VOCAL IMITATION OF PITCH, SPECTRAL SHAPE AND LOUDNESS ENVELOPES

Adib Mehrabi, Simon Dixon, Mark B Sandler
Centre for Digital Music, Queen Mary University of London

ABSTRACT

We conducted a vocal imitation study to investigate the degree to which people can exercise vocal control over multiple feature envelopes simultaneously. Participants were asked to imitate 44 synthesised stimuli that varied in pitch (P), loudness (L) and spectral centroid (C). The envelope shapes applied to these features were ramps (up and down), and modulation (with a rate of 5Hz and 2Hz). The imitations of stimuli with a single feature envelope (e.g. 'P ramp up') were then compared to imitations of stimuli with two feature envelopes combined (e.g. 'P ramp up' with 'L ramp down'). In general the combination of feature envelopes appears to have more of a significant impact on the accuracy of the imitations for P than it does for L and C. Interestingly, the accuracy of envelope imitations for C is generally not affected by envelope combinations.

1. INTRODUCTION

Vocal imitations have been studied for applications in: sound design [3]; sound classification [4]; describing sounds [5]; audio sample retrieval [1] and automatically setting synthesiser parameters [2]. Previous studies address applications of the voice when used to imitate sounds, however there appears to be very little low-level feature based analysis of how accurately people can imitate multiple time varying features. We have conducted this study to create a new dataset of stimuli and imitations, which will allow us to develop understanding of vocal production for applications in vocally controlled music making systems (i.e. query by vocalisation for audio sample retrieval).

2. METHOD

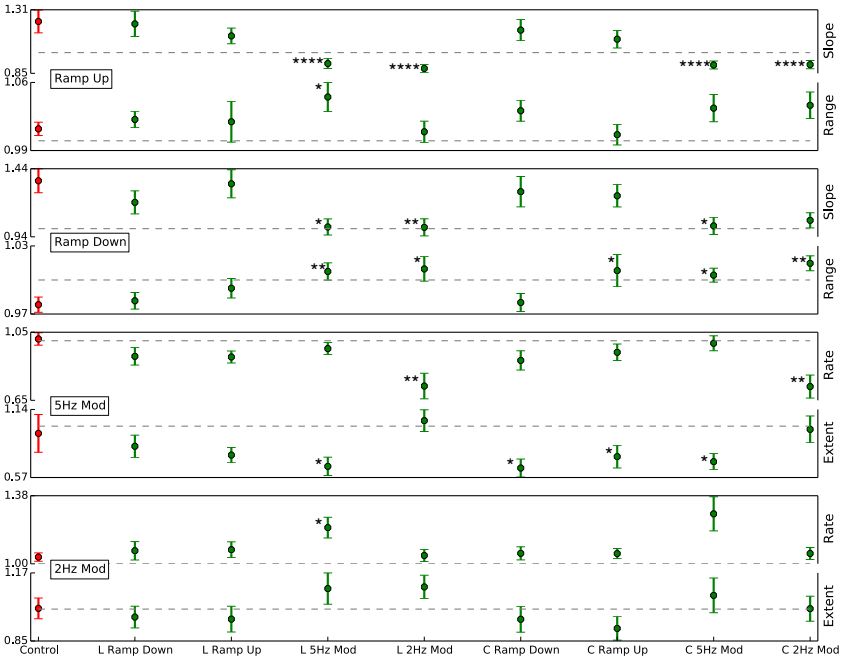
The 4 feature envelopes shown in Figure 1 were applied to create the stimuli. These were all synthesised using a basic subtractive model with a single sawtooth oscillator and three parameters for F0, gain and cutoff frequency of a resonant lowpass filter. 4 feature envelopes were applied to the three parameters, giving 12 control stimuli, each with a single feature envelope. A further 36 treatment stimuli were generated using all pairwise combinations of P-L and P-C feature envelopes. All the envelopes consist of 3 pieces with durations of 500ms, 1000ms and 500ms. This was to give the participants a clear origin and destination value for each of the envelope shapes.

Figure 1: Feature-envelope shapes used for the stimuli.

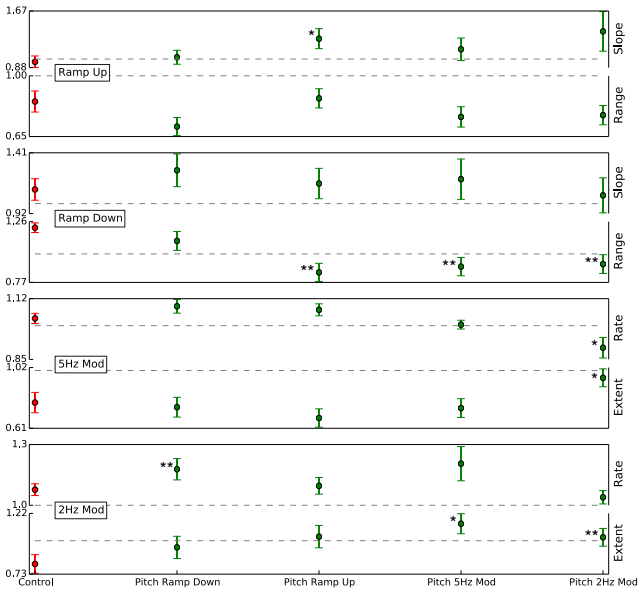
The study was conducted in an acoustically treated listening room. Participants were allowed to listen to each stimulus as many times as they wished before recording their imitation. 19 participants with musical training ($>$ 5 years) took the study, resulting in 836 imitations. Rate and extent parameters were extracted from the modulation imitations using a threshold based peak picking algorithm. Range and slope parameters were extracted from the ramp imitations using a continuous linear piecewise regression model with the following constraints: number of pieces = 3 ($p1, p2, p3$), slope of $p1$ and $p3 = 0$.

3. RESULTS

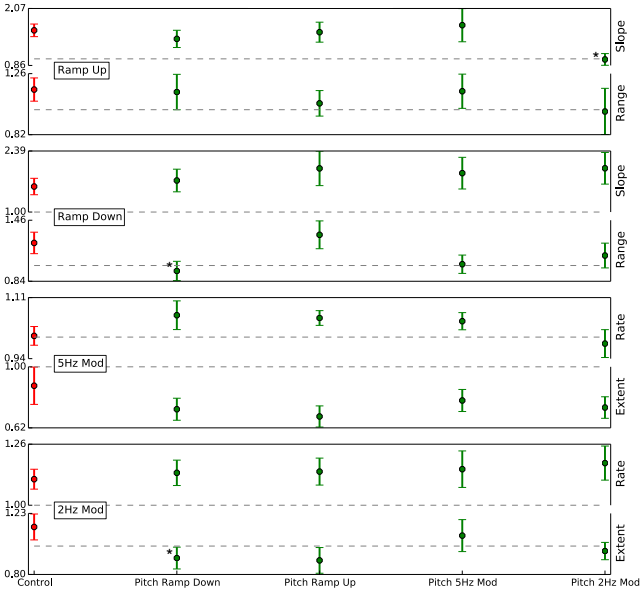
The resulting parameters are expressed as ratios of imitation:stimulus. Significance testing was performed using a Wilcoxon signed rank test. The results are show in Figure 2. P ramp scores for range were reasonably accurate (0.97-1.06), indicating that all participants were able to sing somewhat in tune. P ramp slopes were less accurate than range, due to duration errors in the ramp section of the envelope. This timing is significantly improved when the ramp is combined with modulation envelopes for other features, which may be due to the cycles acting as a time-keeping aid. Accuracy of P modulation rate is significantly lower when the envelope is combined with those for other features that have different rates (e.g. 5Hz P combined with 2Hz L). The opposite effect can be observed for extent (5Hz control). The range for L and C envelopes is generally more accurate than the slope, but considerably less accurate than the ranges for P. This may be due to the lack of a familiar interval based scale for these features. The slopes for L and C tend to be steeper than the stimuli, and the range for L is generally lower. For all features the extent is lower than the stimuli for the 5Hz envelopes, in contrast to the 2Hz envelopes.



(a) Pitch envelopes.



(b) Loudness envelopes.



(c) Spectral centroid envelopes

Figure 2: Each of the 12 controls (stimuli with 1 feature envelope) against the feature-envelope combinations. Parameters are expressed as mean and standard errors of the $imitation$: $stimulus$ ratio for all participants. Significance is indicated by ****, ***, ** and * for $p < 0.000$, $p < 0.001$, $p < 0.01$ and $p < 0.05$ respectively.

4. REFERENCES

[1] David Sanchez Blancas and Jordi Janer. Sound retrieval from voice imitation queries in collaborative databases. In *Proceedings of the Audio Engineering Society 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.

[2] Mark Cartwright and Bryan Pardo. Synthassist: Querying an audio synthesizer by vocal imitation. In *Proceedings of the conference on New Interfaces for Musical Expression*, London, 2014.

[3] Stefano Delle Monache, Stefano Baldan, Davide A Mauro, and Davide Rocchesso. A design exploration on the effectiveness of vocal imitations. In *Proceedings of the International Computer Music Conference*, Athens, Greece, 2014.

[4] Arnaud Dessein, Guillaume Lemaitre, et al. Free classification of vocal imitations of everyday sounds. In *Proceedings of the 6th Conference on Sound and Music Computing (SMC)*, Porto, Portugal, 2009.

[5] Guillaume Lemaitre and Davide Rocchesso. On the effectiveness of vocal imitations and verbal descriptions of sounds. *The Journal of the Acoustical Society of America*, 135(2):862–873, 2014.