst CT. 2021 Registers-The Snake Pit of Voice Pedagogy. Part 2: Mixed voice, vocal tract influences, individual teaching systems. *Journal of Singing* **77**, 345–359.

Ethics

We do not report the results of any experiments in this paper.

Data Accessibility

Supplementary tables, vignettes, datasets, and R scripts for acoustic analysis and synthesis can be downloaded from the Open Science Framework (https://osf.io/gs8u3/,
DOI 10.17605/OSF.IO/GS8U3). These supplemental materials enable full validation and replication of results.

Competing Interests

The authors declare no competing interests.

Authors' Contributions

AA: conceptualization, formal analysis, software, writing – original draft. CTH: conceptualization, writing – review & editing.