emotion review

Emotion Review
Vol. 13, No. 1 (January 2021) 12–24
© The Author(s) 2020
ISSN 1754-0739
DOI: 10.1177/1754073920934544
https://journals.sagepub.com/home/emr

Beyond Correlation: Acoustic Transformation Methods for the Experimental Study of Emotional Voice and Speech

Pablo Arias*, Laura Rachman*, Marco Liuni and Jean-Julien Aucouturier
STMS UMR9912, IRCAM/CNRS/Sorbonne Université, France

Abstract

While acoustic analysis methods have become a commodity in voice emotion research, experiments that attempt not only to describe but to computationally manipulate expressive cues in emotional voice and speech have remained relatively rare. We give here a nontechnical overview of voice-transformation techniques from the audio signal-processing community that we believe are ripe for adoption in this context. We provide sound examples of what they can achieve, examples of experimental questions for which they can be used, and links to open-source implementations. We point at a number of methodological properties of these algorithms, such as being specific, parametric, exhaustive, and real-time, and describe the new possibilities that these open for the experimental study of the emotional voice.

Keywords
emotion, methods, transformation, voice

Introduction

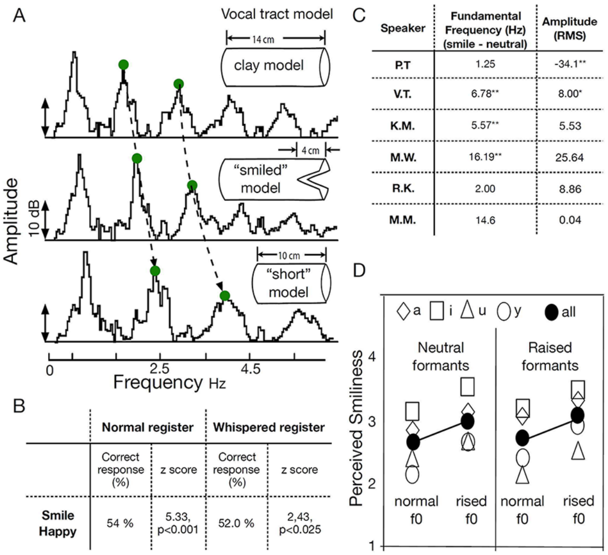
A typical approach to study how humans and other animals communicate emotions vocally uses acoustic analysis to quantify the physical features of vocalizations, such as their fundamental frequency (F0), intensity, or spectrum, and seek how they relate to the affect of the speaker (Bachorowski & Owren, 1995) or listener (for a review, see e.g., Juslin & Laukka, 2003; Scherer, 2003). Several well-known tools exist in the community for this purpose, including PRAAT[1] for speech data (Boersma & Weenink, 2002), OpenSMILE[2] (Eyben, Weninger, Gross, & Schuller, 2013), MIRToolbox[3] for song and musical data (Lartillot & Toiviainen, 2007), and Sound Analysis Pro[4] (Tchernichovski & Mitra, 2004) or Seawave[5] (Sueur, Aubin, & Simonis, 2008) for animal communication. The availability of dedicated software has an important impact on research: it gives access to audio signal-processing techniques such as F0 extraction without needing a technical background; it helps standardize the definition of vocal features by providing reference implementations (when one studies jitter while referencing PRAAT, others know what is meant and how to reproduce the work); and it provides an interdisciplinary interface between the research that creates these tools and the research that uses them, to share new techniques and new research needs.

One limitation of this methodology, however, is that it is intrinsically correlational. Analysing large corpora of speech or vocalizations to establish for example that happy voices have statistically higher F0s, faster rate, and more animated intonations (Banse & Scherer, 1996; for a recent review, see Kamiloglu, Fischer, & Sauter, 2020) does not allow us to conclude that these features are biological signals that are causally involved in the decoding of these emotions. For instance, it is now relatively well described that smiling, the bilateral contraction of the zygomatics facial muscles, has perceivable acoustic consequences on the speaking voice that can be heard, for example, on the phone (Tartter & Braun, 1994). How listeners recognize smiles in speech,

Author note: *Equal contribution. The first, second, and third authors produced the experimental data and software on which this review was based. The fourth author wrote the manuscript, with contributions from all other authors.

Marco Liuni is also affiliated to Alta Voce SAS, Houilles, France.

Corresponding author: Jean-Julien Aucouturier, STMS UMR9912, Institut de Recherche et Coordination Acoustique/Musique (IRCAM), 1 Place Stravinsky, Paris, 75004, France. *Email:* aucouturier@gmail.com



	Normal register		Whispered register	
	Correct response (%)	z score	Correct response (%)	z score
Smile Happy	54 %	5.33, p<0.001	52.0 %	2,43, p<0.025

Speaker	Fundamental Frequency (Hz) (smile - neutral)	Amplitude (RMS)
P.T	1.25	-34.1**
V.T.	6.78**	8.00*
K.M.	5.57**	5.53
M.W.	16.19**	25.64
R.K.	2.00	8.86
M.M.	14.6	0.04

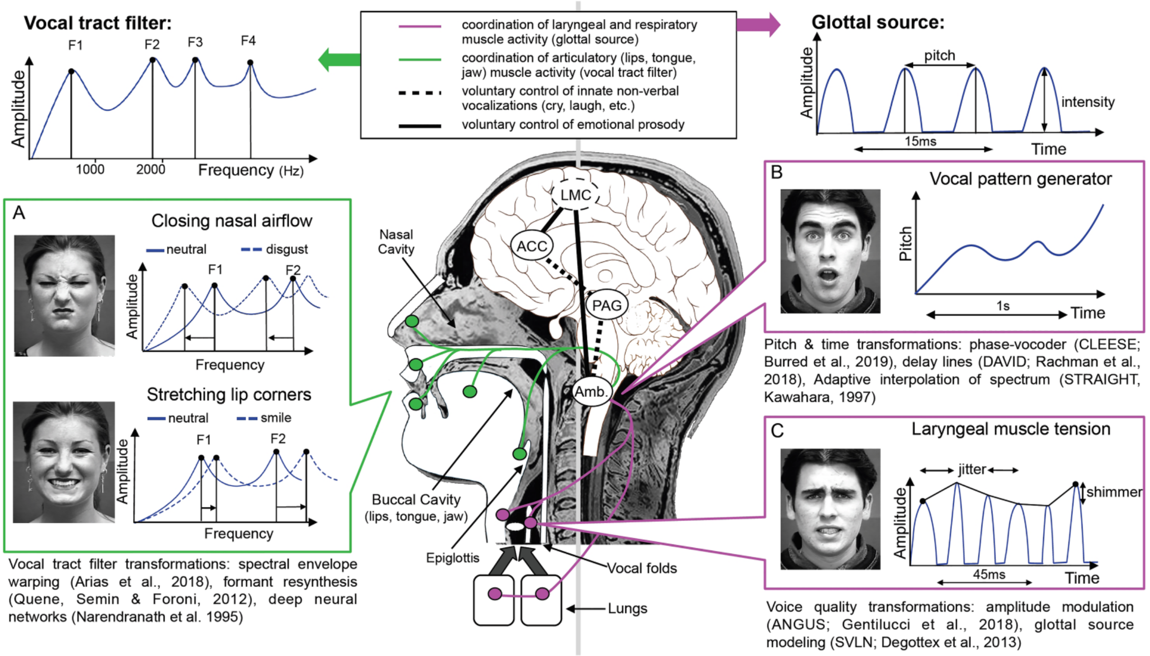
Figure 1. The confusing case of fundamental frequency in smiled speech. Ohala (1980) demonstrated that the resonant frequencies (formants) of a cylindrical clay model of the vocal tract are lowered when the shape of a smile is carved into it (A, middle), and that these new frequencies are similar to that of a non-smiling, but shortened cylinder (A, bottom). He concluded that smiling only has mechanistic consequences on the resonances of the vocal tract filter (A). This is confirmed by the fact that listeners are able to recognize smiled speech in both pitched and non-pitched (whispered) vocalizations (B, adapted from Tartter and Braun, 1994). However, it is typical that speakers simultaneously raise their pitch while smiling (C, adapted from Tartter, 1980), and thus listeners use higher pitch as a cue to identify smiled speech (D, adapted from Lasarcyk and Trouvain, 2008), even though it is not causally implicated in smiling. The sole acoustic analysis of speech corpora therefore cannot fully elucidate what cognitive mechanisms are involved in how smile is communicated vocally.

however, is remarkably complicated. On the one hand, physical models show that labial spreading, which reduces vocal tract length, has no mechanistic consequence on the F0 of the glottal source, but only on formant frequencies (see Figure 1A; Drahota, Costall, & Reddy, 2008; Ohala, 1980); and thus, listeners are able to identify smiled speech in the whisper register (see Figure 1B; Tartter & Braun, 1994), which has no audible F0. On the other hand, corpus analyses typically find strong associations between smiled speech and F0 (see Figure 1C; Barthel & Quené, 2015; Tartter, 1980), which suggests that it is neurologically difficult for speakers to smile without simultaneously raising their pitch; and thus, in the normal register, listeners reliably use pitch as a cue to identify smiled speech, even though it is not causally implicated in its production (see Figure 1D; El Haddad, Dupont, d'Alessandro, & Dutoit, 2015; Lasarcyk & Trouvain, 2008). It is therefore clear, in this field of study like in any other (Casadevall & Fang, 2008), that correlation does not imply causation and that the sole acoustic analysis of what is incidentally present in voice and speech may obfuscate the mechanisms with which emotions are produced, or recognized (see Armstrong, Lee, & Feinberg, 2019, for a similar argument on the signaling of body size by low F0 voices).

Rather than describing them, one would like the ability to manipulate the acoustic factors of interest in stimuli, in order to confirm experimentally that they causally lead to a change of behavior in the predicted direction when they are perceived. The manipulation of acoustic cues provides an approach complementary to corpus analyses, where the latter can establish a relation between two phenomena (e.g., shifted formants when people smile) and the former can be used to build a model and test for their involvement in perception[6] (see also Goldstone & Lupyan, 2016). Yet, while analysis tools are many, experiments that attempt to manipulate acoustic dimensions computationally in complex stimuli such as speech (Scherer, 1972), music (Ilie & Thompson, 2006), or animal vocalizations (Hienz, Jones, & Weerts, 2004) have been, until recently, remarkably rare. Perhaps because acoustic transformation tools are perceived to be too technical or of unsufficient quality, a steady stream of research has even preferred less flexible but more ethologically valid ways to manipulate vocal characteristics, such as immersing animals in heliox (Nowicki, Mitani, Nelson, & Marler, 1989; Rand & Dudley, 1993).

Two lines of research have significantly advanced the quest for acoustic control and causal inference in voice and speech research: vocal morphing and speech synthesis. On the one hand, morphing—an algorithmic method to combine two voices by interpolating their spectral features (Kawahara & Matsui, 2003)—has allowed researchers to describe, for example, how formants are processed to represent vocal identity (Latinus, McAleer, Bestelmeyer, & Belin, 2013), whether averaged voices are perceived as more attractive (Bruckert et al., 2010), or whether vocal emotions are perceived categorically (Laukka, 2005). However, morphing is generally performed between voices that differ in more than one acoustic dimension. For example, two morphed vocal identities that differ in their fundamental frequency, formant dispersion, and harmonic-to-noise ratio (HNR), will inevitably generate experimental conditions where these acoustic features covary (Latinus et al., 2013). On the other hand, speech synthesis—a vast family of methods allowing to create artificial vocal stimuli from scratch by specifying part or all of their physical parameters (for a review, see e.g., Govind & Prasanna, 2013; Malisz et al., 2019)—has been used, for example, to reveal that vocal emotions can be recognized in isolated pitch contours (Scherer & Oshinsky, 1977), to compare the emotional impact of various forms of nonlinearities (Anikin, 2019b), or test the effect of formant frequencies on the recognition of smiled speech (Quené, Semin, & Foroni, 2012). While allowing theoretically unlimited control over the physical properties of the stimuli, synthesis methods have the caveat of decontextualizing acoustic features (e.g., when replacing full speech with isolated pitch contours) and may also suffer from sonic artefacts (e.g., voices sounding robotic and artificial; although see this article's final prospective note on deep-learning techniques).

In our view, voice transformation—the technique to manipulate an original, natural vocal utterance in order to alter a specific acoustic dimension—provides a useful alternative[7] to the morphing and synthesis approaches. By giving experimenters the ability to predict how behavioral, physiological, or neural reactions vary depending on specific acoustic changes while leaving all other features unchanged, transformations are well



Vocal tract filter transformations: spectral envelope warping (Arias et al., 2018), formant resynthesis (Quene, Semin & Foroni, 2012), deep neural networks (Narendranath et al. 1995)

Pitch & time transformations: phase-vocoder (CLEESE; Burred et al., 2019), delay lines (DAVID; Rachman et al., 2018), Adaptive interpolation of spectrum (STRAIGHT, Kawahara, 1997)

Voice quality transformations: amplitude modulation (ANGUS; Gentilucci et al., 2018), glottal source modeling (SVLN; Degottex et al., 2013)

Figure 2. Neurological, anatomical, and acoustic characteristics of the vocal production pathway.

Note. Voice is produced when the expiratory airflow from the lungs sets the vocal folds of the larynx into oscillations (right). This sound wave, or glottal source, is resonated through the vocal tract and filtered by oral articulators such as the tongue and lips (left), and is finally radiated from the mouth and nose. Neural control at different levels of this pathway involves a hierarchy of cortical and subcortical systems, including premotor nuclei of the brainstem such as the nucleus ambiguus (Amb.), periaqueductal gray (PAG), anterior cingulate cortex (ACC), and laryngeal motor cortex (LMC). Voice-transformation techniques exist to parametrically manipulate a prerecorded vocal signal at all levels of the voice production pathway: vocal tract filter (A), pitch and intonation (B), and glottal source timbre (C).

adapted to the nature of hypothesis-driven experimental research. Furthermore, because of recent improvements in vocal transformation algorithms and their increasing availability in the computer-science communities (Mohammadi & Kain, 2017; Stylianou, 2009), it may be time to consider these technologies part of the toolbox of vocal emotion researchers.

In this article, we review the neurological and acoustic bases of emotional voice production, and show that there are in fact well-established, high-quality algorithms that are able to parametrically transform a prerecorded vocal signal at all levels of the voice production pathway. We provide sound examples from our own recent work that illustrate these techniques, examples of experimental questions for which they can be used, and links to open-source software to replicate and extend these studies. Finally, we point at a number of methodological properties of these algorithms, such as being specific, parametric, exhaustive, and real-time, and describe the new possibilities and questions that these entail for experimental research.

Voice Transformations Along the Vocal Production Pathway

Voice is produced when the expiratory airflow from the lungs, generated by thoracic and abdominal muscles, sets the vocal folds of the larynx into oscillations. This sound wave, or glottal source, is resonated through the vocal tract, filtered by oral articulators such as the tongue and lips which amplify certain bands of energy (or formants) in its frequency spectrum, and is finally radiated from the mouth and nose (see Figure 2; Titze, 1994). While the vibration of the vocal folds is a passive process, their oscillatory properties, the airflow that sets them into motion, and the resonance characteristics of the vocal tract are all controlled by over 100 respiratory, laryngeal, and oro-facial muscles (Simonyan & Horwitz, 2011), whose motoneurons originate from the spinal cord and brainstem.

In emotional vocalizations, neural control over these muscles involves a hierarchy of cortical and subcortical systems, including the periaqueductal gray (PAG), anterior cingulate cortex (ACC), and laryngeal motor cortex (LMC). These subcortical and cortical influences on muscle actuators at every stage of the vocal production pathway have different, complementary effects on the final acoustic properties of the vocal signals, and specific voice-transformation techniques exist to reproduce these changes in ecological voice and speech recordings.

Glottal Source Transformations

Changes in the subglottal pressure due to the contraction of thoracic and abdominal muscles, which are controlled from the anterior horn of the spinal cord, primarily lead to modulations of voice intensity. At moderate intensities, such as in happy, aroused voices compared to calm or sad voices, the effect of the modulation is carried linearly through the vocal pathway and can be simulated with a simple scalar multiplication of the recording's root mean square (RMS) intensity (see e.g., Ilie & Thompson, 2006) or, for arbitrary intensity profiles, a piecewise linear function as implemented for example in the reverse-correlation toolbox CLEESE[8] (Burred, Ponsot, Goupil, Liuni, & Aucouturier, 2019).

Increased airflow, such as in pain cries or anger shouts, but also possibly altered neurological control over the laryngeal muscles, such as in stress or anxiety, may drive the vocal folds into nonlinear/chaotic oscillatory regimes and, more generally, change the shape and periodicity of glottal pulses, resulting in audible alterations of sound quality such as roughness, noisiness, or breathiness (see Figure 2C). In voice measurements, such nonlinearities are often analysed in terms of jitter and shimmer (cycle-to-cycle variations in the period and amplitude of glottal pulses, respectively) and harmonic-to-noise ratio (HNR; Boersma & Weenink, 2002). Such modulations of vocal source quality are important in emotional behaviors (Gobl & Chasaide, 2010; Johnstone & Scherer, 1999) and have been related, in listeners, to subcortical processing by the amygdala (Arnal, Flinker, Kleinschmidt, Giraud, & Poeppel, 2015).

Not all glottal source changes are easily simulated with voice transformations. In theory, analysis-resynthesis techniques can be used to, first, estimate the recording's series of glottal pulses (Degottex, Lanchantin, Roebel, & Rodet, 2013) and then, resynthesize the vocal signal from a manipulated series of pulses with artificially varied amplitude and period (Bõhm, Audibert, Shattuck-Hufnagel, Németh, & Aubergé, 2008; Ruinskiy & Lavner, 2008; Verma & Kumar, 2005). However, these techniques rely on an explicit model of pulse variability, which is typically learned from one or several target examples of naturally rough voices (Bonada & Blaauw, 2013), and it is unclear how such predetermined patterns should be selected for arbitrary voices. Moreover, because of the computational complexity of the initial stage of glottal source estimation, these techniques cannot operate in real time. Alternative approaches can also simulate variations in pulse periodicity by overlapping randomly time-shifted copies of the original recording (Loscos & Bonada, 2004) or modulating it at a divider of F0 to create subharmonics, as implemented for example in the ANGUS toolbox[9] (Gentilucci, Ardaillon, & Liuni, 2019). However, these only allow exploring a subset of all possible nonlinearities (e.g., subharmonics, but not biphonation in general), and vocal source spectrum is one area of vocal production for which pure speech-synthesis approaches, in which variability in the glottal shape or periodicity can be specified explicitly (Anikin, 2019a; Brady, 2005), may provide more experimental control than transformations.

> **Sound Example S1:** Female singing voice, first: original; second: manipulated with subharmonics with the ANGUS toolbox (Gentilucci et al., 2019). All sound examples are available as supplemental material.

Note that, because vocal folds or analog anatomical structures are present in a large number of species, source nonlinearities are not unique to human vocalizations but are used as a signal of threat and alarm by a wide range of animals, including primates (Fitch, Neubauer, & Herzel, 2002), but also rodents (Blumstein & Recapet, 2009), canids (Wilden, Herzel, Peters, & Tembrock, 1998), whales (Tyson, Nowacek, & Miller, 2007), and birds (Fee, Shraiman, Pesaran, & Mitra, 1998). The use of glottal source transformations can therefore be extended to the study of animal behavior.

Changes in the oscillatory properties of the vocal folds, linked for example to their length and opening, are mostly controlled by the intrinsic laryngeal muscles, innervated from the vagal nerve originating in the nucleus ambiguus (Amb.) of the medulla, and lead to modulations of vocal source timbre (as aforementioned) but also, and perhaps most importantly, of vocal F0 (see Figure 2B). A range of techniques exist to manipulate F0. Simple algorithms, as used for example in altered auditory feedback research (Hain, Burnett, Larson, & Kiran, 2001) and implemented in the DAVID toolbox[10] (Rachman et al., 2018), are based on resampling or multiple delay lines (a technique that introduces a small delay to an audio signal in order to play it faster/slower, thus raising/lowering its pitch; Dattorro, 1997) and may alter vocal tract filtering or formants unrealistically beyond small parametric changes. State-of-the-art techniques that allow separating source and filter information to avoid such artefacts are based on reconstructions of the signal's short-time Fourier transform (STFT) at nonuniform rates, such as the pitch synchronous overlap and add (PSOLA) method as implemented for example in PRAAT (Boersma & Weenink, 2002); the phase-vocoder method (Moulines & Laroche, 1995) as implemented for example in CLEESE (Burred et al., 2019); or pitch-adaptive analyses techniques such as the adaptive interpolation of weighted spectrum method as implemented in STRAIGHT[11] (Kawahara, 1997). These transformation methods not only allow raising or lowering the mean pitch of a recording, which may correspond to a baseline change of valence (see Figure 3A; see e.g., Ilie & Thompson, 2006), but can also manipulate the difference between the instantaneous and mean F0 to exaggerate or lessen variations, as seen for example in fearful versus sad vocalizations (see Figure 3B; see e.g., Pell & Kotz, 2011); create parametric F0 contours such as vibrato in anxious voices (Figure 3C; see e.g., Bachorowski & Owren, 1995), or local intonations at the start or end of an utterance, as in surprised or assertive speech (see Figure 3D; see e.g., Jiang & Pell, 2017). In addition, most pitch-shifting methods can also be used to manipulate the duration or speech rate of utterances (a process known as time-stretching), producing faster or slower speech in positive or negative emotional states (Scherer & Oshinsky, 1977).

> **Sound Example S2:** Male speech, first: original; second: pitch increased by 50 cents; third: pitch decreased by 50 cents; fourth: pitch modulated with an 8 Hz vibrato. All transformations made with delay lines, using the DAVID toolbox (Rachman et al., 2018).

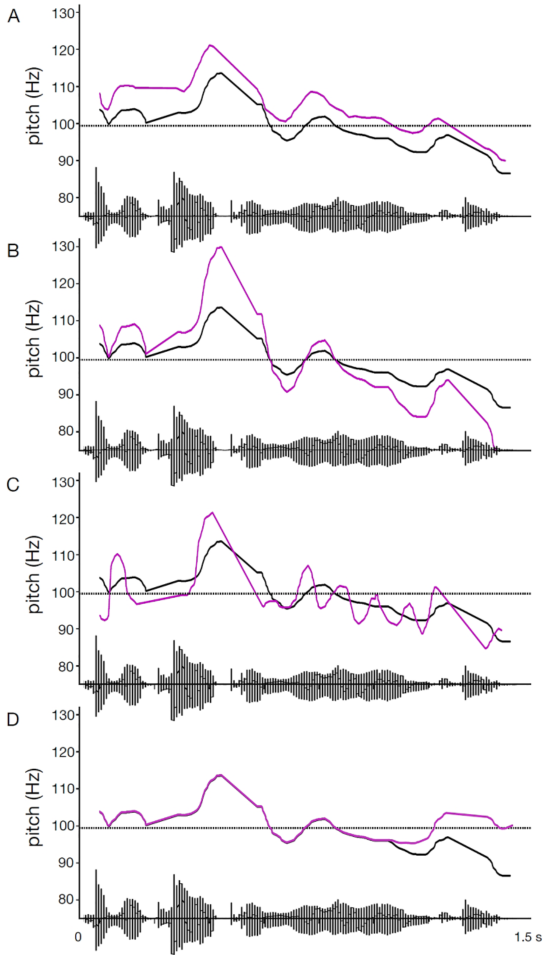


Figure 3. Four examples of pitch transformations on a single recording of the sentence "I would like a new alarm clock," produced by a male English speaker.

Note. The original pitch values are presented in black and the transformed pitch values in magenta. The speech waveform is shown on the x-axis, and the dotted line indicates the mean F0 of the recording. (A) 100-cent upward pitch shift, applied uniformly over time. (B) Exaggerated pitch dynamics by a 15% increase of pitch values with respect to the mean pitch. (C) Vibrato applied with a 100-cent depth and a rate of 8.5 Hz. (D) Increasing pitch at the end of the utterance.

Sound Example S3: Female speech, first: original; second–sixth: random pitch intonations, generated on six successive time-windows, using Gaussian distributions centered at ±0 cents, *SD* = ±200 cents. All transformations made with a phase-vocoder, using the CLEESE toolbox (Burred et al., 2019).

Sound Example S4: Female speech, first: original; second–sixth: random variations of speed rate, generated on six successive time-windows, using Gaussian distributions centered at ±0%, *SD* = ±10% original duration. All transformations made with a phase-vocoder, using the CLEESE toolbox (Burred et al., 2019).

Vocal Tract Transformations

The shape and resonating characteristics of the vocal tract, and thus, the spectral properties of the sound, are modulated by the articulators of the supraglottal region (e.g., lips, tongue, jaw), which are controlled by the oro-facial motoneurons of the facial and trigeminal nuclei of the pons (see Figure 2A), but also by the extrinsic laryngeal muscles, whose motoneurons originate from the hypoglossal nucleus of the medulla and which have the ability to raise or lower the position of the larynx within the neck (Titze, 1994). As for the glottal source, a number of techniques exist to transform vocal tract characteristics without altering other aspects of vocal production. Simple techniques, such as the PRAAT "change gender" method (Boersma & Weenink, 2002), exploit the side effects of the resampling pitch-shifting method to reduce or increase formant dispersion, thus simulating changes in vocal tract length and physical dominance (Boidron, Boudenia, Avena, Boucheix, & Aucouturier, 2016).

Other techniques enabling the manipulation of individual formant frequencies include formant resynthesis methods (Quené et al., 2012), spectral envelope manipulations by frequency warping (Arias, Soladie, et al., 2018), and neural networks (Narendranath, Murthy, Rajendran, & Yegnanarayana, 1995). While vocal tract characteristics have been mainly studied in the context of speaker identity (Mohammadi & Kain, 2017), recent work has suggested they are also actively manipulated by emotional oro-facial gestures such as those involved in the expression of disgust (Chong, Kim, & Davis, 2018) or smiling (Arias, Belin, & Aucouturier, 2018).

Sound Example S5: Female speech, first: original; second–sixth: random variations of vocal tract filter, generated with spectral envelope frequency warping using the CLEESE toolbox (Burred et al., 2019).

Sound Example S6: Female speech, first and third: original; second and fourth: formant shifted upwards to simulate smiling, generated with frequency warping (Arias, Soladie, et al., 2018).

As for glottal source, vocal tract transformations are also relevant beyond human voice and speech, for the study of animal communication. Formant dispersion correlates with body size in *Rhesus macaque* monkeys (Fitch, 1997), and dominant frequency (that of the formant with the highest amplitude) inversely correlates with body size across 91 mammalian species (Bowling et al., 2017). Researchers have used formant manipulations of animal vocalizations in playback experiments to show for example that red deer stags are more attentive and reply more to calls from bigger conspecifics (Reby et al., 2005), or that female koalas spend more time attending to vocalizations simulating larger males (Charlton, Ellis, Brumm, Nilsson, & Fitch, 2012).

Properties of Voice Transformations for Experimental Research

Voice-transformation techniques have a number of methodological properties that make them well suited for experimental

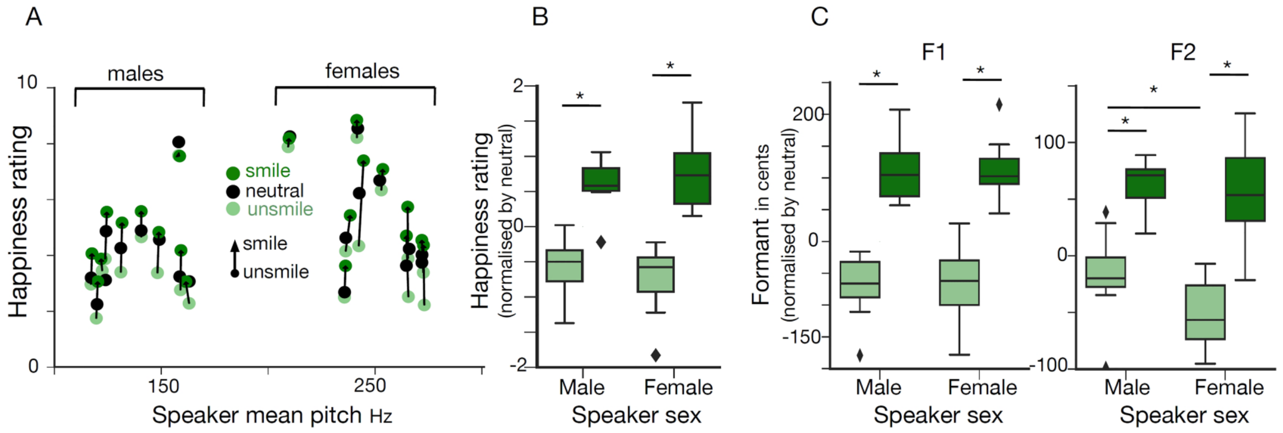


Figure 4. Attributed ratings of speaker happiness to manipulated vocal tract cues in smiled speech.
Note. (A) Mean rating of happiness for 20 male and female recordings as a function of the recording's mean pitch and sound manipulation: smile (dark green), neutral (black), unsmile (light green); each dot represents one separate recording; black arrows link unsmile with smile transformations of the same original recording. (B) Normalized ratings of happiness for male and female recordings manipulated with smile (light green) and unsmile (dark green) transformations; asterisks indicate significant statistical differences between the smile and unsmile distributions, computed with paired *t* tests ($p < .01$); all ratings normalized by the corresponding neutral (nonmanipulated) rating. (C) Difference of formant frequency (in cents) between smile (dark green) and unsmile (smile green) transformations and the corresponding nonmanipulated sound (Formant 1: F1, Formant 2: F2); asterisks indicate significant statistical difference between the distributions, computed with paired *t* tests ($p < .01$).
Data in Figure 4 collected at the Sorbonne-Universit INSEAD Center for Behavioural Sciences.

research, and allow experimental paradigms that would not be otherwise possible with traditional analysis methods or using actor-recorded vocalizations. We list and comment here on a number of these properties, namely those of specificity, parametricity, exhaustivity, and real-time behavior. Despite their advantages for research, assessing whether a given acoustic transformation possesses each of these properties also brings a number of methodological and theoretical difficulties, which we discuss next.

Specificity

Vocal behavior is highly multidimensional and, either for anatomical or neural reasons, it is nearly impossible for speakers to produce for instance only the pitch aspects of a sad expression without simultaneously varying timbre or duration, or timbre aspects without simultaneously varying pitch (see Figure 1). Covariation of features in naturalistic recordings means that one cannot conclusively establish what acoustic property drives listeners' emotional evaluations. The ability to transform voice along specific signal dimensions, such as pitch or timbre, while preserving all other aspects of the original, opens the possibility to create pairs of stimuli (original and transformed) that differ only in one experimental factor. Comparing behavioral or physiological measures within pairs thus allows controlling for covarying factors and identifying causal relations that would otherwise be missed in noise or, worse, spuriously attributed to the wrong factor.

Figure 4A illustrates this situation with recent data from Arias, Belin, and Aucouturier (2018), in which we manipulated 20 original male and female recordings with a vocal tract transformation designed to simulate the acoustic effect of smiling, and asked 35 participants to evaluate their perceived happiness of the speaker. Each recording was transformed into two variants, smiled and unsmiled versions, respectively expanding or compressing formants by the same amount. These two versions are thus two opposite transformations relative to a neutral, nonmanipulated stimulus. Because the transformation did not affect glottal source properties, the F0s of the original recordings were preserved by the two transformations, and ranged from 117 to 300 Hz. While there was a main effect of the smile transformation on ratings of perceived happiness within each triplet of recordings, $\chi^2\,(12) = 55.2$, $p = 1.0\text{e}{-12}$, variation of ratings across triplets was much greater because of the varied phrase content (positive vs. negative semantics), prosody, speaker identity, etc., of the original recordings. These variations would likely mask the smaller yet remarkably consistent effect of vocal tract properties if investigated with another experimental design.

Caveat: Feature specificity is not a given in signal processing, though. For instance, pitch-shifting algorithms based on harmonizers, as used in the altered vocal feedback literature, do not attempt to separate glottal and vocal tract characteristics, and thus have side effects on the signal's formants. Perhaps because of their technical nature, these effects are not always properly acknowledged in the psychological literature even though they may confound results attributed to the manipulation:

> The harmonizer shifts all frequencies, voice F0 as well as formants, and thus the shifted feedback signal sounds like a person's normal voice at a different F0 (details of the pitch-shifting algorithm are a trade secret of the manufacturer and are thus unavailable). (Hain et al., 2001: 2147)

Ensuring the right level of specificity, that is, using traditional acoustic analysis methods such as PRAAT to validate the absence of effect of the transformation on vocal characteristics that could have a confounding impact on the measures of interest, should therefore be high on the agenda for researchers aiming to adopt these methodologies. This will be greatly helped by the availability of open-source, validated software whose possibilities and limitations are well documented.

Parametricity

A common research question aims to compare how emotional expressions are processed across stimulus conditions, testing for example whether listeners perceive emotional speech differently in their native or a foreign language (Pell, Monetta, Paulmann, & Kotz, 2009; Scherer, Banse, & Wallbott, 2001), on familiar on nonfamiliar speakers (Chen, Kitaoka, & Takeda, 2016), male or female voices (Bonebright, Thompson, & Leger, 1996), or even across speech and music (Juslin & Laukka, 2003). When using emotional stimuli produced by actors, differences in decoding performance across groups may arise either because of production or perception differences. For instance, if French listeners have difficulties processing emotional cues spoken by Japanese speakers, it may be because their auditory representations of the Japanese phonetic inventory are poor (a perception effect; Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999), because one does not use the same cues to express joy in Japanese and in French (a production effect; Kitayama, Mesquita, & Karasawa, 2006), or both. Balanced designs, in which stimuli produced by more than two culture groups are presented to all participants cross-culturally, is a useful strategy to disentangle such decoding and encoding differences: encoding differences should result in poor recognition across nonnative groups, while decoding differences should result in impaired recognition only for a given group, but across stimuli from the other groups (Matsumoto, 2002; Yoshie & Sauter, 2019). Once an in-group advantage has been established, however, transformation tools that can parametrically manipulate stimuli then become useful to examine the causal mechanisms that underlie such cross-cultural differences.

In programming language theory, parametricity is the property of a function that can handle input values identically without depending on their type (Pierce & Benjamin, 2002). We use the term here to describe the property of voice transformations having a uniform acoustic effect regardless of the characteristics of the original signal, their speaker, content, or culture. Parametricity opens the possibility to create emotional voice stimuli which utilize exactly the same prosodic cues in exactly the same manner (e.g., a 50-cent pitch increase in the first 2 seconds), and thus to separate the effect of production and perception in intergroup differences.

Data in Figure 4B illustrate this possibility by comparing the effect of the same smile transformation on a variety of recorded sentences by both male and female speakers. Even though the acoustics of male and female utterances widely differ, notably because of pitch dimorphism (Titze, 1994), listeners' ratings of speaker positivity in both smile and unsmile transformations can be normalized with respect to original ratings, which allows us to compare the effect of identical cues between sexes. In the same experimental logics, audio transformations have been used to compare identical cues on speech, music, and environmental sounds (Ilie & Thompson, 2006; Ma & Thompson, 2015).

Caveat: As for specificity, a number of technical considerations may reduce a given algorithm's parametricity. Some may result from algorithmic design choices. For instance, because F0/pitch is ill-defined on unvoiced portions of speech, pitch transformations, such as vibrato, are often designed to only operate on voiced sections of the signal, leaving transients untransformed. Transformations like vibrato therefore rely on the availability of relatively long voiced portions in phonemes (e.g., 250 ms for two cycles of an 8-Hz vibrato to be perceived) and, even if specified with identical parameters, may not have as much physical effect on speakers with a faster rather than slower speech rate, or languages with a larger rather than smaller consonant/vowel ratio (see e.g., the failed generalization to Swedish in Rachman et al., 2018). Other breaches of parametricity may result from algorithmic limitations. Figure 4C illustrates this situation: despite the identical transformation parameters, the smile transformation as implemented in Arias, Belin, and Aucouturier (2018) unintentionally leads to smaller physical changes of formant frequency F2 (as measured by PRAAT) in male voices than in female voices, making the higher ratings of perceived smiliness measured on transformed female voices difficult to attribute to a purely perceptual effect. It is therefore crucial for researchers to develop a technical understanding, as well as the methodological know-how, to critically assess a transformation's parametricity.

Exhaustivity

Although the human vocal apparatus can produce many different vocal sounds, languages generally use only a small subset of these sounds, and these are not uniformly distributed between and within languages (De Boer, 2000). Many psychophysical procedures, such as adaptive staircase methods but also reverse correlation (Jack & Schyns, 2017), require presenting participants with random distributions of stimuli in feature space, which is often impractical using naturalistic vocalizations (Belin, Boehme, & McAleer, 2017). With voice transformations, experimenters can uniformly or adaptively sample a large space of prosodic variations, for example, all vibrato frequencies between 1 and 10 Hz, regardless of how common these may be in actual behavior, and are thus able to draw better inferences about how these features are processed.

For instance, in Ponsot, Burred, Belin, and Aucouturier (2018), we used a phase-vocoder transformation to generate more than 70,000 random prosodic variations from a single recording of the word "*bonjour*" (hello), and used reverse correlation to uncover participants' mental representations of a dominant or trustworthy way to pronounce that word. Similar paradigms can be applied to study how healthy participants and patients mentally represent emotional prosody, and address for example emotion perception deficits in pathologies like autistic spectrum disorder (ASD) or amusia.

Caveat: While their parameters can be explored exhaustively, most transformation techniques do not implement boundaries on what is or is not physiologically possible and thus, beyond a certain parameter range, may not sound like authentic human speech. In addition, even within realistic parameter ranges, transformations may generate artefacts such as unnatural timbres, doubling of F0, or smearing of transients, and give the transformed sound a robotic, artificial tone. Even if the impact of artificiality on emotional judgements remains unclear (Burleigh, Schoenherr, & Lacroix, 2013), one can rightly question the ecological validity of behaviors measured in response to such stimuli. While most artefacts can be avoided with simple heuristics (e.g., basing parameters on the range of variation measured in natural voices, or clipping sampling distributions at ±2 *SD*), quantifying how natural, or easily detectable, a given transformation sounds to participants can be remarkably complicated. First, the acceptance of transformed voice as authentic is heavily dependent on context. For instance, transformations may be more easily detected in situations where variants can be compared with the original sound. Second, judgements of naturalness are multifaceted, and likely incorporate evaluations of biological plausibility ("this doesn't sound human"), agentivity ("this sounds tampered with"), vocal or social typicality ("no one in their right mind would do this"), or even semantics ("it does not make sense to say this with a happy voice"). Future research will benefit from more principled ways to measure transformation naturalness and its impact on participants.

Real-Time Behavior

One typical way to study the role of emotional expressions in social interactions is either to explicitly instruct social partners to display a certain emotion (e.g., Tice, 1992) or to indirectly lead them to express it using a cover story (e.g., van Doorn, Heerdink, & van Kleef, 2012). Because many voice-transformation techniques allow real-time processing (typically, 50–100 ms latency for transformations based on phase-vocoder; E. Lee, Karrer, & Borchers, 2007), they open the possibility to control emotional expression in continuous, real-time interactions (e.g., on the phone) with no experimental demand, and possibly even without participants' awareness of the manipulation. For instance, we have used a vocal tract transformation to manipulate the perceived body size of mock patients calling a medical call center simulator (Boidron et al., 2016), and found that callers whose voice was perceived as indicative of physical dominance obtained a higher grade of response, a higher evaluation of medical emergency, and longer attention from physicians than callers with strictly identical medical needs whose voice signaled lower dominance. Similar paradigms can be used for example to study how congruent or incongruent emotional expressions influence the outcome of group decision-making, or group creativity.

In addition, some voice-transformation techniques, such as those operating in the time domain (Juillerat, Schubiger-Banz, & Arisona, 2008), can be so fast that they can not only manipulate a social partner's voice without disrupting the flow of interaction, but also manipulate the participant's own voice without disrupting their sensorimotor feedback (e.g., with less than 20 ms latency between the original input voice and the manipulated output). This opens the possibility to build altered auditory feedback paradigms and test how, for example, hearing one's voice with a happier or sadder tone influences one's emotional experiences, judgments, or decisions. For instance, we used a time-domain transformation of pitch to modify participants' voices in a happy or sad direction as they read out an emotional neutral text, and found that participants who heard themselves with emotionally manipulated voices reported significantly different moods, as well as elevated levels of skin conductance (Aucouturier et al., 2016). Similar paradigms can be used to study for example whether personal emotional memories can be reencoded with different valence when healthy participants or patients hear themselves narrate them with a transformed tone of voice.

Caveat: Once given the possibility to transform continuous speech in real-time interactions, the immediate next question concerns the contextual adaptation of transformations to speech content. While transformation parametricity is methodologicaly useful, applying the same pitch increase on spoken sentences regardless of, say, their original prosody or stress words (which are also marked with pitch; Pell, 2001) may create unnatural or misadaptative expressions, and it is computationally unclear how to adapt transformations to speech content, especially in a real-time context. In that respect, voice transformations should not be considered experimental materials, but experimental methods. They do not replace actor-produced stimuli nor provide ready-made expressions with which to study the impact of emotions on behavior (in truth, transformed vocalizations may be less intense, less natural, and less well-recognized than natural expressions; see e.g., Rachman et al., 2018). Rather, by manipulating signal properties at all stages of the vocal production pathway, they provide control over the physical properties of the stimuli, bringing unprecedented precision on the neurological, anatomical, and acoustic components of what makes voice and speech emotional, but leaving it to the experimenter to construe how speakers and listeners integrate these components with other aspects of affective and cognitive processing.

A Prospective Note on Deep-Learning Techniques

In recent years, artificial voices have become an integral part of consumer electronic appliances (e.g., smart assistants, car navigators, augmentative and alternative communication), and the amount of funding private companies such as Amazon, Google, or Apple have injected into the speech synthesis community has transformed the field into a fast-paced and competitive domain. In just a few years, neural machine learning (deep neural networks [DNNs]) operating on waveform samples, as opposed to, for example, spectrogram features, has become the de facto standard in voice synthesis (van den Oord et al., 2016).

The DNN architecture, which directly models waveforms using a neural network method trained with recordings of real speech, provides superior acoustic quality for several high-level vocal applications such as text-to-speech (Y. Wang et al., 2017) and speech enhancement (Pascual, Bonafonte, & Serra, 2017), but also to generate expressive voices (Akuzawa, Iwasawa, & Matsuo, 2018; Y. Lee, Rabiee, & Lee, 2017) or convert their emotions (Luo, Chen, Takiguchi, & Ariki, 2017). However, although highly flexible, these systems have so far failed to exhibit the level of parametric control that we argue here is needed for experimental applications. First, machine-learning speech generally emulates vocal patterns learned from large sets of recordings where all vocal features covary, and therefore typically lacks feature specificity and exhaustivity. Second, deep-learning architectures do not allow easy introspection into how information is represented in the network, making it difficult to know what exact vocal features are being manipulated and lacking interpretability. As far as we know, these limitations have so far prevented the application of DNNs for the kind of experimental work reviewed here (although see Sun, Anumanchipalli, & Chang, 2019).

However, progress in the field is fast. One promising line of research uses post hoc, data-driven methods to reveal how stimulus information is encoded into network layers (Hsu, Zhang, & Glass, 2017). In the visual sciences, these methods have been used to compare what visual features human and machine use to achieve face classification (Xu et al., 2018), and similar approaches could be used for speech. Another relevant line of research aims to create DNN-based vocoders (Wu, Hayashi, Tobing, Kobayashi, & Toda, 2019), in which, like in the traditional vocoders reviewed before, speech synthesis can be controled with specific pitch or duration parameters while conserving the acoustic performance of deep-learning models. Finally, generative adversarial networks (GANs), a special class of DNN architecture capable of learning a deterministic mapping from one style of stimulus to another (Goodfellow et al., 2014), are increasingly used to create visual transformations (e.g., smiles; W. Wang et al., 2018) and have also started to be applied to speech transformations. For example, GANs were recently used to transform a voice into its Lombard counterpart (a particular type of vocal effort which makes the voice more intelligible in background noise; Seshadri, Juvela, Alku, & Räsänen, 2019). All such advances open exciting new possibilities to create emotional voice and speech transformations, which will certainly find their way to the community in the upcoming years.

Conclusion

So far, the experimental study of emotional voice and speech has largely relied on acoustic analyses of datasets of natural emotional vocalizations, or the use of these recordings as stimuli in experimental research. While natural vocalizations have the advantage of being realistic and ecological, they often vary in several acoustic dimensions, making mechanistic conclusions difficult to establish. Yet, much is known about the essential anatomic building blocks involved in emotional vocal production (e.g., laryngeal muscle tension, vocal fold oscillatory regime, oro-facial gestures). These anatomic mechanisms are controlled by increasingly well-identified neural structures and have specific, and to some extent independent, acoustic consequences that can be modeled computationally—because they have a physical basis.

In this article, we reviewed a wide range of recent (or not so recent) technologies that allow researchers to manipulate specific acoustic features along the voice production pathway. From the glottal source to the vocal tract, we presented the acoustic consequence of each of the building blocks of the vocal apparatus, as well as corresponding acoustic models and transformation algorithms from the signal-processing literature.

We suggest that using such transformations to control the content of vocal stimuli in experimental research is a promising line of work. This methodology allows researchers to formulate and test computational predictions about the behavioral, physiological, and neural consequences of specific acoustic changes, enabling them to draw strong, causal links between the anatomic mechanisms involved in voice production and their subsequent reactions in listeners. Transformation technologies can easily be shared between research groups, made open source, and deployed across several types of studies (e.g., cross-linguistic, cross-species) and auditory modalities, such as nonverbal behavior, speech, and music. When possible, we gave here links to some of these tools that are available freely, as well as examples of studies that use them, and hope that this list will only be growing.

However, in order for these technologies to be useful in an experimental context, they must deliver the proper type of acoustic control. We identified four of such constraints. First, in our view, transformations should be specific, that is, transform sound in a single acoustic dimension mirroring an isolated anatomic mechanism (e.g., vocal fold saturation or zygomatic muscle contraction). Second, transformations should be parametric, that is, have a uniform acoustic effect regardless of the original signal (e.g., semantic content, age, identity, sex, species), thus allowing comparative studies. Third, transformations should be exhaustive, that is, unconstrained by what speakers usually produce, but rather by what they *can* produce, in order to reduce sampling bias for psychophysical research. Finally, in an era where the study of social interactions is at the top of the cognitive-science research agenda, the community should favor transformations that can operate in real time. We hope that these recommendations can be used as a “check-list” for machine-learning and signal-processing researchers involved in creating new vocal transformations. If new tools, included those emerging from the recent trend of deep-learning research, follow these constraints, they will be more easily used for experimental research.

Acknowledgement

We thank Gilles Degottex for feedback on Figure 1.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Work funded by ERC StG CREAM (335536) and ANR REFLETS (2017-2021) to the fourth author.

ORCID iD

Jean-Julien Aucouturier https://orcid.org/0000-0002-4477-4812

Supplemental Material

Supplemental material for this article is available online.

Notes

1 http://www.praat.org/
2 https://www.audeering.com/opensmile/
3 https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox
4 http://soundanalysispro.com
5 http://rug.mnhn.fr/seewave/
6 It should be noted that transformations do not help to establish the involvement of the cues in the production/expression of emotion; for this, one has to manipulate the actual emotional state of the signalers (e.g., Bachorowski & Owren, 1995).
7 Note that, strictly speaking, there is a technological overlap between voice transformation and certain forms of (re)synthesis techniques, which analyse an original signal down to its generative parameters and then resynthesize a variant of the signal by altering these parameters. We review some of these techniques in the rest of the article under the encompassing term of transformation.
8 http://forumnet.ircam.fr/product/cleese
9 http://forumnet.ircam.fr/product/angus
10 http://forumnet.ircam.fr/product/david
11 https://github.com/HidekiKawahara/legacy_STRAIGHT

References

Akuzawa, K., Iwasawa, Y., & Matsuo, Y. (2018). *Expressive speech synthesis via modeling expressions with variational autoencoder*. Retrieved from arXiv:1804.02135

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*. Advance online publication. https://doi.org/10.1080/09524622.2019.1581839

Arias, P., Belin, P., & Aucouturier, J. J. (2018). Auditory smiles trigger unconscious facial imitation. *Current Biology*, *28*(14), R782–R783.

Arias, P., Soladie, C., Bouafif, O., Robel, A., Seguier, R., & Aucouturier, J. J. (2018). Realistic transformation of facial and vocal smiles in real-time audiovisual streams. *IEEE Transactions on Affective Computing*. Advance online publication. https://doi.org/10.1109/TAFFC.2018.2811465

Armstrong, M. M., Lee, A. J., & Feinberg, D. R. (2019). A house of cards: Bias in perception of body size mediates the relationship between voice pitch and perceptions of dominance. *Animal Behaviour*, *147*, 43–51.

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.

Aucouturier, J. J., Johansson, P., Hall, L., Segnini, R., Mercadié, L., & Watanabe, K. (2016). Covert digital manipulation of vocal emotion alter speakers emotional states in a congruent direction. *Proceedings of the National Academy of Sciences*, *113*(4), 948–953.

Bachorowski, J. A., & Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science*, *6*(4), 219–224.

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

Barthel, H., & Quené, H. (2015). Acoustic-phonetic properties of smiling revised: Measurements on a natural video corpus. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences* (pp. 1–5). Glasgow, UK: The University of Glasgow.

Belin, P., Boehme, B., & McAleer, P. (2017). The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLoS ONE*, *12*(10). https://doi.org/10.1371/journal.pone.0211282.

Blumstein, D. T., & Recapet, C. (2009). The sound of arousal: The addition of novel non-linearities increases responsiveness in marmot alarm calls. *Ethology*, *115*(11), 1074–1081.

Boersma, P., & Weenink, D. (2002). PRAAT, a system for doing phonetics by computer. *Glot International*, *5*(9–10), 341–345.

Bõhm, T., Audibert, N., Shattuck-Hufnagel, S., Németh, G., & Aubergé, V. (2008). Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles. *Journal of the Acoustical Society of America*, *123*(5). https://doi.org/10.1121/1.2935816

Boidron, L., Boudenia, K., Avena, C., Boucheix, J. M., & Aucouturier, J. J. (2016). Emergency medical triage decisions are swayed by computer-manipulated cues of physical dominance in callers voice. *Scientific Reports*, *6*. https://doi.org/10.1038/srep30219

Bonada, J., & Blaauw, M. (2013). Generation of growl-type voice qualities by spectral morphing. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6910–6914). New York, NY: IEEE.

Bonebright, T. L., Thompson, J. L., & Leger, D. W. (1996). Gender stereotypes in the expression and perception of vocal affect. *Sex Roles*, *34*(5–6), 429–445.

Bowling, D., Garcia, M., Dunn, J., Ruprecht, R., Stewart, A., Frommolt, K. H., & Fitch, W. (2017). Body size and vocalization in primates and carnivores. *Scientific Reports*, *7*(41070), 1–11. https://doi.org/10.1038/srep41070

Brady, M. C. (2005). Synthesizing affect with an analog vocal tract: Glottal source. In *Toward Social Mechanisms of Android Science: A CogSci 2005 Workshop*, Stresa, Italy, 25–26 July 2005 (pp. 45–49). Cognitive Science Society.

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., . . . Belin, P. (2010). Vocal attractiveness increases by averaging. *Current Biology*, *20*(2), 116–120.

Burleigh, T., Schoenherr, J., & Lacroix, G. (2013). Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in Human Behavior*, *29*(3), 759–771.

Burred, J. J., 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*(4). https://doi.org/10.1371/journal.pone.0205943

Casadevall, A., & Fang, F. C. (2008). Descriptive science. *Infection and Immunity*, *76*(9), 3835–3836.

Charlton, B. D., Ellis, W. A., Brumm, J., Nilsson, K., & Fitch, W. T. (2012). Female koalas prefer bellows in which lower formants indicate larger males. *Animal Behaviour*, *84*(6), 1565–1571.

Chen, B., Kitaoka, N., & Takeda, K. (2016). Impact of acoustic similarity on efficiency of verbal information transmission via subtle prosodic cues. *EURASIP Journal on Audio, Speech, and Music Processing*, *2016*(1). https://doi.org/10.1186/s13636-016-0097-6

Chong, C. S., Kim, J., & Davis, C. (2018). Disgust expressive speech: The acoustic consequences of the facial expression of emotion. *Speech Communication*, *98*, 68–72.

Dattorro, J. (1997). Effect design. Part 2: Delay line modulation and chorus. *Journal of the Audio Engineering Society*, *45*(10), 764–788.

De Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics*, *28*(4), 441–465.

Degottex, G., Lanchantin, P., Roebel, A., & Rodet, X. (2013). Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis. *Speech Communication*, *55*(2), 278–294.

Drahota, A., Costall, A., & Reddy, V. (2008). The vocal communication of different kinds of smile. *Speech Communication*, *50*(4), 278–287.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1568–1578.

El Haddad, K., Dupont, S., d'Alessandro, N., & Dutoit, T. (2015). An HMM-based speech-smile synthesis system: An approach for amusement synthesis. In *Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (Vol. 5, pp. 1–6). New York, NY: IEEE.

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia* (pp. 835–838). New York, NY: ACM.

Fee, M. S., Shraiman, B., Pesaran, B., & Mitra, P. P. (1998). The role of nonlinear dynamics of the syrinx in the vocalizations of a songbird. *Nature*, *395*, 67–71.

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in Rhesus macaques. *The Journal of the Acoustical Society of America*, *102*(2), 1213–1222.

Fitch, W. T., Neubauer, J., & Herzel, H. (2002). Calls out of chaos: The adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal Behaviour*, *63*(3), 407–418.

Gentilucci, M., Ardaillon, L., & Liuni, M. (2019). Composing vocal distortion: A tool for real-time generation of roughness. *Computer Music Journal*, *42*(4), 26–40.

Gobl, C., & Chasaide, A. N. (2010). 11 voice source variation and its communicative functions. *The handbook of phonetic sciences* 1: 378.

Gobl, C., & Chasaide, A. N. (2012). Voice source variation and its communicative functions. In W. J. Hardcastle, J. Laver & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (Vol. 119, pp. 378–423). John Wiley & Sons.

Goldstone, R. L., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in Cognitive Science*, *8*(3), 548–568.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. In *Annual Conference on Neural Information Processing Systems 2014*, Montreal, Quebec, Canada, 8–13 December 2014 (pp. 2672–2680).

Govind, D., & Prasanna, S. M. (2013). Expressive speech synthesis: A review. *International Journal of Speech Technology*, *16*(2), 237–260.

Hain, T. C., Burnett, T. A., Larson, C. R., & Kiran, S. (2001). Effects of delayed auditory feedback (daf) on the pitch-shift reflex. *The Journal of the Acoustical Society of America*, *109*(5), 2146–2152.

Hienz, R. D., Jones, A. M., & Weerts, E. M. (2004). The discrimination of baboon grunt calls and human vowel sounds by baboons. *The Journal of the Acoustical Society of America*, *116*, 1692–1697.

Hsu, W. N., Zhang, Y., & Glass, J. (2017). *Learning latent representations for speech generation and transformation*. Retrieved from arXiv:1704.04222

Ilie, G., & Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception*, *23*(4), 319–329.

Jack, R. E., & Schyns, P. G. (2017). Toward a social psychophysics of face communication. *Annual Review of Psychology*, *68*, 269–297.

Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication*, *88*, 106–126.

Johnstone, T., & Scherer, K. R. (1999). The effects of emotions on voice quality. In *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, 1–7 August 1999 (pp. 2029–2032). International Phonetic Association.

Juillerat, N., Schubiger-Banz, S., & Arisona, S. M. (2008). Low latency audio pitch shifting in the time domain. In *Proceedings of the IEEE 2008 International Conference on Audio, Language and Image Processing* (pp. 29–35). New York, NY: IEEE.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770–814.

Kamiloglu, R., Fischer, A., & Sauter, D. A. (2020). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin & Review*, *27*, 237–265.

Kawahara, H. (1997). Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. In *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 2, pp. 1303–1306). New York, NY: IEEE.

Kawahara, H., & Matsui, H. (2003). Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 256–259). New York, NY: IEEE.

Kitayama, S., Mesquita, B., & Karasawa, M. (2006). Cultural affordances and emotional experience: Socially engaging and disengaging emotions in Japan and the United States. *Journal of Personality and Social Psychology*, *91*(5), 890–903.

Lartillot, O., & Toiviainen, P. (2007). A MATLAB toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects*, Bordeaux, France, 10–15 September 2007 (pp. 237–244). DAFx.

Lasarcyk, E., & Trouvain, J. (2008). Spread lips + raised larynx + higher F0 = smiled speech? – An articulatory synthesis approach. *Proceedings of the 8th International Seminar on Speech Production* (pp. 43–48). Nancy, France: Institut National Polytechnique de Lorraine.

Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, *23*(12), 1075–1080.

Laukka, P. (2005). Categorical perception of vocal emotion expressions. *Emotion*, *5*(3), 277–295.

Lee, E., Karrer, T., & Borchers, J. O. (2007). An analysis of startup and dynamic latency in phase vocoder-based time-stretching algorithms. In *Proceeding International Computer Music Conference*, Copenhagen, Denmark, 27–31 August. International Computer Music Association.

Lee, Y., Rabiee, A., & Lee, S. Y. (2017). *Emotional end-to-end neural speech synthesizer*. Retrieved from arXiv:1711.05447

Loscos, A., & Bonada, J. (2004). Emulating rough and growl voice in spectral domain. In *Proceedings of the International Conference on Digital Audio Effects*, Naples, Italy, 5–8 October 2004 (pp. 49–52). DAFx.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn–Kanade Dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, San Francisco, CA, USA, 13–18 June 2010 (pp. 94–101). New York, NY: IEEE.

Luo, Z., Chen, J., Takiguchi, T., & Ariki, Y. (2017). Emotional voice conversion with adaptive scales F0 based on wavelet transform using limited amount of emotional data. In *Proceedings of the 2017 Interspeech Conference*, Stockholm, Sweden, 20–24 August 2017 (pp. 3399–3403). International Speech Communication Association.

Ma, W., & Thompson, W. F. (2015). Human emotions track changes in the acoustic environment. *Proceedings of the National Academy of Sciences of the USA*, *112*(47), 14563–14568.

Malisz, Z., Henter, G. E., Valentini-Botinhao, C., Watts, O., Beskow, J., & Gustafson, J. (2019). Modern speech synthesis for phonetic sciences: A discussion and an evaluation. In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, 5–9 August 2019 (pp. 487–491). Australasian Speech Science and Technology Association.

Matsumoto, D. (2002). Methodological requirements to test a possible in-group advantage in judging emotions across cultures: Comment on Elfenbein and Ambady (2002) and evidence. *Psychological Bulletin*, *128*(2), 236–242.

Mohammadi, S. H., & Kain, A. (2017). An overview of voice conversion systems. *Speech Communication*, *88*, 65–82.

Moulines, E., & Laroche, J. (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, *16*(2), 175–205.

Narendranath, M., Murthy, H. A., Rajendran, S., & Yegnanarayana, B. (1995). Transformation of formants for voice conversion using artificial neural networks. *Speech Communication*, *16*(2), 207–216.

Nowicki, S., Mitani, J. C., Nelson, D. A., & Marler, P. (1989). The communicative significance of tonality in birdsong: Responses to songs produced in helium. *Bioacoustics*, *2*(1), 35–46.

Ohala, J. J. (1980). The acoustic origin of the smile. *The Journal of the Acoustical Society of America*, *68*(S1), S33.

Pascual, S., Bonafonte, A., & Serra, J. (2017). *SEGAN: Speech enhancement generative adversarial network*. Retrieved from arXiv:*1703*.09452

Pell, M. D. (2001). Influence of emotion and focus location on prosody in matched statements and questions. *The Journal of the Acoustical Society of America*, *109*(4), 1668–1680.

Pell, M. D., & Kotz, S. A. (2011). On the time course of vocal emotion recognition. *PLoS One*, *6*(11). https://doi.org/10.1371/journal.pone.0027256

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, *33*(2), 107–120.

Pierce, B. C., & Benjamin, C. (2002). *Types and programming languages*. Cambridge, MA: MIT Press.

Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J. J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences of the USA*, *115*(15), 3972–3977.

Quené, H., Semin, G. R., & Foroni, F. (2012). Audible smiles and frowns affect speech comprehension. *Speech Communication*, *54*(7), 917–922.

Rachman, L., Liuni, M., Arias, P., Lind, A., Johansson, P., Hall, L., . . . Aucouturier, J. J. (2018). DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech. *Behavior Research Methods*, *50*(1), 323–343.

Rand, A. S., & Dudley, R. (1993). Frogs in helium: The anuran vocal sac is not a cavity resonator. *Physiological Zoology*, *66*(5), 793–806.

Reby, D., McComb, K., Cargnelutti, B., Darwin, C., Fitch, W. T., & Clutton-Brock, T. (2005). Red deer stags use formants as assessment cues during intrasexual agonistic interactions. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1566), 941–947.

Ruinskiy, D., & Lavner, Y. (2008). Stochastic models of pitch jitter and amplitude shimmer for voice modification. In Proceedings of the 2008 Convention of Electrical and Electronics Engineers in Israel (pp. 489–493). New York, NY: IEEE.

Scherer, K. R. (1972, April). *Acoustic concomitants of emotional dimensions: Judging affect from synthesized tone sequences*. Paper presented at the Eastern Psychological Association Meeting, Boston, MA.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1–2), 227–256.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, *32*(1), 76–92.

Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, *1*(4), 331–346.

Seshadri, S., Juvela, L., Alku, P., & Räsänen, O. (2019). Augmented cyclegans for continuous scale normal-to-Lombard speaking style conversion. In *Proceedings of the 2019 Interspeech Conference*, Graz, Austria, 15–19 September 2019 (pp. 2838–2842). International Speech Communication Association.

Simonyan, K., & Horwitz, B. (2011). Laryngeal motor cortex and control of speech in humans. *The Neuroscientist*, *17*(2), 197–208.

Stylianou, Y. (2009). Voice transformation: A survey. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3585–3588). New York, NY: IEEE.

Sueur, J., Aubin, T., & Simonis, C. (2008). Seewave, a free modular tool for sound analysis and synthesis. *Bioacoustics*, *18*(2), 213–226.

Sun, P., Anumanchipalli, G. K., & Chang, E. F. (2019). *Brain2char: A deep architecture for decoding text from brain recordings*. Retrieved from arXiv:*1909*.01401

Tartter, V. C. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics*, *27*(1), 24–27.

Tartter, V. C., & Braun, D. (1994). Hearing smiles and frowns in normal and whisper registers. *The Journal of the Acoustical Society of America*, *96*(4), 2101–2107.

Tchernichovski, O., & Mitra, P. (2004). *Sound Analysis Pro user manual*. New York, NY: CCNY.

Tice, D. M. (1992). Self-concept change and self-presentation: The looking glass self is also a magnifying glass. *Journal of Personality and Social Psychology*, *63*(3), 435–451.

Titze, I. (1994). *Principles of voice production*. Englewood Cliffs, NJ: Prentice Hall.

Tyson, R. B., Nowacek, D. P., & Miller, P. J. (2007). Nonlinear phenomena in the vocalizations of North Atlantic right whales (Eubalaena glacialis) and killer whales (Orcinus orca). *The Journal of the Acoustical Society of America*, *122*(3), 1365–1373.

Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., . . . Kavukcuoglu, K. (2016). *Wavenet: A generative model for raw audio*. Retrieved from arXiv:1609.03499

Van Doorn, E. A., Heerdink, M. W., & van Kleef, G. A. (2012). Emotion and the construal of social situations: Inferences of cooperation versus competition from expressions of anger, happiness, and disappointment. *Cognition & Emotion*, *26*(3), 442–461.

Verma, A., & Kumar, A. (2005). Introducing roughness in individuality transformation through jitter modeling and modification. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 1–5). New York, NY: IEEE.

Wang, W., Alameda-Pineda, X., Xu, D., Fua, P., Ricci, E., & Sebe, N. (2018). Every smile is unique: Landmark-guided diverse smile generation. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7083–7092). New York, NY: IEEE.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., . . . Saurous, R. A. (2017). *Tacotron: Towards end-to-end speech synthesis*. Retrieved from arXiv:1703.10135

Wilden, I., Herzel, H., Peters, G., & Tembrock, G. (1998). Subharmonics, biphonation, and deterministic chaos in mammal vocalization. *Bioacoustics*, *9*(3), 171–196.

Wu, Y. C., Hayashi, T., Tobing, P. L., Kobayashi, K., & Toda, T. (2019). *Quasi-periodic wavenet vocoder: A pitch dependent dilated convolution model for parametric speech generation*. Retrieved from arXiv:1907.00797

Xu, T., Zhan, J., Garrod, O. G., Torr, P. H., Zhu, S. C., Ince, R. A., & Schyns, P. G. (2018). *Deeper interpretability of deep networks*. Retrieved from arXiv:1811.07807

Yoshie, M., & Sauter, D. A. (2019). Cultural norms influence nonverbal emotion communication: Japanese vocalizations of socially disengaging emotions. *Emotion*. Advance online publication. https://doi.org/10.1037/emo0000580