PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Article submitted to journal

Subject Areas:
BIOLOGY: behaviour, cognition

Keywords:
voice, music, emotions

Author for correspondence:
JJ. Aucouturier
e-mail: aucouturier@gmail.com

Even violins can cry: Specifically-vocal emotional behaviours also drive the perception of emotions in non-vocal music

D. Bedoya[1], P. Arias[1,2], L. Rachman[3], M. Liuni[4], C. Canonne[1], L. Goupil[5], J-J. Aucouturier[6]

[1] Science and Technology of Music and Sound, IRCAM/CNRS/Sorbonne Université, Paris (France).
[2] Dept of Cognitive Science, Lund University, Lund (Sweden)
[3] Faculty of Medical Sciences, University of Groningen, Groningen (NL)
[4] Alta Voce SAS, Houilles (FR)
[5] BabyDevLab, University of East London, London (UK)
[6] FEMTO-ST Institute, Univ. Bourgogne Franche-Comté / CNRS, Besançon (France)

A wealth of theoretical and empirical arguments have suggested that music triggers emotional responses by resembling the inflections of expressive vocalizations, but have done so using low-level acoustic parameters (pitch, loudness, speed) which, in fact, may not be processed by the listener in reference to human voice. Here, we take the opportunity of the recent availability of computational models which allow the simulation of three specifically-vocal emotional behaviours: smiling, vocal tremor, and vocal roughness. When applied to musical material, we find that these three acoustic manipulations trigger emotional perceptions which are remarkably similar to those observed on speech and scream sounds, and identical across musician and non-musician listeners. Strikingly, this not only applied to singing voice with and without musical background, but also to purely instrumental material.

Originally invoked to describe the vocal monodic style of the Florentine Camerata in the 17th century [37], the idea that music expresses emotions by resembling the inflections of expressive speech (the so-called ’speech theory’) has grown into a prominent view in recent psychological [36], neuroscientific [52] and evolutionary [27] accounts of music cognition. This view is notably supported by a wealth of studies showing that music’s expressive acoustic features mirror those used in vocal expression, with e.g. fast pace and high intensity for happy music/voice, and monotonous pitches and dark timbres for sad music/voice [19,34,35,58]. In addition, music and voice processing appear to obey similar innate developmental constraints, as shown e.g. by comparable impairments in congenital amusia [66] or by improvements of prosodic perception after musical training [41].

THE ROYAL SOCIETY
PUBLISHING

It is unclear, however, whether these similarities reveal a genuine cross-domain recycling of cognitive resources developed originally either for voice or music; or whether they reflect a mechanism that is simply more generic than either, and encompasses both. Voice and music cognition are indeed continuous with generic auditory cognition [62], and the majority of acoustic characteristics tested by prior work (e.g., pitch, loudness, speed) carry biologically significant information about a vaster diversity of sound sources than voice or music. For instance, abstract sound sources with increasing loudness and rising pitch may be perceived as gaining energy and moving closer, triggering avoidance reactions and a sense of urgency [49,65]. Similarly, adults, and infants as early as 6-month-old, associate lower pitch with larger and potentially more formidable objects [26]. Accordingly, research has shown that changes in frequency, rate and intensity that are known to support emotional interpretations in speech and music in fact also trigger similar emotional responses when applied to environmental sounds such as rain, thunder or wind [44]. In addition, cross-domain contrasts in brain imaging of speech and music emotion typically do not reveal common sensory representations in temporal voice areas, as would be expected if these were voice-specific effects, but only supramodal emotion representations in the frontal cortices [24,53].

All of this suggests that the perceptual mechanisms so far tested in speech and music studies may not, in fact, be processed by the listener in reference to human voice. It remains unknown whether specifically-vocal expressive cues, such the unstable phonatory muscle control of an anxious voice, the non-linear vocal fold vibration of a scream, or the bright resonating quality of smiled speech, also trigger comparable emotional reactions when they occur in music.

One reason previous research hasn't tested voice-specific cross-domain effects is the lack of tools able to simulate such phenomena in arbitrary audio material. First, typical acoustic manipulations in experimental stimuli have used generic audio processing software such as Audacity (Audacity Team) or ProTools (Avid Technology, Inc.) [34,44], which only allow the transformation of low-level parameters such as pitch, intensity and speed. Second, voice-specific tools such as Praat [15] or SoundGen [1], which are able to model phonatory or articulatory aspects of human voice, do not allow transforming musical excerpts in a way that mirror these characteristics.

Here, we take the opportunity of a series of recent developments in audio transformation technologies [7] which provide novel technical ways to simulate the effect of three voice-specific emotional behaviours (one articulatory, smiled speech [8]; two phonatory, vocal tremor [56] and vocal roughness [42]) identically in matched speech and music stimuli:

(i) Smiling, like other orolabial gestures such as nose wrinkling [18], modify the shape and length of the vocal tract [51], shifting its resonating frequencies (Figure 1-A). These changes can be simulated using frequency warping on the spectral envelope of the sounds, inside a phase vocoder architecture [8]. In listening experiments, English speech samples manipulated with such a transformation were validated to sound more smiling, and generally more positive [6,8]; in production experiments, participants asked to imitate voices manipulated with such changes do so by smiling while they vocalize [6].

(ii) Vocal tremor, which can occur physiologically from cold, fatigue or anxiety, is a rhythmical and involuntary oscillatory movement affecting the vocal folds, thought to result from disturbances in the neurophysiological feedback processes of phonatory muscle control [30,48]. It causes cyclical fluctuations in pitch (vibrato, Figure 1-B) and loudness (tremolo), which can be simulated in recordings as the sinusoidal modulation of a pitch shift effect [56]. In listening experiments, English, French, Swedish and Japanese speech samples manipulated with such a transformation were validated to sound more anxious, negative and aroused [10,56]; in production experiments, participants who heard themselves speak while their auditory feedback was manipulated with tremor reported feeling more negative and more aroused [10].

(iii) Vocal roughness, which occurs when excessive subglottal pressure due to effort or arousal causes nonlinearities in vocal fold vibration, reveals the presence in voice of subharmonics (Figure 1-C) which, along with other nonlinearities such as frequency jumps, broadband noise or chaos, gives voice a rough and noisy quality [28]. Vocal roughness in screams, cries, grunts or moans has an important communicative function in the human expressive repertoire, because it signals aversive states such as fear, pain or distress [3,9]. Vocal roughness can be simulated using pitch-synchronous amplitude modulation to add sub-harmonics in the original signal [42]. In listening experiments, speech samples manipulated with such a transformation were validated to sound more negative and aroused [42].

Using such manipulations designed in clear mechanistic analogy with the human voice is important because it ensures that we only explore a range of acoustic variations which correspond to what voice can do (e.g., smiling operates on the 2-4kHz frequency range, and not, say, at 1kHz or 8kHz), at a level of intensity which conforms to daily "mundane" expressions (e.g. a pitch shift of +25 cents, a quarter of a semitone, and not, say, +3-4 semitones), and avoid broad claims of similarity

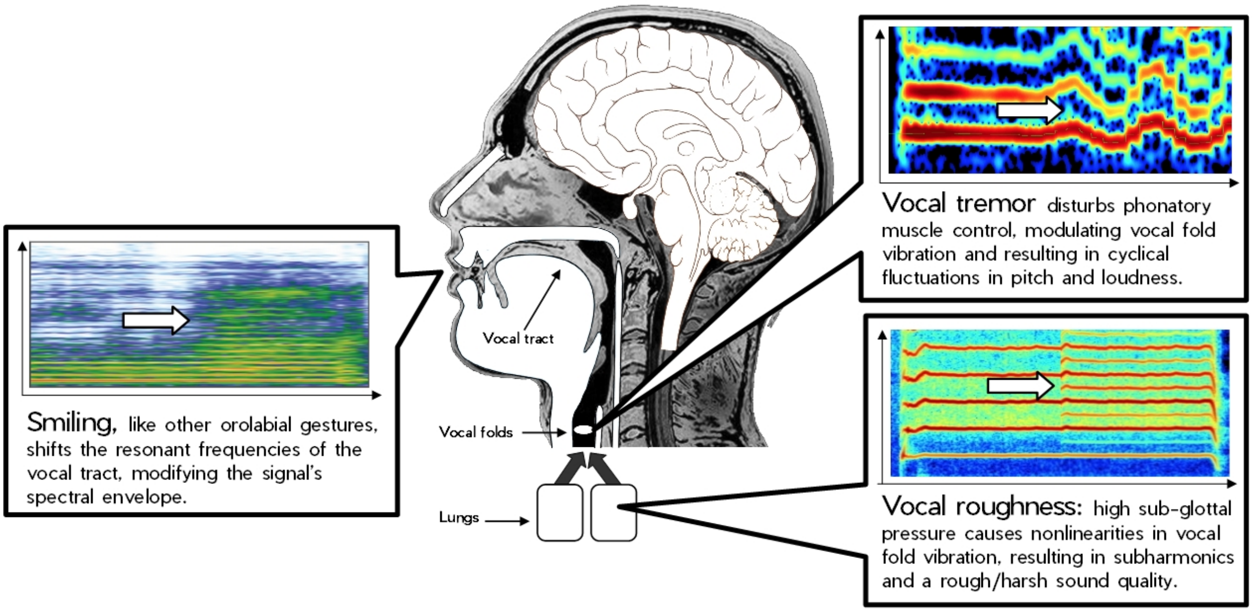
Smiling, like other orolabial gestures, shifts the resonant frequencies of the vocal tract, modifying the signal's spectral envelope.
Vocal tract
Vocal folds
Lungs
Vocal tremor disturbs phonatory muscle control, modulating vocal fold vibration and resulting in cyclical fluctuations in pitch and loudness.
Vocal roughness: high sub-glottal pressure causes nonlinearities in vocal fold vibration, resulting in subharmonics and a rough/harsh sound quality.

Figure 1. Three expressive acoustic changes that have a specifically vocal origin in the physiology of human/mammalian vocal apparatus: (A) smiling, (B) vocal tremor and (C) vocal roughness. All three changes are simulated here by signal processing techniques, which can modulate both speech and music recordings.

based on sound manipulations (e.g. a wholesale +5 semitone applied to a complete orchestral piece) which, in fact, may not be processed by the listener in reference to human voice.

In this work, we applied all three vocal manipulations to matched speech, vocal music and instrumental music extracts. We asked two groups of N=29 musicians and N=31 non-musician listeners to compare pairs composed of the manipulated and non-manipulated variants of each sound using two Likert scales for expressed emotional valence and arousal, and examined whether the manipulations led to similar emotional interpretations when they occurred in speech and music. Ratings of valence and arousal were chosen in order to measure the low-level expression of "core affect" [23], which is more likely to capture affective similarities between speech and music pairs than higher-level categorical constructs such as emotions, which are expected to be more heavily influenced by context such as the presence or absence of lyrics [11] or of a specific musical instrument [33].

Results

(a) Preregistered hypotheses

We tested the impact of the three manipulations (smiling, vocal tremor and vocal roughness) on five types of sounds: two types of non-musical vocal sounds (speech and screams), and three types of musical sounds (singing only, singing + music, violin + music).

In the following, we separately report, for each of the three manipulations, on five-level analyses including all these types of sounds. However, our hypotheses, which we preregistered[1], concerned only a subset of these combinations:

[1] https://aspredicted.org/mc72i.pdf

(i) Smiling and vocal tremor are manipulations originally developed and validated for speech sounds [8,56]. Following these studies, we hypothesized that smiling would increase valence and arousal, and vocal tremor would decrease valence and increase arousal for speech stimuli. We made no hypotheses for how these manipulations would affect the perception of screams.

(ii) Conversely, vocal roughness is a manipulation originally developed and validated for screams [42]. Following this study, we hypothesized that roughness would decrease valence and increase arousal for scream sounds. We made no hypothesis for how vocal roughness would affect the perception of speech.

(iii) Similarly, our hypotheses concerning the transfer of affective qualities from non-musical vocal sounds (speech and screams) to musical sounds concerned speech effects for smiling and vocal tremor (i.e., similar to speech, smiling would increase valence and arousal for musical sounds, and vocal tremor would decrease valence and increase arousal) and scream effects for vocal roughness (i.e., similar to screams, vocal roughness would decrease valence and increase arousal for musical sounds).

(b) The three manipulations worked as intended on vocal sounds

We first validated that the three voice manipulations triggered emotional judgements as intended when occurring on vocal sounds. N=60 participants (among whom N=29 musicians) rated pairs on matched manipulated and non-manipulated sounds on both valence and arousal. As preregistered, we aggregated participant ratings for each type of stimuli and transformation, and analysed the effect of transformation using rm-ANOVAs and paired t-tests.

(i) The effect of applying the smile transformation (smile vs unsmile) to speech stimuli was very large and statistically significant: as predicted, it led to higher perceived valence (M=+1.01, [+0.79, +1.24] scale points, t(59)=9.09, p=8.00e-13, Cohen's *d*=1.92) and perceived arousal (M=+1.27, [1.02 1.53], t(59)=10.08, p=1.89e-14, Cohen's *d*=2.09). Neither of these effects interacted statistically with participants being musicians or not (interaction musician x transformation, valence: F(2,116)=1.23, p=0.30, η_p^2=0.02; arousal: F(2,116)=2.40, p=0.10, η_p^2=0.04; test sensitive to effect size $d \geq 0.28$ at power $1-\beta=0.95$ and $\alpha=.05$)

(ii) The effect of applying the tremor transformation (tremor vs non-manipulated) to speech stimuli was medium and statistically significant. As expected, it decreased perceived valence (M=-0.19, [-0.28 -0.11], t(59)=-4.55, p=2.77e-05, Cohen's *d*=0.59). However, contrary to what we predicted, tremor also decreased perceived arousal (M=-0.19, [-0.27 -0.11], t(59)=-4.88, p=8.56e-06, Cohen's *d*=0.55). Neither of these effects interacted statistically with participants being musicians or not (interaction musician x transformation, valence: F(1,58)=2.62, p=0.11, η_p^2=0.04; arousal: F(1,58)=0.03, p=0.87, η_p^2=0.00; test sensitive to effect size $d \geq 0.31$ at power $1-\beta=0.95$ and $\alpha=.05$).

(iii) The effect of applying the roughness transformation (rough vs non-manipulated) to scream stimuli was very large and statistically significant. As expected, it decreased perceived valence (M=-0.71, [-0.89 -0.53], t(59)=-7.78, p=1.28e-10, Cohen's *d*=1.30) and increased arousal (M=+0.62, [0.45 0.8], t(59)=7.09, p=1.90e-09, Cohen's *d*=1.21). Neither of these effects interacted statistically with participants being musicians or not (valence: F(1,58)=0.94, p=0.34, η_p^2=0.02; arousal: F(1,58)=0.27, p=0.60, η_p^2=0.00; test sensitive to effect size $d \geq 0.31$ at power $1-\beta=0.95$ and $\alpha=.05$)

In sum, the effect of the three manipulations were largely consistent with our predictions for vocal sounds. Descriptively, the effect of smiling on speech was consistent with expressing more positivity and arousal, tremor on speech with expressing more negativity and less arousal (note that previous work associated tremor with increased, rather than decreased, arousal [10,56]) and roughness on screams with expressing more negativity and more arousal.

(c) Extension to non-preregistered vocal modes

Even though we only preregistered hypotheses for smile and tremor on speech, and for roughness on screams (respecting the vocal modes for which the manipulations were originally intended), all three manipulations were also tested for the other vocal mode:

(i) The effect of smiling on screams was consistent with predictions made for speech (valence: M=+0.53, [0.26 0.81], t(59)=3.88, p=.0003, Cohen's *d*=0.77, arousal: M=+1.13 [0.86, 1.39], t(59)=8.37, p=1.32e-11, Cohen's d=1.68).

(ii) Contrary to speech, tremor had no effect on the valence of screams (M=-0.04 [-0.19, 0.12], t(59)=-0.45, p=.65, Cohen's *d*=0.07) and increased their perceived arousal (M=+0.18 [0.07, 0.3], t(59)=3.14, p=.002, Cohen's *d*=0.46; note, prospectively, that the effect of tremor on scream arousal was in an opposite direction to all other sound types) (Figure 2).

(iii) Finally, the effect of roughness on speech was consistent with predictions made for screams, decreasing valence (M=-0.21, [-0.33 -0.09], t(59)=-3.45, p=.001, Cohen's *d*=0.54) and increasing arousal, albeit non-significantly (M=+0.05, [-0.04 0.13], t(59)=1.12, p=.26, Cohen's *d*=0.14).

(d) All voice manipulations had a similar effect on vocal and instrumental musical sounds

The same N=60 participants then rated manipulated pairs of matched musical sounds in three conditions: singing only ('a cappella' recording reproducing the same verbal content as the speech stimuli), singing + music (manipulated singing track, mixed with non-manipulated instrumental background) and violin + music (manipulated violin track recorded to imitate the singing track, mixed with non-manipulated instrumental background).

To avoid demand effects, participants rated the music pairs before rating the speech and scream pairs used for validation above; all three types of musical sounds and three types of transformations were randomized within the music block; participants were unaware of the possibility of algorithmic manipulation; and pairs of identical stimuli were included for control (similar procedure as [44], see Materials and Methods).

All three vocal manipulations triggered emotional judgements on musical stimuli which were strikingly similar to those observed on vocal stimuli (Figure 2):

(i) The 5-level sound-type factor interacted significantly with the effect of smile on valence (F(8,472)=11.58, p=4.60e-15, η_p^2=0.16) and arousal (F(8,472)=15.57, p=2.12e-20, η_p^2=0.21), but all effects were in the same direction. Our prediction for transfer to musical sounds concerned the effect of smiling on speech: Similarly to speech, the smile manipulation increased the perceived valence and arousal when applied to a cappella singing (valence:M=+1.45, [1.14 1.75], t(59)=9.56, p=1.37e-13, Cohen's *d*=2.07; arousal: M=+1.41, [1.14 1.67], t(59)=10.50, p=4.05e-15, Cohen's *d*=2.17), to singing mixed with instrumental background (valence: M=+1.02, [0.76 1.28], t(59)=7.89, p=8.56e-11, Cohen's *d*=1.55; arousal: M=+0.44, [0.23 0.65], t(59)=4.16, p=1.06e-04, Cohen's *d*=0.76), but also when applied to a non-vocal (violin) track mixed with instrumental background (valence: M=+0.57, [0.35 0.8], t(59)=5.13, p=3.43e-06, Cohen's *d*=0.89; arousal: M=0.54, [0.33 0.74], t(59)=5.30, p=1.82e-06, Cohen's *d*=0.93). In short, as for speech, violin made to sound more smiling was perceived as more positive and more aroused.

(ii) The 5-level sound-type factor interacted significantly with the effect of tremor on valence (F(4,236)=3.72, p=5.90e-03, η_p^2=0.06) and arousal (F(4,236)=9.37, p=4.78e-07, η_p^2=0.14) but, again, all effects were in the same direction (except for the non-preregistered case of scream arousal). Our prediction for transfer to musical sounds concerned the effect of tremor on speech: similarly to speech, the tremor manipulation decreased the perceived valence and arousal (the latter, non significantly) when applied to a cappella singing (valence:M=-0.37, [-0.51 -0.22], t(59)=-5.09, p=3.89e-06, Cohen's *d*=0.85; arousal: M=-0.08, [-0.22 0.05], t(59)=-1.20, p=2.37e-01, Cohen's *d*=0.19), decreased both significantly when applied to singing + music (valence: M=-0.26, [-0.39 -0.12], t(59)=-3.86, p=2.87e-04, Cohen's *d*=0.59; arousal: M=-0.19, [-0.3 -0.09], t(59)=-3.80, p=3.41e-04, Cohen's *d*=0.50) and to violin + music (valence: M=-0.28, [-0.41 -0.14], t(59)=-3.99, p=1.84e-04, Cohen's *d*=0.62; arousal: M=-0.19, [-0.31 -0.06], t(59)=-3.04, p=3.48e-03, Cohen's *d*=0.42). In short, as for speech, violin made to sound more trembling was perceived as less positive and less aroused.

(iii) The 5-level sound-type factor interacted significantly with the effect of roughness on valence (F(4,236)=12.70, p=2.25e-09, η_p^2=0.18) and arousal (F(4,236)=13.57, p=5.69e-10, η_p^2=0.19) but, again, all effects were in the same direction. Our prediction for transfer to musical sounds concerned the effect of roughness on screams: similarly to screams, the roughness manipulation decreased valence and increased arousal when applied to a cappella singing (valence:M=-0.85, [-1.08 -0.61], t(59)=-7.24, p=1.05e-09, Cohen's *d*=1.33; arousal: M=+0.24, [0.07 0.41], t(59)=2.77, p=7.49e-03, Cohen's *d*=0.49), and decreased valence when applied to singing + music (valence: M=-0.66, [-0.87 -0.45], t(59)=-6.17, p=6.83e-08, Cohen's d=1.05) and to violin + music (valence: M=-0.49, [-0.68 -0.31], t(59)=-5.27, p=2.02e-06, Cohen's d=0.87). The effect of vocal roughness on arousal for singing + music and violin + music was also in the expected direction, but non-significantly (singing+music: M=+0.13, [-0.02 0.27], t(59)=1.74, p=.09, Cohen's d=0.28; violin + music: M=+0.10, [-0.03 0.23], t(59)=1.51, p=.13, Cohen's d=0.22). In short, as for screams, violin made to sound rougher was perceived as less positive and more aroused.

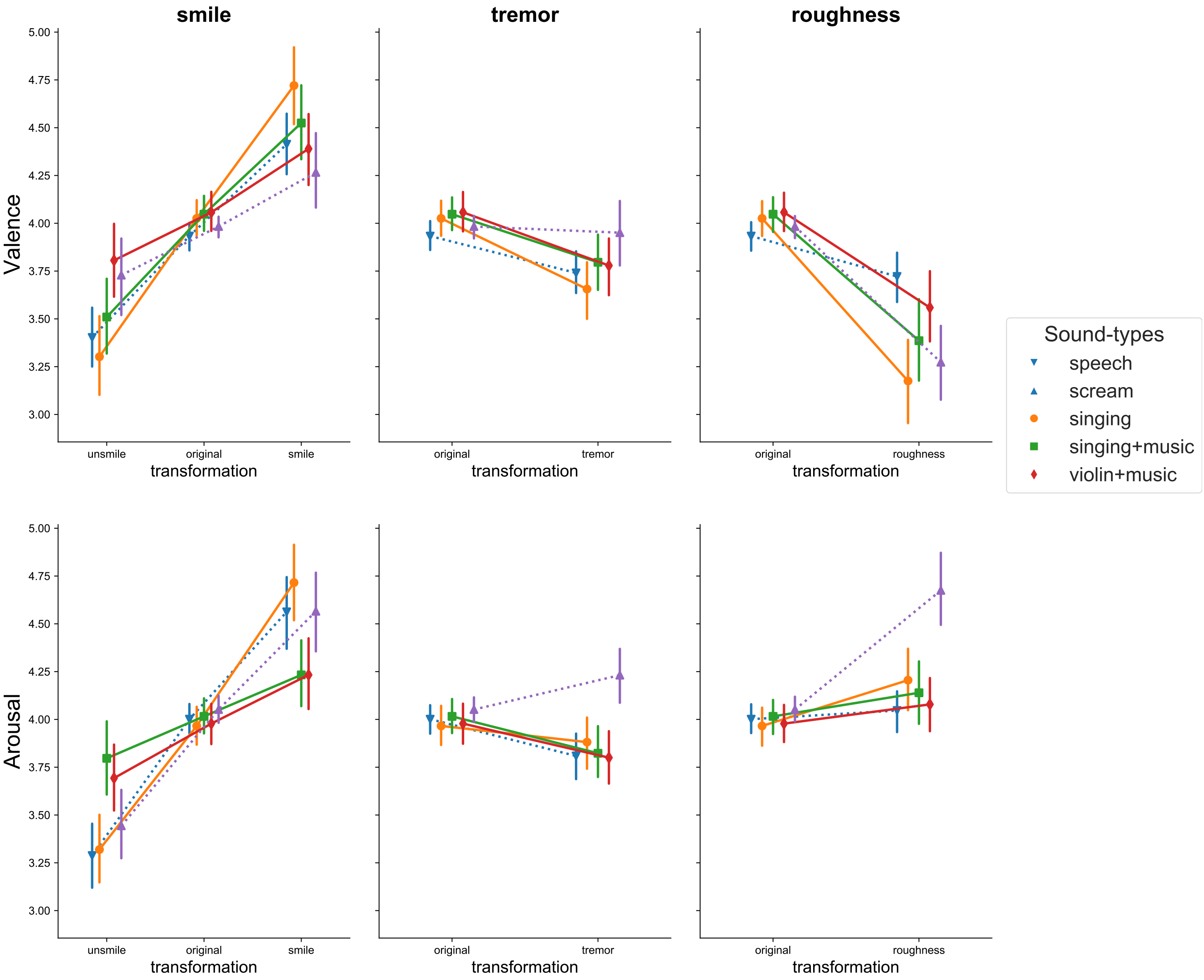

Figure 2. Vocal manipulations of smiling, tremor and roughness trigger similar emotional perceptions in both vocal and non-vocal music. Valence (a) and Arousal (b) ratings for smiling, vocal tremor and vocal roughness manipulations of matched vocal (speech, scream; dotted line) and musical stimuli (solid line). For each manipulation and each sound-type, ratings are given both for manipulated pairs (12-14 pairs consisting of one manipulated sound, evaluated in comparison with its non-manipulated variant; labeled as "smile', 'tremor', etc.) and for control pairs (12-14 pairs consisting of one non-manipulated sound, evaluated in comparison to itself; labeled as 'original') . Error bars indicate 95% confidence intervals on the mean.

(e) Effects were larger on isolated singing than with musical accompaniments

Even though all emotional perceptions in manipulated musical sounds were in the same direction as for vocal sounds, there were differences in the intensity of these perceptions, as indicated by statistical interactions between manipulation and sound type (Figure 3):

(i) The 5-level sound type interacted with the effect of smiling on both perceived valence: F(4,236)=14.93, p=6.83e-11, η_p^2=0.20; and arousal: F(4,236)=21.11, p=6.81e-15, η_p^2=0.26.
For valence, the effect of smiling was larger for speech (*d*=1.92) than screams (*d*=0.77, t(59)=-3.35, p=.001). Within musical sounds, it was maximal for singing voice (Cohen's *d*=2.07), for which it was larger than speech (t(59)=3.23, p=.002) and screams (t(59)=5.44, p<.00001). Compared to singing, the effect of smiling was smaller for singing + music (*d*=1.55; t(59)=-4.17, p<.00001) and smaller again (but remained large) for violin + music (*d*=0.89; t(59)=-6.33,

p<.00001).

For arousal, the effect of smiling did not differ between speech (*d*=2.09), screams (*d*=1.68; t(59)=1.21, p=.23) and singing (*d*=d=2.17; t(59)=0.89, p=.37). It was smaller than singing (but remained large) on singing + music (*d*=0.76; t(59)=-8.87, p<.00001) and violin + music (*d*=0.93; t(59)=-6.60, p<.00001; Figure 3-left).

(ii) The 5-level sound type interacted with the effect of tremor on both perceived valence: F(4,236)=3.72, p=.0059, η_p^2=0.06; and arousal: F(4,236)=9.37, p=4.78e-07, η_p^2, but these interactions were merely driven by the difference between speech and screams (for which tremor had no effect on valence and an opposed effect on arousal).

For valence, the effect of tremor was marginally larger (more negative) for speech (*d*=0.59) than for screams (*d*=0.07; t(59)=1.76, p=.083). Within musical sounds, the valence effect of tremor was maximal (i.e. more negative) for singing (*d*=0.85), for which it was larger than speech (t(59)=2.19, p=.033) and screams (t(59)=2.95, p=.005). Compared to singing, the valence effect of tremor was not significantly smaller for singing + music (*d*=0.59; t(59)=-1.49, p=.14) or for violin + music (*d*=0.62; t(59)=-0.94, p=.35).

For arousal, the effect of tremor was significantly different, and in opposed directions, for speech (less arousal, *d*=0.55) and screams (more arousal, *d*=0.46, t(59)=5.64, p<.00001). Within musical sounds, none of the arousal effects were of significantly different amplitude than for speech (singing: *d*=0.19, t(59)=-1.76,p=.08; singing + music: *d*=0.50, t(59)=-0.05, p=.96; violin + music: *d*=0.42, t(59)=-0.09, p=.93), nor did they differ from one another (all ps>.21). All differed significantly from screams (singing: t(59)=3.12, p=.003; singing + music: t(59)=5.27, p<.00001; violin + music: t(59)=5.17, p<.00001; Figure 3-middle).

(iii) The 5-level sound type interacted with the effect of roughness on both perceived valence: F(4,236)=12.70, p=2.25e-09, η_p^2=0.18; and arousal: F(4,236)=13.57, p=5.69e-10, η_p^2=0.19.

For valence, the effect of vocal roughness was maximum on singing voice (*d*=1.33)) and screams (*d*=1.30; no statistical difference:t(59)=1.20, p=.23). It was smaller than singing (but remained large) for singing + music (*d*=1.05; t(59)=-2.85, p=.006) and for violin + music (*d*=0.87; t(59)=-3.50, p=.001).

For arousal, the effect of vocal roughness was maximum on screams (*d*=1.21), for which it was larger than for speech (*d*=0.14; t(59)=6.36, p<.00001). Within musical sounds, the effect of roughness was smaller than screams for singing (*d*=0.49; t(59)=-3.47, p=.001), singing + music (*d*=0.28; t(59)=-5.17, p<.00001) and violin + music (*d*=0.22; t(59)=-4.84, p<.00001; Figure 3-right)

(f) No effect of musicianship

Finally, to examine whether participant musicianship interacted with the effects, we computed normalized valence and arousal ratings (smile: smile - unsmile; tremor: tremor - original, roughness: rough - original) and averaged over all stimuli per participant and sound type. Whether participants were self-declared musicians (N=29) or non-musicians (N=31) did not interact with the effect of sound type on normalized valence and arousal, for any of the manipulations (all p's $>$ .49, except smiling arousal: F(4,232)=2.24, p=.066, η_p^2=0.04; Figure 4; test sensitive to effect size $d \geq 0.23$ at power $1-\beta=0.95$ and $\alpha=.05$).

Discussion

A wealth of theoretical and empirical arguments have suggested that music triggers emotional reactions by resembling the inflections of expressive vocalizations, but past research focused on low-level acoustic parameters (pitch, loudness, speed) which, in fact, may not be processed by the listener in reference to human voice. Here, we provided a more direct test of the hypothesis by using computational voice-transformation models that simulate of three emotional behaviours linked to specifically-vocal mechanisms of articulation (smiling) and phonation (vocal tremor and vocal roughness). When applied to musical material, we found that these three highly-specific acoustic manipulations trigger emotional perceptions which were remarkably similar to those observed on speech and scream sounds. Strikingly, this not only applied to singing voice with and without musical background, but also to purely instrumental material: even violins can cry, or at least sound more positive and aroused when smiling, more negative and less aroused when trembling, and more negative when screaming (Figure 2).

Importantly, while they can be simulated using inanimate, non-vocal artefacts (e.g., a dented clay cylinder for smile, [51]; a periodically rotating sound source for vocal tremor, [40]), none of the three behaviors tested here have non-vocal ecological equivalents in nature, because they closely depend on the dynamics and physiology of the mammalian larynx: smiling is a dynamic change of resonating frequencies of the vocal tract, vocal tremor is an extrinsic modulation of the vocal folds of

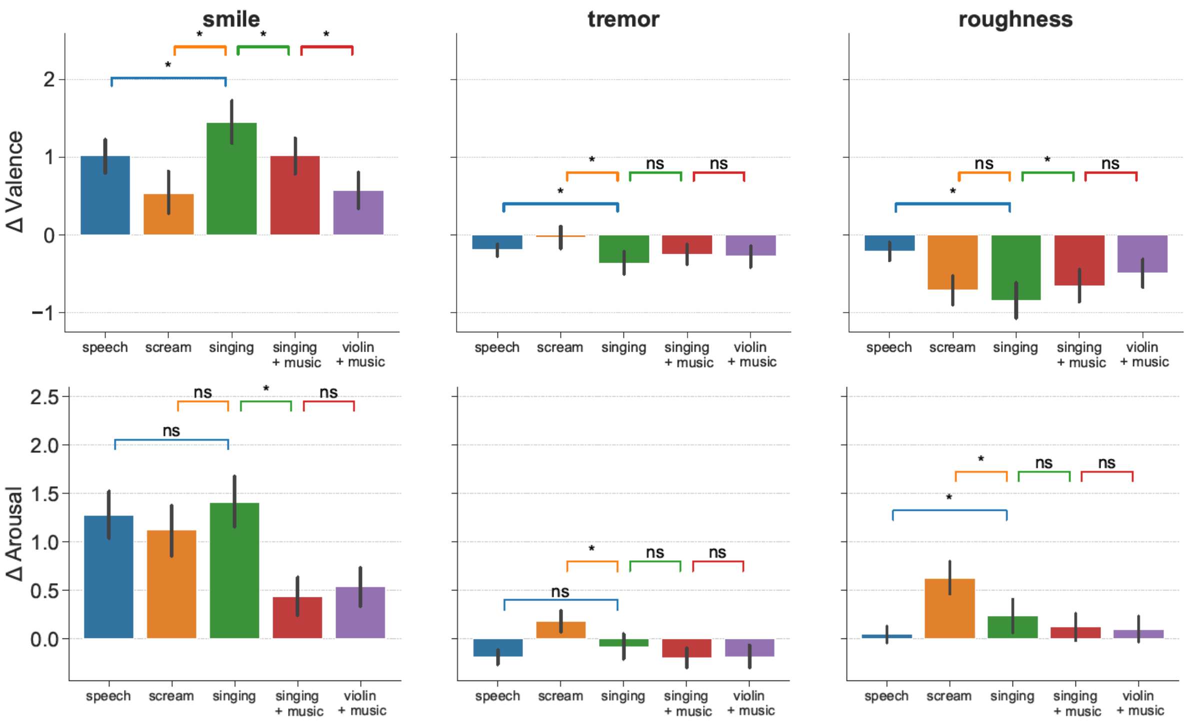

Figure 3. The effect of vocal manipulations was similar or larger than spoken voice (blue, orange) for isolated singing (green), but smaller for instrumental music (red, purple). Normalized ratings (smile: smile - unsmile; tremor: tremor - original, roughness: rough - original) for Valence (top) and Arousal (bottom) of the smiling, vocal tremor and vocal roughness manipulations in each type of stimuli. Asterisks indicate statistical significance of pairwise t-tests at the p<.05 level. Error bars indicate 95% confidence intervals on the mean

muscular-control origins, and vocal roughness is the consequence of a non-linear regime of vocal fold oscillation. If these changes also impart emotional qualities when they occur in music, then these must therefore necessarily be of human (or animal) vocal origin. Our results therefore provide the literal confirmation of Darwin's conjecture that musical emotions can stem from acoustic features which resemble *"the voices of other animals and man's own instinctive cries"* [20].

Even though all emotional perceptions in manipulated musical sounds were in the same direction as vocal sounds, there were differences in the intensity of these perceptions, both among musical and non-musical sounds. Among non-musical sounds (speech and screams), smiling and tremor both had greater effects (resp. positive and negative) on perceived valence in speech than in screams; conversely, vocal roughness had a more negative effect on the perceived valence of screams than speech, and no arousing effect on speech. These differences between speech and screams are likely explained by discrepancies between the emotional valence of the changes and the vocal context in which they occur. For instance, while smiling can signal dominance [59], it is not typically associated with screamed vocalizations and therefore plausibly warrant less univocally positive interpretations in this context than on spoken voice. Similarly, while vocal tremor in vocal registers with low subglottal pressure is typically associated with negative evaluations of e.g. sadness or stress [30,56], the same pitch oscillations when heard in screamed stimuli may be associated to non-linearities due to high subglottal pressure (e.g. pitch jumps) and attributed to higher arousal or intensity rather than lower valence [5]; and, in a similar manner, vocal roughness, while indicative of arousal and aversiveness in screams, may be attributed in the low-pressure register of spoken voice to non-emotional phenomena such as vocal fatigue or hoarseness [2,38]. Finally, it should be noted that the effect of vocal tremor on arousal was in a different direction for speech (negative) than for screams (positive; Figure 2-middle-bottom). That speech effect was the only effect found in a direction which we did not predict. Because the effect was negative both for speech and music, it is plausible that the low-arousal effect of tremor is a genuine effect which transferred from speech to music (our main hypothesis), but it also remains possible that the tremor effect on speech is due to a learning effect carried over from

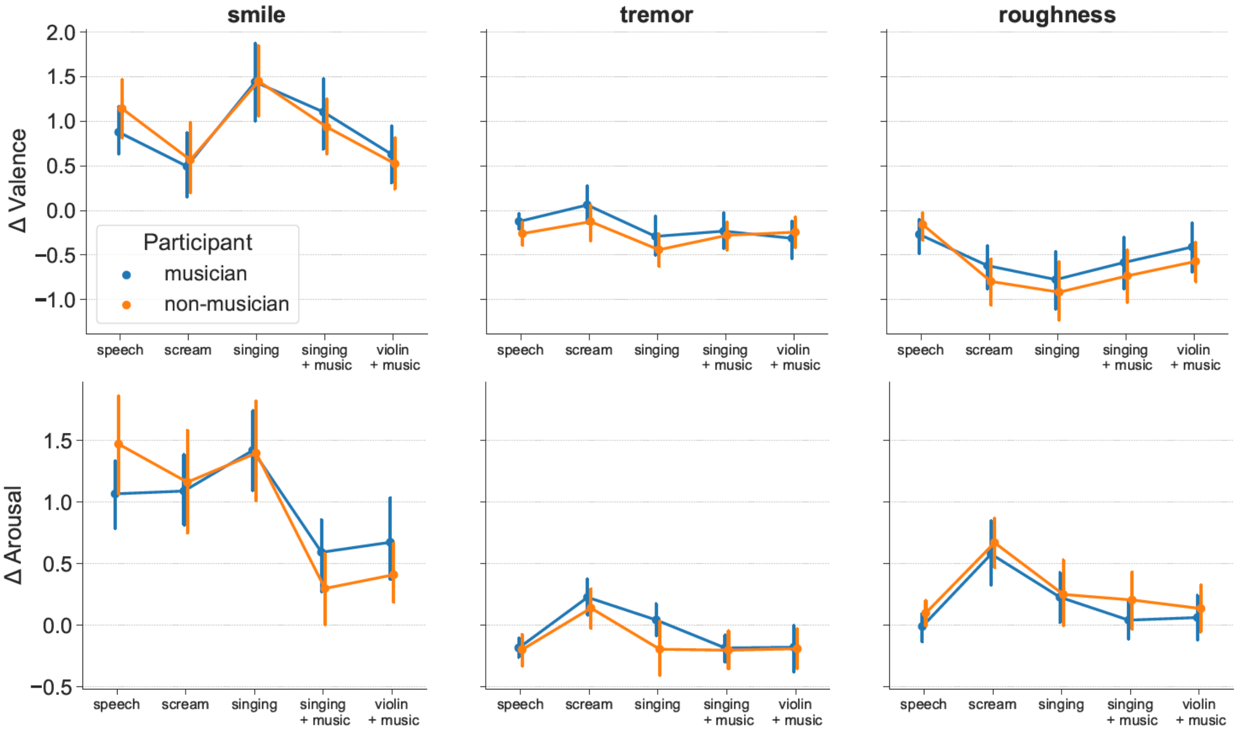

Figure 4. No interaction of musicianship on the effect of sound type on normalized valence and arousal, for any of the manipulations. Normalized ratings (smile: smile - unsmile; tremor: tremor - original, roughness: rough - original) for valence (top) and arousal (bottom), in the musician (blue) and non-musician (orange) groups. Error bars indicate 95% confidence intervals on the mean.

the (previously judged) musical pairs, which would have been evaluated differently had the speech pairs been presented in isolation.

Among musical sounds, the effect of the three manipulations was generally larger for 'a cappella' singing voice than for non-musical vocalizations (speech or scream): this was true for the effect of smile, vocal tremor and, to some extent, vocal roughness on valence (but not on arousal). It is possible that the acoustical properties of singing voice [61] benefit the perception of the three cues used here. For instance, musical melody in the contemporary commercial music genres considered here features discrete and relatively stable pitch series which, as opposed to the continuously changing pitch of speech intonation [69], may facilitate the processing of slowly-changing pitch modulations in vocal tremor. Further, the fact that sung vowels and consonants are typically longer than in their normal occurrence in speech [22] may also allow the faster accumulation of spectral/harmonic information to register changes like smile or vocal roughness. Such an explanation may be conceptually related to the "super-expressive voice hypothesis", a prominent theory of musical emotions stating that, because of their wider pitch and dynamic range, music may be processed as amplified and exaggerated vocal expressions, resulting in more intense emotional reactions [35,36]. It is possible that, even when manipulation intensity is controlled to be strictly identical as for speech, the specific acoustics of singing voice may provide a clearer, more contrasting background for emotional expression than connected speech.

On the other hand, while our three manipulations were qualitatively similar on vocal and instrumental music, they were not perceived as more intense on non-vocal musical instruments than on human voice (if anything, they were even less intense). Among musical sounds, the effect of the three manipulations was indeed greater for 'a cappella' singing than for music with instrumental background. One possible explanation is perceptual, as the additional instrumental background may create masking effects which makes registering the (relatively subtle) changes of the main track more difficult. For instance, smiling is a spectral manipulation mostly manifest in the high-medium frequency range of formants F2-F5 (600-3500 Hz) [54], which is a frequency band likely to be already crowded in the instrumental mixes of the popular music genres tested here. Similarly, the perception of vocal roughness involves the registering of irregularities in the harmonicity of the source (i.e., subharmonics), which may be hindered in the presence of a harmonic musical background [43]. Another possible explanation is psychological, where the emotional quality of the manipulated vocal source may be dampened because of its superposition with a non-manipulated and possibly non emotionally-congruent background. In the present work, participants were instructed to rate the expression perceived in music as a whole, and not e.g. of a specific vocal source while ignoring the background [43], which may have also contributed to these effects. Finally, the explanation may also be technical, due to the possibly limited applicability of the transformation algorithms to non-vocal material. The fact that we did not present participants with a solo-instrument condition (without concurrent musical background) is limiting our ability to arbitrate between these possibilities, and could be considered for future work.

While the fact that singing voices can be expressively smiling, trembling or screaming may not appear surprising from a naturalistic, biological point of view, and is in accordance with comparative acoustic analyses of emotion production in speech and singing [61], it strongly contrasts with an 'artificialistic' view, prevalent for instance in the musicology of the great virtuoso performers of the nineteenth century [21], of singing voice as a disembodied musical instrument bearing no natural relation to the singer's body [68]. The present results suggest, on the contrary, that singing and non-vocal musical sounds can both be processed *as if* they were spoken voice, mobilizing cognitive mechanisms linked to the detection and interpretation of physiological phenomena. The violin stimuli used here were artificially constructed using voice-specific gestures and one may question their ecological validity, i.e. whether musicians can actually manipulate these aspects of their sounds. Many elements suggest they can. First, there are well-described acoustic similarities between the human voice and violin [39,46], which has a similar frequency range and a formant structure exhibiting vowel-like qualities [64], leading many to describe violin playing as sounding either male (*"He had a stroke so sweet, and made it speak like the voice of a man"* [60]) or female (*"There are in the music of the violin — if one does not see the instrument itself [...] — accents which are so closely akin to those of certain contralto voices, that one has the illusion that a singer has taken her place amid the orchestra"* [55]). Second, many traditional violin gestures can be said to ressemble the source-filter parameters manipulated in this work: while violin strings are ordinarily bowed or plucked in the center of the fingerboard, violinists intentionally bow strings at the other positions (e.g. close to the bridge: *sul ponticello*) to create variations in timbre, which may resemble the type of gesture found in smiling, or nasality [32]; vibrato is commonly produced by oscillating the left hand around the position where it stops the string against the fingerboard and, while typically slower, is a clear parent to singing vibrato and vocal tremor [57] (*"It's particularly interesting that it's singing that violin playing has always been said to imitate, with violinists considered the divas of instrumental playing. The ease with which a violinist produces portamento and vibrato is, of course, the main reason*, [39]); finally, in contemporary performance, high bow pressure can be used to create distortion and "scratching" sounds which may resemble vocal roughness [63]. Similar gestures are also found in other instruments, such as controlling brightness in brass instrument by employing slight changes in embouchure, akin to smiling [50], or saturated electrified instruments, which acoustic similarities to rough alarm calls have been studied in the field of animal communication [14]. All these examples suggest that cultural evolution has found ways, by virtue of innovations in organology, performance or repertoire, to map the natural expressive resources of spoken voice to musical parameters, and ritualize them into musical practice.

Furthering this idea, we tested two groups of (self-reported) musicians and non-musicians. A wealth of empirical evidence has shown that musical training enhances auditory and pitch processing [47] and the ability to recognize emotions in music [17], and that these effects transfer to recognizing emotions in speech [25,41,67]. It could therefore be expected that musicians should perform differently from non-musicians, either because of an enhanced ability to perceive subtle vocal cues in complex music mixes, because of greater familiarity with e.g. the instrumental timbre of the violin, or because of a different cultural understanding of cues like vibrato or spectrum. We found no evidence that it was the case: whether participants were self-declared musicians or non-musicians did not interact with the effect of the manipulations, in any of the sound types tested here. This pattern of results reinforces the notion that, when applied to musical material, the three acoustic manipulations considered here do not operate as domain-specific conventions, but are rather founded in natural vocal expression. Note however that it is questionable whether a small, 3-years-of-musical-practice difference between groups can elicit such behavioral variation, and that future work could consider better-controlled measures of musical ability before issuing strong conclusions about individual differences in how vocal expressions are perceived in music.

Finally, the work reported here is purely behavioral, and involves explicit ratings. From this sole comparison of vocal and musical expression, it is difficult to judge the extent to which the two types of processing are similar: they could involve similar sensorimotor representations (in effect hearing smiling violons *as if* they were smiling), or different representations converging at the same evaluation. Further work could attempt to clarify the sensory and cognitive mechanisms involved in the evaluation of specifically-vocal changes on non-vocal sources such as violins using adaptation paradigms with voice-instrument hybrid sources [13,16] or implicit sensorimotor paradigms such as facial mimicry (e.g., does one imitate a smiling violin ? [6]). It is also an open question whether the same sound variations would impart the same emotional effects in non-vocal natural sounds [44]. Even if the acoustic signatures considered here can be found elsewhere and have non-vocal origins (e.g. roughness in the rumble of thunder, or fluctuations of brightness in the colored noise of wind), it is still possible that our multimodal (audiovisual, proprioceptive, etc.) experience of similar signatures in voices gives meaning to these otherwise meaningless sound variations.

It also remains unknown whether the almost transparent transfer of vocal parameters to non-vocal musical sounds demonstrated here applies to all music, or all experiences of music. It is probable that vocal cues only drive expressivity for music that bears some amount of analogy to human vocalization, making it possible to hear it 'as if' it was voice [37]. This is notoriously the case of violin, as already noted, and it would therefore be interesting to test whether these results extend to other musical instruments. It is also possible that some of the present results depend on the specific music genres (contemporary commercial music) used in this study. This may be especially true of vocal tremor, which is found here to be congruent (more negative, less aroused) in both speech and music, while previous research with operatic singers has found discrepancies between the use of speech vibrato associated with sadness (like here) and sung vibrato with anger (unlike here, i.e. greater rather than lower arousal) [61]. More generally, the mechanism identified here is plausibly only one of a plurality of ways by which music can be expressive. Musical emotions are shaped by cultural-evolutionary processes occurring in a great diversity of contexts, which are likely to take biological foundation not only in communicative adaptations such as vocal signaling, but also expressive motion [31], environmental monitoring [44], coalitional interactions and infant care [45], and others. It is now important to understand how these mechanisms interact with each other to shape our emotional musical experiences.

Materials and Methods

Participants

N=60 participants (M=23.1yo, SD=3.2; female: 31) took part in the experiment. N=29 identified as musicians (more than 3 years of formal musical practice) and N=31 as non-musicians (no formal musical practice). All participants reported normal hearing, normal or corrected-to-normal vision and no neurological or psychiatric disorder.

Auditory Stimuli

We selected 14 excerpts from songs of various popular music genres (pop, jazz, rock), available as unmixed, multi-track recordings from the free online resource 'Mixing Secrets For the Small Studio' [2]. For each recording, we selected one full musical phrase (singing + accompaniment) of average duration M = 7 sec.

For each excerpt, we then used the available multi-tracks to create variants in 4 conditions: singing (the lead vocal track, without instrumental accompaniment), singing + accompaniment (the original song, composed of lead vocal track and instrumental accompaniment), violin + accompaniment (the original song in which the lead vocal track was replaced by a violin instrumental track matching the main melody) and speech (a recording of a transcription of the lyrics of the lead vocal track, performed as non-musical speech). None of the accompaniment tracks in conditions 'singing + accompaniment' and 'violin + accompaniment' contained additional background vocals.

The instrumental track in the 'violin + accompaniment' condition was recorded on the violin by a semi-professional musician (*Chœurs et Orchestres des Grandes Écoles*) in overdubbing conditions matching the pitch and phrasing of the original vocal track. Speech tracks in the 'speech' condition were recorded by two native English speakers (one male, one female, matching the gender of the original singer), who performed a spoken, neutral-tone rendition of the lyrics, without knowing nor hearing that these were originally singing material. All recordings were performed in music production studios in IRCAM (Paris, France) by a professional sound engineer (D.B.). In addition, we also selected 12 'scream' stimuli from a previous study [42], which consisted of short, isolated shouts of phoneme /a/, recorded by 6 male and 6 female actors. These resulted in 68 sets of multi-track stimuli, matched in 5 different conditions (Speech: 14; singing: 14; singing + accompaniment: 14; violin + accompaniment: 14; and an unmatched set of 12 screams).

Before mixing, the lead track (vocal in conditions 'speech', 'screams', 'singing', 'singing + accompaniment'; violin in condition 'violin + accompaniment') in each of the multi-track stimuli was then processed with three acoustic manipulations simulating specifically-vocal behaviours: smiling (two levels: *smile* and *unsmile*), vocal tremor (one level: *tremor*) and vocal

[2] http://www.cambridge-mt.com/ms-mtk.htm

roughness (one level: *tension*). Finally, the tracks of each stimulus were mixed by a professional sound engineer (DB), resulting in 68 non-manipulated and 272 manipulated stereo stimuli.

Audio manipulation algorithms

Contrary to previous work, which manipulated the complete music ensemble of their stimuli [34,44], we took advantage of professional multi-track recordings and only applied our acoustic manipulations to the 'lead' track in each stimulus, before mixing it down with their non-manipulated accompaniment. This applied to vocal tracks in the 'speech', 'screams', 'singing', 'singing + accompaniment' conditions, and to violin tracks in the 'violin + accompaniment' condition.

Vocal and violin tracks manipulated in the 'smiling' condition underwent a spectral transformation designed to simulate the effect of stretching lips while talking [8]. The transformation extracts the spectral envelope of each successive time frames of the incoming signal, and uses a technique called 'frequency warping' to stretch the maxima and minima of this envelope in the $[100, 5000]$Hz frequency band, which loosely correspond to the first five formants of a vocal signal [54]. It then reconstructs the original signal using a phase-vocoder algorithm. In previous work, the transformation was validated to be both natural and effective in simulating the impression of a smiling voice [6,8]. Importantly, like the other two transformations, the procedure can be applied to non-vocal sounds without modification, which allows us to compare the effect of the transformation on vocal (conditions 'speech', 'screams', 'singing', 'singing + accompaniment') and non-vocal (condition 'violin + accompaniment') tracks. The intensity of the transformation is controlled by multiplicative parameter α, used to stretch or compress the signal's spectral envelope. We applied the smiling transformation in two levels: 'smile' ($\alpha = 1.25$), which increased the amount of smile compared to the original, non-manipulated stimuli; and 'unsmile' ($\alpha = 0.85$), which decreased the amount of smile.

Vocal and violin tracks manipulated in the 'vocal tremor' condition underwent a cyclical pitch shifting transformation designed to simulate vibrato in afraid/anxious voices (DAVID [56], available open-source at `https://forum.ircam.fr/projects/detail/david/`). Pitch shifting denotes the multiplication of the fundamental frequency (f0) of the original voice signal by a factor β (e.g. + 25 cents, a 1.5% change of f0). Here, we apply a periodic modulation of voice f0, implemented as a sinusoidal modulation of the pitch shift effect with a fixed depth, rate and a small random variation of the rate to increase naturalness. For vocal tremor stimuli in this work, we used a depth of 25 cents, rate of 8Hz and a randomness parameter of 20%. These parameters were validated in previous work to be both natural and effective in simulating the impression of an anxious voice [56]. Like the other two transformations, the procedure can be applied to either vocal or non-vocal sounds without modification.

Finally, vocal and violin tracks manipulated in the 'vocal roughness' condition underwent an amplitude modulation procedure designed to simulate non-linear phenomena in vocal fold vibration (namely, subharmonics) due to high vocal effort and arousal (ANGUS [42], available open-source at `https://forum.ircam.fr/projects/detail/angus`). The transformation operates by multiplying the original signal by a lower-frequency modulating signal synchronized on its fundamental frequency (f0/2), which creates subharmonics at f0+f0/2 and f0-f0/2, high-pass filtering the resulting subharmonics and mixing them together with the original signal with mixing factor $\alpha = 1$. These parameters were validated in previous work to be both natural and effective in simulating the impression of a negatively aroused voice [42] and, like all others, the procedure can be applied to either vocal or non-vocal sounds without modification.

Procedure

Participants were presented with pairs of stimuli composed of matched manipulated and non-manipulated versions of the same recording. There were 4 transformation conditions (68 smile vs non-manipulated pairs; 68 unsmile vs non-manipulated pairs; 68 tremor vs non-manipulated pairs; 68 rough vs non-manipulated pairs) as well as 68 non-manipulated vs non-manipulated control pairs. Presentation order within a pair (manipulated vs non-manipulated, or non-manipulated vs manipulated) was randomized within-participant.

For each pair, participants were asked to evaluate the emotion that was expressed by one recording compared to the other, using a 7-point Likert scale for valence (1= more negative, 4= no difference, 7= more positive) and arousal (1= more calm, 4= no change, 7= more energetic). The order of the comparison within a pair (rating the first recording against the second, or rating the second recording against the first) was fixed within-participant, but counterbalanced between participants. This procedure was the same as [44].

It is to be noted that results obtained with such an explicit pairwise comparison procedure may differ from those obtained e.g. with single-item rating scales [12] or implicit methods such as the Implicit Association Test [4]. By emphasizing the acoustic difference within pairs, the pairwise method allows answering a low-level decoding question ("if forced to focus attention on a given acoustic change, what emotional interpretation would that change result in ?"). Having maximum experimental control over the participant's locus of attention is important because there are well-known individual- and group-level differences in how people attend to elements in music [29]. Conversely, the pairwise methods does not allow to address questions such as "would attention be spontaneously be drawn to that feature in a single (unpaired) presentation, compared to other features of the sound ?". Like rating scales, it is also plagued with demand effects, and cannot establish whether such interpretations would be more spontaneously scored as valence/arousal or other untested and potentially non-emotional constructs. We mitigate these effects here by randomizing trials over all manipulations (i.e. having pairs which differ unpredictably on several possible dimensions) and adding control pairs (i.e. pairs with no stimulus difference).

The experiment was divided into 3 blocks, preceded with a short training block. In the first block participants judged the 3 musical conditions: 'singing', 'singing + accompaniment','violin+ accompaniment'. In this block, all stimulus pairs were randomized across conditions. Participants then rated 'speech' stimuli in the second block and 'scream' stimuli in the third block. The order of these 3 blocks was fixed for all participants. This procedure (non-music vocal sounds last) was adopted to avoid demand effects where a response strategy learned on speech/screams could then transfer artificially to music stimuli. The procedure leaves the converse risk that participants have learned a strategy on music, and then transferred it to speech and screams, but we alleviated the impact of that possibility on our subsequent interpretations of results by having clear, preregistered hypotheses about the impact of the three manipulations on the latter non-musical stimuli, and finding that these predictions were met.

Ethics. All participants tested at the Sorbonne-INSEAD Center for Behavioral Science. The experiment was approved by the Institut Européen d'Administration des Affaires (INSEAD) IRB. All participants gave their informed consent and were debriefed about the purpose of the research immediately after the experiment.

Data Accessibility. Examples of stimuli, experimental data and analysis code provided as supplementary material. All stimuli available at `https://archive.org/details/smiling_violins`.

Authors' Contributions. Authors DB, LG and JJA designed the study. PA, LR, ML designed the vocal transformations and assisted in stimulus generation. DB conducted data collection. DB, LG and JJA analysed the data. DB and JJA wrote the manuscript, with feedback from all others.

Competing Interests. No competing interests.

Funding. Study funded by European Research Council Starting Grant (CREAM 335634), Agence Nationale de la Recherche grant (REFLETS, SEPIA), and Fondation Pour l'Audition (FPA RD-2018-2).

Acknowledgements. The authors thank Rosalie Ollivier for recording the violin stimuli.

References

1. Andrey Anikin.
   Soundgen: An open-source tool for synthesizing nonverbal vocalizations.
   Behavior research methods, 51(2):778–792, 2019.
2. Andrey Anikin.
   A moan of pleasure should be breathy: the effect of voice quality on the meaning of human nonverbal vocalizations.
   Phonetica, 77(5):327–349, 2020.
3. Andrey Anikin.
   The perceptual effects of manipulating nonlinear phenomena in synthetic nonverbal vocalizations.
   Bioacoustics, 29(2):226–247, 2020.
4. Andrey Anikin and Niklas Johansson.
   Implicit associations between individual properties of color and sound.
   Attention, Perception, & Psychophysics, 81(3):764–777, 2019.
5. Andrey Anikin, Katarzyna Pisanski, and David Reby.
   Do nonlinear vocal phenomena signal negative valence or high emotion intensity?
   Royal Society Open Science, 7(12):201306, 2020.
6. Pablo Arias, Pascal Belin, and Jean-Julien Aucouturier.
   Auditory smiles trigger unconscious facial imitation.
   Current Biology, 28(14):R782–R783, 2018.
7. Pablo Arias, Laura Rachman, Marco Liuni, and Jean-Julien Aucouturier.

Beyond correlation: acoustic transformation methods for the experimental study of emotional voice and speech.
Emotion Review, page 1754073920934544, 2020.

8. Pablo Arias, Catherine Soladie, Oussema Bouafif, Axel Robel, Renaud Seguier, and Jean-Julien Aucouturier.
Realistic transformation of facial and vocal smiles in real-time audiovisual streams.
IEEE Transactions on Affective Computing, 2018.
9. Luc H Arnal, Adeen Flinker, Andreas Kleinschmidt, Anne-Lise Giraud, and David Poeppel.
Human screams occupy a privileged niche in the communication soundscape.
Current Biology, 25(15):2051–2056, 2015.
10. Jean-Julien Aucouturier, Petter Johansson, Lars Hall, Rodrigo Segnini, Lolita Mercadié, and Katsumi Watanabe.
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, 2016.
11. Lisa Feldman Barrett, Kristen A Lindquist, and Maria Gendron.
Language as context for the perception of emotion.
Trends in cognitive sciences, 11(8):327–332, 2007.
12. Anja Belz and Eric Kow.
Comparing rating scales and preference judgements in language evaluation.
In *Proceedings of the 6th International Natural Language Generation Conference*, 2010.
13. Patricia EG Bestelmeyer, Julien Rouger, Lisa M DeBruine, and Pascal Belin.
Auditory adaptation in vocal affect perception.
Cognition, 117(2):217–223, 2010.
14. Daniel T Blumstein, Gregory A Bryant, and Peter Kaye.
The sound of arousal in music is context-dependent.
Biology letters, 8(5):744–747, 2012.
15. Paul Boersma.
Praat, a system for doing phonetics by computer.
Glot. Int., 5(9):341–345, 2001.
16. Casady Bowman and Takashi Yamauchi.
Processing emotions in sounds: cross-domain aftereffects of vocal utterances and musical sounds.
Cognition and Emotion, 31(8):1610–1626, 2017.
17. São Luís Castro and César F Lima.
Age and musical expertise influence emotion recognition in music.
Music Perception: An Interdisciplinary Journal, 32(2):125–142, 2014.
18. Chee Seng Chong, Jeesun Kim, and Chris Davis.
Disgust expressive speech: The acoustic consequences of the facial expression of emotion.
Speech Communication, 98:68–72, 2018.
19. Eduardo Coutinho and Nicola Dibben.
Psychoacoustic cues to emotion in speech prosody and music.
Cognition & emotion, 27(4):658–684, 2013.
20. Charles Darwin.
The descent of man and selection in relation to sex.
Murray, 2nd edition, 1874.
21. James Q Davies.
Romantic Anatomies of Performance.
Univ of California Press, 2014.
22. Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang.
The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech.
In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–9. IEEE, 2013.
23. Panteleimon Ekkekakis.
The measurement of affect, mood, and emotion: A guide for health-behavioral research.
Cambridge University Press, 2013.
24. Nicolas Escoffier, Jidan Zhong, Annett Schirmer, and Anqi Qiu.
Emotional expressions in voice and music: same code, same effect?
Human Brain Mapping, 34(8):1796–1810, 2013.
25. Eliot Farmer, Crescent Jicol, and Karin Petrini.
Musicianship enhances perception but not feeling of emotion from others' social interaction through speech prosody.
Music Perception, 37(4):323–338, 2020.
26. Irune Fernández-Prieto, Jordi Navarra, and Ferran Pons.
How big is this sound? crossmodal association between pitch and size in infants.
Infant Behavior and Development, 38:77–81, 2015.

27. W Tecumseh Fitch.
Musical protolanguage: Darwin's theory of language evolution revisited.
Birdsong, speech, and language: Exploring the evolution of mind and brain, 489:503, 2013.
28. W Tecumseh Fitch, Jürgen Neubauer, and Hanspeter Herzel.
Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production.
Animal behaviour, 63(3):407–418, 2002.
29. John M Geringer and Clifford K Madsen.
Focus of attention to elements: Listening patterns of musicians and nonmusicians.
Bulletin of the Council for Research in Music Education, pages 80–87, 1995.
30. Cheryl L Giddens, Kirk W Barron, Jennifer Byrd-Craven, Keith F Clark, and A Scott Winter.
Vocal indices of stress: a review.
Journal of voice, 27(3):390–e21, 2013.
31. Bruno L Giordano, Hauke Egermann, and Roberto Bresin.
The production and perception of emotionally expressive walking sounds: Similarities between musical performance and everyday motor activity.
PloS one, 9(12):e115587, 2014.
32. Knut Guettler.
The violin bow in action: A sound sculpturing wand.
International Center for Mechanical Sciences, pages 1–10, 2006.
33. Julia C Hailstone, Rohani Omar, Susie MD Henley, Chris Frost, Michael G Kenward, and Jason D Warren.
It's not what you play, it's how you play it: Timbre affects perception of emotion in music.
Quarterly journal of experimental psychology, 62(11):2141–2155, 2009.
34. Gabriella Ilie and William Forde Thompson.
A comparison of acoustic cues in music and speech for three dimensions of affect.
Music Perception: An Interdisciplinary Journal, 23(4):319–330, 2006.
35. Patrik N Juslin and Petri Laukka.
Communication of emotions in vocal expression and music performance: Different channels, same code?
Psychological bulletin, 129(5):770, 2003.
36. Patrik N Juslin and Daniel Vastfjall.
Emotional responses to music: The need to consider underlying mechanisms.
Behavioral and brain sciences, 31(5):559, 2008.
37. P. Kivy.
Sound sentiment: An essay on the musical emotions.
Temple University Press, 1989.
38. Anne-Maria Laukkanen, Irma Ilomäki, Kirsti Leppänen, and Erkki Vilkman.
Acoustic measures and self-reports of vocal fatigue by female teachers.
Journal of Voice, 22(3):283–289, 2008.
39. Daniel Leech-Wilkinson.
The changing sound of music: Approaches to studying recorded musical performances.
Centre for the History and Analysis of Recorded Music London, UK, 2009.
40. Donald J Leslie.
Apparatus for imposing vibrato on sound, December 23 1952.
US Patent 2,622,692.
41. César F Lima and São Luís Castro.
Speaking to the trained ear: musical expertise enhances the recognition of emotions in speech prosody.
Emotion, 11(5):1021, 2011.
42. Marco Liuni, Luc Ardaillon, Louise Bonal, Lou Seropian, and Jean-Julien Aucouturier.
Angus: Real-time manipulation of vocal roughness for emotional speech transformations.
arXiv preprint arXiv:2008.11241, 2020.
43. Marco Liuni, Emmanuel Ponsot, Gregory A Bryant, and Jean-Julien Aucouturier.
Sound context modulates perceived vocal emotion.
Behavioural processes, 172:104042, 2020.
44. Weiyi Ma and William Forde Thompson.
Human emotions track changes in the acoustic environment.
Proceedings of the National Academy of Sciences, 112(47):14563–14568, 2015.
45. Samuel A Mehr, Max M Krasnow, Gregory A Bryant, and Edward H Hagen.
Origins of music in credible signaling.
Behavioral and Brain Sciences, pages 1–41, 2020.

46. David Milsom.
Theory and practice in late nineteenth-century violin performance: an examination of style in performance, 1850-1900.
Ashgate Publishing Ltd., 2003.
47. Sylvain Moreno, Carlos Marques, Andreia Santos, Manuela Santos, São Luís Castro, and Mireille Besson.
Musical training influences linguistic abilities in 8-year-old children: more evidence for brain plasticity.
Cerebral Cortex, 19(3):712–723, 2009.
48. Cleopatra Christina Moshona.
On the psychoacoustics of vocal tremor: identifying severity predictor variables.
Master's thesis, Institut für Sprache und Kommunikation, Technische Universität Berlin, 2018.
49. John G Neuhoff.
An adaptive bias in the perception of looming auditory motion.
Ecological Psychology, 13(2):87–110, 2001.
50. Lisa Norman, JP Chick, DM Campbell, Arnold Myers, and Joël Gilbert.
Player control of'brassiness' at intermediate dynamic levels in brass instruments.
Acta Acustica united with Acustica, 96(4):614–621, 2010.
51. John J Ohala.
The acoustic origin of the smile.
The Journal of the Acoustical Society of America, 68(S1):S33–S33, 1980.
52. Aniruddh D Patel.
Music, language, and the brain.
Oxford university press, 2010.
53. Marius V Peelen, Anthony P Atkinson, and Patrik Vuilleumier.
Supramodal representations of perceived emotions in the human brain.
Journal of Neuroscience, 30(30):10127–10134, 2010.
54. E. Ponsot, P. Arias, and JJ. Aucouturier.
Uncovering mental representations of smiled speech using reverse correlation.
Journal of the Acoustical Society of America (submitted), 2018.
55. Marcel Proust.
Swann's Way: In Search of Lost Time, Volume 1.
Yale University Press, 1913 (ed. 2013).
56. Laura Rachman, Marco Liuni, Pablo Arias, Andreas Lind, Petter Johansson, Lars Hall, Daniel Richardson, Katsumi Watanabe, Stephanie Dubal, and Jean-Julien Aucouturier.
David: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech.
Behavior Research Methods, pages 1–21, 2017.
57. Lorraine A Ramig and Thomas Shipp.
Comparative measures of vocal tremor and vocal vibrato.
Journal of Voice, 1(2):162–167, 1987.
58. Deborah Ross, Jonathan Choi, and Dale Purves.
Musical intervals in speech.
Proceedings of the National Academy of Sciences, 104(23):9852–9857, 2007.
59. Magdalena Rychlowska, Rachael E Jack, Oliver GB Garrod, Philippe G Schyns, Jared D Martin, and Paula M Niedenthal.
Functional smiles: Tools for love, sympathy, and war.
Psychological science, 28(9):1259–1270, 2017.
60. William Sandys and Simon Andrew Forster.
The history of the violin: and other instruments played on with the bow from the remotest times to the present.
JR Smith, 1864.
61. Klaus R Scherer, Johan Sundberg, Lucas Tamarit, and Gláucia L Salomão.
Comparing the acoustic expression of emotion in the speaking and the singing voice.
Computer Speech & Language, 29(1):218–235, 2015.
62. Philippe Schlenker.
Outline of music semantics.
Music Perception: An Interdisciplinary Journal, 35(1):3–37, 2017.
63. Patricia Strange and Allen Strange.
The contemporary violin: Extended performance techniques, volume 7.
Scarecrow Press, 2003.
64. Hwan-Ching Tai, Yen-Ping Shen, Jer-Horng Lin, and Dai-Ting Chung.
Acoustic evolution of old italian violins from amati to stradivari.
Proceedings of the National Academy of Sciences, 115(23):5926–5931, 2018.

65. Ana Tajadura-Jiménez, Aleksander Valjamae, and Daniel Vastfjall.
Emotional bias for the perception of rising tones.
Journal of the Acoustical Society of America, 123(5):3245, 2008.

66. William Forde Thompson, Manuela M Marin, and Lauren Stewart.
Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage hypothesis.
Proceedings of the National Academy of Sciences, 109(46):19027–19032, 2012.

67. William Forde Thompson, E Glenn Schellenberg, and Gabriela Husain.
Decoding speech prosody: Do music lessons help?
Emotion, 4(1):46, 2004.

68. Holly Watkins and Melina Esse.
Down with disembodiment; or, musicology and the material turn.
Women and Music: A Journal of Gender and Culture, 19(1):160–168, 2015.

69. Robert J Zatorre and Shari R Baum.
Musical melody and speech intonation: Singing a different tune.
PLoS Biol, 10(7):e1001372, 2012.