Larrouy-Maestri, P., Poeppel, D., & Pfordresher, P. Q. (in press). Pitch units in music and speech prosody. In R. Wiese & M. Scharinger (Eds). *How language speaks to music: prosody from a cross-domain perspective*. Berlin: de Gruyter.

Pitch units in music and speech prosody

Pauline Larrouy-Maestri,[1] David Poeppel,[1,2,3] & Peter Q. Pfordresher[4]

1. Neuroscience Department, Max-Planck-Institute for Empirical Aesthetics, Germany
2. Psychology Department, New York University, USA
3. Center for Language, Music, and Emotion, New York, USA
4. Department of Psychology, University at Buffalo, State University of New York, USA

Researchers have long been intrigued by the degree to which language and music share acoustic properties, and whether our brain processes input from each domain using similar or distinct resources (Jackendoff, 2008; Patel, 2008). However, the comparison of these two domains is challenging, as has been extensively discussed by Fritz, Poeppel, Trainor, et al. (2013). One key point of that chapter, focused on the neurobiology of language, speech, and music, is that while the 'units' or 'primitives' that define perceptually separable events in each domain (phonemes, syllables versus notes, melody, etc.) might be different in kind, many cognitive or computational operations (e.g. segmentation, concatenation, hierarchy formation, categorization, etc.) might be similar.

A first key step towards such cross-domain investigation and comparison consists, on this perspective, in identifying the relevant units and processes to be compared in each domain. In this chapter, we focus on one feature: fundamental frequency (f_0) dynamics, associated with melodies in music and with intonation in speech. f_0 is important both in human communication and in conveying musical melodies. A next step consists of testing different hypotheses regarding the integration of f_0 units of different timescales. Theoretically, one hypothesis is that auditory processing is *statistical* in nature: listeners' perception is based on a summary statistic (say mean f_0) across a larger unit, taking into account the dynamic pitch information contained in the smaller one(s). A different hypothesis is that auditory processing is *teleological*. On this view, listeners perceive small units based on their relationship to larger units, assuming that the larger unit constitutes the endgoal point for a (motor) plan. This form of listening focuses on

the sound sequence trajectory formed by larger-scale units as opposed to summary statistics within a unit. Note that these two hypotheses are not necessarily exclusive, but could be combined or varied, depending on the function of the process. In this chapter we focus in particular on small f_0 units in the context of singing performances. In addition to shedding light on the processes involved in the perception of pitch when listening to music, we aim to lay the ground for a meaningful comparison between the music and language domains.

Role of pitch in speech prosody and music

Listeners' sensitivity to pitch variation is usually examined with pitch discrimination tasks (see, e.g., the tasks in the toolbox of Soranzo & Grassi, 2014). Thresholds are generally far lower than necessary to process pitch changes in music. This is particulary true for listeners with formal musical training, and they vary slightly from one individual to the other (e.g., Micheyl, Delhommeau, Perrot, & Oxenham, 2006; Moore, 1973). For larger contexts (i.e., non isolated sounds), the literature on dynamic changes highlights the perception of glides between pure tones (Lyzenga, Carlyon, & Moore, 2004) and glides at the end of pure tones (Wang, Tan, & Martin, 2013), confirming the perceptual relevance of quick pitch movements. Such dynamic changes influence the perception of tone sequences (Kerivan & Carey, 1976). However, it should be emphasized that not every pitch change is perceived. Indeed, the length and slope of the variation need to reach a threshold to be perceived. For instance, twenty milliseconds of signal are necessary to identifying the sweep direction in frequency-modulated signals (Gordon & Poeppel, 2002; Luo, Boemio, Gordon, & Poeppel, 2007), and the rate of frequency rise (or fall) should be more than $0.16/T^2$, where T is the duration of the glide ('t Hart, Collier, & Cohen, 1990).

Besides the discrimination abilities of listeners and the f_0 characteristics of a stimulus, the perceptual relevance of pitch obviously depends on the context. For instance, Warrier and Zatorre (2002) observed that small pitch manipulations are better perceived in melodic contexts compared to intervals. In fact, the perceptual relevance of pitch (and its role) is highly sensitive to the domain under consideration.

Pitch in speech prosody

Speech conveys linguistic information through the combination of lexical elements, while also carrying paralinguistic information, such as information about speaker gender (Hillenbrand & Clarke 2009; Latinus & Belin, 2011; Latinus, McAleer, Bestelmeyer, & Belin,

2013), age (see Gamba, 2014, for a review; Shigeno, 2016), size (Fitch & Giedd, 1999; Rendall, Vokey, & Nemeth, 2007; Smith, Patterson, Turner, Kawahara, & Irino, 2005), or personality (McAleer, Todorov, & Belin, 2014). Importantly, prosody also carries information about the speakers' emotions (Banse & Scherer, 1996), which is commonly refered to as emotional prosody. Emotional prosody is a crucial ingredient of human communication, since affective communication has an adaptive value, social function, and influence on the behavior of the communication partner (Bandstra, Chambers, McGrath, & Moore, 2011; Fischer & Manstead, 2008; Keltner & Haidt, 1999; Shariff & Tracy, 2011; Sinaceur, Kopelman, Vasiljevic, & Haag, 2015; Wubben, De Cremer, & van Dijk, 2011).

As summarized by Bänziger, Hosoya, and Scherer (2015), emotional prosody can be examined from two different angles: acoustic parameters of stimuli or recognition of emotion by listeners. Numerous studies using filtered, foreign, or pseudo-language material have confirmed that emotions can be identified well above chance level on the basis of acoustic cues alone (e.g., Bänziger & Scherer, 2005; Bryant & Barret, 2008; Dromey, Silviera, & Sandor, 2005; Pell, Monetta, Paulmann, & Kotz, 2009; Pell & Skorup, 2008; Scherer, Banse, & Wallbott, 2001; Wildgruber, Riecker, Hertrich, Erb, Grodd, Ethofer, & Ackermann, 2005). The prevalent features associated with emotional prosody include pitch-related features, formant frequencies, timing, voice-quality parameters, and articulation parameters (e.g., Juslin & Laukka, 2003).

Among these acoustic features, pitch has been repeatedly reported as crucial (already by Fairbanks & Pronovost, 1939). By directly comparing listeners' ratings and acoustic analyses of the same speech materials, Banse and Scherer (1996) highlighted the role of mean f_0 for the distinction of the emotional states of actors pronouncing meaningless sentences. In other words, the global pitch height provides information about the emotional state of a speaker. Such summary statistics have also been associated with social relations. For instance, Zoghaib (2019) observed that speakers with high-pitched voices were perceived as the most competent, whereas Pavela Banai, Banai, and Bovan (2016) report that political "winners" (according to presidential election outcomes) had lower-pitched voices. People spontaneously modulate their vocal pitch, which influenced listeners' perception of their 'rank' (Cheng, Tracy, Ho, & Henrich, 2016), supporting that pitch has a key role in human communication.

Pitch in music

In contrast to speech, the communicative aspect of music is not clearly defined. However, the organization of pitch in time provides information that allows listeners to

recognize, evaluate, and enjoy a musical performance. In different musical styles, tones are associated with various symbols that have been developed since the late middle ages (Cohen, 2002). Tones are usually defined as the smallest discrete unit in Western music and their organization along the melody defines its tonality that plays an important role in the expectations of the listeners of a specific culture. There exist different musical systems, each with its scales, rules, and grammars (Cross, 2001; Krumhansl, 1979; Lerdahl & Jackendoff, 1983; Ringer, 2002; Thompson, 2013).

Western musical culture is generally based on the equal temperament system, arguably the most common tuning system for the past few hundred years (Parncutt & Hair, 2018). The smallest theoretical distance in tempered systems between the tones of a scale is called a semitone (assumed to be equal), and musical intervals can be described in terms of number of semitones. Studies exploring the perception of intervals confirmed the relevance of semitones as discrete categories (Burns & Ward, 1978; Zarate, Ritson, & Poeppel, 2012), which is a much larger difference than the pitch discrimination thresholds observed in a general population. In other words, pitch perception in music is linked to the musical system and cannot be explained only by pitch discrimination abilities. In principle, the identity of a musical piece should remain invariant when other acoustic dimensions such as timbre or tempo are altered, as long as the tones fall within the right semitone category. For instance, "Jingle Bells" played on a keyboard, a trombone, or sung by a young child, at a slow or fast tempo remains "Jingle Bells" as long as the relationship between tones that compose this melody is preserved. Note that a melody transposed is still recognized (Stalinski & Schellenberg, 2010), which supports the notion that the general pitch height is not relevant as long as the size of the intervals between consecutive notes is preserved.

In addition to the relevance of pitch in melody recognition, pitch manipulations affect the correctness of a musical performance. By growing up in a specific culture, listeners develop an internal representation of what is 'correct' in terms of pitch accuracy (Larrouy-Maestri, Lévêque, Schön, Giovanni, & Morsomme, 2013; Larrouy-Maestri, Magis, Grabenhorst, & Morsomme, 2015). Larrouy-Maestri (2018) reported a series of experiments designed to clarify the amount of pitch deviation that causes listeners to evaluate a musical sequence as correct or incorrect. Listeners' tolerance with regard to mistuning (i.e., the range of pitch variability heard as correct) was examined by presenting parametrically manipulated melodies to a large group of listeners with different musical expertise and asking them to identify each version as in-tune or out-of-tune. In four experiments, listeners' tolerance with regard to mistuning was compared

across melodies to examine the effect of interval size, direction, familiarity. Taken together, the results support the existence of a 'tolerance zone' when evaluating the correctness of a melody and show that listeners accept pitch deviations that are both larger than the typical pitch discrimination thresholds and smaller than a semitone.

Listeners' perception of correctness in naturalistic music of different genres can now be quantified with the newly developed Mistuning Perception Test (MPT, Larrouy-Maestri, Harrison, & Müllensiefen, 2019). Interestingly, in Larrouy-Maestri (2018) we also showed that the notion of correctness is slightly different when the known melody has been mostly previously heard as slightly 'deviant,' such as the rarely accurately performed "Happy Birthday" song - suggesting the role of previous exposure in the shaping of a correctness category. Even if a direct comparison of correctness judgment and preferences/liking would be necessary to confirm the relevance of pitch in non-technical judgments, it seems reasonable to assume that difficult-to-recognize or totally out-of-tune performances would rarely be the most appreciated.

Definition of pitch units

Units in speech prosody

Banse and Scherer's (1996) foundational study underscores that the average of f_0 over sentences is a parameter that contributes to the perception of different emotions such as hot anger, panic fear, anxiety, desperation, elation, boredom, and contempt. The pitch height of entire sentences conveys information about an intended (and perceived as such) emotional state. In other words, listeners might process pitch information at the **sentence level**. However, such large units cannot be the only ones. In the domain of speech comprehension, the multi-time resolution hypothesis suggests that dynamic changes in small time windows are relevant to listeners (Poeppel, 2003; Teng, Tian, Poeppel, 2016). Linguistic elements of different sizes, such as vowel, segments, syllables, words, phrases, and sentences, are concurrently tracked and temporally integrated (Ding, Melloni, Yang, Wang, Zhang, & Poeppel, 2017; Ding, Melloni, Zhang, Tian, & Poeppel, 2015; Keitel, Gross, & Kayser, 2018). With regard to emotional prosody, it seems reasonable to hypothesize that units of different size exist and are integrated over time (Jiang, Paulmann, Robin & Pell, 2015; Pell & Kotz, 2011; Waaramaa, Laukkanen, Airas, & Alku, 2010). Whereas the most obvious carrier of emotional prosody is the intonation contour over entire phrases or sentences, words carry prosodic information (also linguistic information in the case of tone languages since it generates different lexical items). Likewise,

syllables and even single phonological segments might carry prosodic information as well, which support the need of studies investigating units of very different levels of organization and abstraction.

Several arguments suggest the existence of **smaller units** for emotional prosody comprehension. For instance, the modulation of f_0 in sentences seems to play a role, as shown by the significant contribution of the standard deviation of the f_0 in the recognition of hot anger, anxiety, desperation, and contempt (Banse & Scherer, 1996). Also, previous research has focused on other units such as segments (e.g., Shami & Kamel, 2005), syllables (e.g., Agrima, Farchi, Elmazouzi, Mounir, & Mounir, 2019), phonemes (e.g., Bitouk, Nenkova, & Verma, 2009), or selected vowels (e.g., Goudbeek, Goldman, & Scherer, 2009). The fact that emotional states can be described or recognized from these temporally shorter units indicates that the information is not equally distributed over the full sentence. Pell and Kotz (2011) as well as Nordström and Laukka (2019) reached the same conclusion with a different approach. They used a gating paradigm and showed that the size of units in speech affected the identification of specific emotions. Recognition improved over time and thus depended on the amount of information accumulated (i.e., gate duration), which suggests that emotional prosody might be a dynamic signal constituted of small units (i.e., smaller than the full sentence) containing critical acoustic information, including f_0, potentially integrated over time.

The hypothesis of the relevance of small windows in prosody has also been supported by recent studies using the reverse correlation method which permits access to mental representations without presenting the actual/stereotypical target. For instance, Ponsot, Burred, Belin, and Aucouturier (2018) presented pairs of randomly manipulated pitch contours of single words to listeners and observed specific contours associated with attitudes such as 'dominance,' as well as great individual differences. Using morphing methods, other studies confirmed the role of intonation (see Belin, Boehme, & McAleer, 2017, about the perception of trustworthiness; Sammler, Grosbras, Anwander, Bestelmeyer, & Belin, 2015, about the perception of questions/statements). Altogether, these studies provide evidence for the perceptual and social relevance of pitch contours. The exact nature of pitch movements over time is under study. For example, van Rijn, Poeppel, & Larrouy-Maestri (in prep.) developed new features describing slopes, general shapes, and pitch changes of spoken sentences. The addition of these features to an existing standard feature set (eGemaps of Eyben, Scherer, Schuller, et al., 2016) significantly improves the classification of emotions. The preliminary results offer promising perspectives with regard to the identification of small units in speech prosody.

Units in music

Documented units

Several models (e.g., Pearce, 2005; Temperley, 2013) have been proposed to describe the organization of tones within the more general structure of musical phrases, thereby quantifying listeners' expectations grounded both in Gestalt-like principles and in statistical learning (Morgan, Fogel, Nair, & Patel, 2019). Whereas the exact cause of listeners' expectations remains to be clarified, it appears that lay listeners use tone units and implicitly develop knowledge about the music system of their culture (Bigand & Poulin-Charronnat, 2006; Hannon & Trainor, 2007; Larrouy-Maestri, 2018; Marmel, Tillmann, & Dowling, 2008; McDermott, Schultz, Undurraga, & Godoy, 2016). It has also been repeatedly shown that irregularities with regard to the music system are detected in the brain responses of listeners (e.g., Koelsch & Friederici, 2003). In addition to behavioral arguments about melodic recognition and evaluation, these findings again support the organization of tones in larger structures following specific rules. Taken together, studies on music perception have repeatedly confirmed the perceptual relevance of two units of different size: tones and melodies.

Whereas tones are commonly described as the smallest discrete units that allow for the construction of structured melodies in the tonal system of Western culture, one might wonder whether smaller units, such as pitch fluctuations within the tones themselves, have relevance in music. It has been shown that pitch alterations such as vibrato rate and extent are important features in pitch perception (e.g., van Besouw & Howard, 2009) and in music evaluation (e.g., Larrouy-Maestri, Morsomme, Magis, & Poeppel, 2017). Considered as a vocal quality developed through training (Mürbe, Zahnert, Kuhlisch, & Sundberg, 2007), vibrato is appreciated in highly trained voices (Garnier, Henrich, Castellengo, Sotiropoulos, & Dubois, 2007). Even if vibrato is not limited to operatic voices, it remains something special, and is rarely associated with the singing of occasional/untrained singers. As illustrated in Figure 1A, a series of sung tones can be described as a series of flat lines (f_0) in a spectrogram, but zooming in (see Figure 1B) reveals the presence of unsteady parts. In order to examine widely present pitch fluctuations, we focus here on pitch movements happening at the start and end of tones: 'scoops.'

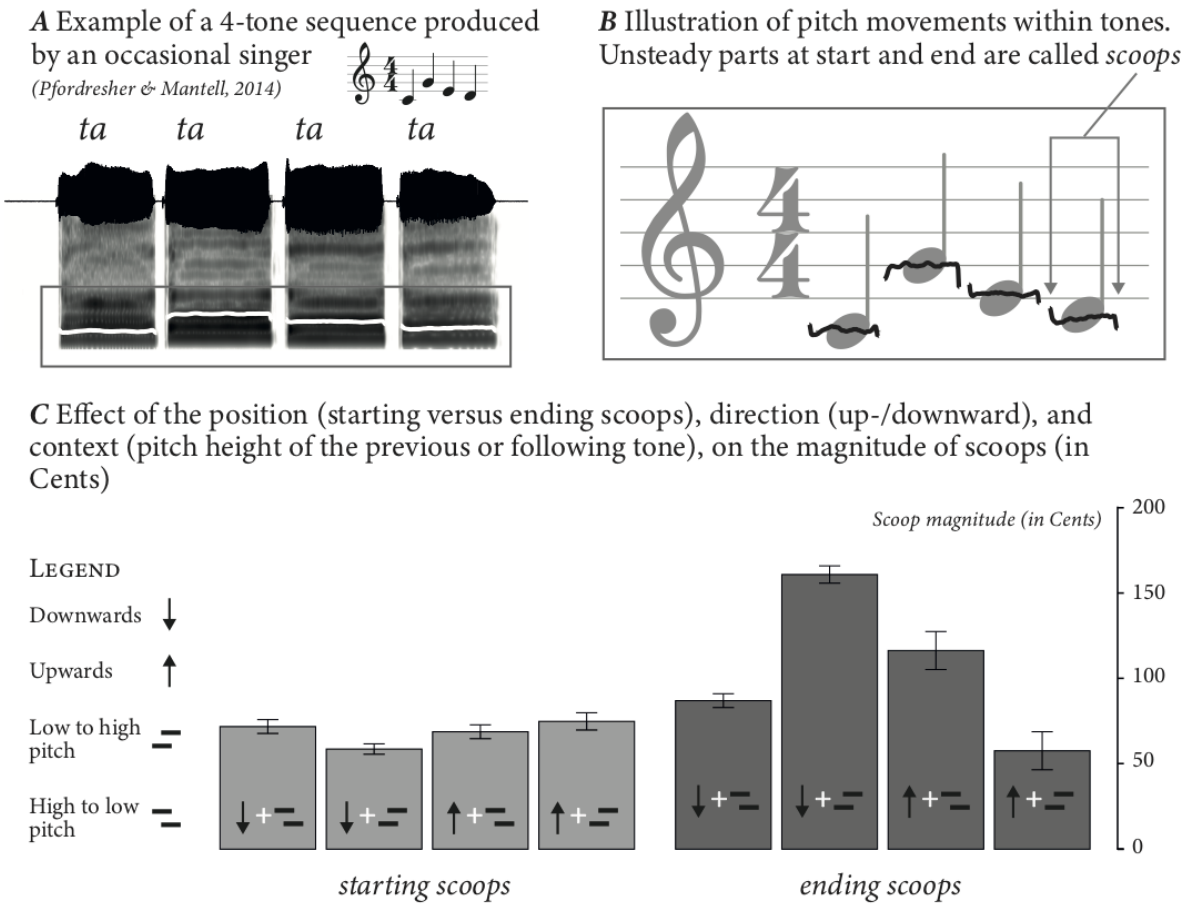

Figure 1. **A.** Illustration of a 4-tone sequence produced with the syllable /ta/ by an occasional singer (data from Pfordresher & Mantell, 2014). Beneath the musical notation, the wave form of each syllable as well as their spectrum are represented, the white lines (in the frame) show the position of the relatively stable fundamental frequency (f_0). **B**. Superposition of the f_0 and the musical notation illustrating that the f_0 of each tone is not perfectly straight. **C**. Average (and SEM) of the magnitude of scoops (in Cents) of 1461 tones analyzed (i.e., not aggregated per singer), according to the position (starting or ending), the direction (up-/downwards), and the pitch height of the tone adjacent to the scoop (higher or lower than the target) in the melody.

Definition of scoops

Singing requires fine control of the vocal instrument (e.g., Sundberg, 2013; Titze, 2000), leading to unavoidable motor adjustments (Hutchins, Larrouy-Maestri, & Peretz, 2014). Even highly trained singers produce pitch transitions toward or away from a target pitch (Hutchins & Campbell, 2009; Mori, Odagiri, Kasuya, & Honda, 2004; Saitou, Unoki, & Akagi, 2005; Stevens & Miles, 1928). Besides the physiological constraints of singing, these pitch movements may be used expressively. In fact, scoops do not only reflect mere noise in the motor system. Note that in the psychophysical literature, pitch movements in small time windows are often referred to as "glides" or "sweeps." In line with the terminology proposed by Larrouy-Maestri and Pfordresher (2018), we use the term "scoop" to refer to dynamic changes at the start or end of sung tones, whether produced voluntarily or not.

Scoops have been recently investigated by examining the singing productions of occasional singers varying greatly in terms of singing accuracy (see Appendix A of Larrouy-Maestri & Pfordresher, 2018). The main contribution of this study consisted in describing the characteristics of scoops at the start and end of musical sung tones in order to synthesize realistic material for the listening task of the main experiment. For this purpose, this study focused on the amplitude and rate of scoops before and after the stable middle part of the tone (i.e., the asymptote). The f_0 values of a total of 1,874 single tones of about 1 s length were selected from recordings described by Pfordresher and Mantell (2014). The analysis adapted the model of Large, Fink, and Kelso (2002), originally designed to quantify variations in temporal perturbation, in order to predict how f_0 changes at the beginning and ending of a sung tone (see Equation A2 and Figure A1, of Larrouy-Maestri & Pfordresher, 2018). The amplitude of the scoop was one of the fixed properties of the data, corresponding to the difference between the median of the stable part of the tone and the pitch at the start or end. The fitted parameters concerned the rate of the scoop (time to reach the stable part or to end the tone), and the oscillation around this part. Each tone was fitted with a least-squares approximation using the optimization toolbox in Matlab. A conservative threshold for goodness of fit was chosen and led to the selection of 1,461 tones (78% of the available tones). Parameters, in particular the magnitude and rate of scoops from these fits, were used to describe characteristics of scoops in the general population. As visible in Figure 1C, the means of scoop magnitude ranged from 50 to 150 cents.

Variability of scoops

Besides estimating the scoop magnitude of occasional singers, the sung material of Pfordresher and Mantell (2014) offered the possibility of testing the effect of direction (upward vs. downward), position (start vs. end), and context (adjacent tone higher or lower than the target tone) on scoop characteristics. As illustrated in Figure 1C, the magnitude of scoops (i.e., difference from center of tone) varied considerably based on these factors. In addition to the main effects of the direction (larger magnitude for downward scoops, $p = .006$, $\eta_p^2 = .249$) and of the position (larger magnitude for ending scoops, $p = 001$, $\eta_p^2 = .440$), there was a significant three-way interaction among position, direction, and context ($p = .001$, $\eta_p^2 = .564$). In other words, the magnitude of the scoop varied systematically depending on the position, direction, and surrounding musical context. By contrast, no significant effects of these variables were found for the rate of scoops or oscillation.

In addition to this information about scoop variability, a closer look at this dataset informs us about the variability according to singer quality. Indeed, in Pfordresher and Mantell (2014), participants were categorized as accurate or inaccurate (see Dalla Bella, 2015, or Pfordresher & Larrouy-Maestri, 2015, for discussion of cutoff criteria). As a consequence, we were able to compare the scoops of two contrasting groups: accurate singers (n = 12) and poor-pitch singers (n = 19). As illustrated in Figure 2A, poor-pitch singers performed with a greater scoop magnitude than accurate singers, $t(15.78) = 2.54$, $p = .022$, with a mean about 121 cents and 89 cents respectively.

Since the magnitude of scoops varied greatly even within groups, we also examined the relationship between the magnitude of the scoops and the amplitude of the global deviation of the asymptote of the tones analyzed (Figure 2B). The Spearman coefficient correlation did not reach significance level ($r(28) = .218$, $p = .254$). Note that the same profile was observed for accurate (black circles) and inaccurate singers (grey circles) when the analysis was performed separately ($p > .05$), supporting the lack of direct relationship between mistuning of the stable part and magnitude of scoops even if inaccurate singers show generally more deviation in both compared to accurate singers. Altogether, the additional analyses performed on the dataset described in Larrouy-Maestri and Pfordresher (2018) support the claim that scoops greatly depend on both the musical material and the singing abilities of the performer.

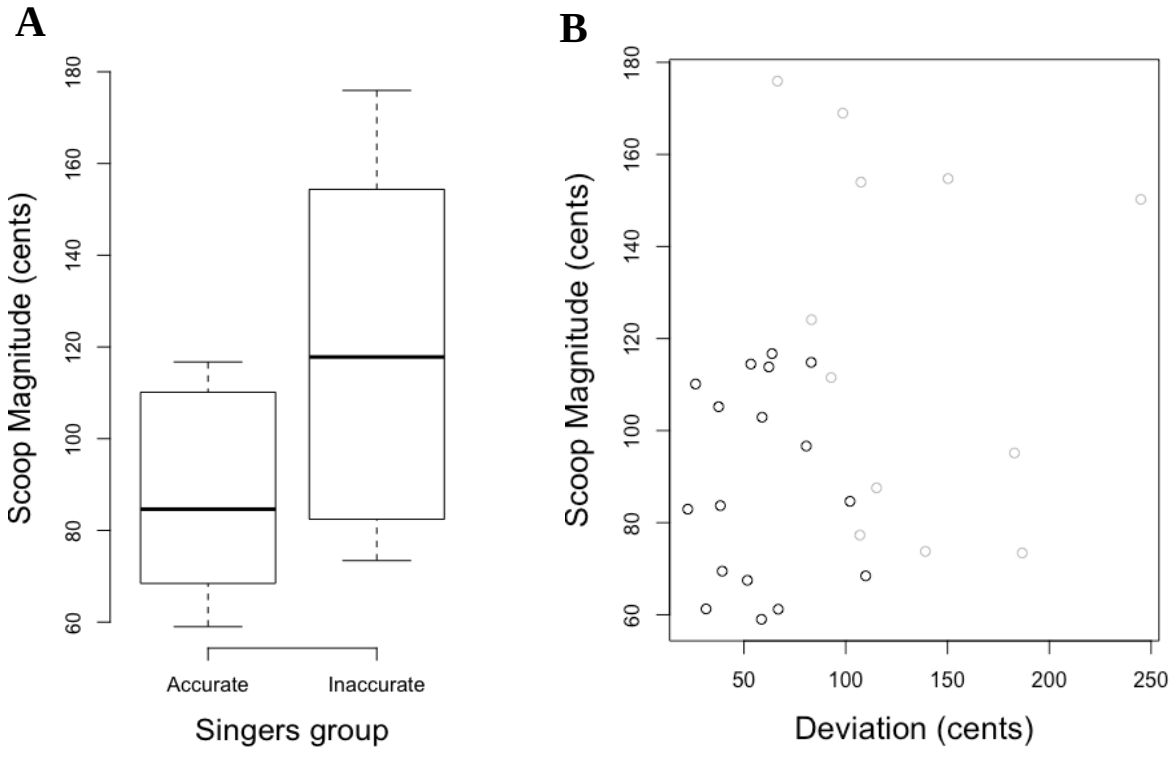

Figure 2. **A** Summary of the distribution of magnitude of scoops (in cents) in the two groups (accurate and inaccurate singers) recorded by Pfordresher and Mantell (2014). The bottom and top of the boxes represent the 25th and 75th percentiles (lower and upper quartiles), with a line at the median. Error bars: lowest and highest scores within a 1.5 interquartile range (IQR). **B**. Relationship between the global pitch deviation of the stable middle part from the target and the absolute value of the scoops magnitude. Black circles represent the accurate singers and grey circles represent the inaccurate singers of Pfordresher and Mantell (2014).

Perceptual relevance of scoops

The scoop magnitudes and rates discussed above (cf. Figure 1) were above the perceptual thresholds proposed so far. Therefore it was assumed that scoops would be heard when listening to music. In three experiments proposed by Larrouy-Maestri and Pfordresher (2018), 110 participants were asked to evaluate the intonation of 4-tone melodies in which the third tone's tuning could vary within the stable part of tones, or by virtue of scoops at the beginning and/or end of the tone. As expected, it was observed that listeners' ratings of pitch accuracy were affected by the presence of scoops. Dynamic changes affected evaluation, particularly when they occured at the end of tones (Experiments 1, 2, and 3 of Larrouy-Maestri & Pfordresher, 2018), or when they were present at both the start and the end of tones (Experiment 2 of the same study). Note that units of small size are not limited to pitch dynamics at the start and/or end of tones but can also take place in the form of frequency modulations (Gockel, Moore, & Carlyon, 2001) or vibrato (Larrouy-Maestri et al., 2017; van Besouw & Howard, 2009) and should be specifically examined. Nevertheless, the findings of Larrouy-Maestri and Pfordresher (2018) confirmed that scoops are treated as informative when listening to music, supporting the claim that the auditory system processes units that are considerable smaller than tones.

Processing small units: the case of scoops in music perception

Background and hypotheses

Since a strong advance has been made regarding the identification of small units in music perception (i.e., scoops) - whereas the ongoing research is still in progress in the speech prosody domain - this section focuses on the musical domain. As discussed, Larrouy-Maestri and Pfordresher's (2018) study confirmed the perceptual relevance of scoops. In addition, they tested two hypotheses about the process behind the influence of scoops in correctness judgments in order to obtain a better understanding of the potential cognitive operations relative to scoop perception:

- The *statistical* hypothesis assumes that listeners' perception is based on the average f_0 across a sung tone. As a consequence, listeners would perceive melodic performances as more in tune when the addition of a scoop would lead to a smaller pitch deviation of the entire tone. For instance, if the middle stable part of a tone is too sharp and the starting scoop goes upward and/or the ending scoop goes downward, the average f_0 of the sung tone will be less sharp than if the scoops go in the other direction (making the average f_0 of the tone even sharper), and thus will be perceived as less out-of-tune. In other words, scoops compensating for the deviation of the sung tone would be perceived as leading to a more correct pitch than scoops that failed to compensate.
- The *teleological* hypothesis assumes that listener's perception of scoops depends on the general trajectory of the melody, the singers' goals corresponding to the middle stable part of the sung tone. This hypothesis treats scoops as distinct from the way in which a scoop influences the average f_0 across the entire tone, but interprets them in relation to the surrounding context of the sung tone manipulated. Concretely, scoops can either enhance *continuity* across successive discrete tones, or *anticontinuity* between adjacent tones.

To summarize, Larrouy-Maestri and Pfordresher (2018) highlighted the coexistence of two hypothesized mechanisms behind the perception scoops: statistical and teleological. Listener perception of scoops is based both on the relationship of the scoop to the tuning of the associated tone (higher scores for stimuli in which the scoops compensated the pitch deviation of the tone) as well as on the relationship of the scoop to the broader melodic context (higher scores for stimuli in which the scoops were not continuous between the two adjacent tones). These findings clarify an important point in the perception of correctness. However, they might not be generalizable to more typical auditory tasks. Indeed, as listeners commonly attend to musical performances for pleasure and appreciation, rather than to judge their technical correctness, the sensitivity to dynamic change in small time windows might vary depending on the question under study. For instance, the relevance of scoops might be lessened or enhanced when the judgment refers to melodic recognition or aesthetic/beauty judgments rather than to pitch accuracy (as tested in Larrouy-Maestri & Pfordresher, 2018). Also, if scoops are relevant, whatever the question, the ratings might still differ (e.g., preference for continuity for aesthetic ratings and the opposite for pitch accuracy ratings). Finally, the processes themselves (statistical or teleological) might differ. As a consequence, it seems challenging to generalize the present findings when focusing only on a specific type of judgment.

The direct comparison of different types of judgments, examined with similar methods was proposed as a follow-up experiment to address the effect of the task on the perception of scoops. Concretely, the study reported here aimed at 1) replicating the findings of Larrouy-Maestri and Pfordresher (2018) with a new sample of participants, and 2) extending the examination of scoops to a more natural listening condition: preference judgment.

Methods

Participants

Fifty-two students at the State University of New York at Buffalo (29 females), ranging from 18 to 23 years old (M = 19.02, SD = 1.33), participated in the experiment in exchange for course credit. Participants reported normal hearing abilities, and a few of them reported a limited amount of formal music training (up to 8 years, M = 0.94, SD = 1.85). The general musical sophistication scores estimated with the Gold-MSI self-report questionnaire (Müllensiefen, Gingras, Musil, & Stewart, 2014) ranged from 48 to 91, with a mean of 72.04, which stand within the range of the scores examined by Müllensiefen et al. (2014), meaning that they were comparable to the scores of a general population.

Material

The stimuli were identical to the set used in Larrouy-Maestri and Pfordresher (2018, Experiment 1). Four tones were synthesized using a male timbre (Vocaloid, Zero-G Limited, Okehampton, England) and arranged in two melodies: C3 (131 Hz) - E3 - D3 - G3 and G3 - D3 - E3 - C3, using equal temperament. Each tone was 900 milliseconds long and articulated with the syllable /da/. The characteristics of the third tone of each melody were manipulated at two different levels: the central portion of the tone (correct, flattened, or sharpened by 50 cents from ideal equal tempered tuning, 100 cents = 1 semitone), and/or the scoop at the start or end of the tone. The scoop magnitude was chosen according to the analyses of pitch fluctuations (Figure 1C) in order to propose realistic stimuli to the listeners. The direction of scoops was defined relative to the middle part: an upward scoop at the start of the tone starts lower than the middle part whereas an upward scoop at the end of the tone ends higher than the middle part.

Procedure

Similar to the study by Larrouy-Maestri and Pfordresher (2018), we used a pairwise comparison procedure. Each participant was exposed to two sets of 15 variants of a melody (n = 105 pairs per melody) and was asked to select one performance of each pair according to the question “Which performance is more in tune?” or “Which performance do you prefer?”. The

question depended on the block. In addition to the within-subjects variables (tone deviation and scoop), we manipulated the order of the questions / melodies between participants, with half the participants listening to variations of Melody 1 first, the other half listening to variations of Melody 2 first. For each group defined by which melody was heard first, half started with the block dedicated to the "correctness" question and the other half with the block dedicated to the "preference" question. All participants switched to the other melody and the other question half way through the experiment. All possible stimulus pairs were evaluated for each combination of melody and question.

Aggregated ratings of preference and correctness were computed for a participant in the following way: For each stimulus and each condition (correctness versus preference judgment), the initial score was set to zero and was increased by one if that condition was selected as more correct/preferred relative to the other stimulus of the trial (a score of 0.5 was recorded if neither stimulus was selected). The final score for each condition was computed by accumulating points over trials, ranging from 0 (i.e., stimulus never selected) to 14 (i.e., stimulus always selected). This rating procedure enables the ranking of the manipulated melodies from the most out-of-tune to the most in-tune or from the least to the most preferred, depending on the block. The highest scores were thus associated to correct or preferred stimuli.

Results and discussion

First, we focused on the correctness judgments and observed the replication of the findings reported in the reference study (Larrouy-Maestri and Pfordresher 2018) summarized above. The ANOVA performed on the new data confirmed the main effect of scoops ($p < .001$, $\eta_p^2 = .28$). Stimuli without scoops were rated significantly higher (i.e., more correct) than stimuli with scoops at the start or end of the tone. We also observed an interaction between scoops and deviation of the asymptote ($p = .004$, $\eta_p^2 = .05$) as well as with the melody ($p < .001$, $\eta_p^2 = .07$). Figure 3A illustrates high reliability between the ratings provided by two independent groups of participants (reference study and new data). These results confirm the salience of small-timescale vocal scoops on judgements of "correct" intonation. Then, we directly compared the ratings provided in the correctness and preference blocks. Figure 3B illustrates the strong relationship between the two types of ratings. The high degree of similarity between the two rankings supports the claim that in-tune stimuli are preferred. Because it is possible that a carry-over effect increased these associations, we also analyzed the correlation between correctness and preference ratings within the first half of the experiment (i.e., before participants transitioned to the other task). This association was still highly significant though smaller in magnitude, $r^2 = .71$.

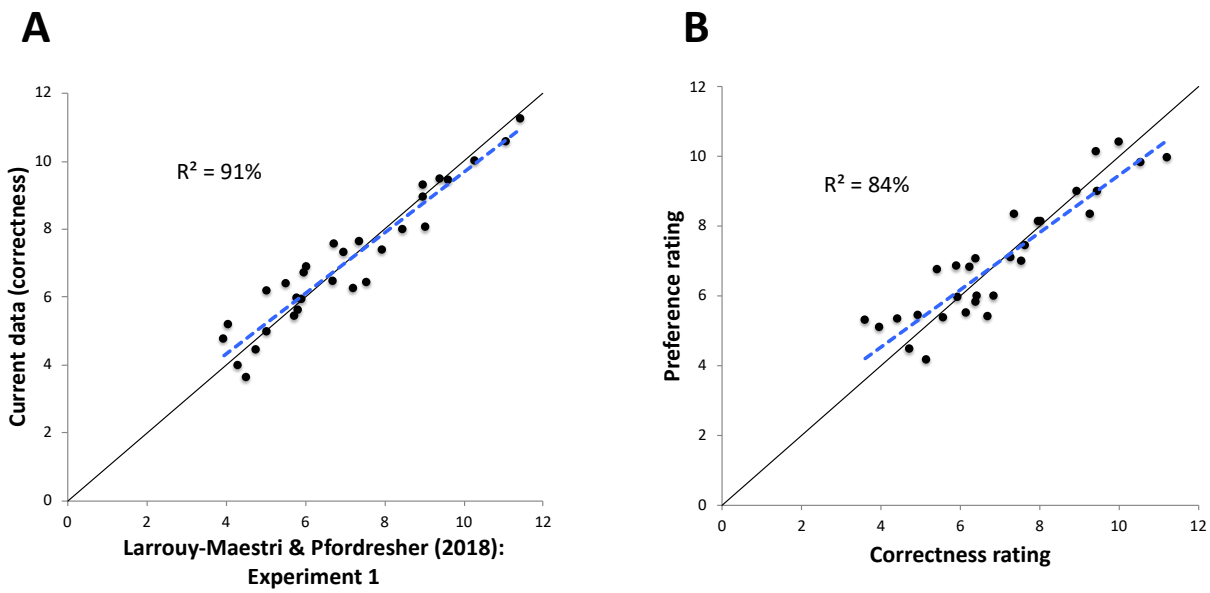

Figure 3. Reliability of judgements for vocal scoops. **A**. Relation between correctness judgements obtained from the participants of Larrouy-Maestri and Pfordresher (2018, Experiment 1) and correctness judgements from the participants of the current data. **B**. Relation between correctness and preference judgements from the current study. Dashed lines represent least-squares linear regression, plotted in comparison to the unity line. Both $p < .001$.

This new dataset replicated the findings reported in Larrouy-Maestri and Pfordresher (2018), confirming the perceptual relevance of scoops when judging the correctness of sung performances, while also providing a way to compare different types of judgments: technical evaluation versus listeners' preferences. As shown in Figure 4, we observed the same pattern of results whatever the type of judgment, with the higher scores for the anti-continuity condition (Figure 4A) and for the compensation condition (Figure 4B). This means that *listeners prefer scoops when they enhance the discreteness of tones* and *listeners also integrate scoops as part of the tone*, whatever the question asked (correctness judgment in grey and preference judgment in white). In other words, for better or worse, our findings support both the statistical or teleological hypotheses when listening to melodies.

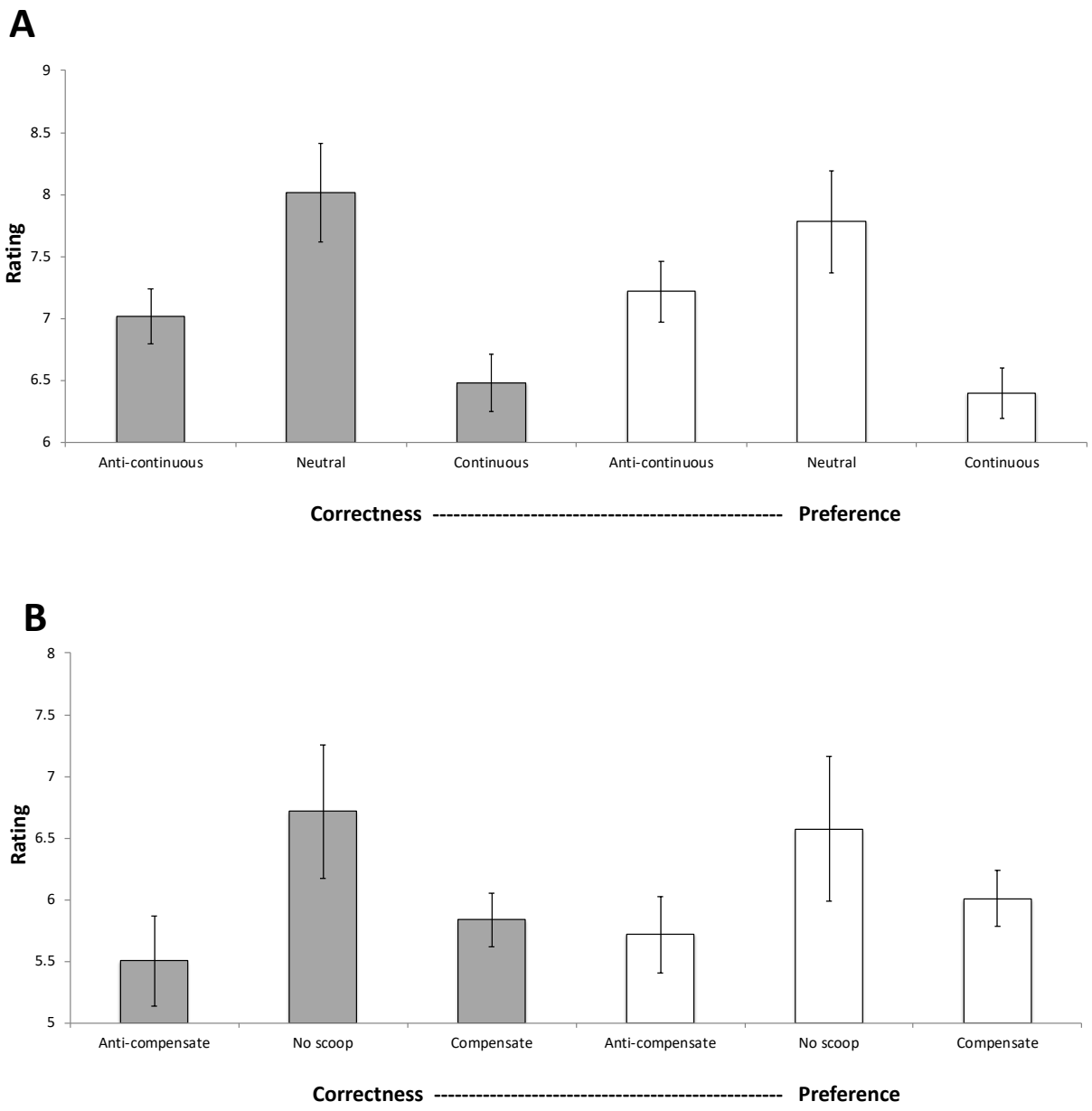

Figure 4. Parallel between correctness and preference ratings. Mean ratings (95% CI) given to stimuli for which scoops favor continuity (**A**) and compensation (**B**), in the two tasks (grey bars correctness judgment; white bars preference judgment).

Conclusion - Outlook

Pitch has different functions in music and language, but the mechanisms underlying these processes might be partially shared across domains. The identification of these mechanisms paves the way for cross-domain comparisons and consequently for a better understanding of auditory sequence processing in general. As highlighted by Fritz et al. (2013), examining similarities and differences across domains requires knowledge about the respective units and processes. In music, it is now clear that pitch is an important feature for judging the

technical (i.e., correctness) and aesthetic quality of a performance and that units of different timescales are used in combination. More specifically, both the data reported in Larrouy-Maestri and Pfordresher (2018) and the data reported in this chapter highlight the unique contribution of *scoops*, whatever the type of judgment, and support the coexistence of two distinct mechanisms: averaging pitch across the duration of the tone and processing fluctuations in relation to the surrounding context. Research in speech prosody is advancing considerably and we believe that the direct comparison of pitch perception in both domains promises to shed light on the long-standing and nevertheless still open debate about similarities between music and speech.

References

't Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press. doi:10.1017/S0022226700015395

Agrima, A., Farchi, A., Elmazouzi, L., Mounir, I., & Mounir, B. (2019). Emotion recognition from Maroccan dialect speech and energy band distribution. *IEEE.* doi:10.1109/WITS.2019.8723775

Bandstra, N. F., Chambers, C. T., McGrath, P. J., & Moore, C. (2011). The behavioural expression of empathy to others' pain versus others' sadness in young children. *Pain, 152*(5), 1074-1082. doi:10.1016/j.pain.2011.01.024

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614-636. doi:10.1037/0022-3514.70.3.614

Bänziger, T., Hosoya, G., & Scherer, K. R. (2015). Path models of vocal emotion communication. *PLoS ONE, 10*(9), e0136675. doi:10.1371/journal.pone.0136675

Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication, 46*(3-4), 252-267. doi:10.1016/j.specom.2005.02.016

Belin, P., Boehme, B., & McAleer, P. (2017). The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLoS ONE, 12*(10), e0185651. doi:10.1371/journal.pone.0185651

Bigand, E., & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. *Cognition, 100*(1), 100-130. doi:10.1016/j.cognition.2005.11.007

Bitouk, D., Nenkova, A., & Verma, R. (2009). Improving emotion recognition using class-level spectral features. Paper presented at Interspeech conference, Brighton.

Bryant, G. A., & Barrett, H. C. (2008). Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture, 8*(1-2), 135-148. doi:10.1163/156770908X289242

Burns, E. M., & Ward, W. D. (1978). Categorical perception - phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. *Journal of Acoustical Society of America, 63*(2), 456-468. doi:10.1121/1.381737

Cheng, J. T., Tracy, J. L., Ho, S., & Henrich, J. (2016). Listen, follow me: Dynamic vocal signals of dominance predict emergent social rank in humans. *Journal of Experimental Psychology: General, 145*(5), 536-547. doi:10.1037/xge0000166

Cohen, D. (2002). Notes, scales, and modes in the earlier Middle Ages. In T. Christensen (Ed.), *The Cambridge History of Western Music Theory* (The Cambridge History of Music, pp. 305-363). Cambridge: Cambridge University Press. doi:10.1017/CHOL9780521623711.013

Cross, I. (2001). *Music, cognition, culture, and evolution.* Annals of the New York Academy of Sciences, 930: 28-42. doi:10.1111/j.1749-6632.2001.tb05723.x

Dalla Bella, S. (2015). Defining poor-pitch singing: A problem of measurement and sensitivity. *Music Perception, 32*(3), 272-282. doi:10.1525/MP.2015.32.3.272

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Frontiers in Human Neuroscience, 11*, 481. doi:10.3389/fnhum.2017.00481

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2015). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience, 19*(1), 158-164. doi:10.1038/nn.4186

Dromey, C., Silveira, J., & Sandor, P. (2005). Recognition of affective prosody by speakers of English as a first or foreign language. *Speech Communication, 47*(3), 351-359. doi:10.1016/j.specom.2004.09.010

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., . . . Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing, 7*(2), 190-202. doi:10.1109/taffc.2015.2457417

Fairbanks, G., & Pronovost, W. (1938). Vocal pitch during simulated emotion. *Science, 88*(2286), 382-383. doi:10.1126/science.88.2286.382

Fischer, A. H., & Manstead, A. S. R. (2008). *Social functions of emotions.* In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), Handbook of emotions (pp. 456-468). New York - London: The Guilford press.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of Acoustical Society of America, 106*(3), 1511-1522.

Fritz, J., Poeppel, D., Trainor, L., Schlaug, G., Patel, A. D., Peretz, I., . . . Parsons, L. M. (2013). The reurobiology of language, speech, and music. In M. A. Arbib (Ed.), *Language, music, and the brain*. Cambridge, MA: MIT Press.

Gamba, M. (2014). Vocal tract-related cues across human and nonhuman signals. *Reti, saperi, linguaggi, Italian Journal of Cognitive Sciences, 1*, 49-65, doi: 10.12832/77496

Garnier, M., Henrich, N., Castellengo, M., Sotiropoulos, D., & Dubois, D. (2007). Characterisation of voice quality in Western lyrical singing: from teachers' judgements to acoustic descriptions. *Journal of interdisciplinary music studies, 1*(2), 62-91.

Gockel, H., Moore, B. C. J., & Carlyon, R. P. (2001). Influence of rate of change of frequency on the overall pitch of frequency-modulated tones. *Journal of the Acoustical Society of America, 109*(2), 701-712. doi:10.1121/1.1342073

Gordon, M., & Poeppel, D. (2002). Inequality in identification of direction of frequency change (up vs. down) for rapid frequency modulated sweeps. *Acoustics Research Letters Online - Acoustical Society of America, 3,* 29–34. http://dx.doi.org/10.1121/1.1429653

Goudbeek, M. B., Goldman, J. P., & Scherer, K. R. (2009). Emotion dimensions and formant position. In M. Uther, R. Moore, & S. Cox (Eds.), *Proceedings of Interspeech 2009: 10th Annual Conference of the International Speech Communication Association* (pp. 1575-1578). Brighton, UK: ISCA.

Hannon, E., & Trainor, L. (2007). Music acquisition: effects of enculturation and formal training on development. *Trends in Cognitive Sciences, 11*(11), 466-472. doi:10.1016/j.tics.2007.08.008

Hillenbrand, J. M., & Clark, M. J. (2009). The role of f0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics, 71*(5), 1150-1166. doi:10.3758/APP.71.5.1150

Hutchins, S., & Campbell, D. (2009). Estimating the time to reach a target frequency in singing. *Annals of the New York Academy of Sciences, 1169,* 116-120. doi:10.1111/j.1749-6632.2009.04856.x

Hutchins, S., Larrouy-Maestri, P., & Peretz, I. (2014). Singing ability is rooted in vocal-motor control of pitch. *Attention, Perception, & Psychophysics, 76*(8), 2522-2530. doi:10.3758/s13414-014-0732-1

Hutchins, S., Roquet, C., & Peretz, I. (2012). The vocal generosity effect: How bad can your singing be? *Music Perception, 30*(2), 147-159. doi:10.1525/mp.2012.30.2.147

Jackendoff, R. (2009). Parallels and nonparallels between language and music. *Music Perception, 26*(3), 195-204. doi:10.1525/mp.2009.26.3.195

Jiang, X., Paulmann, S., Robin, J., & Pell, M. D. (2015). More than accuracy: Nonverbal dialects modulate the time course of vocal emotion recognition across cultures. *Journal of Experimental Psychology. Human Perception and Performance, 41*(3), 597-612. doi:10.1037/xhp0000043

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. doi:10.1037/0033-2909.129.5.770

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology, 16*(3), e2004473. doi:10.1371/journal.pbio.2004473

Keltner, D., & Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition and emotion, 13*(5), 505-521. doi: 10.1080/026999399379168

Kerivan, J. E., & Carey, B. J. (1976). Pattern identification of pure tones and frequency glides by untrained listeners. *Perception & Psychophysics, 20*(6), 489-492. doi:10.3758/BF03208287

Koelsch, S., & Friederici, A. D. (2003). Toward the neural basis of processing structure in music. *Annals of the New York Academy of Sciences, 999*, 15-28. doi:10.1196/annals.1284.002

Krumhansl, C. L. (1979). The psychological representation of musical pitch in a tonal context. *Cognitive Psychology, 11*(3), 346-374. doi:10.1016/0010-0285(79)90016-1

Large, E. W., Fink, P., & Kelso, J. A. S. (2002). Tracking simple and complex sequences. *Psychological Research, 66*, 3-17. doi:10.1007/s004260100069

Larrouy-Maestri, P. (2018). "I know it when I hear it": On listeners' perception of mistuning. *Music and Science, 1*, 1-17. doi:10.1177/2059204318784582

Larrouy-Maestri, P., Harrison, P. M. C., & Müllensiefen, D. (2019). The mistuning perception test: A new measurement instrument. *Behavioral Research Methods*, *51*(2), 663-675. doi:10.3758/s13428-019-01225-1

Larrouy-Maestri, P., Lévêque, Y., Schön, D., Giovanni, A., & Morsomme, D. (2013). The evaluation of singing voice accuracy: a comparison between subjective and objective methods. *Journal of Voice, 27*(2), 259 e251-259 e255. doi:10.1016/j.jvoice.2012.11.003

Larrouy-Maestri, P., Magis, D., Grabenhorst, M., & Morsomme, D. (2015). Layman versus professional musician: Who makes the better judge? *PLoS ONE, 10*(8), e0135394. doi:10.1371/journal.pone.0135394

Larrouy-Maestri, P., Morsomme, D., Magis, D., & Poeppel, D. (2017). Lay listeners can evaluate the pitch accuracy of operatic voices. *Music Perception, 34*(4), 489-495. doi:10.1525/mp.2017.34.4.489

Larrouy-Maestri, P., & Pfordresher, P. Q. (2018). Pitch perception in music: Do scoops matter? *Journal of Experimental Psychology: Human Perception and Performance, 44*(10), 1523-1541. doi:10.1037/xhp0000550

Larrouy-Maestri, P., Poeppel, D., & Pell, M.D. (in prep.). The sound of emotions: Cracking the code of emotional speech prosody.

Latinus, M., & Belin, P. (2011). Human voice perception. *Current biology, 21*(4), 143-145. doi:10.1016/j.cub.2010.12.033

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. doi:10.1016/j.cub.2013.04.055

Lerdahl, F., & Jackendoff, R. (1983/1984). An overview of hierarchical structure in music. *Music Perception, 1*(2), 229-252. doi:10.2307/40285257

Luo, H., Boemio, A., Gordon, M., & Poeppel, D. (2007). The perception of FM sweeps by Chinese and English listeners. *Hearing Research, 224*(1-2), 75-83. doi:10.1016/j.heares.2006.11.007

Lyzenga, J., Carlyon, R. P., & Moore, B. C. J. (2004). The effects of real and illusory glides on pure-tone frequency discrimination. *Journal of the Acoustical Society of America, 116*(1), 491-501. doi:10.1121/1.1756616

Marmel, F., Tillmann, B., & Dowling, W. J. (2008). Tonal expectations influence pitch perception. *Perception & Psychophysics, 70*(5), 841-852. doi:10.3758/pp.70.5.841

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say "Hello"? Personality impressions from brief novel voices. PLoS ONE, 9(3), e90779. doi: 10.1371/journal.pone.0090779

McDermott, J. H., Schultz, A. F., Undurraga, E. A., & Godoy, R. A. (2016). Indifference to dissonance in native Amazonians reveals cultural variation in music perception. *Nature, 535*, 547-550. doi:10.1038/nature18635

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research, 219*(1-2), 36-47. doi:10.1016/j.heares.2006.05.004

Moore, B. C. J. (1973). Frequency difference limens for short-duration tones. *Journal of the Acoustical Society of America, 54*(3), 610-619. doi:10.1121/1.1913640

Morgan, E., Fogel, A., Nair, A., & Patel, A. D. (2019). Statistical learning and Gestalt-like principles predict melodic expectations. *Cognition, 189*, 23-34. doi:10.1016/j.cognition.2018.12.015

Mori, H., Odagiri, W., Kasuya, H., & Honda, K. (2004, April). *Transitional characteristics of fundamental frequency in singing*. Paper presented at the 18th International Congress on Acoustics, Kyoto, Japan.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PLoS ONE, 9*(2), e89642. doi:10.1371/journal.pone.0089642

Mürbe, D., Zahnert, T., Kuhlisch, E., & Sundberg, J. (2007). Effects of professional singing education on vocal vibrato-a longitudinal study. *Journal of Voice, 21*(6), 683-688. doi:10.1016/j.jvoice.2006.06.002

Nordström, H., & Laukka, P. (2019). The time course of emotion recognition in speech and music. *Journal of the Acoustical Society of America*, 145(5), 3058. doi:10.1121/1.5108601

Parncutt, R., & Hair, G. (2018). A psychocultural theory of musical interval. *Music Perception, 35*(4), 475-501. doi:10.1525/mp.2018.35.4.475

Patel, A. D. (2008). *Music, language and the brain*. Oxford: Oxford University Press.

Pavela Banai, I., Banai, B., & Bovan, K. (2016). Vocal characteristics of presidential candidates can predict the outcome of actual elections. *Evolution and Human Behavior, 38*(3), 309-314. doi:10.1016/j.evolhumbehav.2016.10.012

Pearce, M. T. (2005). The construction and evaluation of statistical models of melodic structure in music perception and composition. Doctoral dissertation, Department of Computing, City University, London, UK.

Pell, M. D., & Kotz, S. A. (2011). On the time course of vocal emotion recognition. *PLoS ONE, 6*(11), e27256. doi: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. doi:10.1007/s10919-008-0065-7

Pell, M. D., & Skorup, V. (2008). Implicit processing of emotional prosody in a foreign versus native language. *Speech Communication, 50*(6), 519-530. doi:10.1016/j.specom.2008.03.006

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*(1), 245-255. doi:10.1016/s0167-6393(02)00107-3

Pfordresher, P. Q., & Larrouy-Maestri, P. (2015). On drawing a line through the spectrogram: How do we understand deficits of vocal pitch imitation? *Frontiers in human neuroscience, 9*. doi:10.3389/fnhum.2015.00271

Pfordresher, P. Q., & Mantell, J. T. (2014). Singing with yourself: evidence for an inverse modeling account of poor-pitch singing. *Cognitive Psychology, 70*, 31-57. doi:10.1016/j.cogpsych.2013.12.005

Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J.-J. (2018). Cracking the social code of speech prosody using reverse correlation. *PNAS, 115*(15), 3972-3977. doi:10.1073/pnas.1716090115

Rendall, D., Vokey, J. R., & Nemeth, C. (2007). Lifting the curtain on the Wizard of Oz: Biased voice-based impressions of speaker size. *Journal of Experimental Psychology: Human Perception and Performance, 33*(5), 1208-1219. doi:10.1037/0096-1523.33.5.1208

Ringer, A. L. (2002). *Melody: Definition and origins.* In The New Grove Dictionary of Music Online, L. Macy Editor, Macmillan Online Publishing, http://www.grovemusic.com.

Saitou, T., Unoki, M., & Akagi, M. (2005). Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis. *Speech Communication, 46*(3-4), 405-417. doi:10.1016/j.specom.2005.01.010

Sammler, D., Grosbras, M. H., Anwander, A., Bestelmeyer, P. E., & Belin, P. (2015). Dorsal and ventral pathways for prosody. *Current Biololgy, 25*(23), 3079-3085. doi:10.1016/j.cub.2015.10.009

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. doi:10.1177/0022022101032001009

Shami, M. T., & Kamel, M. S. (2005). Segment-based approach to the recognition of emotions in speech. *IEEE International Conference on Multimedia and Expo*, Amsterdam, 2005, doi: 10.1109/ICME.2005.1521436

Shariff, A. F., & Tracy, J. L. (2011). Emotion expressions: On signals, symbols, and spandrels−a response to Barrett (2011). *Current Directions in Psychological Science, 20*(6), 407-408. doi:10.1177/0963721411429126

Shigeno, S. (2016). Speaking with a happy voice makes you sound younger. *International Journal of Psychological Studies, 8*(4), 71-76. doi:10.5539/ijps.v8n4p71

Sinaceur, M., Kopelman, S., Vasiljevic, D., & Haag, C. (2015). Weep and get more: When and why sadness expression is effective in negotiations. *Journal of Applied Psychology, 100*(6), 1847-1871. doi:10.1037/a0038783

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *Journal of the Acoustical Society of America, 117*(1), 305-318. doi:10.1121/1.1828637

Soranzo, A., & Grassi, M. (2014). PSYCHOACOUSTICS: a comprehensive MATLAB toolbox for auditory testing. *Frontiers in Psychology, 5*, 712. doi:10.3389/fpsyg.2014.00712

Stalinski, S. M., & Schellenberg, E. G. (2010). Shifting perceptions: Developmental changes in judgments of melodic similarity. *Developmental Psychology, 46*(6), 1799-1803. doi:10.1037/a0020658

Stevens, F. A., & Miles, W. R. (1928). The first vocal vibrations in the attack in singing. *Psychological Monographs, 39*(2), 200–220. doi :10.1037/h0093347

Sundberg, J. (2013). Perception of singing. In D. Deutsch (Ed.), *The psychology of music* (p. 69–105). Elsevier Academic Press. doi:10.1016/b978-0-12-381460-9.00003-1

Temperley, D. (2013). Computational models of music cognition. *Psychology of Music*, 327-368. doi:10.1016/b978-0-12-381460-9.00008-0

Teng, X., Tian, X., & Poeppel, D. (2016). Testing multi-scale processing in the auditory system. *Scientific Reports, 6*, 34390. doi:10.1038/srep34390

Thompson, W. F. (2013). Intervals and Scales. In D. Deutsch (Ed.), *The Psychology of Music* (3rd ed., pp. 107-140). (Cognition and Perception). London; New York: Elsevier. doi :10.1016/B978-0-12-381460-9.00004-3

Titze, I. R. (2000). *Principles of voice production* (2nd edition). Iowa City, IA: National Center for Voice and Speech.

van Besouw, R. M., & Howard, D. M. (2009). Effects of carrier and phase on the pitch of long-duration vibrato tone. *Musicae Scientiae, 8*(1), 139-161. doi:10.1177/102986490901300100 6

van Rijn, P., Poeppel, D., & Larrouy-Maestri, P. (in prep.). Measures of pitch over time improve classification of emotional speech.

Waaramaa, T., Laukkanen, A. M., Airas, M., & Alku, P. (2010). Perception of emotional valences and activity levels from vowel segments of continuous speech. *Journal of Voice, 24*(1), 30-38. doi:10.1016/j.jvoice.2008.04.004

Wang, W.-J., Tan, C.-T., & Martin, B. A. (2013). Auditory evoked responses to a frequency glide following a static pure tone. *Journal of the Acoustical Society of America, 133*(5), 3429-3429. doi:10.1121/1.4806040

Warrier, C. M., & Zatorre, R. J. (2002). Influence of tonal context and timbral variation on perception of pitch. *Perception & Psychophysics, 64*(2), 198-207. doi :10.3758/BF03195786

Wildgruber, D., Riecker, A., Hertrich, I., Erb, M., Grodd, W., Ethofer, T., & Ackermann, H. (2005). Identification of emotional intonation evaluated by fMRI. *NeuroImage, 24*(4), 1233-1241. doi:10.1016/j.neuroimage.2004.10.034

Wubben, M. J. J., de Cremer, D., & van Dijk, E. (2011). The communication of anger and disappointment helps to establish cooperation through indirect reciprocity. *Journal of Economic Psychology, 32*(3), 489-501. doi:10.1016/j.joep.2011.03.016

Zarate, J. M., Ritson, C. R., & Poeppel, D. (2012). Pitch-interval discrimination and musical expertise: is the semitone a perceptual boundary? *Journal of the Acoustical Society of America, 132*(2), 984-993. doi:10.1121/1.4733535

Zoghaib, A. (2019). Persuasion of voices: The effects of a speaker's voice characteristics and gender on consumers' responses. *Recherche et Applications En Marketing (English Edition), 34*(3), 83–110. doi:10.1177/2051570719828687

Acknowledgments

We are grateful to Pol van Rijn for his help with Figure 1, to Shi En Gloria Huan for helping with the data collection of preference judgments, and to Madita Hörster and Carmel Raz for edits. This research was supported in part by NSF Grant BCS-1848930 awarded to Peter Q. Pfordresher and by the Max Planck Society.