Vocal imitation of synthesised sounds varying in pitch, loudness and spectral centroid

Adib Mehrabi,[a] Simon Dixon, and Mark B. Sandler
Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom

(Received 4 February 2016; revised 27 December 2016; accepted 6 January 2017; published online 13 February 2017)

Vocal imitations are often used to convey sonic ideas [Lemaitre, Dessein, Susini, and Aura. (2011). Ecol. Psych. **23**(4), 267–307]. For computer based systems to interpret these vocalisations, it is advantageous to apply knowledge of what happens when people vocalise sounds where the acoustic features have different temporal envelopes. In the present study, 19 experienced musicians and music producers were asked to imitate 44 sounds with one or two feature envelopes applied. The study addresses two main questions: (1) How accurately can people imitate ramp and modulation envelopes for pitch, loudness, and spectral centroid?; (2) What happens to this accuracy when people are asked to imitate two feature envelopes simultaneously? The results show that experienced musicians can imitate pitch, loudness, and spectral centroid accurately, and that imitation accuracy is generally preserved when the imitated stimuli combine two, non-necessarily congruent features. This demonstrates the viability of using the voice as a natural means of expressing time series of two features simultaneously.

[http://dx.doi.org/10.1121/1.4974825]

[JFL] Pages: 783–796

I. INTRODUCTION

The voice is a powerful and expressive means of communicating non-verbal sounds, and is particularly useful when the sounds are unidentifiable or artificial, such as synthesised sweeps and tones (Lemaitre and Rocchesso, 2014). In the collaborative music production process this method of communication allows for fluid transfer of ideas between the people making the music. For example a musician or producer might vocalise a target kick drum or synthesiser sound, in order to describe salient characteristics of the sound they are trying to create. Vocalisation may also be used in sound design to describe sounds that cannot otherwise be described verbally. These vocalisations can be used to assist in the task of searching for sounds in large sample libraries, speeding up a tedious and time consuming part of a production process. Audio based query by example (QBE) systems allow for search and retrieval of sound files using an audio input as the query (Helén and Virtanen, 2007; Xue *et al.*, 2008). Query by vocalisation (QBV) is a particular case of QBE, where the input is a vocalisation, or imitation of the sound being searched for (Blancas and Janer, 2014; Roma and Serra, 2015). However, bridging the gap between the sonic spaces of the voice and a large sample library is not a trivial task, and ideally requires *a priori* knowledge of the feature level accuracy at which the "vocalist" can imitate sounds in order to set proper error thresholds in the vocalisation search space. For example, typical vocal ranges of different acoustic features might be mapped to relative ranges in the sample library space, and feature-specific tolerances may be included for similarity metrics.

Related work on non-verbal vocal imitations has primarily focused on classification tasks (Dessein and Lemaitre, 2009; Lemaitre *et al.*, 2011; Rocchesso and Mauro, 2014; Zhang and Duan, 2015). Imitation classification is the process of identifying the class of sounds a vocal imitation belongs to, or identifying the target stimuli for a given imitation. While imitation accuracy can to some extent be inferred from classification results, it is difficult to identify whether any observed effect can be directly attributed to them. For example, in a classification study of 4429 vocal imitations, participants were asked to select the correct stimulus (audio file or label) for a given imitation in a 10 way forced choice test (Cartwright and Pardo, 2015). The authors found large differences in the results for each category of sounds, with correct stimulus selection scores ranging from 42% for commercial synthesisers to 80% for everyday sounds. This indicates that some types of sounds may be easier to imitate accurately than others. However, it is worth noting that the within-category variety of sounds and number of sounds was not the same across categories, so these results cannot be solely attributed to imitation accuracy.

There has been increasing research in the related field of vocal controlled synthesis systems. Typically these work by extracting audio features from the voice and mapping them to parameters on a synthesiser. Although the types of feature vary across studies, they generally include pitch based (e.g., F_0) and timbre based (e.g., spectral centroid) features (Janer, 2005; Stowell, 2010; Cartwright and Pardo, 2014). These studies present interesting novel methods and applications

[a]Electronic mail: a.mehrabi@qmul.ac.uk

for the voice, however they do not address the question of how well people might be able to control the features that are being used for the mapping.

The ability to vocally imitate sounds is limited by both physical and perceptual constraints. The vibration rate of the vocal folds, physical dimensions of the vocal tract, and air flow limit the dynamic range, frequency range, and types of sounds that can be produced. In addition to these limitations, overlapping sound events and sounds requiring fast utterances (such as coins falling on a plate) can be difficult, if not impossible, to imitate accurately with the voice (Lemaitre and Rocchesso, 2014). In terms of vocal control, there has been significant research on pitch range (Zraick *et al.*, 2000; DeLeo LeBorgne and Weinrich, 2002), rate of pitch change (Sundberg, 1979; Dromey *et al.*, 2003; Xu and Sun, 2000), and sound intensity level range (Colton, 1970; Coleman *et al.*, 1977). However, there are two major gaps in current research: (1) much of the literature on vocal control is from the fields of singing voice and speech research, which although relevant, is not always applicable to vocal imitations in general; (2), this literature mainly focuses on single features, with the exception of studies on phonetograms such as DeLeo LeBorgne and Weinrich (2002). There is very little work that has investigated imitation accuracy at the acoustic feature level when people try to exercise control over multiple time varying features. Here we address this issue by conducting an experiment with experienced musicians and music producers to test the effect of stimuli containing pitch, loudness, and spectral centroid envelopes on the accuracy of vocal imitations.

In addition to the physical aspect of vocalising sounds, studies on loudness and pitch have highlighted perceptual biases related to the temporal envelopes of these features. For example, there is evidence of perceptual asymmetries between ascending and descending ramps: people tend to be more accurate at identifying the end pitch for ascending ramps compared to descending (d'Alessandro *et al.*, 1998); and there is a tendency to overestimate the range of a ramp that increases in loudness compared to one that decreases (Neuhoff, 1998, 2001). These perceptual biases may influence someones' ability to vocalise a sound (or even a sonic idea), if there is a difference between what they think they are vocalising and the actual acoustic properties of the vocalisation. It is important to note that in the present study we are not concerned with testing the physical limits of the vocal system or perception of different feature envelopes of sounds. For this reason the stimuli used here have been selected to be comfortably within both the physically producible and perceivable limits in terms of the range and rate of change of the features.

In a study with similar motivations to the one presented here, the accuracy of vocal imitations with respect to pitch, tempo, sharpness, and onset features was investigated (Lemaitre *et al.*, 2016). The authors found that participants were able to accurately imitate pitch and tempo in absolute terms and sharpness in relative terms, with onset imitated least accurately out of the four features. Here we investigate similar features: pitch; loudness (related to onset); spectral centroid (related to sharpness). Instead of using constant (flat) temporal envelopes for pitch and sharpness, we applied four envelopes [ramp up (*RU*), ramp down (*RD*), 2 Hz modulation (*MS*), 5 Hz modulation (*MF*)] to each of the features, and included all pitch based pairwise combinations of these feature envelopes: pitch and loudness; pitch and spectral centroid. This design allows us to study the effect of features and envelope shapes on imitation accuracy independently, as well as test for pitch-loudness and pitch-spectral centroid interactions.

In the present study we use ramp and modulation envelope shapes because they represent a base group of shapes from which a wide variety of more complex shapes can be constructed (arguably all non-static sounds are made up of various combinations of ascending and/or descending acoustic features), yet are relatively simple, obviously perceptible, and easily differentiable with respect to one another. We focus on pitch and loudness because they are fundamental features of singing and music, and we expected musically trained participants to be able to exercise some degree of control over these. We include spectral centroid because it serves as an important timbral feature, and we expected participants to have some control over this through physical manipulation of the vocal tract.

There are a number of ways to measure the accuracy of a vocal imitation, such as self-assessment of imitations (Cartwright and Pardo, 2015), classification of imitations as described above (Rocchesso and Mauro, 2014; Zhang and Duan, 2015), and feature level accuracy (Lemaitre *et al.*, 2016). The purpose of the present study is to test for the effect of single and double feature envelopes on imitation accuracy; therefore we require a metric to compare differences between the feature time series of an imitation and its corresponding stimulus. To achieve this we measure imitation accuracy using parameters for each envelope that capture information about both the range of feature values and the temporal pattern. These are modulation rate, modulation extent, ramp range, and ramp slope. Synthesised stimuli are generated with target envelope shapes using parameter controls that correspond to the extracted features (F_0 for pitch, gain for loudness, and a low pass filter for spectral centroid). We then extract the same features from both the stimuli and vocal imitations. To extract the ramp parameters, we fit each ramp feature vector to a piecewise linear model that is representative of the stimuli ramps. Modulation parameters are extracted using a low pass filtering and peak picking method taken from previous studies on vibrato (Prame, 1994; Xu and Sun, 2002; Ferrante, 2011). This approach gives us comparable imitation-stimulus values with which to measure the accuracy of each imitation. Details and results of the modeling and modulation parameters are discussed in Sec. II E.

This paper is laid out as follows: In Sec. II we describe the stimuli, experimental procedure, and parameter extraction methods; a statistical analysis of the results is presented in Sec. III, followed by a discussion in Sec. IV. Summary conclusions and suggestions for future work are discussed in Sec. V.

II. METHOD

A. Stimuli

The stimuli were generated using a sawtooth oscillator with pitch (*P*), gain (*L*), and cutoff frequency (*C*) parameters

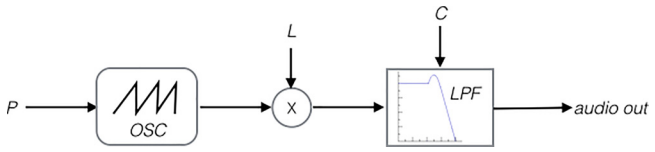

FIG. 1. (Color online) Block diagram of the synthesis model used to generate the stimuli. P = pitch, OSC = sawtooth oscillator, L = gain, LPF = second order IIR low pass filter, C = cutoff frequency. The parameters relate to the vocal features of interest: F_0, loudness, and spectral centroid.

(Fig. 1). These were scaled in semitones (st), decibels (dB), and linear Hz, respectively. The four envelope shapes (Fig. 2) RD, RU, MF, and MS were separately applied to each of the three parameters on the synthesiser, giving 12 control stimuli with a single feature envelope applied: *PRD*, *PRU*, *PMF*, *PMS*, *LRD*, *LRU*, *LMF*, *LMS*, *CRD*, *CRU*, *CMF*, and *CMS* (where the first letter indicates the parameter and the last two letters indicate the envelope shape). A further 32 stimuli were then generated by combining the eight L and C stimuli with the four P stimuli in a pairwise manner, shown in Table I. This design gives 12 control stimuli which can be compared to the 32 double-feature stimuli to test for the effect of different envelope combinations on imitation accuracy.

Each stimulus is 2 s in duration, and each of the flat sections in the envelope shapes are 0.5 s. These flat sections were included to give the participants a clear start and destination value for each feature envelope.

1. Parameter selection

For this study we are not concerned with testing the limits of perception or vocal ranges with respect to pitch, loudness, and spectral centroid, therefore the stimuli parameters are within generally producible and perceivable ranges. The base fundamental frequency (F_0) of the stimuli is 110 Hz for male participants and 220 Hz for female, in line with the typical F_0 ranges for speech (Baken and Orlikoff, 2000; Fitch and Holbrook, 1970; Brown *et al.*, 1993).

The F_0 parameter range of the *PRD* and *PRU* envelopes are also different for male and female participants: 110–220 Hz for males and 220–440 Hz for females, each corresponding to 12 st. Previous studies show these values to be within typical F_0 ranges (Kent *et al.*, 1987). The loudness parameter is set based on the dynamic range of the voice. This is dependent on phonation frequency, and is approximately 50 dB at normal F_0 values (Colton, 1970; Coleman *et al.*, 1977). The range of the *LRU* and *LRD* envelopes is within this range, at 24 dB. The cutoff frequency range used

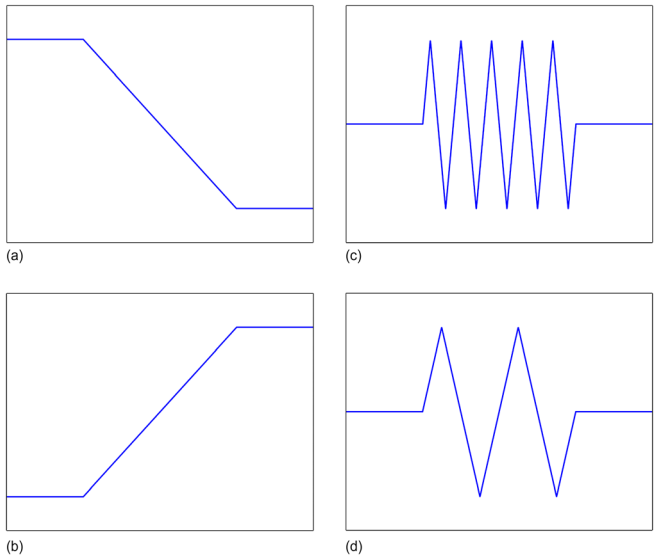

FIG. 2. (Color online) Temporal envelope shapes used for the stimuli. All envelopes are made up of two 0.5 s sections at the start and end, with a 1 s middle section. (a) *RD*, (b) *RU*, (c) *MF*, and (d) *MS*.

for the *CRU* and *CRD* envelopes is 300 Hz to 1.3 KHz. This corresponds to a range of spectral centroid in the stimuli of approximately 600 Hz.

The modulation rates for the *MF* and *MS* envelopes are 5 and 2 Hz, respectively. The 5 Hz rate is comfortably producible for pitch modulation, as it is in the region of a natural vibrato rate (Hakes *et al.*, 1988; Sundberg, 1994; Prame, 1994). Maximum rate of amplitude change has been studied with pitch modulation, and the two effects are generally considered to be closely coupled (Sundberg, 1989). We therefore expected participants to be able to produce amplitude modulations at 5 Hz. In terms of spectral centroid, the sound of a modulating cutoff frequency is somewhat similar to a "wah-wah" sound. It was therefore conceivable that people might use the phonemes [a] and [u] to create this effect. It has been shown that this diphthong glide (such as in the word "bout") can be voiced at moderate and fast speaking rates in 112 and 98 ms, respectively (Gay, 1968). At a modulation rate of 5 Hz the duration of a complete diphthong glide is 100 ms, and at 2 Hz it is 250 ms: both are comfortably within the producible range. For the *PMF* and *PMS* envelopes the modulation extent is 3 st, which equates to maximum excursion rates of 30 st/s at 5 Hz and 12 st/s at 2 Hz. Previous studies have shown these pitch excursion rates to be within typically producible ranges (Sundberg, 1973; Xu and Sun, 2000). The extent for *LMF* and *LMS* envelopes is

TABLE I. Identifiers for the 32 double-feature stimuli. These are each of the four pitch (P) control stimuli combined with each of the loudness (L) and spectral centroid (C) control stimuli.

	L Controls				C Controls			
P Controls	RD	RU	MF	MS	RD	RU	MF	MS
RD	*PRD + LRD*	*PRD + LRU*	*PRD + LMF*	*PRD + LMS*	*PRD + CRD*	*PRD + CRU*	*PRD + CMF*	*PRD + CMS*
RU	*PRU + LRD*	*PRU + LRU*	*PRU + LMF*	*PRU + LMS*	*PRU + CRD*	*PRU + CRU*	*PRU + CMF*	*PRU + CMS*
MF	*PMF + LRD*	*PMF + LRU*	*PMF + LMF*	*PMF + LMS*	*PMF + CRD*	*PMF + CRU*	*PMF + CMF*	*PMF + CMS*
MS	*PMS + LRD*	*PMS + LRU*	*PMS + LMF*	*PMS + LMS*	*PMS + CRD*	*PMS + CRU*	*PMS + CMF*	*PMS + CMS*

±6 dB (total extent of 12 dB), and for *CMF* and *CMS* it is ±500 Hz (total extent of 1 kHz).

Finally, it is worth mentioning here that pitch, loudness, and spectral centroid all interact when producing vocal sounds. For example, producing a vocal vibrato will also create modulations in loudness due to the physical properties of the vocal tract (Sundberg, 1989). This means that it is unreasonable to expect anyone to perfectly imitate a sound containing pitch modulations without modulating other acoustic features such as loudness and spectral centroid. To what extent these features interact is one of the questions of this study and is discussed in Sec. IV.

B. Participants

Nineteen participants took part in the study. Of these 16 were male and 3 were female. All the participants had experience in computer based music production and over five years' experience playing an instrument. Two participants were aged 18–25, 13 aged 26–35, and 4 aged 36–45.

C. Procedure

The study took place in an acoustically treated room. The recording chain was an AKG C414 (AKG Acoustics, Austria) microphone (cardioid polar pattern, low cut disabled, no pad engaged) and an Apogee Duet 2 (Apogee Electronics, CA) audio interface (microphone preamp and analogue to digital converter). The monitoring chain was an Apogee Duet 2 interface (digital to analogue conversion), Audient ASP 510 (Audient, England) monitor controller, and PMC AML (PMC, England) monitors. All audio was recorded at a sample rate of 44.1 KHz and bit depth of 24.

The participants were seated at a computer and presented with a basic interface for auditioning the stimuli and recording their imitations. They were advised that the aim of the study was to establish how accurately they could imitate the sounds with regards to pitch, loudness, and spectral envelope. The instructor then gave an overview of the interface and left the room for the duration of the study, to remove any potential influence on the participants.

For the workflow, each stimulus could be auditioned as many times as the participant wanted. The imitation could then be practised and recorded when ready. Participants were not able to listen back to their recordings, however if they were not happy with their performance they were able to re-record it as many times as they wished. Participants were advised that the final recording of each sound would be used for the analysis. The stimuli were split into two sets: controls and double-feature stimuli. The order of the stimuli within each set was randomised.

D. Feature extraction

The imitation files were manually edited in Apple Logic Pro to remove sections of silence/noise floor. The Sonic Annotator Vamp host (Cannam *et al.*, 2010) was then used to batch extract F_0, loudness, and spectral centroid features. The autocorrelation Yin based method by Mauch and Dixon (2014) was used to calculate F_0. Spectral centroid and loudness were extracted using the LibXtract Vamp plugins (Bullock, 2007): spectral centroid was calculated as the barycenter of the spectrum, using the definition given by Peeters (2004); loudness was calculated in sones, based on an implementation of the loudness model by Moore *et al.* (1997), described by Peeters (2004). All features were extracted with a 1024 sample window size and 256 sample window increment. This gives one frame-wise feature vector for each of the control imitations and two for each of the double-feature imitations.

E. Parameter extraction

To compare the imitations to the stimuli, envelope parameters were first extracted for each imitation. These are mean modulation rate and extent for the *MF* and *MS* envelopes, and range and slope of the ramp for the *RU* and *RD* envelopes. The methods for each of these processes are given in this section. Range and extent are measured in st for pitch, Hz for spectral centroid, and a ratio of max:min value in sones for loudness: for pitch and loudness these parameters are independent of the absolute value that the participant vocalises.

1. Modulation rate and extent

To extract rate and extent parameters we use methods that have previously been applied to vibrato parameter extraction. The initial steps are similar to the method used by Ferrante (2011), as follows:

(1) Low pass filter using a zero-phase sixth order IIR filter with a cutoff of 10 and 5 Hz for imitations of the *MF* and *MS* envelopes, respectively.
(2) Locate local maxima using a peak-picking algorithm.
(3) Interpolate the maxima positions (quadratic) to improve the rate calculation accuracy.
(4) Remove any neighbouring maximum within the minimum period threshold (0.1 s for 5 Hz and 0.2 s for 2 Hz), keeping the greater maximum.
(5) Find the minima between the maxima and interpolate the values (quadratic).
(6) Find the modulation area (first and last cycle with an extent $>1/6$ of the extent in the stimulus). This is to remove any flat start and end sections in the imitation.
(7) Calculate the per cycle rate (Fig. 3): Inverse of distance between two maxima/minima (Prame, 1994; Dromey *et al.*, 2003; Ferrante, 2011).
(8) Calculate the per cycle extent (Fig. 3): For pitch and spectral centroid this is the absolute difference between the highest and lowest values in each cycle (Hakes *et al.*, 1988; Xu and Sun, 2002), measured in st and Hz, respectively. For loudness this is measured as the ratio between the highest and lowest sone values in a cycle.
(9) Calculate the mean rate and extent for each imitation.

The detected minima and maxima were manually checked and adjusted where necessary [after step (6) above]. In 24 of the 722 feature envelope imitations there were no modulation cycles where the extent was above our minimum

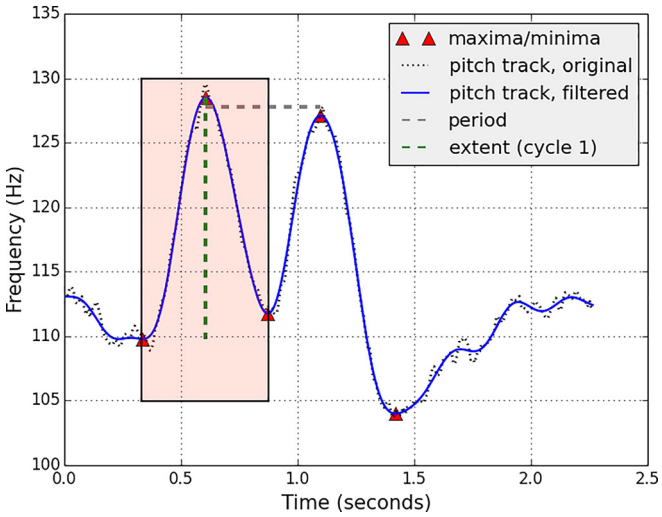

FIG. 3. (Color online) F_0 of one participant's imitation of the *PMS* envelope. The modulation rate is calculated as the inverse of the distance between two maxima. Extent is the difference between the highest and lowest values in a cycle (measured in st). The shaded area highlights a single cycle.

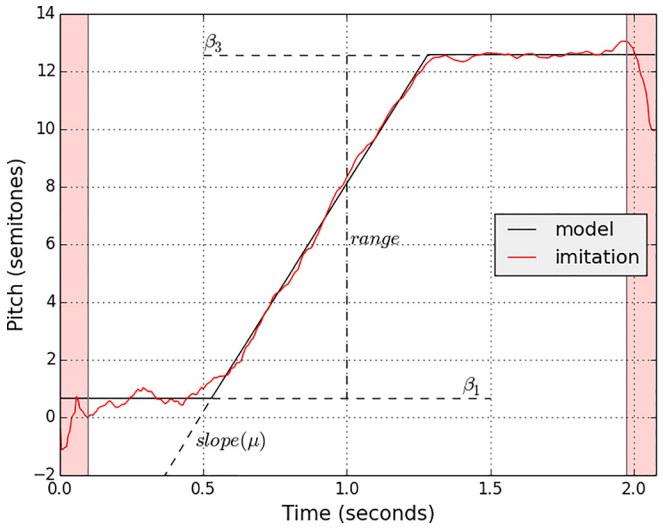

FIG. 4. (Color online) Pitch track (in st) of one participant's imitation of the *PRU* envelope, overlaid with the fitted model. The shaded sections (first and last 5%) are ignored for the model fitting as they tend to have a large error due to variation as people settle on a pitch and end a vocalisation.

threshold, i.e., the participant had failed to vocalise a suitable modulation. These cases were removed from the analysis.

2. Ramp slope and range

There are a number of ways to measure imitation accuracy for the ramp envelopes (*RD* and *RU*). These include cross correlating the imitation with the stimulus, using dynamic time warping to find the least cost alignment path, or simply measuring the error of the imitation with respect to the stimulus by testing the goodness of fit between the two. However, for this analysis we are particularly interested in the range and slope parameters of the imitated ramp, therefore we require a model that can be fitted to each imitation with certain constraints to provide the parameters of interest. The ramp envelopes in the stimuli are piecewise linear functions (Fig. 2), therefore it is reasonable to fit the feature vector of each imitation to such a function to determine the range and slope parameters, as follows.

We first remove the start and end 5% of the vector, as we are only interested in the parameters of the middle section of the envelope where the ramp exists and these sections can contain a lot of variation (see Fig. 4). Next we fit a continuous piecewise model that consists of 2 knots (k_1 and k_2), and where the slope for pieces 1 and 3 is 0. This model is given by

$$y = \begin{cases} \beta_1 + \epsilon(\chi), & \chi < k_1 \\ \beta_2 + \mu\chi + \epsilon(\chi), & k_1 \geq \chi \leq k_2 \\ \beta_3 + \epsilon(\chi), & \chi > k_2, \end{cases} \quad (1)$$

where β_1, β_2, and β_3 are the intercepts for each piece, μ is the slope of piece 2, ϵ is the squared error, and χ is the frame number. The best fit is found by iterating through every possible combination of χ positions for k_1, k_2, where $k_1 < k_2$, $\beta_1 = (\beta_2 + \mu k_1 + \epsilon_2(k_1))$ and $\beta_3 = (\beta_2 + \mu k_2 + \epsilon_2(k_2))$, minimising $\sum_{\chi=N*0.05}^{\lfloor *N*0.95 \rfloor} \epsilon(\chi)$, where $N =$ number of feature frames for a given imitation. See Fig. 4 for an illustration of this process. Once a best fit is found, the slope and range of the imitation ramp can be extracted from the model. For pitch and spectral centroid, the slope is given by β_2 and range by $|\beta_1 - \beta_3|$. For loudness we measure range and slope as values relative to the loudness of the vocalisation. The range is therefore given by $\max(\beta_1, \beta_3)/\min(\beta_1, \beta_3)$, and slope is taken as the range divided by the duration of piece 2. To our knowledge this approach has not been previously applied to ramp-based parameter extraction from acoustic feature vectors, however this is not surprising as the approach is tailored to our particular problem, where we want to extract the range and slope parameters from imitations of three piece continuous linear functions.

This method is based on the assumption that participants did indeed imitate a linear function. To test this we first visually inspected each imitation feature vector plotted over its respective model (as shown in Fig. 4). We then tested the linearity of the data using the Pearson product-moment correlation, and found a strong indication of linearity (mean across all feature vectors: $|r| = 0.79$). Of the resulting 722 pairs of parameters, 56 had either a middle ramp section duration <0.2 s, or the slope was in the opposite direction to that in the stimulus. These cases were removed from the analysis.

III. STATISTICAL ANALYSIS

Extent, range, and slope parameters were also extracted from the stimuli (as described in Sec. II), giving comparable parameters in the same units as the imitations (st, sone ratio, and Hz). The temporal parameters (rate and pieces) were taken from the synthesiser parameters, to remove the chance of any small errors that may be introduced from the frame-level averaging of the feature extraction. The parameters for each imitation were then compared to the respective stimuli

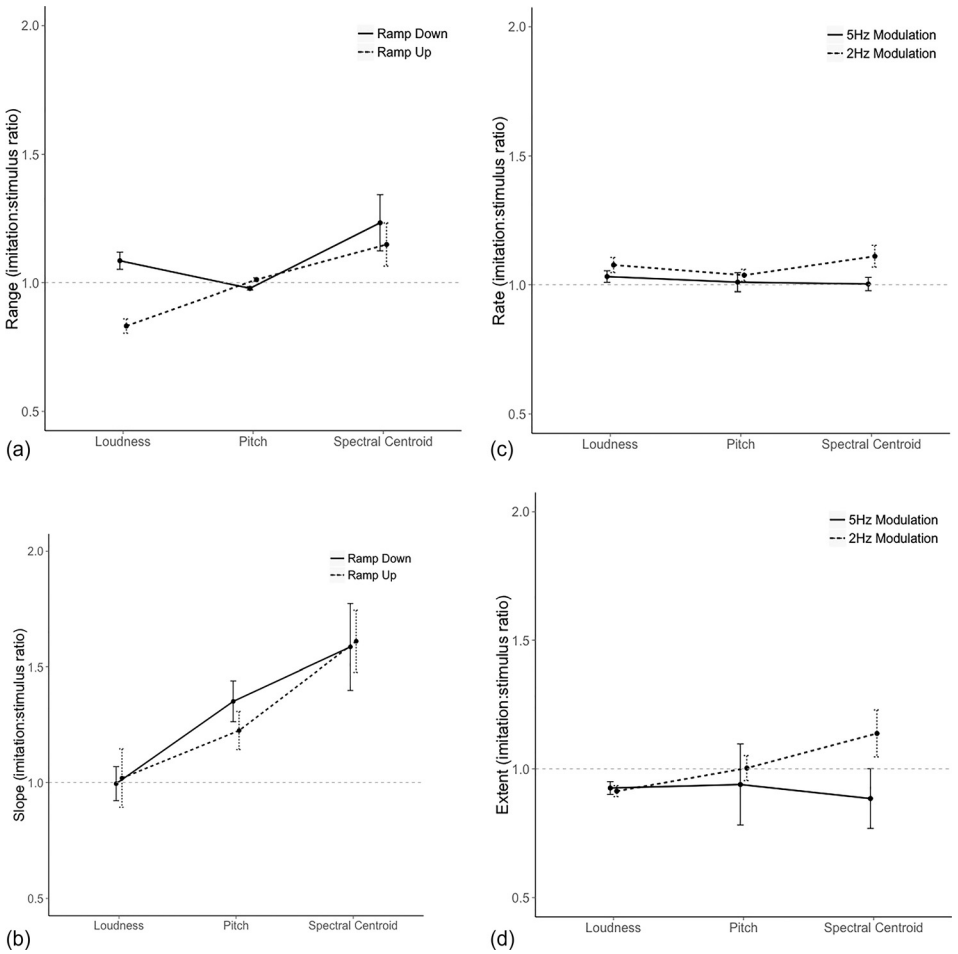

FIG. 5. Range (a), slope (b), rate (c), and extent (d) of accuracy for imitations of the 12 control stimuli (2 ramp envelopes and 2 modulation envelopes for each of the 3 features), across all participants. Values are means across participants with standard error bars.

to give a ratio of imitation:stimuli. This section is split into two parts: In Sec. III A we analyse imitations of the single-feature stimuli, followed by an analysis of the double-feature imitations in Sec. III B.

A. Single feature imitations

We tested for the effect of two factors on imitation accuracy: envelope and feature, using linear mixed effect (LME) regression. This was used because it is suited to a factorial analysis for within-participant repeated measures, controls for variance due to random effects, and can handle missing data (from the failed imitation cases) more effectively than fixed effect models.

Separate LME models were built for each parameter, with feature and envelope as fixed effects (with interaction terms), and a random intercept for each participant. Normality and homoscedasticity of the residuals were checked for each model by visual inspection. In cases where these assumptions were not clearly met we ran robust models (Koller, 2013). In these cases we found no major differences in parameter estimates or their variances between robust and non-robust approaches. All the models were built using the lme4 package (Bates *et al.*, 2015) for R (R Development Core Team, 2008). The effect of each factor was tested using type III analysis of variance (ANOVA) with Satterthwaite's degrees of freedom approximation from the lmerTest package (Kuznetsova *et al.*, 2016), with all p-values adjusted using the Benjamini & Hochberg false discovery rate correction from the *p.adjust*() function in R.

1. Ramp envelopes

A full factorial ANOVA was conducted on the range and slope LME models, testing for the fixed effects of feature (pitch, loudness, and spectral centroid), envelope (*RU* and *RD*), and interactions between the factors. For imitation range, there is a significant interaction between the feature and envelope factors [$F(2, 93) = 5.1$, $p_{adj} = 0.024$]. A significant interaction between factors means that it is not reasonable to analyse this model in terms of main effects (Nelder, 1977), therefore we conducted a *post hoc* analysis of interaction contrasts using the phia package for R (De Rosario-Martinez, 2015). This showed a significant contrast between loudness/pitch features and *RU*/*RD* envelopes [$\chi^2(1) = 9.9$, $p = 0.005$] and a smaller but marginally significant contrast between loudness/spectral centroid features and *RU*/*RD* envelopes [$\chi^2(1) = 4.1$, $p = 0.066$]. This effect is shown in Fig. 5(a), where the relatively large difference between *RU* and *RD* envelopes for loudness does not exist for pitch and spectral centroid.

Participants tended to imitate a larger loudness range for descending ramps than for ascending, with mean ranges of 1.09 (*LRU*) and 0.83 (*LRD*). The imitation ranges are generally larger than the stimuli range, except in the case of *LRU* and *PRD*, where it is very close to 1 (0.98). Participants tended to overshoot the range for *PRU* (1.01) whereas they undershot for *PRD* (0.98). Imitations of pitch range are more accurate and have a much lower variance than for loudness and spectral centroid.

In terms of ramp slope [Fig. 5(b)], we found no significant interaction between the envelope and feature factors,

and no significant effect of envelope on imitation accuracy. There is a significant and large effect of feature [$F(2, 92) = 17.2$, $p_{adj} < 0.001$]: slope means are most accurate for loudness (1.01) followed by pitch (1.29) and spectral centroid (1.59). The slopes of the imitations are steeper than the stimuli slopes for all features and envelopes.

2. Modulation envelopes

As with the ramp envelopes, a full factorial ANOVA was conducted on the rate and extent LME models, with the same factors of feature and envelope, but levels of *MF* and *MS* for the envelope factors (instead of *RU* and *RD*). The most striking finding here is the relative consistency of the modulation rate results across all features, compared to the other parameters. In general participants managed to imitate the rate with a high level of accuracy, with mean rates only slightly above the target for all stimuli [Fig. 5(c)]. There is a significant effect of envelope on modulation rate [$F(1, 113) = 6.4$, $p_{adj} = 0.025$]: imitation rates are higher than the stimuli for 2 Hz envelopes compared to 5 Hz. This effect is observed for all features, but is largest for spectral centroid.

It is worth noting that an alternative, and perhaps a more reasonable way to measure imitation accuracy for rate is to take the error in Hz instead of using the imitation:stimulus ratio. For example, a ratio of 1.5 at 2 Hz equates to an error of 1 Hz, whereas a ratio of 1.5 at 5 Hz equates to an error of 2.5 Hz. Conceivably these errors are therefore not the same in real terms. To test this we repeated the analysis using error in Hz instead of ratio and found that the effect of envelope disappears.

For modulation extent there is no interaction between factors or effect of feature, but there is a significant effect of envelope [$F(1, 94) = 7, 0.9$, $p_{adj} = 0.024$]. This is not observed for loudness, where there is little difference between fast and slow modulation rates and extent for the imitations is consistently lower than for the stimuli [Fig. 5(d)]. However, for pitch and spectral centroid a slower modulation rate appears to lead to a larger imitation extent. Overall, participants performed best when imitating the extent for *PMS*, indicating a positive effect of a slower rate for pitch. This effect is not observed for spectral centroid or loudness.

B. Double-feature imitations

In this section we report how the accuracy of each single feature envelope (i.e., *PRU*) changes when it is combined with envelopes of another feature (i.e., each of the spectral centroid and loudness envelopes). We perform the analysis by modeling the imitation accuracy for each feature separately, using LME regression. This gives eight LME models for each feature: four for each set of controls by two parameters. Each LME model has a fixed effect of stimulus type, with a random intercept for each participant.

The factor of stimulus type has three levels for pitch (pitch, pitch+loudness, pitch+spectral centroid), and two levels for loudness and spectral centroid (loudness and loudness+pitch, spectral centroid and spectral centroid+pitch). We average over the double-feature stimuli for each level, allowing us to test for the effect of the different feature combinations on each of the controls. This is tested by submitting each LME to a one way type III ANOVA using Satterthwaite's approximation for denominator degrees of freedom, with a factor of stimulus type. All *p*-values for each feature were adjusted using the Benjamini & Hochberg false discovery rate correction from the *p.adjust*() function in R.

1. Pitch

One way ANOVA was conducted on each of the eight LME models for pitch, testing for the effect of stimulus type as a factor with three levels (pitch, pitch+loudness, pitch+spectral centroid). We found no significant effect of the double-feature stimuli on accuracy of the range, rate, or extent parameters (Table II). There is however a significant effect of the double-feature stimuli on slope accuracy for both the *PRU* [$F(2, 152) = 5.6, p_{adj} = 0.017$] and *PRD* [$F(2, 150) = 7.2, p_{adj} = 0.008$] envelopes.

A Tukey *post hoc* analysis of the *PRU* slope model showed significant differences between the control and both pitch+loudness ($z = -3.1, p_{adj} = 0.003$), and pitch+spectral centroid ($z = -3.3, p_{adj} = 0.003$) stimulus types. This effect is also observed for the *PRD* slope model, with significant differences between the control and both pitch+loudness ($z = -3.7, p_{adj} < 0.001$), and pitch+spectral centroid ($z = -3.385, p_{adj} = 0.001$) stimulus types. For both models we found no significant differences between the loudness and spectral centroid double-feature stimulus types.

Figures 6(a) and 6(b) show the effect of each stimulus on slope accuracy for *PRU* and *PRD*: There is an improvement in slope accuracy when the *PRD* and *PRU* envelopes are combined with modulation envelopes of loudness or spectral centroid, particularly so for *PRD*. This effect may be due to the loudness and spectral centroid modulation cycles serving as a time-keeping aid for the pitch ramp stimuli, however it is not observed when loudness or spectral centroid ramp envelopes are combined with pitch modulation envelopes (see the Appendix, Tables VI, VII, and VIII).

Although there are no significant effects of the double-feature stimulus types for the range, rate, and extent

TABLE II. Means (and standard errors) of pitch imitation accuracy for pitch vs the double-feature stimulus types (pitch+loudness, pitch+spectral centroid). Bold values indicate a significant effect of stimulus type (e.g., single vs double feature) on imitation accuracy.

	Range		Slope		Rate		Extent	
Stimulus Type	*PRD*	*PRU*	*PRD*	*PRU*	*PMF*	*PMS*	*PMF*	*PMS*
Pitch (Control)	0.98 [0.01]	1.01 [0.01]	**1.35 [0.09]**	**1.23 [0.08]**	1.01 [0.04]	1.04 [0.02]	0.94 [0.16]	1.00 [0.05]
Pitch + Loudness	1.00 [0.00]	1.02 [0.01]	**1.14 [0.04]**	**1.03 [0.03]**	0.87 [0.03]	1.10 [0.02]	0.83 [0.04]	1.03 [0.03]
Pitch + Sp. Centroid	1.00 [0.00]	1.03 [0.01]	**1.15 [0.04]**	**1.02 [0.03]**	0.88 [0.03]	1.11 [0.03]	0.77 [0.05]	0.98 [0.03]

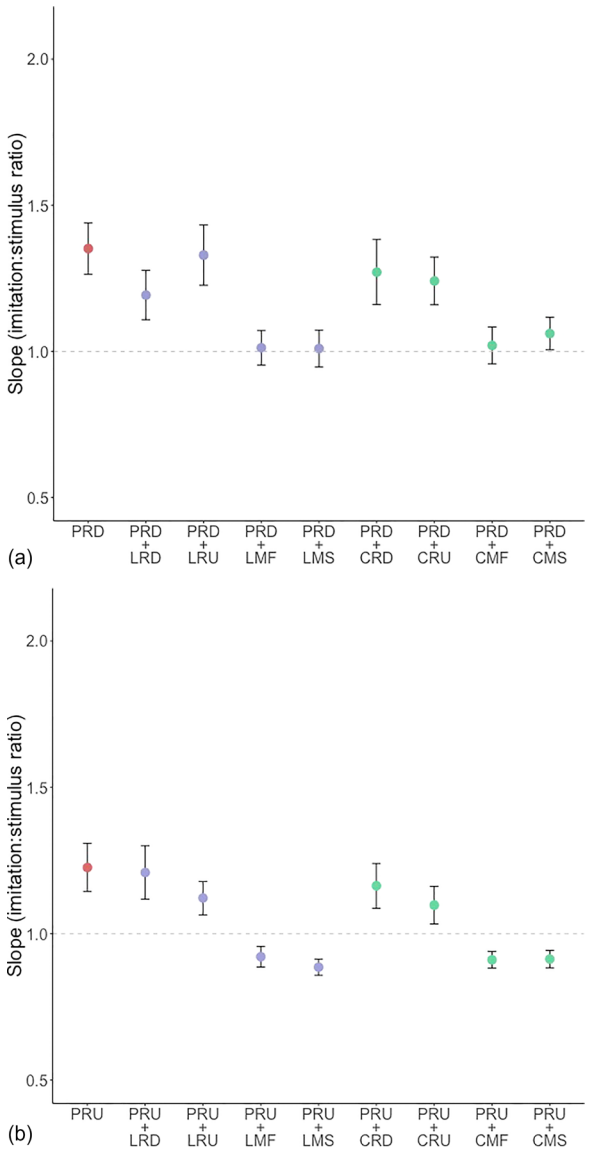

FIG. 6. (Color online) Pitch slope accuracy for controls *PRD* (a) and *PRU* (b) as single feature envelopes and when combined with each of the loudness and spectral centroid envelopes. Values are means across participants with standard error bars.

parameters, we note the following observations: Accuracy for the range parameter is very high compared to the other parameters (with max/min 95% confidence intervals of 0.97/1.04 across all stimulus types), indicating that participants were able to imitate the target ranges of the ramps even when they were imitating the double-feature stimuli. The accuracy of modulation rate for double-feature stimuli is lower than for the single-feature stimuli, however the direction of error is different for the 5 and 2 Hz envelopes: with the *PMF* envelope the rate is lower for imitations of double-feature stimuli than for single-feature, whereas for the *PMS* envelope the opposite trend is observed. Modulation extent is below the target extent for the 5 Hz pitch envelopes (*PMF*), and the double-feature stimuli appear to have a larger effect on extent accuracy for the *PMF* envelope than for *PMS*.

2. Loudness

We conducted a one way ANOVA on each of the eight LME models for loudness, testing for the effect of stimulus type as a factor with two levels (loudness, pitch+loudness). We found a significant effect of the double-feature stimulus type on accuracy of range for the *LRD* envelope [$F(1,73) = 14.6, p_{adj} < 0.001$], as can be seen in Table III. Figure 7(a) illustrates how this effect is driven by an asymmetry in the error between the *LRD* envelope and all the double-feature envelopes except *LRD+PRD*: For *LRD* and *LRD+PRD* participants tended to imitate a larger loudness range, whereas for the other double-feature envelopes the imitation range is smaller than the stimulus. This effect is not observed for *LRU*, where there is very little difference between the stimulus types.

There is also a significant effect of double-feature envelopes on accuracy of loudness extent for the *LMS* envelopes [$F(1,71) = 10.7, p_{adj} = 0.006$]. Here there is a small but notable improvement in imitation accuracy when the *LMS* envelope is combined with any of the pitch envelopes, as can be seen in Fig. 7(b).

There are no statistically significant differences between stimulus types for the other loudness envelope parameters, however there is a notable difference between accuracy of the ramp slope for the controls compared to the double-feature stimuli. Participants tend to imitate a steeper slope when the loudness ramps are combined with pitch envelopes, for both ascending and descending ramps. This is the case for all double-feature stimuli (see the Appendix, Table VII).

3. Spectral centroid

A one way ANOVA on each of the eight LME models for spectral centroid showed no significant effect of stimulus type on imitation accuracy for any of the envelope parameters (Table IV). Interestingly, there is notably a lower variance for both the range and extent parameters when the spectral centroid envelopes are combined with other pitch envelopes, and the lack of statistical significance for this effect is likely due to the large variance in the single-feature imitations.

TABLE III. Means (and standard errors) of loudness imitation accuracy for loudness vs pitch+loudness stimulus types. Bold values indicate a significant effect of stimulus type (e.g., single vs double feature) on imitation accuracy.

	Range		Slope		Rate		Extent	
Stimulus Type	*LRD*	*LRU*	*LRD*	*LRU*	*LMF*	*LMS*	*LMF*	*LMS*
Loudness(Control)	**1.09 [0.03]**	0.83 [0.03]	1.00 [0.07]	1.02 [0.13]	1.03 [0.02]	1.08 [0.03]	0.93 [0.02]	**0.91 [0.02]**
Pitch + Loudness	**0.92 [0.02]**	0.81 [0.01]	1.17 [0.06]	1.28 [0.09]	1.01 [0.02]	1.13 [0.03]	0.92 [0.01]	**0.99 [0.01]**

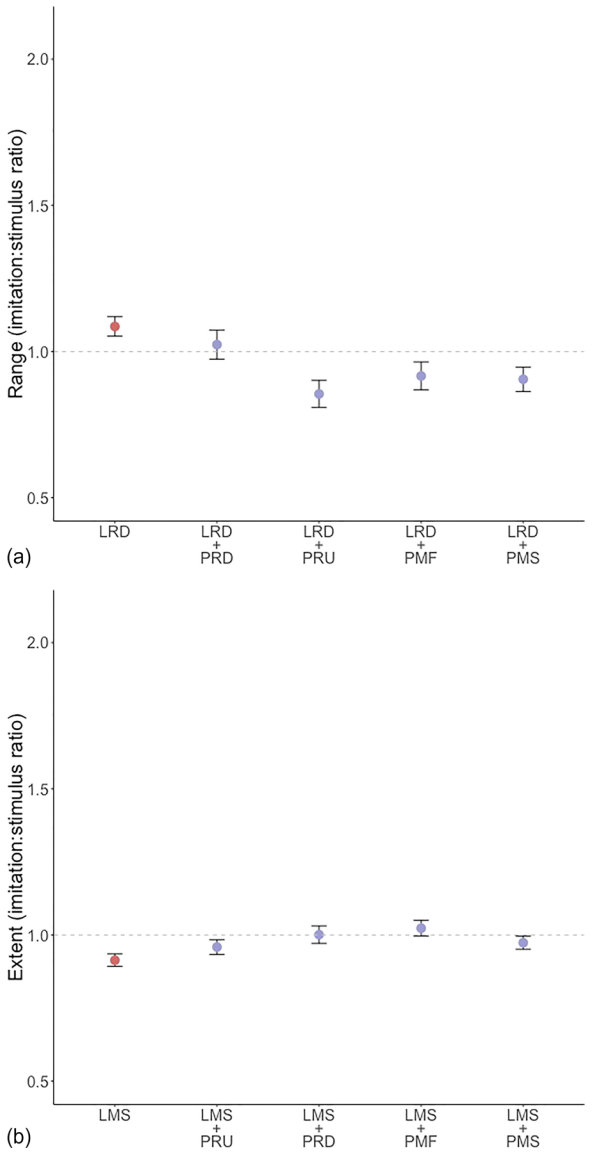

FIG. 7. (Color online) Loudness range (a) and extent (b) accuracy for loudness controls *LRD* (a) and *LMS* (b) as single feature envelopes and when combined with each of the pitch envelopes. Values are means across participants with standard error bars.

IV. DISCUSSION

A. How accurately can people imitate the temporal envelopes of pitch, loudness, and spectral centroid?

To address this question we focus the discussion on imitations of the control stimuli, and consider the ramp and modulation envelopes separately. Regarding the ramp envelopes, pitch was clearly the most accurate feature and had the lowest variance in terms of range, with mean ratios of 0.98 for descending ramps and 1.01 for ascending. This result is somewhat expected as there is a well-established relative scale for pitch, giving a concrete reference point for start and destination values that may not exist for loudness and spectral centroid.

There is an asymmetry in the accuracy of pitch range, with a clear effect of ramp direction. Perceptual accuracy of pitch ramp extreme values has been shown to be more accurate at the higher extremities (d'Alessandro *et al.*, 1998). This may explain why imitation range is more accurate for ascending ramps, if the participants were better able to perceive the correct ramp end pitch for upwards ramps. Interestingly, these results contrast those in the case of imitating a pitch interval, where it has been shown that both good and poor pitch singers tend to compress the interval, irrespective of direction (Pfordresher and Brown, 2007).

In terms of pitch error, the ratios of 0.98 for descending and 1.01 for ascending equate to errors of −34 cents and +12 cents, respectively, with a mean absolute error of 23 cents across both ramp envelopes. These results are not directly comparable to the many studies on singing voice pitch accuracy. Such studies tend to measure pitch interval error using melodies or intervals with discrete notes (our stimuli are based on a ramp between two notes). Nonetheless, previous studies on pitch interval accuracy show higher interval errors for non-musician adults: Pfordresher *et al.* (2010) report mean error of 87 cents for a 5 note melody task; Pfordresher and Brown (2007) report approximate mean error of 80 cents for good singers, and 155 cents for poor pitch singers in an interval task. In contrast, our results are similar to those in Mürbe *et al.* (2004), where professional singers exhibited a mean interval error of 19 cents when singing a slow, legato triad.

There does not appear to be any effect of ramp direction on spectral centroid range, however there is a clear asymmetry in the loudness imitations: participants exceeded the target range for descending ramps, and did not reach it for ascending, with mean ratios of 1.09 and 0.83, respectively. There are two factors at play here: ramp direction and autophonic loudness. Autophonic loudness is the perceived loudness of a sound that one produces with their own voice. Lane *et al.* (1961) show that autophonic loudness resembles a power function with an exponent of 1.1 (slope on a log-log scale of dB sound pressure level and autophonic loudness). Subsequent studies have validated the presence of this effect, with autophonic loudness slopes of 1.2 (Lane *et al.*, 1970) and 1.3 (Yadav and Cabrera, 2016) for the phoneme [a]. Ectophonic loudness is the perceived loudness of sounds external to the body (Yadav and Cabrera, 2016), which also

TABLE IV. Means (and standard errors) of spectral centroid imitation accuracy for spectral centroid vs pitch+spectral centroid stimulus types.

	Range		Slope		Rate		Extent	
Stimulus Type	*CRD*	*CRU*	*CRD*	*CRU*	*CMF*	*CMS*	*CMF*	*CMS*
Sp. Centroid (Control)	1.23 [0.11]	1.15 [0.08]	1.58 [0.19]	1.61 [0.13]	1.00 [0.03]	1.11 [0.04]	0.88 [0.12]	1.13 [0.09]
Pitch + Sp. Centroid	1.08 [0.06]	1.08 [0.06]	1.89 [0.17]	1.48 [0.13]	1.03 [0.02]	1.15 [0.03]	0.74 [0.03]	0.96 [0.04]

resembles a power function but with a slope of 0.6. This means that autophonic stimuli (i.e., one's own voice) will sound louder than ectophonic stimuli with equivalent loudness. In accordance with this power law, one would expect a vocalist to overestimate the actual loudness they produce, stopping short of the target destination loudness for an ascending ramp and surpassing it for descending. Our results show this effect, however we must also consider the perceptual bias of ramp direction.

Neuhoff (1998, 2001) shows that people tend to overestimate the loudness of rising sounds compared to falling. It is therefore conceivable that participants may have overestimated the ectophonic loudness range for ascending ramps and underestimated it for descending, when listening to the stimuli. This has the opposite effect of autophonic loudness. Our results indicate that the effect size of the autophonic loudness response counteracts the perceptual bias for rising tones. This effect is consistent across the participants [see standard error bars in Fig. 5(a)].

The fact that spectral centroid and sharpness both correlate with brightness (Schubert and Wolfe, 2006; Ilkowska and Miśkiewicz, 2006) allows us to compare the spectral centroid imitations of our participants to those of Lemaitre *et al.* (2016), where participants imitated the sharpness of sounds (amongst other features). The authors define sharpness as "*the sensation that distinguishes sounds on a continuum ranging from dull to sharp (or bright)*," which is calculated using the acum descriptor of Fastl and Zwicker (2006). Lemaitre *et al.* (2016) found a strong correlation between sharpness in the stimuli and imitations, with all participants producing sharpness levels around 30% higher than in the stimuli. We also found that participants tended to imitate sounds with greater spectral centroid values (and ranges) than in the stimuli. In our study the stimuli spectral centroid ranges are approximately 300–900 Hz for males and 400 Hz–1 kHz for females. These appear to be comfortably within the producible ranges for speech (Přibil and Přibilová, 2012), indicating that this finding is not due to physical limitations on upper or lower bounds of spectral centroid in vocalisations. We also note that participants did not produce upper spectral centroid values near those given in Přibil and Přibilová (2012). This is likely due to the fact that they were producing voiced phonemes: speech will typically have higher spectral centroid values due to the presence of unvoiced phonemes.

The results for slope accuracy are somewhat surprising. Even without a clear relative scale, such as we have with pitch, we would expect the rate of change to be similar across features (given equal level of control over each feature). In fact we see a clear and large effect of feature, with the slopes imitated remarkably well for loudness (1.0 descending, 1.02 ascending), followed by pitch (1.35 descending, 1.23 ascending) and spectral centroid (1.58 descending, 1.61 ascending), and no effect of ramp direction. The high accuracy of pitch range means that we can attribute the slope error to elongation of the ramp envelopes (participants imitated the correct range over a longer period). We also observe high slope values for spectral centroid. This may be due to participants trying to vocalise the correct duration for the stimuli: as the ranges tend to be larger, so the slopes must be steeper for the duration to be accurate. The steep slopes for spectral centroid ramps may be due to the unfamiliar process of vocalising a diphthong (as in "wah") slowly: this is normally spoken at a natural, relatively fast rate compared to the ramps in the stimuli.

In general, accuracy was high for the modulation rate, with mean ratios for each stimulus ranging from 1.00 (*CMF*) to 1.11(*CMS*). As noted in Sec. III A 2, when measured as a ratio the modulation rate error is higher for the 2 Hz stimuli than for 5 Hz, across all features. This shows that for the two rates we have tested here, relative error appears to be inversely proportional to modulation rate. This is likely to be influenced by two factors: First, the 5 Hz rate is well within the producible range, particularly for pitch change (Sundberg, 1973; Xu and Sun, 2000) and also at a natural vibrato rate (Hakes *et al.*, 1988; Sundberg, 1994; Prame, 1994); second, 2 Hz is such a slow modulation rate that slight deviations in timing would cause a relatively large error compared to the 5 Hz stimuli.

As with ramp range, modulation extent is considerably more accurate for pitch than loudness and spectral centroid, with ratio scores corresponding to mean errors of 1 cent at 2 Hz and −18 cents at 5 Hz (target extent for both rates was 3 st). The difference between modulation rates indicates that participants were more able to imitate the target range at 2 Hz; an effect that is not observed for loudness or spectral centroid. The difference in accuracy between pitch, loudness, and spectral centroid is again likely due to the existence of a well-established relative scale for pitch, and the below-target loudness extent is likely due to the effect of autophonic response (Lane *et al.*, 1961), as discussed above.

B. What happens to imitation accuracy when people are asked to imitate multiple feature envelopes simultaneously?

In general imitation accuracy was not significantly different between the double- and single-feature stimuli. Imitation accuracy of ramp range is not significantly improved for double-feature envelopes of the same shape, nor adversely affected for double-feature envelopes with inverse shapes (e.g., pitch ramp down with loudness ramp up). This is surprising as previous studies have identified interactions between pitch, loudness, and formants. For example, phonetogram studies have shown positive correlations of pitch and loudness for speech (Gramming *et al.*, 1988; Gramming, 1991; Alku *et al.*, 2002). This has also been shown to exist in singing (Sundberg *et al.*, 1993) and the sustained vowel [a] (Huber *et al.*, 1999) (Huber *et al.* also identified an increase in first formant frequency with intensity). In addition to these findings, it is clear that an increase in pitch would naturally produce an increase in spectral centroid. This suggests that for us to find no significant change in imitation accuracy for double-feature envelopes, the participants demonstrated an ability to control multiple features simultaneously, at least within a similar level of error to when they were required to control a single feature.

Pitch slope accuracy is improved when pitch ramp envelopes are combined with modulation envelopes for other features. We believe that this is due to the modulation cycles

acting as a time keeping aid, which combined with accurate pitch range will naturally bring the slope closer to the target. The effect is not observed for loudness or spectral centroid ramps. This is interesting because pitch rate is adversely affected by double-feature stimuli, whereas loudness and spectral centroid rate are not. Therefore it appears that participants are not able to retain control over pitch modulation as well as they are for the other two features.

There is some indication that combining feature envelopes may introduce conformity amongst how participants imitate the sound. In most cases there is a lower or equal variance in the imitations for the double-feature stimuli compared to those with single features. This effect is unexpected if we consider double-feature envelopes to be more difficult to imitate than single features: intuitively one would expect across participant variation to increase with difficulty.

Finally, when imitating stimuli containing two modulation envelopes with different rates, participants tended to find a rate somewhere between 2 and 5 Hz (for example, the pitch rate accuracy ratio for both *PMF+LMS* and *PMF+CMS* is 0.73, which equates to 3.65 Hz). This indicates an inability to accurately vocalise multiple feature envelopes with different modulation rates, as might be expected.

C. Effects of singing and sex

The effects of singing experience and sex are not within of the scope of this study, and not required to answer our research questions, therefore we did not control for these when recruiting the participants. We did however ensure that the stimuli parameters for pitch were suitably differentiated for male ($n = 16$) and female ($n = 3$) participants with regards to range and extent (see Sec. II A 1 for details).

In terms of singing training, participants were asked if they play an instrument or sing, and if so for how many years they had spent doing this. Of the 19 participants, 6 responded as having been a singer for 5 yrs or more. We tested for the effect of both singing training and sex on the imitation accuracy of each parameter using LME models with participant as a random effect and the following fixed effects: feature, envelope, singing training, and sex. A full factorial ANOVA on the LME models indicated no significant effects of either singing experience or sex on the imitation accuracy of any of the four parameters. It is worth mentioning that this does not mean that singing training and or sex have no effect on a persons' ability to imitate the stimuli used in this study: the lack of a significant effect may be due to a number of factors such as the limited sample size and ambiguity about what constitutes singing experience. To establish the effect of these factors we would recommend conducting a follow up study where sex and singing training are controlled for and suitable sample sizes used.

D. Participant feedback

The participants completed a short feedback questionnaire following the study. A breakdown of the responses is shown in Table V. All participants reported that they were able to detect which features were changing in each sound, however only 14/19 felt that they were able to vocalise the features with regards to timing, and 10/19 with regards to depth/extent. This indicates that participants felt that they could always hear and perceive what was happening in the stimuli; however, they were not always confident in the accuracy of their vocalisations. There was also more uncertainty ("Neither" response) in the imitation accuracy of depth/extent, with 6/19 participants unsure of whether they were able to imitate it accurately (compared to 0/19 for timing). This feedback is partially reflected in the results, where timing (rate) accuracy for modulation envelopes is generally higher than extent accuracy for the 5 Hz envelopes, however it is not the case for 2 Hz envelopes. Most of the participants (17/19) felt that it was more difficult to imitate the double-feature envelopes than the controls. This is interesting given that results show that for most double-feature envelopes the control imitations are not significantly more accurate than the respective double-feature envelopes.

V. CONCLUSIONS AND FUTURE WORK

The findings of this study complement previous work on vocal imitations by studying the interactions of three features central to voice quality: pitch, loudness, and spectral centroid, when applied to a foundation set of envelope shapes. This knowledge is useful for the design of QBV and vocally controlled synthesis systems. For example, a QBV system might accommodate for asymmetries in the vocalisation of ascending vs descending ramps for loudness and pitch, or

TABLE V. Participant responses from the post study questionnaire. The responses were recorded on a seven point Likert scale, which is summarised here on a three point scale.

	# Responses		
	Disagree	Neither	Agree
"I was able to detect which features were changing in each sound"	0	0	19
"I managed to accurately vocalise the features with regards to timing"	5	0	14
"I managed to accurately vocalise the features with regards to depth/extent"	3	6	10
"It was more difficult to imitate two features changing simultaneously than one"	0	2	17
"I felt comfortable using my voice in this way [as required for the study]"	3	4	12
"I have good vocal control of pitch"	2	6	11
"I have good vocal control of loudness"	3	7	9
"I have good vocal control of timbre"	2	5	12
"If I have a sound in my head, I can describe it using my voice (without using words)"	2	4	13
"When making music with other people, I sometimes use my voice to describe sounds (non-verbally)"	3	3	13

the under-estimation of modulation extent for faster modulation rates.

In general participants performed remarkably well at imitating pitch range and modulation extent, which is likely due to their musical training. This indicates that musicians can exercise a high level of control over pitch and perform vibrato, irrespective of their singing experience. Most importantly though, the results of this study suggest that the participants were able to exercise control over two feature envelopes simultaneously, at least as well as they were able to imitate single feature envelopes. In addition, there is a small but consistent effect of double-feature stimuli on across-participant variation (it is lower for double-feature stimuli than for single-feature). The main findings are summarised as follows:

(1) In most cases, combining two features envelopes does not have a significant effect on imitation accuracy.
(2) There is asymmetry in the accuracy and direction of error for both pitch and loudness ramps. For pitch, participants tended to overshoot the target range for ascending ramps, and these were imitated more accurately. The opposite effect is observed for loudness.
(3) Ramp range accuracy is highest for pitch, with considerably less variation compared to loudness and spectral centroid range. There is also a significant effect of feature on slope accuracy: loudness is most accurate, followed by pitch and spectral centroid.
(4) Participants generally imitated modulation rates of 2 and 5 Hz with high accuracy for all features.
(5) Modulation extent is more accurate for pitch than for loudness and spectral centroid.
(6) There are clear effects of modulation rate (2 vs 5 Hz) on both rate and extent accuracy: higher (and overestimated) imitation rates occur at 2 Hz for all features; and larger extents are also observed at 2 Hz for pitch and spectral centroid (but not loudness).
(7) Slope accuracy tends to improve when the ramp envelope is combined with a modulation envelope of another feature, if the modulation rate is reasonably accurate.
(8) Double-feature envelopes containing modulation envelopes at different rates tend to reduce rate accuracy for both features, to a rate somewhere between the two rates.

We analysed the imitations using only two parameters for each shape. This makes it possible to compare performance across different features, however it is difficult to compare results across all the stimuli (ramp vs modulation) using these shape-specific measures. Future research on vocal imitations could pursue the development of a standard imitation accuracy score that includes the parameters used here along with additional features such as segment duration, start and end values, and morphological features that are suited to a wider range of envelope shapes (Marchetto and Peeters, 2015).

Finally, we have measured imitation accuracy using computational methods, which we propose should be complemented by human listener evaluations. Results of a perceptual analysis of imitation accuracy for each stimulus would inform us about the relevance of the parameters used here. This could also be combined with an analysis of our dataset using the morphological features described by Marchetto and Peeters (2015), to derive a feature set that correlates with human perception of imitation accuracy. Measures of imitation accuracy could then be compared to classification accuracy results for vocal imitations such as in Dessein and Lemaitre (2009), Rocchesso and Mauro (2014), and Zhang and Duan (2015), providing new metrics for evaluating such systems. We encourage the research community to exploit the vocal imitation data that have been collected and analysed in this study, which is available online.[1]

ACKNOWLEDGMENTS

The authors would like to thank Zia Mehrabi, György Fazekas, and the anonymous reviewers for their invaluable comments. This work was supported by the Media and Arts Technology programme, EPSRC Doctoral Training Centre EP/G03723X/1.

APPENDIX

TABLE VI. Results (mean and standard error) for imitations of the four pitch envelopes, both individually (*Control*) and when combined with each of the loudness and spectral centroid envelopes. For details on stimuli labels see Sec. II.

		Double-feature stimuli (combined with control)							
	Control	*LRD*	*LRU*	*LMF*	*LMS*	*CRD*	*CRU*	*CMF*	*CMS*
PRD									
Slope	1.35 [0.09]	1.19 [0.08]	1.33 [0.10]	1.01 [0.06]	1.01 [0.06]	1.27 [0.11]	1.24 [0.08]	1.02 [0.06]	1.06 [0.06]
Range	0.98 [0.01]	0.98 [0.01]	0.99 [0.01]	1.01 [0.01]	1.01 [0.01]	0.98 [0.01]	1.01 [0.01]	1.00 [0.01]	1.01 [0.01]
PRU									
Slope	1.23 [0.08]	1.21 [0.09]	1.12 [0.06]	0.92 [0.04]	0.89 [0.03]	1.16 [0.08]	1.10 [0.06]	0.91 [0.03]	0.91 [0.03]
Range	1.01 [0.01]	1.02 [0.01]	1.02 [0.02]	1.05 [0.01]	1.01 [0.01]	1.03 [0.01]	1.01 [0.01]	1.03 [0.01]	1.04 [0.01]
PMF									
Rate	1.01 [0.04]	0.91 [0.05]	0.90 [0.04]	0.95 [0.04]	0.73 [[0.08]	0.88 [[0.06]	0.93 [[0.05]	0.99 [[0.04]	0.73 [[0.07]
Extent	0.94 [0.16]	0.83 [0.09]	0.76 [0.06]	0.66 [0.08]	1.05 [0.09]	0.65 [0.08]	0.74 [[0.09]	0.70 [[0.07]	0.97 [[0.11]
PMS									
Rate	1.04 [0.02]	1.07 [0.05]	1.08 [0.04]	1.20 [0.06]	1.05 [0.03]	1.06 [0.04]	1.06 [[0.03]	1.28 [[0.10]	1.06 [[0.03]
Extent	1.00 [0.05]	0.96 [0.05]	0.95 [0.06]	1.10 [0.07]	1.10 [0.05]	0.95 [0.06]	0.91 [[0.06]	1.06 [[0.08]	1.00 [[0.06]

TABLE VII. Results (mean and standard error) for imitations of the four loudness envelopes, individually (*Control*) and when combined with each of the pitch envelopes.

		Double-feature stimuli (combined with control)			
	Control	*PRD*	*PRU*	*PMF*	*PMS*
LRD					
Slope	1.00 [0.07]	1.18 [0.11]	1.22 [0.14]	1.19 [0.13]	1.10 [0.14]
Range	1.09 [0.03]	1.02 [0.05]	0.85 [0.05]	0.92 [0.05]	0.90 [0.04]
LRU					
Slope	1.02 [0.13]	1.15 [0.12]	1.32 [0.16]	1.23 [0.17]	1.43 [0.25]
Range	0.83 [0.03]	0.77 [0.03]	0.87 [0.03]	0.80 [0.03]	0.80 [0.02]
LMF					
Rate	1.03 [0.02]	1.09 [0.03]	1.07 [0.03]	1.00 [0.02]	0.90 [0.05]
Extent	0.93 [0.02]	0.92 [0.02]	0.90 [0.02]	0.90 [0.02]	0.97 [0.02]
LMS					
Rate	1.08 [0.03]	1.18 [0.05]	1.10 [0.04]	1.21 [0.08]	1.04 [0.03]
Extent	0.91 [0.02]	0.96 [0.03]	1.00 [0.03]	1.02 [0.03]	0.97 [0.02]

TABLE VIII. Results (mean and standard error) for imitations of the four spectral centroid envelopes, individually (*Control*) and when combined with each of the pitch envelopes.

		Double-feature stimuli (combined with control)			
	Control	*PRD*	*PRU*	*PMF*	*PMS*
CRD					
Slope	1.58 [0.19]	1.72 [0.26]	1.99 [0.39]	1.89 [0.36]	2.00 [0.36]
Range	1.23 [0.11]	0.94 [0.1.0]	1.31 [0.14]	1.01 [0.09]	1.10 [0.13]
CRU					
Slope	1.61 [0.13]	1.42 [0.18]	1.57 [0.21]	1.72 [0.35]	0.99 [0.12]
Range	1.15 [0.08]	1.13 [0.13]	1.05 [0.09]	1.13 [0.12]	0.99 [0.17]
CMF					
Rate	1.00 [0.03]	1.06 [0.04]	1.05 [0.02]	1.04 [0.02]	0.98 [0.04]
Extent	0.88 [0.12]	0.74 [0.07]	0.69 [0.07]	0.79 [0.07]	0.75 [0.07]
CMS					
Rate	1.11 [0.04]	1.14 [0.05]	1.14 [0.06]	1.15 [0.08]	1.18 [0.07]
Extent	1.13 [0.09]	0.92 [0.08]	0.90 [0.09]	1.07 [0.11]	0.96 [0.06]

[1]Audio files of the stimuli and acoustic features of the imitations are available at www.adibmehrabi.com/vocal_imitation_dataset.

Alku, P., Vintturi, J., and Vilkman, E. (**2002**). "Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation," Speech Commun. **38**(3), 321–334.

Baken, R. J., and Orlikoff, R. F. (**2000**). *Clinical Measurement of Speech and Voice*, 2nd ed. (Singular Thomson Learning, San Diego, CA), 175 pp.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Software **67**(1), 1–48.

Blancas, D. S., and Janer, J. (**2014**). "Sound retrieval from voice imitation queries in collaborative databases," in *Proceedings of the 53rd Audio Engineering Society Conference*, London, England, pp. 2–8.

Brown, W., Morris, R. J., Hicks, D. M., and Howell, E. (**1993**). "Phonational profiles of female professional singers and nonsingers," J. Voice **7**(3), 219–226.

Bullock, J. (**2007**). "Libxtract: A lightweight library for audio feature extraction," in *Proceedings of the International Computer Music Conference*, Copenhagen, Denmark, pp. 25–28.

Cannam, C., Sandler, M., Jewell, M. O., Rhodes, C., and d'Inverno, M. (**2010**). "Linked data and you: Bringing music research software into the semantic web," J. New Mus. Res. **39**(4), 313–325.

Cartwright, M., and Pardo, B. (**2014**). "Synthassist: Querying an audio synthesizer by vocal imitation," in *Proceedings of the Conference on New Interfaces for Musical Expression*, London, United Kingdom, pp. 363–366.

Cartwright, M., and Pardo, B. (**2015**). "Vocalsketch: Vocally imitating audio concepts," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul, Korea, pp. 43–46.

Coleman, R. F., Mabis, J. H., and Hinson, J. K. (**1977**). "Fundamental frequency-sound pressure level profiles of adult male and female voices," J. Speech, Lang., Hear. Res. **20**(2), 197–204.

Colton, R. H. (**1970**). "Vocal intensity in the modal and falsetto registers," J. Acoust. Soc. Am. **47**(1A), 105.

d'Alessandro, C., Rosset, S., and Rossi, J. P. (**1998**). "The pitch of short-duration fundamental frequency glissandos," J. Acoust. Soc. Am. **104**(4), 2339–2348.

DeLeo LeBorgne, W., and Weinrich, B. D. (**2002**). "Phonetogram changes for trained singers over a nine-month period of vocal training," J. Voice **16**(1), 37–43.

De Rosario-Martinez, H. (**2015**). phia: Post-Hoc Interaction Analysis (R package version 0.2-1).

Dessein, A., and Lemaitre, G. (**2009**). "Free classification of vocal imitations of everyday sounds," in *Proceedings of the 6th Conference on Sound and Music Computing*, Porto, Portugal, pp. 213–218.

Dromey, C., Carter, N., and Hopkin, A. (**2003**). "Vibrato rate adjustment," J. Voice **17**(2), 168–178.

Fastl, H., and Zwicker, E. (**2006**). *Psychoacoustics: Facts and Models* (Springer: Science and Business Media, Heidelberg, Berlin), Vol. 22, pp. 239–243.

Ferrante, I. (**2011**). "Vibrato rate and extent in soprano voice: A survey on one century of singing," J. Acoust. Soc. Am. **130**(3), 1683–1688.

Fitch, J. L., and Holbrook, A. (**1970**). "Modal vocal fundamental frequency of young adults," Arch. Otolaryngology **92**(4), 379–382.

Gay, T. (**1968**). "Effect of speaking rate on diphthong formant movements," J. Acoust. Soc. Am. **44**(6), 1570–1573.

Gramming, P. (**1991**). "Vocal loudness and frequency capabilities of the voice," J. Voice **5**(2), 144–157.

Gramming, P., Sundberg, J., Ternström, S., Leanderson, R., and Perkins, W. H. (**1988**). "Relationship between changes in voice pitch and loudness," J. Voice **2**(2), 118–126.

Hakes, J., Shipp, T., and Doherty, E. T. (**1988**). "Acoustic characteristics of vocal oscillations: Vibrato, exaggerated vibrato, trill, and trillo," J. Voice **1**(4), 326–331.

Helén, M., and Virtanen, T. (**2007**). "Query by example of audio signals using Euclidean distance between Gaussian mixture models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, Hawaii, Vol. 1, pp. 1–225.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., and Johnson, K. (**1999**). "Formants of children, women, and men: The effects of vocal intensity variation," J. Acoust. Soc. Am. **106**(3), 1532–1542.

Ilkowska, M., and Miśkiewicz, A. (**2006**). "Sharpness versus brightness: A comparison of magnitude estimates," Acta Acust. Acust. **92**(5), 812–819.

Janer, J. (**2005**). "Feature extraction for voice-driven synthesis," in *Proceedings of the 118th Audio Engineering Society Convention*, Barcelona, Spain.

Kent, R. D., Kent, J. F., and Rosenbek, J. C. (**1987**). "Maximum performance tests of speech production," J. Speech Hear. Dis. **52**(4), 367–387.

Koller, M. (**2013**). "Robust estimation of linear mixed models," Ph.D. thesis, ETH, Zurich.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2016**). "lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package)," http://CRAN.R-project.org/package=lmerTest (Last viewed 12/11/16).

Lane, H., Catania, A. C., and Stevens, S. S. (**1961**). "Voice level: Autophonic scale, perceived loudness, and effects of sidetone," J. Acoust. Soc. Am. **33**(2), 160–167.

Lane, H., Tranel, B., and Sisson, C. (**1970**). "Regulation of voice communication by sensory dynamics," J. Acoust. Soc. Am. **47**(2B), 618–624.

Lemaitre, G., Dessein, A., Susini, P., and Aura, K. (**2011**). "Vocal imitations and the identification of sound events," Ecol. Psych. **23**(4), 267–307.

Lemaitre, G., Jabbari, A., Houix, O., Misdariis, N., and Susini, P. (**2016**). "Vocal imitations of basic auditory features," J. Acoust. Soc. Am. **137**(4), 2268–2268.

Lemaitre, G., and Rocchesso, D. (**2014**). "On the effectiveness of vocal imitations and verbal descriptions of sounds," J. Acoust. Soc. Am. **135**(2), 862–873.

Lemaitre, G., Susini, P., Rocchesso, D., Lambourg, C., and Boussard, P. (**2014**). "Non-verbal imitations as a sketching tool for sound design," in *Lecture Notes in Computer Sciences: Sound, Music, and Motion*, edited by

M. Aramaki, O. Derrien, R. Kronland-Martinet, and S. Ystad (Springer, Berlin, Heidelberg, Germany), pp. 558–574.
Marchetto, E., and Peeters, G. (**2015**). "A set of audio features for the morphological description of vocal imitations," in *Proceedings of the 18th International Conference on Digital Audio Effects* (*DAFx*), Trondheim, Norway, pp. 207–214.
Mauch, M., and Dixon, S. (**2014**). "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 659–663.
Moore, B. C., Glasberg, B. R., and Baer, T. (**1997**). "A model for the prediction of thresholds, loudness, and partial loudness," J. Audio. Eng. Soc. **45**(4), 224–240.
Mürbe, D., Pabst, F., Hofmann, G., and Sundberg, J. (**2004**). "Effects of a professional solo singer education on auditory and kinesthetic feedback—A longitudinal study of singers' pitch control," J. Voice **18**(2), 236–241.
Neuhoff, J. (**1998**). "Perceptual bias for rising tones," Nature **395**(6698), 123–124.
Neuhoff, J. (**2001**). "An adaptive bias in the perception of looming auditory motion," Ecol. Psychol. **13**(2), 87–110.
Nelder, J. A. (**1977**). "A reformulation of linear models," J. R. Stat. Soc. A **140**(1), 48–77.
Pfordresher, P., and Brown, S. (**2007**). "Poor-pitch singing in the absence of tone deafness," Music Percept. **25**(2), 95–115.
Pfordresher, P., Brown, S., Meier, K., Belyk, M., and Liotti, M. (**2010**). "Imprecise singing is widespread," J. Acoust. Soc. Am. **128**(4), 2182–2190.
Peeters, G. (**2004**). "A large set of audio features for sound description (similarity and description) in the Cuidado project," CUIDADO I.S.T. Project Report.
Prame, E. (**1994**). "Measurements of the vibrato rate of ten singers," J. Acoust. Soc. Am. **96**(4), 1979–1984.
Přibil, J., and Přibilová, A. (**2012**). "Comparison of complementary spectral features of emotional speech for German, Czech, and Slovak," in *Cognitive Behavioural Systems* (Springer, Berlin Heidelberg), pp. 236–250.
R Development Core Team (**2008**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria.
Rocchesso, D., and Mauro, D. (**2014**). "Self-organising the space of vocal imitations," in *Proceedings of the Colloquio di Informatica Musicale Conference*, Rome, Italy, pp. 124–128.
Roma, G., and Serra, X. (**2015**). "Querying freesound with a microphone," in *Proceedings of the First Web Audio Conference*, Paris, France.
Schubert, E., and Wolfe, J. (**2006**). "Does timbral brightness scale with frequency and spectral centroid?," Acta Acust. Acust. **92**(5), 820–825.
Stowell, D. (**2010**). "Making music through real-time voice timbre analysis: Machine learning and timbral control," Ph.D. thesis, Queen Mary University of London, London, England.
Sundberg, J. (**1973**). "Data on maximum speed of pitch changes," Speech Trans. Lab. Qtr. Prog. Status Rep. **4**, 39–47.
Sundberg, J. (**1979**). "Maximum speed of pitch changes in singers and untrained subjects," J. Phon. **7**(2), 71–79.
Sundberg, J. (**1989**). *Science of the Singing Voice* (Northern Illinois University Press, Illinois), pp. 164–166.
Sundberg, J. (**1994**). "Acoustic and psychoacoustic aspects of vocal vibrato," Speech Trans. Lab. Qtr. Prog Status Rep. **35**(2–3), 45–68.
Sundberg, J., Titze, I., and Scherer, R. (**1993**). "Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source," J. Voice **7**(1), 15–29.
Xu, Y., and Sun, X. (**2000**). "How fast can we really change pitch? Maximum speed of pitch change revisited," in *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, pp. 666–669.
Xu, Y., and Sun, X. (**2002**). "Maximum speed of pitch change and how it may relate to speech," J. Acoust. Soc. Am. **111**(3), 1399–1413.
Xue, J., Wichern, G., Thornburg, H., and Spanias, A. (**2008**). "Fast query by example of environmental sounds via robust and efficient cluster-based indexing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, pp. 5–8.
Yadav, M., and Cabrera, D. (**2016**). "Autophonic loudness of singers in simulated room acoustic environments," J. Voice (published online).
Zhang, Y., and Duan, Z. (**2015**). "Retrieving sounds by vocal imitation recognition," in *Proceedings on the 25th IEEE International Workshop on Machine Learning for Signal Processing*, Boston, MA, pp. 1–6.
Zraick, R. I., Nelson, J. L., Montague, J. C., and Monoson, P. K. (**2000**). "The effect of task on determination of maximum phonational frequency range," J. Voice **14**(2), 154–160.