Vocal imitation for query by vocalisation

Adib Mehrabi

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

School of Electronic Engineering and Computer Science
Queen Mary University of London
United Kingdom

April 2018

Statement of Originality

I, Adib Mehrabi, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged and my contribution indicated. Previously published material is also acknowledged herein.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third partys copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: A. Mehrabi
Date: 17/04/2018

For my wife and children

Abstract

The human voice presents a rich and powerful medium for expressing sonic ideas such as musical sounds. This capability extends beyond the sounds used in speech, evidenced for example in the art form of beatboxing, and recent studies highlighting the utility of vocal imitation for communicating sonic concepts. Meanwhile, the advance of digital audio has resulted in huge libraries of sounds at the disposal of music producers and sound designers. This presents a compelling search problem: with larger search spaces, the task of navigating sound libraries has become increasingly difficult. The versatility and expressive nature of the voice provides a seemingly ideal medium for querying sound libraries, raising the question of how well humans are able to vocally imitate musical sounds, and how we might use the voice as a tool for search. In this thesis we address these questions by investigating the ability of musicians to vocalise synthesised and percussive sounds, and evaluate the suitability of different audio features for predicting the perceptual similarity between vocal imitations and imitated sounds.

In the first experiment, musicians were tasked with imitating synthesised sounds with one or two time–varying feature envelopes applied. The results show that participants were able to imitate pitch, loudness, and spectral centroid features accurately, and that imitation accuracy was generally preserved when the imitated stimuli combined 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.

The second experiment consisted of two parts. In a vocal production task, musicians were asked to imitate drum sounds. Listeners were then asked to rate the similarity between the imitations and sounds from the same category (e.g. kick, snare etc.). The results show that drum sounds received the highest similarity ratings when rated against their imitations (as opposed to imitations of another sound), and overall more than half the imitated sounds were correctly identified with above chance accuracy from the imitations, although this varied considerably between drum categories.

The findings from the vocal imitation experiments highlight the capacity of musicians to vocally imitate musical sounds, and some limitations of non–verbal vocal expression. Finally, we investigated the performance of different

audio features as predictors of perceptual similarity between the imitations and imitated sounds from the second experiment. We show that features learned using convolutional auto–encoders outperform a number of popular heuristic features for this task, and that preservation of temporal information is more important than spectral resolution for differentiating between the vocal imitations and same–category drum sounds.

Acknowledgments

This thesis and the work presented herein would not have been possible without the academic, financial, and personal support of many people.
Firstly, I would like to thank my supervisors, Simon Dixon and Mark Sandler. Simon made this journey an enjoyable one, and kept me sane in times of doubt. His advice, enthusiasm, and support has been invaluable over the last 3 years. Mark provided encouragement and words of wisdom throughout my time in the Centre for Digital Music.
I would like to thank my wonderful colleagues and friends in QMUL, many of whom have supported me in one way or another, whether it was providing technical help, offering academic advice, or simply allowing me to chew their ear off with my many thoughts and ramblings! So, thank you Robert Jack, Alo Allik, Delia Fano, Olsen Wolf, Marcus Pearce, George Fazekas, Thomas Wilmering, Kuenwoo Choi, Florian Thalmann, Sebastian Ewert, Johan Pauwels, Dan Stowell, Mike Terrell, Elio Quinton, Giulio Moro, Chris Harte, Ben White, Mariano Mora, Dave Moffat, and everyone else in C4DM and MAT.
Thanks to all the participants who took in the studies, many of whom provided excellent feedback and thoughts about my research.
I would also like to thank my brother Zia, for the many long conversations about music, science, technology, academia, and putting the world to rights!
Lastly, I would like to thank my wife, Catherine, for supporting (i.e. tolerating) me throughout this journey. It's been tough at times, and this would have never happened without her support and encouragement.

This work was supported by the Engineering and Physical Sciences Research Council (EP/G03723X/1).

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Adam	Adaptive momentum estimation
AE	Auto–Encoder
AIC	Akaike's Information Criterion
ANOVA	ANalysis Of VAriance
CAE	Convolutional Auto–Encoder
CDBN	Convolutional Deep Belief Network
CI	Confidence Interval
CNN	Convolutional Neural Network
CQT	Constant Q Transform
dB	DeciBels
DNN	Deep Neural Network
FDR	False Discovery Rate
FFR	Fundamental Frequency Range
IPA	International Phonetic Alphabet
IQR	Inter–Quartile Range
LAT	Log Attack Time
LMER	Linear Mixed Effect Regression
LPC	Linear Predictive Coding
MDS	Multi–Dimensional Scaling
MFCCs	Mel Frequency Cepstral Coefficients
MIR	Music Information Retrieval
MRR	Mean Reciprocal Rank
MSE	Mean Squared Error
MUSHRA	MUltiple Stimuli with Hidden Reference and Anchor
NN	Neural Network
PCA	Principle Component Analysis
QBV	Query By Vocalisation
QBE	Query By Example
ReLU	Rectified Linear Unit

RMS	Root Mean Squared
SAE	Stacked Auto–Encoder
SC	Spectral Centroid
SFF	Speaking Fundamental Frequency
SGD	Stochastic Gradient Descent
SPL	Sound Pressure Level
ST	SemiTones
STFT	Short–Time Fourier Transform
SVM	Support Vector Machine
TC	Temporal Centroid

Chapter 1

Introduction

1.1 Motivation

A music producer is working on a piece of music. Most of the fundamental components are in place (rhythm section, harmonic structure and melody) but the producer is not entirely satisfied with some of the sounds. The electronic hi–hat lacks the acoustic nuances of a real instrument. The kick drum needs a long, slower decay. The move from verse 2 to the chorus could do with an exciting transitory sweeping sound. She has a good idea of the sounds she wants, and a large collection of audio samples. Yet despite this she is struggling to find the sounds she hears so clearly in her head.

This is a common scenario for musicians and producers. From the novice to seasoned professional, there are many situations where the search for a particular sound is hampered, obstructed and diverted as a consequence of the technology being used. Searching for audio samples, particularly drum sounds, is a core part of the electronic music making process, yet has been identified as a frustrating and time consuming task [Andersen and Grote, 2015], presenting a key area for future technological development. Despite a rapid evolution in the way electronic instruments and sound libraries are used to produce music, common approaches to searching for sounds remain elementary and limited. These typically involve browsing lists of badly labelled files, relying on file names such as 'big_kick' or 'hi-hat22'. Such methods for browsing sound libraries limit the users' ability to efficiently find the sounds they are looking for, and most notably, makes no use of the rich information

content available in the audio.

One major issue with sound search interfaces is enabling the intuitive expression of sonic ideas. The constituent parts of a particular sound may be specified as low level audio features related to semantic descriptors such as pitch, dynamics, timbre and timing. Expressing widely varying combinations of these components along with continuous temporal evolution requires an interface that offers a very high level of control with precision, accuracy and most importantly, ease of use. Audio based search systems should ideally encapsulate these qualities.

The voice is an attractive medium for this as it can be used to express timbral, tonal and dynamic temporal variations [Sundberg, 1989]. It is arguably one of the most versatile tools at our disposal for expressing sonic ideas. From India to Ghana, Cuba, and America, this extraordinary versatility is exploited to communicate, teach and perform non–verbal musical sounds [Atherton, 2007]. Most people have spent their entire lifetime developing the control of their vocal tract, and are able to exhibit a high degree of control for vocalising non–verbal sounds. For example, studies on non–verbal vocalisations have demonstrated the effectiveness of vocal imitations for describing and communicating sounds [Lemaitre and Rocchesso, 2014; Lemaitre et al., 2011], and highlighted the ability of musicians to accurately imitate sounds with respect to basic acoustic features [Lemaitre et al., 2016b], however to date there has been limited research focussed on the vocal imitation accuracy and imitability of musical sounds.

If we are to apply the voice for audio search purposes, i.e. query by vocalisation (QBV) for musical sounds, then we require knowledge of how accurately people can imitate the salient acoustic features in such sounds. If people are indeed able to accurately imitate these features, it validates the potential for using the voice for audio search, whilst furthering understanding of the capabilities of the voice. This application also requires understanding of the perceptual similarity between the domains of the voice and searchable sounds, in order to ascertain which audio features are best suited to the task of mapping between these domains. In this thesis we address these requirements by means of a series of vocal production and listening experiments, culminating in an analysis of heuristic and learned features for measuring similarity between vocal imitations and imitated sounds[1].

[1] the sounds being imitated (i.e. the 'target' sound), as opposed to the imitation

1.2 Research questions and approach

The aim of this thesis is to further understanding of how accurately people are able to vocalise musical sounds, in terms of both acoustic features and perceptual similarity between vocal imitations and imitated sounds. In addition, we aim to establish which audio features can be used to best represent the perceptual similarity between vocalisations and the musical sounds, in particular percussion sounds. Rather than adopting a hypothesis driven approach, we will address these objectives in an exploratory manner, in order to probe the relationship between vocal imitations and imitated sounds. Here we define the fundamental research questions, outline our approach to answering them, and clarify the scope of the work.

1. How accurately can people vocalise salient acoustic features in musical sounds?

    a) How accurately can people imitate sounds with time–varying acoustic features?

    b) What happens to this accuracy when people are asked to imitate two feature envelopes simultaneously?

    We address question 1a by conducting a vocal production experiment where participants were asked to vocally imitate sounds with controlled parameters for three important acoustic features: pitch, loudness, and spectral centroid. To answer question 1b we combine each of the sounds varying in pitch with each of those varying in spectral centroid and loudness. The acoustic features were extracted from the vocalisations and compared to those from the stimuli, providing a measure of feature–level accuracy for each of the vocal imitations.

2. Can people vocally imitate percussion sounds such that listeners can identify the sounds being imitated?

    This question is addressed by a two part experiment consisting of a vocal production task and a listening test. In the vocal production task, participants were asked to vocally imitate percussion sounds. In the listening test participants were asked to rate the perceptual similarity between the vocal imitations and percussion sounds. The percussion sounds were limited to five categories (kick drum, snare drum, cymbal,

hi–hat, and tom–tom), and the similarity ratings were limited to same–category sounds (e.g. the similarity between a vocal imitation of a kick and real kick drum sounds).

3. Which audio features best predict the perceptual similarity between vocalisations of sounds and actual sounds?

   We address this question by conducting an evaluation of audio features for predicting the similarity ratings from the listening test. Specifically, we compare a large number of heuristic audio features typically used for music information retrieval (MIR) tasks and the analysis of vocal imitations, and suitable subsets thereof, to features learned using convolutional neural networks (CNNs). In addition, we compare a range of CNN architectures and the resulting encoded layers (i.e. feature representations) in terms of size (number of features) and shape (temporal vs. spectral resolution).

1.3 Thesis structure

Chapter 2 introduces the existing research and themes upon which the work of this thesis is based. We first present the human vocal apparatus, in terms of its physiology and acoustic characteristics. This is followed by a discussion on the various modes of vocal expression: speech, singing, vocalisation of non–verbal, non–singing sounds, and how such sounds might be communicated by means of vocal imitation. We identify the problem space and common methods for addressing the issue of searching for sounds, followed by an examination of the literature specific to the research strands presented in each of the subsequent chapters. These include vocal control of acoustic parameters, vocalisation of percussion sounds, quantifying perceptual similarity between sounds, and audio features for the analysis of non–verbal, non–singing vocalisations.

Chapter 3 presents a vocal production experiment investigating the accuracy with which musicians can vocally imitate synthesised sounds in terms of 3 salient acoustic characteristics: pitch, loudness, and spectral centroid. In particular we explore the imitation accuracy of sounds with 1 or 2 features varying over time, and the effects of these features on imitation accuracy. Methods are proposed for quantifying the accuracy of modulation and ramp

envelopes, upon which the experimental results are based.

Chapter 4 investigates vocal imitation of percussion sounds in terms of the perceptual similarity between imitations and imitated sounds. This is explored by means of a 2 part experiment, consisting of a production and perception task. In the first part musicians were tasked with imitating 30 percussion sounds. In the second part listeners were asked to rate the similarity between the imitations and percussion sounds.

Chapter 5 explores the suitability of a number of audio features for predicting the similarity ratings from Chapter 4. We compare a large set of heuristic features and subsets of spectral and temporal features, along with descriptors used in the MPEG–7 standard for describing percussion sounds, and a spectrogram based method that has been shown to be highly correlated with the perceptual similarity between percussion sounds. In addition to the comparison between these features, we demonstrate the effectiveness of CNNs for learning features that represent the similarity between vocal imitations and percussion sounds, with networks trained on a dataset of vocal imitations, percussion sounds, synthesised sounds, and musical instruments. In a similar vain to the heuristic feature comparison, we explore the relative importance of temporal vs. spectral information for this task.

Chapter 6 concludes the preceding work, drawing on the potential impact of our findings for the application of query by vocalisation, and more generally, for the fields of non–verbal vocal analysis and vocal imitation research. We end this thesis by considering some potential paths for future research in these fields.

1.4 Contributions

The main contributions of this thesis are:

- Results of the vocal production task in Chapter 3, demonstrating the degree to which musicians can exercise simultaneous control over different acoustic characteristics of their voice.
- The dataset[2] of vocal imitations from Chapter 3, including the extracted audio features and annotated parameters for each of the feature envelopes.

[2] `https://zenodo.org/record/1215802`

- Results from the vocal production and perception tasks in Chapter 4, investigating the communicability of percussion sounds via vocal imitation. In doing so we determine which types of drum sounds are more imitable than others, and identify some of the imitation strategies that might contribute to this.

- The dataset[3] of vocal imitations, percussion sounds, and listener rating data from the experiments of Chapter 4.

- A comprehensive comparison of heuristic features for predicting the perceptual similarity between vocal imitations and percussion sounds.

- Application of convolutional neural networks for feature learning from percussion sounds and vocal imitations, and comparison of different network architectures that place emphasis on the detail of either spectral or temporal learned representations. This includes a novel evaluation of learned features using perceptual similarity measures.

The vocal production and listening experiments that were conducted as part of this work were approved by the research ethics committee of Queen Mary University of London, under reference numbers (QMREC) 1413, 1491, and 1717a. All participants gave written consent to take part in the experiments and were free to withdraw at any time.

1.5 Associated publications

Much of the work presented in this thesis has been published or is currently under review in the following publications.

[1] **A. Mehrabi**, S. Dixon, and M. Sandler, "Vocal imitation of synthesised sounds varying in pitch, loudness and spectral centroid.", *The Journal of the Acoustical Society of America*, 141(2):783–796, 2017.

[2] **A. Mehrabi**, S. Dixon, and M. Sandler, "Vocal imitation of percussion sounds: on the perceptual similarity between imitations and imitated sounds." (in review).

[3] **A. Mehrabi**, K. Choi, S. Dixon, and M. Sandler, "Similarity mea-

[3] `https://zenodo.org/record/804262`

sures for vocal–based drum sample retrieval using deep convolutional auto–encoders," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, Canada, 2018.

[4] **A. Mehrabi**, S. Dixon, and M. Sandler, "Towards a comprehensive dataset of vocal imitations of drum sounds," *Proceedings of the 2nd AES Workshop on Intelligent Music Production*, London, England, 2016.

[5] **A. Mehrabi**, S. Dixon, and M. Sandler, "Vocal imitation of pitch, spectral shape and loudness envelopes," *16th International Society for Music Information Retrieval Conference (Late Breaking Demo)*, Malaga, Spain, 2015.

An early version of the preliminary results from Chapter 3 was presented at the *International Society for Music Information Retrieval Conference* [5], which was subsequently extended, with full details of the method and further analysis, and published in the *Journal of the Acoustical Society of America* [1]. The method and experimental design from Chapter 4 was presented at the *AES Workshop on Intelligent Music Production* [4], and has been extended with a statistical analysis of the results for a full journal article (currently in review). Part of the content from Chapter 5 (focussing on the evaluation of convolutional neural networks for feature learning and predicting the perceptual similarity between vocalised imitations and percussion sounds) has been accepted for publication in the 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Simon Dixon contributed to each of the publications in a supervisory role by advising on the experimental design, commenting on the manuscripts, and providing helpful discussions on the work. Mark Sandler contributed to [1–5] by advising on the direction of the work, and commented on the manuscript in [1]. Kuenwoo Choi contributed to the work in [3] by advising on the design of the convolutional neural networks and commenting on the parts of the manuscript related to deep learning methods.

In addition, the article below was not directly related to the work in this thesis, but completed during the same period and influenced by the methods and findings herein.

[6] B. White, **A. Mehrabi**, and M. B. Sandler, "An archival echo: Recalling the public domain through real-time query by vocalisation," *Proceedings*

of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences, London, England, 2017.

Chapter 2

Background

The topic of this thesis spans multiple disciplines, including vocal production, audio analysis, audio perception, information retrieval, music production, and machine learning. In this chapter we discuss the related background research from each of these domains, and how this may inform the directions taken in the subsequent chapters. We first review the human voice in Section 2.1, in terms of the physical properties of the vocal apparatus, and how this can be used to control the acoustic output for both speech and musical expression such as singing and extended singing techniques. Beyond these domains, the application of the voice for communicating sonic ideas is discussed in Section 2.2. In Section 2.3 we set the scene for the application of the work in this thesis: query by vocalisation. In Section 2.4 we review the literature specific to the strands of research in each of the following chapters. This is divided into vocal control of key vocal acoustic characteristics in Section 2.4.1, vocalised percussion sounds in Section 2.4.2, measuring perceptual similarity between sounds in Section 2.4.3, and finally, audio features for representing vocalisations of sounds in Section 2.4.4.

2.1 The human voice

> It is difficult to overstate the importance of the human voice. Of all the members of the animal kingdom, we alone have the power of articulate speech [...] In addition, the human voice is our oldest musical instrument [Rossing et al., 2001, p. 335].

The human voice is capable of producing a vast array of sounds for communication (speech), music creation (singing), and emotional expression (laughter, crying etc.). The physical constraints of the vocal tract create a well-defined 'vocal sound space', which as humans we utilise, explore and push the limits of far more than any other vocal–producing animal. In this section we describe how sound is produced by the human voice, and what types of sounds are used for speech, singing, and vocal imitation of non–verbal sounds.

2.1.1 The vocal production system

A functional model of the vocal tract is given in Figure 2.1. The vocal production system can be considered as a source–filter model, where voiced sound is produced at the vocal folds before being filtered by the subsequent sections of the vocal tract. Voiced sound (or phonation) is created by vibration of the vocal folds according to the myoelastic–aerodynamic theory of phonation [Titze, 1980]. First, air is forced up through the vocal tract from the lungs. If the vocal folds are adducted (brought together from a relaxed position), when the air passes the folds, air flow on the outside of the pathway is forced to travel a longer path than that in the middle (where a direct path through the vocal folds exists). This difference in air flow between the outside and central point in the larynx creates a Bernoulli force, forcing the vocal folds together and closing the glottis [Sundberg, 1989, p. 12]. The pressure produced by the lungs then forces the vocal folds open again, and the process repeats until the muscles attached to the vocal folds (the laryngeal musculature) are no longer contracted. This cycle causes a periodic vibration at the oscillation rate (F_0, given in Hz), and voiced sound is created. The phonation frequency is largely determined by the laryngeal musculature, which controls the length, tension and mass of the vocal folds, and to a lesser extent the subglottal pressure from the lungs, which is primarily responsible for the phonation intensity [Sundberg, 1989, p. 16]. The F_0 range (FFR) of the voice varies between males and females, and is dependent on a number of factors including body size and the laryngeal musculature, however for adults this typically spans up to 3 octaves [Kent et al., 1987].

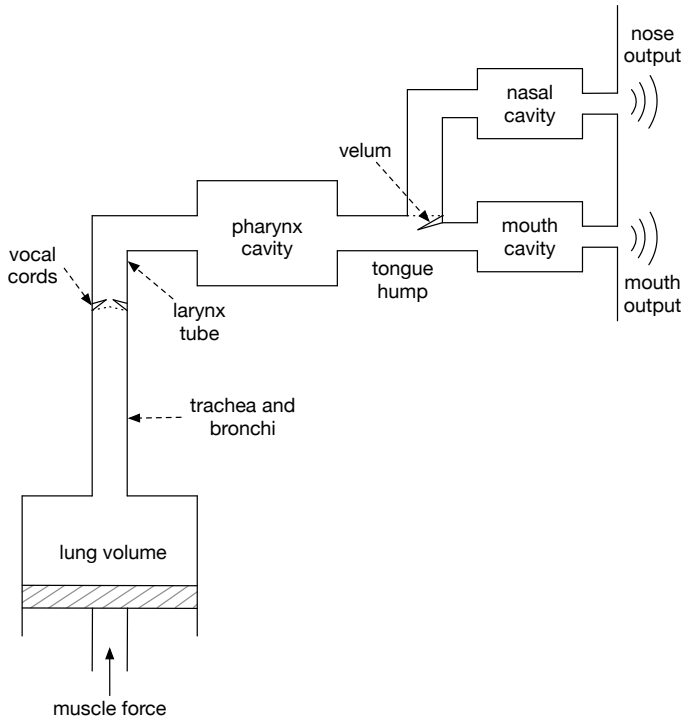


Figure 2.1: Functional components of the vocal tract, after Flanagan [1965].

2.1.2 Voice quality

Beyond the frequency and intensity of phonation, it may also differ in terms of voice quality, for example the difference between normal speech and a whisper. Voice quality can be described in terms of *phonation modes*, which are determined by the amplitude of vocal fold vibration, i.e. glottal adduction [Sundberg, 1994a], and the way in which the folds vibrate. Interestingly, phonation modes are typically described in categorical terms, some of which we lay out below, but as noted by Ladefoged [1971] they actually represent a continuum, from voiceless, through varying degrees of voicing, to full glottal closure. There is some ambiguity in the literature on phonation modes, both on the number of categories and how they are defined. For our purposes we are not so concerned with the division of categories or their boundaries, but rather the general space of phonation modes and the types of sounds that can be made by manipulating the vocal fold apparatus.

Modal voice (sometimes referred to as 'normal' voice) is the most common phonation mode in speech and singing. It occurs when the vocal fold vibration is regular and periodic, with no fricative airflow through

the vocal folds [Hewlett and Beck, 2006, p. 274].

Breathy voice is created when the vocal folds are tense but part of the glottis (i.e. the space between the vocal folds) remains slightly open during phonation, allowing air to flow past. This airflow causes turbulence, generating a fricative noise, or a breathy sound, such as is heard in the /h/[1] in *aha* [Ladefoged, 1993, p. 139]. This is closely related to **whisper voice**, which occurs when the vocal folds are held rigid such that they are not able to vibrate, yet with a small opening in the glottis that allows air to flow through. As a consequence, whisper voice occurs without voicing, whereas breathy voice tends to occur as an augmentation of modal voice [Hewlett and Beck, 2006, p. 279].

Creaky voice occurs the vocal folds are constricted but sub glottal pressure is low, resulting in an irregular F_0. This creates a creak that is audible as discrete events of the glottal opening and closing, typically at low F_0 values of 7Hz–78Hz [Gerratt and Kreiman, 2001], with sub–harmonics at half F_0. Whilst people rarely speak whole phrases in creaky voice, in English speech it is often used at the end of words or phrases with falling intonation contours [p .277 Hewlett and Beck, 2006; Ladefoged, 1993, p. 141].

Falsetto (or head voice) is created by stretching the vocal folds such that they are thinner than with the modal voice, giving rise to higher F_0 values [Hewlett and Beck, 2006, p. 274]. It is more commonly used in singing than speech, to achieve higher F_0 values than are possible using the modal voice [Sundberg, 1989, p.50], although it can also be used to imitate female or young characters, or to voice non–verbal expressions of emotion.

Harsh voice is caused by irregular vibration of the vocal folds in terms of either frequency or amplitude, often as a result of adduction of the ventricular folds (see below), and typically occurs as a modification of modal or falsetto voice [Hewlett and Beck, 2006, p. 278].

Ventricular fold vibration. The ventricular folds (also called the false vocal folds) are small ligament structures situated just above the vocal folds. They can be made to vibrate harmonically with the vocal folds,

[1] throughout this thesis phonemes will be notated in line with the International Phonetic Alphabet (IPA)

typically creating overtones at double F_0. This effect is rarely used in speech, but exists as a core part of certain singing styles, such as Asian throat singing and Mediterranean folk polyphony [Bailly et al., 2010].

As mentioned earlier, both the acoustic and physiological space of phonation modes is actually a continuum, and as such this list is by no means exhaustive. For example, there are several types of creaky voice, each characterised by both acoustic and physiological markers [Keating et al., 2015], and the *National Center for Voice and Speech*[2] identifies no less than 25 categories of voice quality. Nonetheless, these categories highlight the wide range of sounds possible in this space, which can be produced by relatively minor adjustments of the vocal folds. Phonation modes are used in English speech to convey emotion or mood [Gobl et al., 2003], but in general they are not used for linguistic contrasts (other than the voiced/voiceless contrast). However, many other languages make use of modes such as creaky and breathy to distinguish between otherwise similar vowels (see Gordon and Ladefoged [2001] for an excellent review on this topic). Furthermore, as noted by Stowell [2010], vocal artists such beat boxers make use of phonation modes to imitate different types of musical sounds, for example employing falsetto for sound effects or ventricular voice for bass lines.

2.1.3 Articulation

After being generated in the glottis, voiced sound is filtered by the articulatory components of the vocal tract (all components above the vocal folds in Figure 2.1), which can be considered as a resonating tube. By tuning this tube–like structure, a wide range of sounds can be sculpted out of the voiced sound, as is required to produce vowel phonemes in speech. Although all parts of the vocal tract can be somewhat tuned, the oral cavity is arguably the most important for producing different voiced phonemes [Rossing et al., 2001, p. 342]. Varying the configuration of the oral cavity gives rise to different resonant frequencies, called formants, which exist as peaks in the spectrum of voiced sounds. The configuration of the oral cavity and lips may remain constant, to produce constant vowel sounds (monophthongs), or varied over time to produce vowels which shift between two or three vowels (diphthongs and triphthongs). Formants are numbered from F1 – F5, where typically F1 – F3 contribute most to differentiating between the voiced phonemes in speech. Typical ranges for

[2] http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/quality.html

F1, F2, and F3 are between 250Hz – 1000Hz, 600Hz – 2500Hz, and 1700Hz – 3500Hz respectively [Sundberg, 1989, p. 23]. Finally, the radiation efficiency is controlled by the lips, where higher frequencies are radiated most efficiently (as the wavelength is more aligned with the size of the mouth opening).

In addition to filtering voiced sounds, the oral cavity (most notably the tongue) and lips can be manipulated to produce a wide range of consonant sounds, such as clicks, stops, trills, and fricatives. In phonetics, the articulations required to produce these sounds are categorised based on the place and manner of articulation (i.e. the position and motion of the tongue in relation to the oral cavity and lips), the source of airflow (from the lungs [pulmonic] or from the glottis [non–pulmonic]), the degree of stricture (airflow restriction), and the path of airflow (through the centre of the oral cavity [central] or around the sides of a blocking articulator [lateral]). The manners of articulation are typically divided into the below mentioned categories [Ladefoged, 1993, p. 172], however, Ladefoged notes that there are further considerations required to describe consonants fully, including the primary place of articulation and secondary articulations such as rounding of the lips.

Stops: these occur when a part of the vocal tract is fully blocked as a result of the complete closure of articulators. This blockage can be followed by the release of pressure either outwardly (i.e. ejectives) or inwardly (i.e. implosives). They include glottal stops (closing of the glottis), oral stops (closing of the lips or the oral cavity by bringing the tongue against the roof of the mouth), and nasal stops (where the oral cavity is blocked and air is released through the nasal cavity).

Fricatives: where two articulators are brought together, causing partial restriction of airflow and therefore turbulence.

Approximants: where two articulators are brought together, but not close enough to cause turbulence.

Trills, taps and flaps: trills occur when an articulator vibrates periodically, such as rolling of the tongue, whereas taps and flaps occur when the tongue strikes the roof of the mouth or another part of the oral cavity (i.e. a single cycle of a trill).

This section serves to illustrate the core mechanisms behind how the voice can be used to produce speech, and highlight the vast array of sounds that

might be created. Much of this literature is based on research on speech, which is relevant to our work because it has been shown that the same mechanisms used to produce speech can be employed to vocalise non–verbal sounds such as percussion [Proctor et al., 2013]. Beyond speech, phonation and articulation can be viewed as a multi–dimensional continuum in which the space between categories may be explored to vocalise sounds. This is important to bear in mind when asking people to vocalise non–verbal sounds:

2.1.4 Singing vs. speech

Many of the aforementioned characteristics of speech also apply to singing. We are not particularly concerned with differentiating between speech, singing and vocal imitation of non–speech, non–singing sounds, however, in order to imitate musical sounds we expect people to employ techniques from these two predominant applications of the voice. As such there are some notable characteristics specific to the singing voice that do not necessarily apply to speech, and are worthy of mention here.

The most notable differences between speech and singing are in the use of pitch range and vowel duration. In tonal languages pitch is used to determine lexical meaning, and in many non–tonal languages pitch is continuously varied to convey linguistic information (such as a rising pitch to indicate a question), and non–linguistic information such as the speaker's emotional state. The production of music and melody often necessitates that singers make full use of their vocable range (this is discussed further in the following paragraphs), whereas the pitch range of spoken English is generally much lower, at less than an octave [Andreeva et al., 2014]. In addition, professional (particularly opera) singers tend to produce vibrato: a periodic modulation of F_0, which is not normally used in speech [Sundberg, 1989, pp. 163–176]. Singing also typically requires that vowels are somewhat aligned to the rhythmic pattern of the music, and therefore the vowel duration can be extended beyond that used in speech [Reigado and Rodrigues, 2017]. Perhaps the most interesting and less obvious difference between speech and singing is that professional singers (especially male) tend to modify the formant frequencies of certain vowels, in particular F3, F4 and F5 to be in the range of 2.5kHz–3kHz, in order to add brilliance and power to the voice, a phenomenon known as the 'singer's formant' [Sundberg, 1974].

2.1.5 Singing voice quality

We may also view voice quality from a singing perspective, in terms of vocal registers, singing phonation modes and vocal timbre. These are all fuzzy terms in the literature, in that they are generally ill–defined. However, collectively they serve to quantify the variety of ways the voice might sound, beyond simply describing sung notes in terms of pitch and loudness.

Vocal register is closely related to phonation mode, which we discussed in 2.1.2, and indeed there is some overlap in terminology for certain modes, namely modal and falsetto. However, although the concept of phonation modes is generally well understood and agreed upon, there is a considerable lack of clarity on the definition of vocal register [Sundberg, 1989, pp. 49–50]. Hollien [1974] defines vocal register as *" a series or range of consecutive vocal frequencies which can be produced with nearly identical voice quality ... [with] little or no overlap in fundamental frequency (F_0) between adjacent registers."*. The pioneering work of Garcia [1854] in the 19th century investigated vocal register in terms of mechanical principles, as opposed to perceptual acoustic voice qualities. Using elementary instruments such as mirrors placed in the throat of singers, Garcia identified 3 vocal registers, increasing in F_0: *chest*, *falsetto* and *head*, which he differentiated by musculatory configurations of the larynx. These two approaches to understanding registers are somewhat at odds, because changes in musculatory configurations do not necessarily lead to timbral variations [Henrich, 2006]. Indeed, professional singers are often trained to reduce the acoustic effect of a change in the laryngeal musculature, in order to eliminate timbral variation between registers and increase the F_0 range in which their voice timbre remains constant [Sundberg, 1989, p. 51]. To quote Henrich [2006]:

> If a vocal register is defined as a series of consecutive tones produced by the same laryngeal mechanism, the human voice can be characterized by four different voice registers, i.e., laryngeal registers such as Hollien's pulse, modal, loft registers, and the whistle register. If a vocal register is defined as series of consecutive tones produced with similar voice quality, other registers exist in the human singing voice, such as head, belting, middle, upper, voix mixte,...

Vocal registers can be described in terms of where the sound appears to

radiate from or resonate (chest, middle, head, upper) the nature of the produced sound (pulse, falsetto), or even musical instruments (bell, flute, whistle, flageolet). Perhaps the most unambiguous labelling form for labelling registers is the number system described by Henrich [2006], where labels M0–M3 define four laryngeal mechanisms (vocal fold configurations) that are associated with increasing phonation frequency. M0 produces very low registers such as pulse, vocal fry and strohbass, where the voicing is distinguishable by a series of short pulses (similar to the creaky voice phonation mode discussed in Section 2.1.2). M1 is the normal laryngeal mechanism, used for producing modal, chest, male head and belting registers. M2 is used for registers that occur towards the mid–high part of the vocal range, such as falsetto, loft, female head, and upper. Finally, M3 is the laryngeal mechanisms in which the highest possible F_0 values can be produced, in the bell, whistle, flute, and flageolet registers. M3 is not commonly used in classical singing or speech. In summary, the abundance of terms and categories used to define registers highlight a common thread that is relevant to the studies in this thesis, namely that we cannot expect people to produce sounds across their entire phonation range without there being noticeable changes in the vocal fold configuration, and consequently the timbral characteristics of the produced sound.

Another measure of voice quality is singing phonation mode. Confusingly, this is conceptually similar to the phonation modes commonly referred to by phoneticians, however the categories are primarily based on two parameters: sub–glottal pressure and glottal airflow [Sundberg, 1989, p. 80]. Sundberg defines four singing modes within this 2D space: neutral, pressed, breathy, and flow, where flow and breathy have higher glottal airflow than pressed and neutral, and flow and pressed have higher sub–glottal pressure than neutral and breathy. In terms of usage, Proutskova et al. [2013] note that pressed and breathy voice are used in some popular music styles for affect, whereas some styles and performers try to only use a single mode (such as operatic baritone singers using mostly flow), and some styles (such as classical Ottoman singing) make use of all four modes. Acoustically, changing phonation mode changes the spectrum of the sung sounds; for example increasing glottal airflow can substantially increase the amplitude of F_0 (without affecting the formant amplitudes) [Sundberg, 1989, p. 80], and with breathy mode there will be considerable fricative noise introduced as a result of the airflow through a small opening in the glottis during phonation. Finally, voice quality might be described in terms of timbre. Timbre is a poorly defined concept in the

singing literature, where it can be interchangeably used to describe phonation mode, registers, singer sex, and age [Sundberg, 1989].

2.1.6 Extended singing techniques

Much of the above mentioned research into singing voice has focussed on Western classical and popular music. We note that beyond these singing styles a much broader range of vocal techniques exist, some of which may bring the definition of singing voice into question. We will now discuss some notable examples of such practices.

Firstly, in terms of pitch, we note that an equal tempered 12 tone scale is not always used. For example, in classical Indian singing the 12 tone scale is often extended as a type of ornamentation technique [Krishnaswamy, 2003]. In this case singers may voice inflexions above or below the standard 12 intervals in the musical scale, exploring the space in between the standard intervals. Another use of pitch not typically found in Western music styles is in yodelling, where rapid and large changes in pitch are accentuated by concurrent changes in register [Echternach and Richter, 2010]. This results in distinct timbral changes within 'melodies' with large pitch intervals, where a note change may not only differ in pitch, but also voice quality. A particularly unconventional use of pitch comes from Tuvan throat singing, particularly the Mongolian Kargyraa singing. It is thought that this singing style originates from the desire to imitate natural phenomenon such as animal calls or sounds echoing off a cliff [Levin and Edgerton, 1999]. Singers of this style are famous for their ability to create vocalisations that give the impression of two sources with different F_0 values, an effect that is achieved by forcing the ventricular folds to oscillate at a different F_0 to the vocal folds [Bailly et al., 2010; Levin and Edgerton, 1999].

Beyond pitch, vocal styles and music genres can be (at least in part) defined by voice quality. For example, different voice qualities have been observed across music styles (pop, rock, soul, Swedish dance band), predominantly exhibited as an effect of spectrum shift by slight adjustment of the F1–F4 formant frequencies [Borch and Sundberg, 2011]. At the extreme end, these differences are apparent when comparing the typically pure, clean tone of vocal performance in western classical music with that of punk and rock styles for example, where growls, grunts and screams are vocalised [Jahn, 2013, p. 355].

These vocalisations are typically achieved by a tightening of the larynx, restriction of the glottal airflow and vibration of the supra-glottal structure [Borch et al., 2004], creating a noisy sound with many inharmonic partials [Kato and Ito, 2013; Tsai et al., 2010]. In addition to musical applications of the voice, the Irish mourning vocal practices of keening and other cultural funerary practices employ vocal techniques such as voiced inhalation, creaky voice, exhalatory gasps and amplitude–frequency–modulation to generate a series of vocal cries and weeps that have a voice quality not often heard in music [Harvey, 1993].

Many music–related vocal practices do not use words (as with the grunting and screaming discussed above), and instead use nonsensical vocalisations, relying solely on the pitch, rhythm and voice quality for musical effect. Such practices have been observed in African pygmy and bushmen vocal performances, which use vocal effects including hoots, calls, shouts, grunts and screams [Frisbie, 1971]. Voice artists including Phil Minton[3] and Mikhail Karikis[4] also make use of extended vocal techniques in musical performance, including voiced inhalations and exhalations, groans, grunts, and heaves. This kind of non–verbal voiced inhalation and exhalation is also used in the vocal practice of *eefing*, that originated in Tennessee around 100 years ago [Sharpe, 2006], however in eefing the inhalation and exhalation is performed in a rapid, rhythmic manner, typically as an accompaniment to music or 'hand slapping' percussion.

Although it is beyond the scope of this thesis, we note that the singing voice may also be used as a controller for synthesised sound, either of the voice or other musical instruments. Such applications include parameter control [Cartwright and Pardo, 2014; Janer, 2008; Loscos and Aussenac, 2005; Santacruz et al., 2016], vocal morphing [Young, 2014] and mapping vocal timbre to percussion samples [Stowell, 2010]. Typically these methods work by extracting audio features from the voice and mapping them to parameters on a synthesiser. The types of audio features used vary across studies, however they generally include pitch (e.g. F_0) and timbre based (e.g. spectral centroid) features [Cartwright and Pardo, 2014; Janer, 2008; Stowell, 2010], along with amplitude and markers for the start and end of notes. The work of Young [2014] explores extreme transformations of the voice using granular synthesis, time–stretching and filtering methods. In addition to research on this topic,

[3] `https://youtu.be/yXg3sr16Xww`

[4] `https://youtu.be/qND-XvZvOJo`

commercial products such as *Vochlea*[5] and *The Mouth*[6] exist, both of which enable vocal control of synthesiser parameters, sound effects, and triggering of drum samples. These examples present interesting and novel applications of 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.

2.1.7 Summary

In this section we have deliberately avoided addressing vocalised percussion sounds such as beatboxing and vocal imitation of musical instruments in the north Indian tradition of vocally imitating percussion (*tabla bols* and *konnakol*), because we will discuss these topics in Section 2.4.2. What the research addressed in this section highlights most is that humans, particularly singers (or more generally, vocal performers), are able to apply a range of vocal techniques that involve sculpting and manipulation of the vocal tract, to generate a remarkable array of sounds. This illustrates the multi–dimensional nature of voice quality, which clearly stretches far beyond simply pitch and dynamics, to include elements of noisiness, inharmonicity, formant structure, and spectral slope to identify but a few. Vocal performers might change their timbre using any combination of the techniques discussed here, in order to imitate specific characteristics of sound.

2.2 Vocal imitation and communication of sonic ideas

Vocal imitation can be described, at least from a cognitive perspective, as "*vocal re–enactment of previously experienced auditory events*" [Mercado III et al., 2014]. This plays an important role in early language development, for example when learning how to pronounce vowel sounds [Kuhl and Meltzoff, 1996] and prosodic contours [Gratier and Devouche, 2011]. Indeed, the faculty of language would not exist without imitation of spoken language. The way in which humans have utilised the voice for speech is remarkable, yet perhaps even more compelling is that spoken language represents only a small subset of sounds that humans are able to vocalise. Verbally describing sounds can be difficult, particularly if the source of the sound is unknown or when trying

[5] `http://www.vochlea.co.uk/`
[6] `https://youtu.be/zzyr66QhrOw`

to describe the differences between two sounds from similar sources. When suitable words or onomatopoeia do not exist, the expressive nature of the voice can be harnessed to effectively communicate sonic concepts. Recent studies have shown that in such cases vocal imitations can be more effective than verbal descriptions for conveying a sonic concept and identifying the source of a sound [Lemaitre and Rocchesso, 2014; Lemaitre et al., 2011], and that when imitating environmental sounds people make use of articulatory mechanisms that do not exist in their native spoken language [Helgason, 2014; Helgason et al., 2016]. In this section we will briefly review the related literature on the role of vocal imitation for communicating sonic ideas. This research is generally concerned with imitations of everyday and environmental sounds, however many of the constraints and mechanisms behind producing them are applicable to vocalisations of musical, synthesised, and percussion sounds.

Lemaitre et al. [2011] investigated whether listeners were able to identify the imitated sound from an imitation, and the acoustic correlates that allowed listeners to do so (we discuss the latter objective in Section 2.4.4). The imitations were all of kitchen sounds, categorised according to their source (electrical appliances, solids, gases, and liquids), and listeners were asked to sort the imitations to groups on the basis of what was being imitated. The results showed that most of the imitations were grouped into clusters that coincided with their source categories (i.e. the sound concept was successfully communicated via the imitations), except for the liquid based sounds. The authors suggest that this may have been due to either different classification strategies adopted by the listeners, or that the imitators were simply not able to suitably imitate the liquid sounds (characterised by short chirps) due to physical limitations of the vocal apparatus. In a subsequent study, Lemaitre and Rocchesso [2014] compared vocal imitations to verbal descriptions of mechanical and synthesised sounds, in terms of whether listeners could identify the cause of the sounds from the respective representations. They found that when the source of the sound was easily identifiable, vocal imitations were equally as effective as verbal descriptions for this task, yet when the source was not clear or unidentifiable (such as for artificial synthesised sounds), the vocal imitations were much more effective for communicating the referent sound. However, as with the previous study, not all the referent sounds were equally imitable. In particular, the sound of coins falling on a plate was not recognized from the imitations, due to the difficulty of vocalising such a high density of rapid (sometimes overlapping) short events, and the (inaccurate) imitations were

often confused with sounds that contained a more similar temporal profile. The same effect was not observed for imitations that were inaccurate in terms of spectral distribution, indicating that temporal information may be more important than spectral for communicating a sonic concept.

In a similar vain to the aforementioned studies, Edmiston et al. [2017] collected vocal imitations of environmental sounds from 4 categories (zipper, water, tear, glass), and asked listeners to identify the imitated sound from the imitations. However, they also collected imitations of imitations, in a form similar to the 'Telephone' style game, and tested for the effect of imitation 'generation' (i.e. how many imitations from the original sound) on identification accuracy. Interestingly, they found that although identification accuracy started with above chance accuracy for the first generation of imitations, it decreased almost linearly with subsequent generations. This is somewhat surprising given that we might expect a mismatch between a real sound and a vocal imitation due to the physical limitations of the vocal apparatus, however this mismatch should be smaller for imitations of imitations, where the sound sources are the same. The fact that the imitations became less and less identifiable indicates that each imitator focussed their attention on different salient characteristics of the sound to be imitated (whether the original sound or an imitation thereof). This highlights that the communicability of sound sources via vocal imitations may depend as much on the imitator and listener identifying the same salient characteristics for a given sound than on the ability of humans to imitate non–vocal sounds.

Instead of focussing on the sound source, Perlman and Lupyan [2017] investigated the identifiability of vocalisations in terms of 'iconicity', i.e. how iconic vocal imitations are of a given *meaning*. The vocalisations were produced in response to 30 meaning terms from 3 categories: actions (e.g. eat, gather, sleep); nouns (e.g. child, rock, deer); and properties (e.g. big, this, many). Instead of identifying the imitated sound (or source) from an imitation, listeners were presented with a vocalisation and asked them to select the referent meaning from a set of (within or between category) labels. The results show that on average for each condition (within and between category), the meanings were correctly identified with above chance accuracy for all 3 categories. Vocalisations of the action labels were most frequently correctly identified, followed by the nouns, and properties. The authors note imitations of 'tiger', 'eat', and 'many' afforded a high iconicity, whereas properties such as 'this' and 'that' did not. This highlights the strength of verbal communication as a

system for labelling abstract concepts that are not easily represented sonically. These properties can be easily differentiated using words, and even visually and physically using gestures, but it is hard to imagine what sounds, vocal or otherwise, might best represent them. Finally, we note very high variance in identification accuracy across imitators, and that the best performing imitators by this measure were affiliated with academic institutions in linguistics or a related field. This indicates that the ability to imitate sounds effectively may be subject to training and experience in phonetics (it is common on many linguistics–based academic programs to practice phoneme–level vocalisation and analysis).

The work discussed in this section demonstrates that beyond speech, vocal imitation of non–verbal sounds presents a powerful and effective means to communicate sonic ideas. This effectiveness appears to be dependent on a number of factors, namely *i*) whether the sounds, in particular the temporal characteristics are able to be sufficiently produced with the vocal apparatus (this appears to be less important for spectral characteristics), *ii*) the degree of training with respect to linguistics and phonetics, although this may extend to musical training or other means of developing critical listening skills, and *iii*) concordance between imitator and listener regarding the salient characteristics of the sound or concept. In Chapter 4 we investigate whether this effectiveness applies equally to vocal imitations of percussion sounds as it does to the everyday sounds discussed here, and explore the potential for applying vocal imitation to the problem of searching for sounds.

2.3 The problem of searching for sounds

Searching for sounds is a fundamental part of the music production process, particularly for electronic music production where numerous short samples of sounds can be arranged to form a piece of music. In this section we briefly discuss how sounds are typically searched in this domain (text–based search), following by a review of content–aware search methods. We note that our focus is on searching for sounds, as opposed to pieces of music or sequences. The basic idea behind content–aware search is to identify, rank, or group similar sounds in a library, based on some model of similarity between sounds. This presents an enticing alternative to the text–based method, and the application of this thesis – *query by vocalisation* (QBV) – falls into this category of search

methods.

The traditional, and arguably still the most common method of searching for sounds is using textual descriptors. Such descriptors may exist in the form of file names, or meta–data tags associated with sound files. This can be an effective search method if the file names are relevant, meta–data are of a high quality and the user possesses expert knowledge of the sound library. This may be the case for many professional sound libraries; however for crowd–sourced collections such as *freesound*[7] the file names may not best describe the sound file, and meta–data are often inconsistent, if any exists at all. In addition, typically music producers will have collected multiple sound libraries, each with different labelling schema. This makes it extremely difficult to memorise what types of tags are used in each library and where all of the sounds exists in a file system.

Instead of searching through alphabetical lists of sounds, sound libraries may be navigated in a more exploratory manner. One common such approach is to map sounds onto a navigable low dimensional space, where similar sounds are located close to one another. Coleman [2007] proposed such a system for navigating all short segments in a music library, using spectral and temporal audio features mapped onto a 2D space. This concept has since been applied to navigating collections of short audio samples that might be used in music production [Font and Bandiera, 2017; Fried et al., 2014; Heise et al., 2009; Turquois et al., 2016]. The feature–space representations are typically based on acoustic descriptors of timbre such as Mel frequency cepstral coefficients (MFCCs) – as are all four aforementioned systems – however, any audio descriptors may be used: for example, Font and Bandiera [2017] present a system where the user is able to choose between a timbre–based search using MFCCs or a tonality based search using harmonic and pitch related features. The dimensionality of these feature–spaces can be reduced using techniques such as Student-t Stochastic Neighbour Embedding (as used by Turquois et al. [2016] and Font and Bandiera [2017]), Self-Organising Maps (as used by Heise et al. [2009]) and kernalised sorting (as used by Fried et al. [2014]). This approach to search lends itself well to free exploration of sound libraries, however often it is not clear how well the measurement of similarity between sounds relates to the users' understanding of similarity. Despite research highlighting the perceptual relevance of MFCCs to perceived timbral similarity [Terasawa et al., 2005], evaluations of these types of systems tend to focus on the user

[7] `freesound.org`

experience of exploring the 2D space, and not on the perceptual merit of the similarity measures.

An alternative search paradigm that is more relevant to the topic of this thesis is *query by example* (QBE) – indeed, as we will discuss in Section 2.4.4, QBV is a special case of QBE. QBE allows a dataset to be queried using an example of the item one wishes to retrieve [Wold et al., 1996]. For example, a sample of an electric guitar may be used to search for all electric guitar sounds in a sample library. Applying QBE to audio requires a means of measuring similarity between example sounds and the sounds in the dataset to be queried. As noted by Mitrović et al. [2010] and Slaney [2011], measuring the similarity between sounds is an ill–posed problem, in that typically the aim is to estimate similarity or class labels using model parameters such as acoustic features and distance measures. Given the lack of a universal model for human perception of sound (indeed, the concept of what constitutes the perceptual attribute of 'timbre' is still poorly understood [Siedenburg and McAdams, 2017]), no unique solution to this problem exists. Nonetheless, the concept of sound similarity is central to many audio–based music information retrieval (MIR) tasks: ultimately, in such tasks one will want to measure the similarity between sounds at some level and in terms specific to the problem. This is highlighted in that a keyword search on the proceedings of the International Society for Music Information Retrieval Conference (ISMIR)[8] from 2000–2017 reveals 130 articles that include the term 'similarity' in the title. Feature–based measures such as those described in this section use audio features extracted from the query sound, which can then be compared to the audio features of the sounds in the dataset, with distance between the sounds w.r.t. the descriptors taken as a measure of similarity [Aucouturier and Pachet, 2002, 2008; Herrera et al., 2003]. The results can then be presented as a ranked list of all sounds, or a list of sounds in the same class as the example. The two core questions when designing such systems are 1) which audio features to use, and 2) how to compare sounds based on the audio features. These questions are not trivial to answer, and have been the focus of much of the research into MIR.

The suitability of different features and similarity metrics will depend on the types of sounds in the sample library. A common and relatively basic approach is to use the Euclidean distance between MFCCs of the query example and each of the sounds in the library. As well as being used for mapping sounds on the 2D spaces described above, this method has also been applied

[8] http://www.ismir.net/proceedings/

to querying short segments of music [Spevak and Favreau, 2002]. Alternative approaches have been applied to classifying example sounds as either music instruments, speech, noise, and environmental sounds using the Euclidean distance between Gaussian mixture models of MFCCs and spectral descriptors [Helén and Virtanen, 2007], or applying hidden Markov models to model sounds based on raw spectrograms [Casey, 2001] and MFCCs [Helén and Lahti, 2006; Wichern et al., 2007]. Although much of this work has focussed on spectral descriptors, possibly including local differences between frames, one cannot ignore the importance of the global temporal evolution of sounds. Esling and Agon [2013] and Parekh et al. [2016] investigated similarity measures for QBE based on the morphological profile of spectral and loudness features respectively. However, we note that this type of approach is most fruitful when the sounds to be queried are expected to exhibit notably different morphological profiles such as rising, falling, impulsive [Parekh et al., 2016], because morphological features struggle to discriminate between sounds with similar profiles [Esling and Agon, 2013]. Finally, we note that measuring the similarity between a query and all sounds in a library can be computationally expensive, and pre–processing or clustering of the sounds in a library can greatly reduce the search time [Helén and Lahti, 2007; Xue et al., 2008; Zhang and Kuo, 1999].

As we have seen in this section, QBE and other audio search systems typically rely on similarity measures derived from audio features. However, we might also consider the *perceptual* similarity between sounds, which can be derived from similarity ratings provided by listeners [International Telecommunication Union, 2003; Scavone et al., 2001; Wickelmaier et al., 2009] (we provide an overview of methods for quantifying perceptual similarity in Section 2.4.3). There is considerable overlap between feature–based and rating–based similarity measures: the former can be evaluated using perceptual listening tests (see, for example Terasawa et al. [2005] and Pampalk et al. [2008]), and conversely acoustic descriptors are often investigated to explain perceptual ratings [Berenzweig et al., 2004; Elliott et al., 2013; Freed, 1990; Grey, 1977; Gygi et al., 2007; McAdams et al., 1995]. As such, the two paradigms are not independent and we might argue that any measure of similarity for QBE should incorporate both.

2.4 Research context

2.4.1 Vocal control of pitch, loudness, and spectral shape

As we identified in Sections 2.1 and 2.2, the ability to vocalise sounds is determined by physical constraints. The limits of vocal fold vibration rate, physical dimensions of the vocal tract and air flow dictate the dynamic range, frequency range and types of sounds that can be produced. In terms of vocal control, there has been significant research on pitch range, rate of pitch change, sound intensity level range, and speed of phoneme transitions. In this section we will discuss this literature and highlight some of the findings that are relevant to the aims of this thesis concerning the accuracy with which people can vocalise salient acoustic features in musical sounds. In doing so we aim to identify where further research is required to establish if people, in particular musicians, are able to control the types of acoustic characteristics that might be required to imitate musical sounds, and inform the design of sounds that we will ask people to imitate in Chapter 3.

2.4.1.1 Base pitch and range

There are two vocal characteristics of interest here: speaking F_0 (SFF) and F_0 range (FFR). In terms of SFF, Baken and Orlikoff [2000] present a comprehensive overview of the literature for both reading and spontaneous speech. They show that for participants aged 18–62, mean SFF varies between 100–129Hz for males, and 189–224Hz for females. Fitch and Holbrook [1970] recorded the speech of 100 male and 100 female participants aged between 17.5 and 25.5. They report mean SFF of 117Hz for males and 217Hz for females, in line with Baken and Orlikoff [2000] and several related studies. Fitch and Holbrook [1970] also report SFF range values of 85–155Hz for males and 165–255Hz for females. In a study of 57 male singers and non-singers aged between 20–55, Morris et al. [1995] reported an average SFF of 128Hz (with average FFR of 85–822Hz). In another study where the same method was applied to female singers and non–singers, the same authors report average SFF (and FFR) of 203.5Hz (129–1340Hz) [Brown et al., 1993].

Kent et al. [1987] present an excellent review of 7 studies into FFR. Ignoring the results for children and elderly adults, the presented mean FFR values span from 26.6 semitones (ST) to 37.9ST. The largest FRR was presented

by Hollien et al. [1971], who reported mean FFRs of 78–698Hz for males and 139–1108Hz for females. It is worth noting the variance in results presented by Kent et al. [1987] may be due to a number of factors including the stimuli and measurement methods used. Additionally, FFR does not remain stable over time and has been shown to exhibit mean variation of $+/-$ 2 semitones of the lowest frequency within a day, and up to 6ST over a 4–6 week period [Gelfer, 1986]. Finally, in a study of 30 female participants, Zraick et al. [2000] found mean FFR to be approximately 1kHz (equating to 34ST), and notably, he found no significant difference between the FFR when participants produced their full pitch range in either discrete steps or glissando.

2.4.1.2 Speed of pitch change and vibrato

In the experiment presented in Chapter 3 we investigate the ability of people to vocalise vibrato–like effects (pitch modulations), however when considering the extremities of production ability we can cannot rely on vibrato studies, because many are concerned with the natural vibrato rates of singers as opposed to limits of their range. Nonetheless, there appears to be some consensus across studies that the singing average vibrato rate is in the region of 5–7Hz [Sundberg, 1994b]. This is in agreement with the range considered to be 'musically useful' by Martens et al. [2006]. Hakes et al. [1988] reported data on the extreme rates for a study of 10 singers, at 4.81Hz and 6.77Hz. This is slightly less than the extreme rates reported by Prame [1994], at 4.6Hz and 7.4Hz, (although these values are for individual vibrato cycles and not means as in Hakes et al. [1988]). In both of these studies the participants were not asked to maximise or minimise their vibrato rate, although in Hakes et al. [1988] the singers were asked to extend their vibrato to the maximum depth, and report depths ranging from 1.01ST - 3.6ST (mean within singers). In this case the depth was measured as the difference between the minimum and maximum F_0 values within a cycle.

It has been shown that trained singers are able to elicit control over both vibrato rate and depth [Dromey et al., 2003; King and Horii, 1993], although to our knowledge no such studies have been conducted on non–singers. King and Horii [1993] showed no effect of base pitch on the ability of singers to imitate the rate and extent of pitch modulations, and that rate can be controlled much more accurately than depth even when the target differences in depth were within producible and perceivable ranges. They state that target rates of 3Hz

and 5Hz were matched more accurately than 7Hz, although the accuracy was good for all rates (on average the targets were matched to within 13% (0.6Hz) of the target rate).

The maximum speed of pitch change has been studied for both singers and non–singers [Ohala and Ewan, 1973; Sundberg, 1973], however it should be noted that these studies consider the transition area of a pitch change to be the middle 75% of the curve from starting to target pitch, not the entire transition period. It is only relatively recently that the time taken to complete 100% of a pitch change has been measured [Xu and Sun, 2000, 2002]. Xu and Sun [2002] present results from a study where 36 participants aged 18–45 were asked to vocalise two types of carrier sounds (sustained vowel and a syllable sequence) with 3 excursion sizes (4, 7 and 12ST) at a rate of 6Hz. The base pitch was selected by the participants, and two patterns were used, starting on either ascending or descending trajectories. The mean excursion sizes for the intervals were actually 3.8, 4.7 and 6.6ST, indicating that we cannot realistically expect people to produce excursion sizes much greater than 6ST at this rate. Is it worth noting that these interval sizes are much greater than those presented by Hakes et al. [1988]. This could be due to the fact that this study was not measuring pitch modulation in a singing context, whereas the subjects in Hakes' study were probably trying to reach extreme intervals whilst maintaining a suitably musical singing output. It could also be due to the methods used to determine excursion size. Xu and Sun [2002] found that excursion speed and size have a linear relationship, where excursion speed increases with interval size. In other words, larger intervals take longer, but people also perform them with a faster rate of pitch change (so a 4ST interval will not take twice as long as a 2ST interval).

It should also be noted that the rate and depth of vocal vibrato interacts with amplitude and timbre. Due to resonances in the vocal tract, changes in phonation frequency will also change the relationship between harmonic partials and formant frequencies, causing amplitude modulations at specific frequencies [Sundberg, 1989]. This effectively means that it is unreasonable to expect anyone to reproduce a target sound containing pitch modulations without introducing timbral and amplitude modulations. In the experiment of Chapter 3 we ask participants to imitate sounds with pitch, amplitude and spectral modulations. We can therefore expect participants to be limited in their ability to imitate sounds containing modulations of more than one feature, although to our knowledge the effect of the interaction between these

features has not been previously studied.

2.4.1.3 Effects of pitch scaling on imitation accuracy

In addition to the aforementioned limits of pitch production, it is important that we consider the scaling of pitch envelopes that we might expect people to imitate, particularly whether to use linear or logarithmic pitch scales. In terms of imitating pitch ramps, we may refer to studies on singing of glissando. There is some evidence that the change in frequency is linear over time, at least below 400Hz [Henrich et al., 2005; Hoppe et al., 2003], but the results from other studies on glissando are ambiguous on this point. In an example presented by Roubeau et al. [2009], pitch production appears to be logarithmic above 500Hz but linear below 500Hz. Fujisaki [1983] shows curves that do not look definitively logarithmic or linear. However, it is well understood that we perceive pitch on a non–linear scale, and for the study of pitch range it seems sensible to use a musical scaling (i.e. pitch in ST), as per d'Alessandro et al. [1998].

2.4.1.4 Control of voiced intensity

The dynamic range of the voice is dependent on phonation fundamental frequency, and is approximately 50dB at normal SFF values [Coleman et al., 1977; Colton, 1970]. We can expect singers to have a larger dynamic range than non–singers [Sulter et al., 1995], but there does not appear to be any effect of vocal training on the upper limit of vocal intensity, only minimum producible intensity values [DeLeo LeBorgne and Weinrich, 2002]. The maximum rate of amplitude change has been studied as part of pitch modulation, and the two effects are generally considered to be closely coupled [Sundberg, 1989, pp. 164–166]. As a result it might be reasonable to apply the same modulation constraints for pitch (see Section 2.4.1.2) to amplitude. However, it is worth considering how people might achieve amplitude modulation without pitch modulation. This could be using physical gestures (such as thumping on the chest), for which the maximum rate would be limited by motor control of the arm/hands. It could also be achieved by modulating lips and size/shape of the oral cavity. If so, this would have an effect on the vocalised timbre, or vowels, and we could expect the modulation rate to be similar to that for diphthong durations.

2.4.1.5 Control of spectral shape

Although we can expect pitch and voice quality to have an effect on the spectral shape of the vocalised sounds, spectral shape is arguably most determined by articulation (as discussed in Section 2.1.3). As such, changes in spectral shape of voiced sounds, particularly vowels, may be produced by continuous formant changes as are used to produce diphthongs. The spectral shape of vocalised sound can vary from almost broadband noise with no harmonic relationship between the partials and (achieved for example using the post–alveolar fricative, /ʃ/) to pure, pitched vowels, for which the spectral shape will depend on the formant frequencies. The speed with which the spectral shape can be changed is determined by the speed at which people can control the articulatory musculature, primarily the tongue, jaw, and lips, all of which are manipulated to produce different vowel sounds. Therefore we can consider the speed of vowel changes to be a good indicator of how fast people might be able to change their vocalised spectral shape (as a lower bound - it is conceivable that people may produce faster utterances beyond speech). The upper limits for this motion in speech been previously demonstrated by asking people to utter sentences made up of words containing diphthongs at different speaking rates (slow, moderate, fast) [Gay, 1968]. In doing so, Gay observed durations of 123–172ms at slow speaking rates for the diphthongs /aɪ/ and /eɪ/oʊ/ respectively, and 84–98ms at fast speaking rates for the diphthongs /eɪ/ and /aʊ/ respectively. As such, it seems reasonable to consider the maximum speed of changes in spectral shape of voiced sounds to be (at least) between ~100–200ms.

2.4.1.6 Biases regarding ramp directions of pitch and intensity

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 the ability to vocalise a sound (or even a sonic idea), if there is a difference between what one thinks

they are vocalising and the actual acoustic properties of the vocalisation. It is important to note that in this thesis we are not concerned with testing the extremities of the vocal system or perception of different feature envelopes of sounds. For this reason, when designing sounds for people to imitate in Chapter 3, we use features and parameters that are comfortably within both the physically producible and perceivable limits in terms of the range and rate of change of the features.

2.4.1.7 Summary

The literature discussed in this section provides a solid grounding for specifying the parameters of the stimuli that we might ask participants to imitate, yet regarding vocal control of pitch, loudness, and spectral shape we note two major gaps in current research: *i*) 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. In addition, this tends to focus on vocal ranges, for example of pitch and intensity; *ii*) this literature mainly focusses on single features, with the exception of studies on phonetograms [DeLeo LeBorgne and Weinrich, 2002; Sulter et al., 1995]. 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 related to pitch, dynamics, and spectral shape. In a study with similar motivations to our own, the accuracy of vocal imitations with respect to pitch, tempo, sharpness, and onset features was investigated [Lemaitre et al., 2016b]. The authors found that participants were able to accurately imitate pitch and tempo in absolute terms and sharpness in relative terms, with onset (i.e. attack time) imitated least accurately out of the four features. In this thesis we investigate similar features: pitch; loudness (related to onset); spectral centroid (related to sharpness). As we will discuss in Chapter 3, instead of using constant (flat) temporal envelopes for pitch and spectral shape, we will investigate the accuracy with which people can imitate ramp and modulation envelopes for each of the features independently, and the extent to which people can control interaction between pitch and loudness, or pitch and spectral shape.

2.4.2 Vocalising percussion sounds

In Chapters 4 and 5 we turn the focus from vocal imitation of synthesised sounds to that of percussion sounds. One of the most prominent cases of vocalised percussion sounds is the art form of beatboxing – a vocal performance technique where the performer imitates percussion sounds and rhythmic patterns. As we identified in Section 2.3, the application of QBV is of particular interest for musicians and music producers, who will not necessarily be proficient beatboxers. For this reason we did not specifically recruit beatboxers for the vocal production task in Chapter 4, but rather recruited musicians, most of whom did not have previous experience of vocal imitation practice. However, the similarities between beatboxing and vocal imitation of percussion sounds warrant a brief discussion of the related literature.

Beatboxing originates in the hip hop music culture, and many of the vocal techniques were developed to imitate the types of percussion sounds typically used in the music of this genre, such as electronic drum machines. We note some important distinctions between beatboxing and the work presented in Chapters 4 and 5. Beatboxing is a performance practice, with a focus on rhythm as much as imitation accuracy of a given percussion sound. This is notable because when imitating patterns certain vocal techniques might be adopted to enable fast repetition of short sounds, and this constraint does not exist when vocalising a single sound as might be done in the QBV use case. In addition, beatboxed sounds are often convincing imitations of actual percussion sounds, however beatboxers will typically be primed with a set of go–to techniques to produce different 'standard' types of percussion sounds and effects (see Stowell and Plumbley [2008] for an overview of such sounds and techniques). Nonetheless, it is also likely that an experienced beatboxer will possess a much wider repertoire of vocal percussion sounds than a non–beatboxer.

In the experiments of Chapters 4 and 5 we are particularly interested in the perceptual similarity between vocalised percussion sounds and the imitated sounds. To our knowledge there have been no such studies comparing beatboxed sounds and their real–world counterparts, however Lederer [2005] compared spectral and temporal acoustic metrics such as rise time, fade rate and resonant frequency of (professionally) beatboxed versions of 6 popular electronic drum sounds. The author found that beatboxed sounds contained more partials than their electronic drum counterparts and that fade rate was

not accurately imitated, particularly for transient sounds such as clicks, where the vocalist had little or no control over the decay portion of the sound. In general, electronic sounds were identifiable from the vocalisations and it was noted that the more complex sounds were imitated less accurately than simple ones (i.e. hi–hats were imitated less accurately than clave clicks). In terms of vocal technique, Proctor et al. [2013] examined the mechanisms that a professional hip hop vocalist and beatboxer used to imitate percussion sounds, using real–time magnetic resonance imaging. The beatboxer demonstrated use of articulation and air stream mechanisms found in speech, and also used articulation patterns that did not exist in their native language. More recently, Blaylock et al. [2017] conducted a similar study using 5 beatboxers, and found that the beatboxers used non–linguistic articulations and air stream methods, such as lingual egressive and pulmonic ingressive airstreams, indicating that when imitating percussion sounds people may use vocal techniques beyond those found in any known language.

As noted by Atherton [2007], the pedagogical practice of vocalising percussion sounds is commonplace in some musical cultures such as Cuban conga and northern Indian tabla drumming. In such practices the prototypical sounds and rhythmic patterns of the drums are memorised as vocalisations that have some symbolic relationship to the actual sounds of the instruments. Patel and Iversen [2003] investigated this relationship by conducting both acoustic and perceptual analysis of vocalised tabla sounds, or *bols*. They found evidence of onomatopoeia being used to represent the sounds of the tabla, with strong acoustic correlates between bols and their respective tabla sounds for spectral centroid, decay time, F_0, and the duration between consonants in clusters. In the perceptual test they asked listeners who were unfamiliar with tabla playing to match bols to tabla sounds in a forced–choice test with vocable pairs (similar to minimal pairs in phonetics), and found that for 3 out of 4 vocable pairs listeners were able to match the bols to the correct tabla sounds.

The work discussed in this section demonstrates the ability of humans to imitate a broad range of percussion sounds using their voice, by both employing onomatopoeia and using sounds and vocal techniques not encountered in speech production. Furthermore, there is some evidence of the perceptual relevance of certain acoustic features used to vocalise prototypical drum sounds, and the ability of lay listeners to identify representative sounds from vocalisations. However, excepting only the study by Patel and Iversen [2003], which was limited in both the number and type of sounds used, to our knowledge

there is no work that has focussed on the vocal imitation accuracy and imitability of percussion sounds, in terms of perceptual similarity between imitations and imitated sounds, and whether listeners are able to identify imitated sounds from the imitations.

2.4.3 Methods for measuring the perceptual similarity between sounds

As we discussed in Section 2.3, quantifying the similarity between sounds is a core aspect of many MIR and audio signal processing tasks, including QBV, where we are interested in the similarity between vocalisations and sounds to be searched. In the experiments of Chapter 4 we are specifically interested in the subjective similarity between imitations and imitated sounds from a listener perspective. This raises the important question of how best to measure perceptual similarity, therefore in this section we will review common methods for collecting and analysing such measures.

Forced–choice tests are commonly used in psychology for testing personality traits (often referred to as *ipsative* measurement). However, they can also be used in audio perception tasks, where typically a listener is asked to select a stimulus sound from a set of 2 or more sounds. The task may be to identify the actual stimulus in the set (i.e. the same sound), or to select the sound in the set that is most similar to the stimulus [Ellis et al., 2002]. In the latter case, it can be assumed the selected sound is the most similar to the stimulus, and subject to collecting comparisons of all sounds in the set, similarity between the sounds can be quantified using a similarity matrix of the frequencies with which each sound is selected for a given stimulus. Forced–choice tests appear as a popular choice for vocal imitation studies, where the stimulus is typically a vocal imitation, which is compared to a set of referent sounds, one of which is the imitated sound [Cartwright and Pardo, 2015; Lederer, 2005; Lemaitre and Rocchesso, 2014; Lemaitre et al., 2011; Patel and Iversen, 2003]. In addition to the similarity matrix, this provides a measure of imitation accuracy, in terms of how often an imitated sound is identified from its respective vocal imitations.

Pairwise comparison tasks require the listener to rate the similarity between a pair of sounds on a scale from from very similar to very dis-

similar [Caclin et al., 2005; Grey, 1977; McAdams et al., 1995]. As with forced choice tests, the similarity between all sounds in a test set can be quantified in a similarity matrix. This method is commonly used in timbre perception studies, where dimensionality reduction techniques such as multi–dimensional scaling (MDS) can be applied to the similarity matrix in order to find an n–dimensional Euclidean space that preserves the perceptual similarity between sounds [Kruskal, 1964] (such spaces are useful for visualising the distances between sounds and investigating acoustic correlates of the salient dimensions). One notable drawback of pairwise comparison and forced–choice tasks is the number of trials that must be conducted in order to construct a full similarity matrix of N sounds (N^2 if all presentation orders are included). This can quickly lead to more comparisons than is reasonable for a single listener to make, meaning that often the number of sounds in a test set must be kept small (16–18 sounds were used in the above mentioned studies, resulting in 256–324 ratings per listener).

Sorting tasks involve asking listeners to freely sort sounds into categories based on similarity [Dessein and Lemaitre, 2009; Gygi et al., 2007; Parizet and Koehl, 2012; Scavone et al., 2001] or according to representative sounds from each category [Rocchesso et al., 2016a]. In the free sorting scenario, listeners are free to choose both the number of categories and the number of sounds per category. Although listeners do not explicitly provide information about the similarity between sounds within each category or the similarity between categories, as noted by Parizet and Koehl [2012], a full similarity matrix may be constructed by averaging the co–occurrence matrices of all sounds based on the frequency of being grouped together, which can then be subjected to MDS. In addition, hierarchical clustering may be conducted using the similarity matrix, as in [Dessein and Lemaitre, 2009]. This approach permits the salient perceptual attributes for each cluster to be learned, informing, for example whether people group sounds based on the source or acoustic properties, and the acoustic correlates for the clusters may be investigated for the purposes of automatic classification of sounds.

Odd–one–out tasks require listeners to identify the least similar sound (or song) from a triplet (set of 3 sounds) [Wolff and Weyde, 2014], or identify the most and least similar pairs from a triplet [Novello et al., 2006]. In the first case this suggests the 2 remaining sounds are more similar to

one another than they are to the eliminated sound, whereas the second case provides a 3–way pairwise comparison. As with the above mentioned methods, from the 3–way comparison a similarity matrix can be constructed by assigning values to each pair of conditions (most similar, least similar, and 'middle–similar') [Novello et al., 2006]. In the least–similar odd–one–out scenario, a similarity matrix may be constructed by first learning the distances between sounds using metric learning or neural networks [Wolff and Weyde, 2014]. For a complete (i.e. fully balanced) design the number of trials required is actually greater than for the pairwise comparison method: using the formula from Novello et al. [2006], a stimulus set consisting of 10 sounds would require 120 trials. This makes the method infeasible for large stimulus sets if a balanced design is required, however the number of trials may be reduced if the number of comparisons including the same pairs is limited.

Multiple comparison rating tasks may be used where is it desirable to compare 2 or more sounds to a reference sound using continuous rating scales. Perhaps the most widely used version of this method is the Multiple Stimuli with Hidden Reference and Anchor standard (MUSHRA) [International Telecommunication Union, 2003]. Typically the MUSHRA format requires that a single known reference sound is compared to up to 14 test sounds, which include a hidden reference and (optional) hidden anchor. When used to judge audio quality (for which the standard was originally intended), the listener rates the audio quality of each test sound in relation to the reference. The hidden reference and anchor serve to ensure the listener is able to identify an obvious 'best' and 'worst' case example and uses the full range of the rating scale. Although not specified in the standard, similarity (or indeed any attribute of interest) may be assigned to the scale, instead of audio quality. Whilst the MUSHRA standard specifies the use of expert listeners, it has recently been shown that for assessment of source separation audio quality, lay listeners can provide comparable results to expert listeners [Cartwright et al., 2016].

In summary, the choice of method for measuring perceptual similarity between sounds is dependent on *i*) the types of comparisons that are to be made, *ii*) what kind of response data is required for subsequent analysis or modelling, and *iii*) to a certain extent, the difficulty of the task presented to listeners (for example, a pairwise comparison task may be less cognitively

challenging than a MUSHRA test, but may also require many more individual tests for the listener, potentially making them more susceptible to fatigue, whereas the odd-one–out approach can be easily incorporated into a game–like test [Wolff and Weyde, 2014]). Overall the MUSHRA method provides a much richer source of information compared to the alternative above mentioned methods, namely: *a*) individual similarity ratings between each test sound and the reference sound; *b*) identification of the 'most–similar' and 'least–similar' sounds to the reference sound; and *c*) an inherent ranking of and pairwise comparison between the test sounds (with respect to the reference sound) [Sporer et al., 2009]. This makes it an attractive method for the listening task presented in Chapter 4, where we are primarily interested in both the most similar drum sound to a given imitation, and the relative similarity ratings between an imitation and set of drum sounds.

2.4.4 Audio features for vocal imitation analysis and QBV

Typically, in QBV systems audio features are extracted from a vocal example of a target sound and compared to the features of sounds in an audio library, to return a ranked list of similar sounds [Roma and Serra, 2015; Zhang and Duan, 2015]. As such, the audio features used to map between vocalisations and sounds in a sample library are a core part of any QBV system. In this section we review the literature on both QBV, and more generally, analysis of vocal imitations, with a specific focus on the types of audio features that might be useful for QBV applications. We will revisit these features in Chapter 5 in order to investigate their suitability for predicting the similarity between vocal imitations and percussion sounds. The following review is separated into heuristic (Section 2.4.4.1) and learned (2.4.4.2) features.

2.4.4.1 Heuristic features for vocal imitation analysis

Heuristic (or hand–crafted) features are based on knowledge of the acoustic properties of sound, in particular the temporal and spectral characteristics. They are typically deterministic, in that the feature values depend only on the signal from which they are extracted. Although we are particularly interested in representations for vocalised percussion sounds, we also include related work on features for vocal imitations of non–speech, non–singing sounds, regardless of whether the imitations are percussion–specific. In this section we will not

cover the definitions of all the features discussed, but in Chapter 5 we provide the sources and descriptions for each of the features used in our experiments (further details and definitions of many of the features discussed herein are given by Bullock [2008]; Peeters [2004]; Peeters et al. [2011], amongst others).

An overview of heuristic features from the literature on vocal imitations and QBV is given in Table 2.1. These features can be computed over an entire signal, providing a summary (or global) feature value for a whole sound, or alternatively, many can be computed in a frame–wise manner. In the frame–wise approach the signal is split into frames and the feature computed for each frame. Frame–wise features may be used to compute summary statistics (mean, variance, etc.) to represent the evolution of a particular feature over time. For the sake of simplicity, where multiple terms are used in the literature to describe closely related features, such as *relative/absolute/effective* duration, *loudness/energy/rms*, and *spread/variance* of the spectrum, we have assigned a single term in Table 2.1. Some study–specific features have been intentionally excluded from this list if they are not relevant to the types of percussion sounds and imitations used in the experiments of this thesis (in Chapters 4 and 5), such as measures of amplitude modulation and jitter used by Lemaitre et al. [2017]. Finally, we note that the term *morphological* features is used line with Marchetto and Peeters [2015] and the description given in Section 2.3, to denote any feature that describes the trajectory or profile of a low–level feature, such as loudness.

Features for vocalised percussion and musical sounds

As we discussed in Section 2.4.2, there have been a number of studies on vocalised percussion sounds and beatboxing, although these generally did not include use of audio features, and where they did the features were not evaluated according to any criteria (with the exception of Patel and Iversen [2003]). However, in addition to the aforementioned research on beatboxing there have been a number of studies concerned with classification of beatboxed sounds into drum categories such as kick, snare etc. [Hazan, 2005; Kapur et al., 2004; Nakano et al., 2004; Ramires, 2017; Sinyor et al., 2005], for which the audio features used may transfer well to the task of measuring similarity between vocalisations and percussion sounds. Kapur et al. [2004] compared individual and sets of features, and reported the highest performance using zero–crossing rate alone or a set of linear predictive coding (LPC) coefficients. The authors

	Percussion specific							**Non–percussion specific**											
	Patel and Iversen [2003]	Kapur et al. [2004]	Nakano et al. [2004]	Hazan [2005]	Sinyor et al. [2005]	Stowell [2010]	Ramires [2017]	Dessein and Lemaitre [2009]	Lemaitre et al. [2011]	Cartwright and Pardo [2014]	Roma and Serra [2015]	Marchetto and Peeters [2015]	Mauro and Rocchesso [2015]	Lemaitre et al. [2016b]	Lemaitre et al. [2016a]	Baldan et al. [2016]	Rocchesso et al. [2016a]	Rocchesso et al. [2016b]	White et al. [2017]	Lemaitre et al. [2017]
(log) attack time (LAT)														✓						
duration								✓	✓						✓					
temporal crest factor						✓														
zero crossing rate		✓		✓	✓	✓	✓	✓	✓						✓	✓				
decay time	✓																			
energy/loudness					✓		✓			✓			✓		✓	✓	✓	✓		
morphological features								✓				✓	✓		✓		✓			
pitch/F_0	✓					✓			✓	✓				✓	✓	✓		✓		✓
pitch clarity/strength						✓			✓	✓					✓					✓
noisiness															✓					
inharmonicity										✓										
roughness													✓				✓			
spectral centroid	✓	✓		✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓		✓
spectral rolloff		✓		✓	✓	✓	✓													
spectral crest factor						✓														
spectral slope						✓	✓													
spectral spread					✓	✓	✓			✓						✓		✓		
spectral kurtosis							✓			✓								✓		
spectral flatness						✓	✓						✓				✓	✓		
spectral skewness							✓						✓				✓	✓		
spectral entropy													✓				✓			
spectral compactness					✓															
strongest frequency					✓															
spectral flux					✓	✓	✓						✓				✓	✓		
band–specific energy				✓		✓														
LPC coefficients		✓																		
MFCCs/ΔMFCCs		✓	✓			✓	✓				✓								✓	
wavelet coefficients		✓																		

Table 2.1: Heuristic features used in previous studies for the analysis of percussion specific and non–percussion specific vocal imitations.

also included a set of wavelet coefficients in their analysis, but reported higher accuracy with both LPC and MFCC feature sets. The other above mentioned studies all report reasonably high classification accuracy results (between 82–96%), indicating the usefulness of the features highlighted in Table 2.1, however none include comparisons of the features in terms of their contribution to classification accuracy, and in particular do not consider their perceptual rele-

vance. Indeed, excepting the study by Patel and Iversen [2003] there has been little work that considers the perceptual relevance of features for vocalised percussion sounds.

Although not focussed on percussion or vocal sounds, Stowell [2010] presents an analysis of the perceptual relevance of a large set of features (see Table 2.1) using similarity ratings between musical instruments. This was tested using the MDS–derived timbre spaces and stimuli sounds from 3 previous studies [Grey, 1977; Grey and Gordon, 1978; McAdams et al., 1995], and with the exception of spectral centroid and the 95^{th} percentile of the spectrum, there were no features that correlated highly with the dimensions across all 3 timbre spaces. This indicates that the features reported in the original studies as most perceptually relevant for each of the 3 MDS spaces may not be generalisable to sounds beyond those tested, as has since been confirmed by Siddiq et al. [2015]. The author also compared the features in terms of robustness to degraded vocal signals, including beatboxed sounds, using the information theory based measure of *mutual information* between features extracted on clean and degraded signals. Interestingly, ΔMFCCs and MFCCs were reported as being particularly poor by this measure, indicating that these features may not be suitably robust for vocal signals that are not recorded in acoustically clean, quiet environments.

Features for vocal imitations of everyday sounds

Beyond percussion, Del Piccolo and Rocchesso [2016] provide an excellent meta–review of research on non–speech vocal sonic interaction, including the types of acoustic features commonly used for QBV of generic sounds and vocal–based synthesiser parameter control. In terms of synthesiser control, pitch and loudness (or power/energy) appear to be the most frequently used features, such as are used by Rocchesso et al. [2016b], whereas for QBV, they note the common use of low level features such as those presented in Table 2.1. This is understandable given that for many QBV tasks pitch and loudness features may actually be secondary, particularly when searching for un–pitched sounds or those with a similar loudness profile, such as percussion sounds. Any measures that rely on F_0, including pitch, clarity, noisiness, and inharmonicity will be less meaningful for discriminating between these sounds, whereas timbral or temporal descriptors may prove more useful. As with percussion sounds, the literature on features for analysing non–percussion

vocal–imitations tends to be focussed on classification tasks.

Creating meaningful clusters of imitations is useful for 3 reasons: *i*) different audio features can be compared to human–derived clusters, providing understanding of what features are perceptually relevant for this task [Dessein and Lemaitre, 2009; Lemaitre et al., 2011]; *ii*) automatic clustering can be a first step in a QBV system where it may be useful to apply cluster–specific search models, or alternatively for vocally–controlled synthesis, it may be desirable to select different synthesiser models depending on the type of sounds that are being vocalised [Baldan et al., 2016; Rocchesso et al., 2016b]; and *iii*) imitations can be represented as clusters in low dimensional spaces for navigation and exploration of the sounds [Mauro and Rocchesso, 2015; Rocchesso et al., 2016a]. Dessein and Lemaitre [2009] and Lemaitre et al. [2011] asked listeners to freely cluster vocal imitations of everyday sounds, and compared how well a number of different audio features (shown in Table 2.1) could be used to predict the listener–provided clusters of each imitation using simple binary decision tree rules. They found that (loudness weighted) spectral centroid was useful for discriminating between voiced and unvoiced imitations, and by combining this with a measure of modulation amplitude (of the energy envelope), they were able to reliably separate the imitations of gas and electrical–item based sounds, along with identifying whether or not imitations in each class included repetitive elements. In addition they used temporal descriptors (duration) and morphological descriptors (temporal increase of the energy envelope) to perfectly discriminate imitations of gas sounds based on the nature of the attack (brutal or smooth) and duration (long or short).

Marchetto and Peeters [2015] conducted a more in depth study of descriptors for categorising imitations based on their morphological profile (such as up/down, impulse, repetition, stable). They report high classification accuracy (84%) using descriptors based on the signal trend (direction), and the nature of active regions in the imitation. We note that whilst these types of descriptors may be useful for categorising everyday or environmental sounds, they are not suitable for discriminating between sounds that share a similar profile, such as percussion sounds. Mauro and Rocchesso [2015] and Rocchesso et al. [2016a] present a method for exploring a geometric space of vocal imitations using the features listed in Table 2.1. They apply dimensionality reduction, namely principle components analysis (PCA), to produce a 2D space in which like–sounds are grouped together (similar to the exploratory

search methods discussed in Section 2.3). PCA is a projection method that performs an orthogonal transform on a set of correlated variables into a set of uncorrelated variables (principle components) [Jackson, 1991]. As such, by submitting the audio features of all imitations to a PCA, one can obtain the first n dimensions on which the observed data exhibits the maximum variation. PCA can be used to identify the latent components that best describe a large set of audio features for visualisation purposes [Mauro and Rocchesso, 2015; Rocchesso et al., 2016a; Stowell, 2010], or simply to reduce a large number of features to a lower number of components to use as predictors in a classification or regression model, as we will see in Chapter 5.

Features for QBV

Although all of the above work on classification of vocal imitations may be applied to QBV, much of the previous research on QBV systems tends to focus on small numbers of hand–selected features that are specific to the task. For example, Cartwright and Pardo [2014] apply energy, pitch and spectral features for querying a large set of sounds from a single synthesiser in order to tune the parameters, and Roma and Serra [2015] and White et al. [2017] use MFCCs to query crowd–sourced sound collections on freesound[9] and snippets of sounds from 1960s popular music recordings respectively. An exception to this is presented by Blancas and Janer [2014], who use a support vector machine (SVM) classifier to identify vocal imitations of cat, dog, car and drum sounds. They extract the full set of features from Peeters et al. [2011] (many of which are listed in Table 2.1) and select subsets of features using a correlation based feature selection method from Hall [1999]. They report improved classification accuracy using the feature–subsets rather than all features, and report that spectral crest and spectral variation appeared in the subsets for every category, yet the authors do not provide details of all the features selected. Whilst this this research has helped highlight the potential for using the voice as a query medium, there has thus far been little consideration for the effect of different types of features for QBV (i.e. temporal, spectral), and no in depth, formal comparison of heuristic features for this task.

Furthermore, with the exception of Dessein and Lemaitre [2009] and Lemaitre et al. [2011], none of the above mentioned studies consider the perceptual relevance of the features used. Lemaitre et al. [2016a] addressed this by investi-

[9] www.freesound.org

gating whether acoustic features could be used to predict listener–based classification accuracy of vocal imitations of everyday sounds. Listeners were asked to classify vocal imitations in a two–way forced choice experiment (i.e. given a vocalisation, state whether it is an imitation of a 'fridge' or 'blender'). Two sets of acoustic features were compared: one based on Euclidean distance between sounds in a feature space derived using morphological descriptors from Marchetto and Peeters [2015], and one based on the alignment cost between the spectrograms of two sounds (using dynamic time warping). The spectrogram alignment cost outperformed the morphological descriptors in terms of predicting listener classification accuracy. Notably, the authors highlight that the correlation of distance (in terms of alignment cost) over classification accuracy varied considerably across the different families of imitated sounds (such as impulsive, stationary, complex), suggesting that the suitability of acoustic descriptors for predicting the similarity between a vocal imitation and sound class may be specific to the type of sound being imitated.

Summary

The work discussed in this section shows that a wide range of temporal, morphological, pitch based, and spectral features have been applied with varying degrees of success to identify, classify and predict similarity between vocal imitations and imitated sounds. We have seen that for these tasks, features tend to be either hand–selected, or derived using feature selection and dimensionality reduction methods. Yet there has been little focus on comparing different types of features in terms of both classification accuracy (i.e. QBV performance) and perceptual relevance of the features used. For this reason in the experiments of Chapter 5 we evaluate the full set of features from Table 2.1 and suitable subsets thereof, in terms of their perceptual relevance and suitability for predicting similarity between vocalisations and percussion sounds.

2.4.4.2 Feature learning for QBV

Instead of using domain knowledge to specify heuristic features, and selecting feature subsets either heuristically or using dimensionality reduction methods such as PCA, we may apply deep learning methods to learn the features from the audio data automatically. Recent developments in deep learning have highlighted that learned features can outperform heuristic features such MFCCs

for many audio related tasks, including speech coding [Deng et al., 2010], music tagging [Hamel et al., 2011], genre classification [Choi et al., 2017], and more relevant to our task, for QBV of environmental and instrument sounds [Zhang and Duan, 2015, 2016a,b]. Feature (or representation) learning involves automatically learning some representation of the given data that is useful for a particular task [Goodfellow et al., 2016, p. 525]. This may be in the context of a supervised classification model, where the 'representation learning' layers are trained to provide an optimal input to the last layer, typically a linear classifier [Goodfellow et al., 2016, p. 525]. However, where labelled data is sparse or non–existent, it can be applied in an unsupervised setting, to learn some representation of the data such that a desired output may be reconstructed from the representation. In practice, similar methods are used to learn the features in both settings, and are typically based on the *auto–encoder*: a type of neural network (NN) that can be considered the "*quintessential example of a representation learning algorithm*" [Goodfellow et al., 2016, p. 4]. In this section we will discuss the application of auto–encoders for feature learning and related work on feature learning for QBV.

Auto–encoders

An auto–encoder (AE) is a type of NN that consists of 2 parts: an encoder and a decoder. The first part encodes an input, x, using some function, $h = f(x)$. The decoder produces a reconstructed version of the input, y, using some function, $g(h)$ [Goodfellow et al., 2016, p. 499]. Typically the model is designed with constraints such that the AE does not simply directly map the input to the output, i.e. $g(f(x)) = x$. This can be achieved by ensuring the encoded representation, h is smaller than x, adding regularization to the cost function, or introducing non–linearities into the encoding and decoding functions. An example of a general case auto–encoder is given in Figure 2.2.

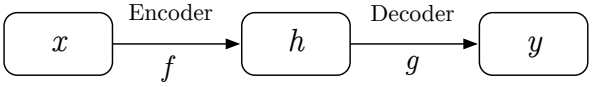


Figure 2.2: General case of an auto–encoder, after [Goodfellow et al., 2016, p. 500]

General overview of auto–encoders

In the minimal form, an AE will have 3 layers: an input layer, a hidden layer, and an output layer. The input and output layer nodes are populated with feature vectors, which in the case of audio could be a vectorised time–frequency representation such as a spectrogram or Mel spectrogram. The hidden layer consists of 'nodes' (or neurons), the number of which determines the size of the layer. Typically the network is 'fully connected' (or 'dense'), meaning that all the nodes of each layer are connected to all the nodes of the proceeding and preceding layers. The inputs to each node of the hidden layer are multiplied by a weighting, then summed and optionally transformed with a non–linear transfer (or activation) function. In the training stage, the weights of the hidden layer are tuned such that the network learns functions h and g that minimise some loss. Mean squared error (MSE) is commonly used as a loss function, and the network is typically trained with stochastic gradient descent (SGD) using back–propagation. For computational efficiency and faster training times it can often be desirable to train the network in batches of inputs using mini–batch SGD [Goodfellow et al., 2016, pp. 274–280]. Finally, a trained AE can be used to extract features from new data by extracting the encoded representation, h.

Interestingly, when an AE has only one hidden layer, there is no non–linear activation function, and MSE is used as a loss function, the AE will learn an orthogonal transformation that is equivalent to PCA, with the number of principle components determined by the number of nodes in the hidden layer [Bengio et al., 2009; Bourlard and Kamp, 1988]. It is often desirable to train a network with more than 1 hidden layer, as it has been shown that increasing the number of hidden layers beyond 1 reduces the reconstruction error (i.e. loss), when the number of network parameters remains equivalent [Hinton and Salakhutdinov, 2006]. In addition, we know from the universal approximation theorem [Hornik et al., 1989] that an AE with a sufficiently large number of nodes in the hidden layer can represent any continuous bounded function, yet a sufficiently large network will probably have too many parameters to train and may not generalise well [Goodfellow et al., 2016, p. 195]. As such, it is often desirable to increase the depth (i.e. the number of hidden layers) of a network beyond 1, to make a 'deep' or 'stacked' auto–encoder (SAE). This has a number of benefits, namely to *i*) reduce the number of parameters required to achieve the same level of representational complexity that would be possible with a larger, single–layer network, *ii*) improve generalisability,

and *iii*) reduce the amount of training data required [Goodfellow et al., 2016, p. 506].

Application to feature learning for QBV

Despite the popularity of AEs and SAEs for feature learning, to date there has been very little research on applying these methods for QBV. The only related work that we are aware of is a series of studies by Zhang and Duan [2015, 2016a,b]. In these experiments the authors adopted the VocalSketch dataset from Cartwright and Pardo [2015]. This dataset consists of 4429 vocal imitations of 240 stimuli, from 185 unique participants. The stimuli are split into four categories: acoustic instruments (n=40); everyday sounds (n=120); commercial synthesisers (n=40); single synthesisers (n=40) (the single synthesiser category contains sounds generated by the authors using a 15-parameter subtractive synthesiser). The imitations were either produced in response to an audio example (n=2418) or a label (n=2011). In Zhang and Duan [2015] MFCC and ΔMFCC features were compared to those extracted using an SAE with 2 hidden layers, comprising 500 and 100 nodes for the first and second layers respectively. The network was trained on 525ms patches of the vocal imitations, which were taken from 20 consecutive frames of a constant Q transform (CQT) with 26.25ms hop size and 72 frequency bins. The authors evaluated the features in terms of same–category classification accuracy using a support vector machine (SVM), and found that for all categories the learned features outperformed MFCCs (albeit with marginal improvements for some categories). Similar findings were reported when comparing MFCCs to learned features with the same AE architecture [Zhang and Duan, 2016a] and a large set of heuristic features (similar to those in Table 2.1), to those from an AE with 1000 and 600 nodes in the hidden layers [Zhang and Duan, 2016b] in an unsupervised classification scenario, based on distance between imitations and imitated sounds in each of the feature spaces.

One notable issue with dense SAEs such as those used in the above experiments is that the features extracted for a given test sample are not time or frequency–invariant. For example, assuming the features are extracted from input representations of an entire drum sound, the distance between 2 drum samples in the learned feature spaces will vary considerably depending on how well aligned the sounds are in terms of the attack and decay portions. In Zhang and Duan [2016b] the authors circumvent this issue by training and testing the SAE on patches of sounds, then calculating 6 summary statistics for each of

the extracted features (min, max, mean, interquartile range (IQR), and standard deviation) to represent an entire sound. However, with 1200 features for a given patch, and 6 statistics, each sound is represented with 7200 features. Due to the large number of dimensions, an infeasibly large number of examples would be required to sufficiently sample the resulting 7200–D Euclidean space if we are interested in finding groups of similar sounds. If the number of samples available is small then it will be desirable to learn much lower dimensional representations of the sounds, ideally using features that are invariant to both dimensions in a time–frequency representation. In addition, we know from research in MIR that features captured from such short–time analysis often do not capture the high level structure of music [Humphrey et al., 2013], and although the imitations and imitated sounds from the VocalSketch dataset are much shorter than typical music, a similar sentiment may also apply here. Fortunately, there is a version of the AE that maybe more suited to feature learning for QBV: the convolutional auto–encoder.

Convolutional auto–encoders

A convolutional auto–encoder (CAE) is an AE based on a convolutional neural network (CNN). The general model of AEs also applies to CAEs, following the example given in Figure 2.2, however as opposed to the AEs discussed thus far, in a CAE the hidden layers are not dense (i.e. with fully connected nodes), but are a set of convolutional filters. DNNs that make use of convolutional layers have 3 inherent characteristics that can be useful in many machine learning applications: sparse interactions, parameter sharing, and, as briefly mentioned in the previous section, equivariant representation [Goodfellow et al., 2016, pp. 329–335]. In this section we will *i*) briefly describe how CAEs (or more generally CNNs) work, *ii*) explain how the aforementioned characteristics are useful for representing percussion sounds (and imitations thereof), *iii*) discuss some considerations and inherent constraints for designing and training a CAE/CNN model, and *iv*) present related work that has used CAEs/CNNs for feature learning in audio based tasks.

General overview of convolutional auto–encoders

CNNs operate on an input grid, or matrix. This type of input is commonly found in image processing (where the input may be a matrix containing the pixel values of an image), and equally for audio when a 2D time–frequency

representation is used, such as spectrograms or Mel spectrograms. As already mentioned, a convolutional layer is made up of filters, which are matrices that are convolved with the input at equally spaced locations across the input matrix, creating a *feature map*. The filters can be thought of as masks that are swept across the input, creating high activations when the shape of the mask is similar to the shape of the input that it is covering (such as lines, or edges in the image). The height and width of the filters is typically much smaller than the input dimensions. Consequently, the number of parameters required to train the model is much less than for dense, fully connected layers. This behaviour, where only a subset of input units interact with a subset of output units, is known as *sparse interaction*, and is a characteristic of CNNs that greatly reduces the computational time and memory required to train the model [Goodfellow et al., 2016, p. 330]. Because each filter is normally convolved over the entire input, the parameters are shared across input units, further reducing the memory requirements for model training, compared to dense layers where every input unit has a unique connection to every node [Goodfellow et al., 2016, p. 333].

As previously mentioned, perhaps the most attractive characteristic of CNNs for our application is that of equivariant representation, meaning the learned features are invariant to spatial shifts in the input matrix [Goodfellow et al., 2016, p. 334]. In other words, if a filter is trained to detect a particular line at a given angle, then it will detect this feature regardless of where the line exists in the input image. This type of invariance occurs as a result of the parameter sharing across the input units, although there is also another source of space–invariance in CNNs, which is due to pooling. To give an example of how pooling works, we may first give an example of convolution without pooling. For a given 2D input matrix, I ($H = 128, W = 128$), we may apply a 3x3 filter, F, centred at every possible location on I. For each location the convolution operation is applied, the first step of which is to conduct an element–wise multiplication of the matrix F and the location on I. The resulting values are then summed, giving a single value 'feature' for that location on I. In this case the resultant feature map would also be 128x128, but it is often desirable for the feature map to be smaller than the input (particularly in the case of a CAE where we wish to learn a compact representation of the input).

This can be achieved by including a pooling layer after the convolutional layer. Pooling layers are used to down–sample the feature map based on some

property of interest such as the maximum or mean values (e.g. for each 2x2 'pooling region' on the feature map one might only take the maximum value as the feature for further processing). As mentioned, in addition to reducing the feature map size, pooling can be used to ensure the learned features (i.e. filter weights) are space–invariant within the pooled region [Scherer et al., 2010]. This can be useful when the approximate location of a feature is more important than the exact location, or where it is desirable to focus more or less on a particular dimension of the input matrix. For example, certain audio related tasks such as onset detection might require high temporal but low spectral resolution [Schlüter and Böck, 2013]. In this case the window of information in each dimension can be determined by the filter shape (e.g. wide in time and narrow in frequency), and the resolution by the shape of the pooling region (e.g. narrow in time and wide in frequency). This allows for the less important dimension to be 'smoothed' over. An attractive alternative to pooling is to shift F across I in steps > 1. The size of the shift is often referred to as the 'stride', and in our example a 2x2 stride would result in a 64x64 feature map. Following from the pooling example, using strides wide in frequency and narrow in time will enforce greater temporal resolution. This 'strided' convolution can be implemented within the convolution layer, removing the need for a pooling layer and reducing the complexity of the model, without compromising on the performance of the trained network or the space–invariance of the learned features [Springenberg et al., 2014].

Considerations and network design

The main parameters to select when designing a CAE architecture, particularly when one is concerned with space–invariance and the importance of temporal vs. spectral features, are arguably the filter shape and size, number of filters and shape of the stride. These parameters, along with the number of layers in the encoder will determine the size and shape of the encoded representation. However, there are additional considerations that apply more generally to designing DNN models, such as the choice of activation function, regularisers, and how the network is trained (data and type of optimiser). Activation functions are typically applied in DNNs following each hidden layer, to transform the output of the layer based on some non–linear (and optionally bounded) function, introducing non–linearities between layers thus enabling the modelling of complex data. The rectified linear unit (ReLU) function has been shown to work well for training DNNs because it addresses the 'vanishing

gradient' problem inherent when using sigmoid and hyperbolic tangent functions on networks with many layers [Glorot et al., 2011]. The ReLU function is defined as being linear for all input values > 0 and 0 elsewhere.

Another consideration is that of regularisation, which is defined as "*any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error*" [Goodfellow et al., 2016, p. 117]. Essentially, regularisation serves to prevent overfitting the network to the training data and can be achieved in a number of ways, although popular methods include *i*) introducing penalty terms to the objective function (preventing the weights from becoming too large), *ii*) randomly ignoring ('dropping out') some units of the layers during each training round, and *iii*) using a 'hold–out' validation set to monitor how the trained network performs on unseen data after each training round (providing a means of identifying when the network starts overfitting).

We briefly mentioned optimisation methods in relation to AEs, and the general approach of SGD using mini–batches applies equally to CNNs. To delve into the pros and cons of different optimisation methods is beyond the scope of this work, and there is no generally agreed upon 'best' optimiser [Goodfellow et al., 2016, p. 306] (although we refer the interested reader to an excellent review of commonly used optimisation methods [Ruder, 2016]). Perhaps one of the most current methods in use today is adaptive momentum estimation (Adam) [Kingma and Ba, 2014], which is an extension to SGD that updates the learning rate (i.e. step size of the weight updates) based on the first and second moments (mean and variance) of the gradient, making use of previous updates. This essentially applies an exponential moving average to the weight updates, reducing the magnitude of any oscillations.

Finally, although it may sound obvious, the choice of training data will determine the usefulness of the learned features. Ideally this data should be sufficiently large and truly representative of the type of data that the trained network is intended to be used for, if one wants the network to generalise beyond the examples used for training. Additionally, the type of input representation should be considered. If the aim is to reconstruct audio from the intermediate layers of a network (as might be used for auralising the learned features [Choi et al., 2016b] or the decoded output from a CAE), then an invertible representation such as a power spectrogram might be used (assuming phase information is available), whereas if the time–frequency representation

should be more aligned with how humans perceive loudness and frequency, one might apply equal loudness contours and frequency scaling using Log, ERB, Mel, or Bark scales.

Application to audio based tasks and QBV

Some of the earliest work on learning audio features using CNNs is presented by Lee et al. [2009]. The authors present a CNN for feature learning based on a convolutional deep belief network (CDBN): a type of unsupervised feature learning network that is based on stacked restricted Boltzmann machines. Whilst there are some notable differences between a CDBN and a CAE, both use convolutional layers for learning and extracting features from the input data. The authors investigated the performance of the learned features against MFCCs for many tasks, including speaker identification, phoneme recognition, gender classification, music artist identification, and genre classification. They found that learned features outperformed MFCCs in all tasks except phoneme recognition (where combining the MFCCs and learned features gave best results), providing a strong case for using CNN based features.

More recently, CNNs have successfully furthered the state–of–the–art in many MIR tasks, including chord recognition [Humphrey and Bello, 2012], onset detection [Schlüter and Böck, 2013], boundary (i.e. verse and chorus) detection [Ullrich et al., 2014], singing voice detection [Schlüter and Grill, 2015], denoising and audio source separation [Grais and Plumbley, 2017], genre classification [Costa et al., 2017], and audio tagging [Choi et al., 2016a], outperforming many previous systems based on heuristic features such as MFCCs and others presented in Table 2.1. Many of these examples do not use CAEs (with the exception of Grais and Plumbley [2017]) but apply CNNs for both feature learning and classification (with the latter achieved using dense layers following the convolutional layers). As such it is not always possible to attribute performance improvements solely to the convolutional part of the system. Nonetheless, the notable progress that has been achieved using CNN–based features in MIR confirms the proposals put forward by Humphrey et al. [2013], namely that feature learning using DNNs presents an attractive means to overcome the limitations of many, more traditional heuristic audio features.

One useful characteristic of automatically learned CNN–based features is that they may offer some insight into what types of features are important

for particular tasks. This can be achieved by extracting the features from the convolutional layers of a CNN for a given input. Synonymous with the idea of 'seeing' what the network learns in computer vision [Zeiler and Fergus, 2014], the output from a given convolutional layer can be deconvolved into short–time Fourier transform (STFT) based spectrograms, which can then be inverted and auralised. Choi et al. [2016b] presents an example of such auralisation from a CNN that was trained for genre classification, consisting of 5 convolutional layers followed by 2 dense layers (for classification). They show that the first layers learned to represent vertical and horizontal lines and suppressors (performing onset detection and harmonic component selection), whereas the deeper layers learned more high-level textures and distributions in the spectrograms. Dieleman and Schrauwen [2014] compared the performance of a CNN trained on a music tagging task using either raw audio (i.e. time–domain samples) or Mel spectrograms for training. They found that CNNs trained on the Mel spectrograms outperformed those trained on raw audio, and interestingly, the filters trained using raw audio learned to represent individual (and groups of) frequency components from an input, somewhat similar to an STFT.

In terms of QBV, to our knowledge there exists only one experiment on CNN–based features, by Zhang and Duan [2017]. The authors present a single–network QBV system based on a CNN implemented in a semi-Siamese network structure, consisting of 2 identical but separate CNNs, one of which is trained to learn the features for imitations and the other for imitated sounds. The CNNs are then joined (the features are concatenated) and followed by 3 dense layers that are used to match input vocalisations to audio samples (i.e. perform classification). The convolutional layers are trained to learn feature representations from CQT spectrograms of vocal imitations and the imitated sounds from the VocalSketch dataset. The system shows promising results, outperforming the systems using SAEs from Zhang and Duan [2016a] in terms of how highly the imitated (i.e. target sound) is ranked out of all retrieved sounds. This work highlights the potential performance increase from using CNN–based features compared to dense SAEs, however we note 2 downsides to this approach: i) in the general case, QBV systems require efficient, deployable querying. This method requires each sample in a sound library to be compared to a given vocal query, meaning a dataset with N data samples requires N forward–pass computations of the network, which is computationally demanding, for example compared to nearest neighbour search in a feature vector space. As such,

it may be more desirable to learn the features in the same way, but search for sounds in the learned feature–space. *ii*) To train this type of CNN classifier requires a large number of training examples. Nonetheless, if training data is scarce one can make use of CNN–based feature learning by training a CAE on a large dataset of unlabelled data that is representative of the types of sounds that might be vocalised and queried, as we will demonstrate in Chapter 5.

Summary

In this section we have reviewed the application of deep neural networks for learning audio features. In general, it is apparent that learned features outperform heuristic features for QBV and many other audio based tasks, including chord recognition, onset detection, and source separation. The go–to models for these tasks appear to be dominated by one network type: CNNs. These are attractive for a number of reasons, including the ability to take time–frequency representations of audio data as input (arguably the most commonly used representation in audio–based MIR), the potential to auralise the learned features at each layer of the trained network (providing insight to what the network is learning at each layer), and the time–frequency invariance of learned features. However, whilst some of the literature discussed in this section draws comparisons between learned features and heuristic features, these tend to be limited in scope.

In Section 2.4.4.1 we reviewed a large number of heuristic features that have been previously applied to analysis of vocal imitations and for QBV. To date there has been no comprehensive comparison of how these features compare to those learned from CNNs, for any of the audio related tasks discussed, let alone QBV. In addition, whilst both SAE and CNN approaches show promising performance in terms of retrieving an *imitated* sound from a set of audio samples, none of the aforementioned feature learning based QBV methods consider the *perceptual similarity* between the query and retrieved sounds. Central to the evaluation of these approaches is the assumption that the target sound is indeed the sound that was imitated, and the task is to match the imitations and imitated sounds accordingly. As such the data labels are treated as a proxy for similarity between sounds. In Chapter 5 we will consider a use case in which the query is not necessarily an imitation of a sound in the database, and investigate which features correlate well with the perceptual similarity between an imitation and a set of audio samples, comparing CNN based features to the full list of heuristic features discussed in Section 2.4.4.1.

Chapter 3

Vocal imitation of synthesised sounds

To establish QBV as a viable search method we must first consider whether people are actually able to vocalise the relevant acoustic features that exist in the types of sounds that might be searched, specifically where these features evolve over time. In Section 2.1 we reviewed the literature on physiological and acoustic analysis of sounds produced in speech and singing, and in doing so illustrated that differences in vocal features such as pitch, loudness, and voice quality (or timbre) can be produced by controlling the airflow, laryngeal musculature and articulatory components of the vocal tract. However, there is a notable lack of research into acoustic analysis of vocalisations for non–verbal, non–singing sounds. In this chapter we address this by presenting the results of a vocal production experiment where musically trained participants were asked to vocally imitate a range of synthesised sounds with different time–varying acoustic features.

This chapter is laid out as follows: We first give the scope of the study and identify the core research questions in Section 3.1. In Section 3.2 we describe the stimuli, experimental procedure, and methods used to extract the parameters of interest from the vocalisations. A statistical analysis of the results is presented in Section 3.3, followed by a discussion in Section 3.4. Finally, summary conclusions and implications of the results are presented in Section 3.5.

3.1 Research questions and scope

The main goal of this work is to establish the level of control with which people can vocalise sounds containing different time–varying acoustic features. We consider this in 2 scenarios. First, we investigate whether people can vocalise target temporal envelopes for single features. We then investigate the case of sounds made up of 2 time–varying features, both congruently and in opposition.

We limit the scope of stimuli to sounds generated using a subtractive synthesiser controlling for pitch, amplitude, and spectral shape (defined as the cutoff frequency on a resonant low pass filter). These parameters are varied over time according to particular envelopes, to control 3 acoustic features: fundamental frequency (F_0), loudness, and spectral centroid. The temporal envelopes include ramps (up and down) and periodic modulations (2Hz and 5Hz). A detailed description of the method used to generate these sounds is discussed in depth in Section 3.2.1. 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 be able to exert control over this through manipulation of the articulatory components of the vocal tract. By asking people to vocally imitate these sounds, we address the following research questions:

1. How accurately can people imitate single features within sounds where only 1 feature is changing over time?

2. When asked to imitate sounds where 2 features are changing, what is the effect of the type of feature envelope, and type of feature?

    (a) Do the type of envelopes being combined have an effect, and are some more accurate than others (i.e. are ramp and modulation combinations more accurate than modulation combinations, and does

ramp direction have an effect)?

(b) Do the type of features being combined have an effect (i.e. for the same envelope combinations, are pitch and amplitude combinations more accurate than pitch and spectral shape combinations)?

3.2 Method

3.2.1 Stimuli

The stimuli were generated using the basic source-filter model depicted in Figure 3.1. The sawtooth oscillator is ideal for our task because it is harmonically simple yet contains enough harmonic content for the effect of the low pass filter to be well perceived, whilst also not having any inharmonics, giving high pitch clarity. The synthesiser time–varying parameters for pitch (P), gain (L) and cutoff frequency (C) are scaled in semitones (ST), decibels (dB) and linear Hz respectively. The 4 envelope shapes (Figure 3.2) ramp down *RD*, ramp up *RU*, 5Hz modulation *MF*, and 2Hz modulation *MS* were separately applied to each of the 3 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*, *CMS*. Each stimulus name indicates the feature and envelope shape; for example *PRD* contains a pitch ramp down. Linear rates of change in the envelope parameters (for ramps and modulations) correspond with linear rates of change in ST, dB, and spectral centroid in the resultant stimuli. A further 32 stimuli were then generated by combining the 8 L and C envelopes with the 4 P envelopes in a pairwise manner, shown in Table 3.1. 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 2s in duration, and each of the flat sections in the envelope shapes are 0.5s. These flat sections were included to give the participants a clear start and destination value for each feature envelope.

3.2.2 Parameter selection

Regarding pitch, the literature discussed in Section 2.4.1 indicates that the differences in SFF due to sex is large enough to warrant different base pitches

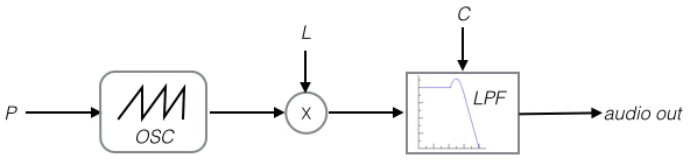
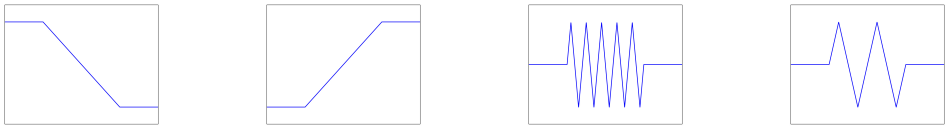


Figure 3.1: Block diagram of the synthesis model used to generate the stimuli. P = pitch, OSC = sawtooth oscillator, L = gain, LPF = 2nd order IIR low pass filter, C = cutoff frequency. The parameters relate to the vocal features of interest: F_0, loudness and spectral centroid.



(a) (b) (c) (d)

Figure 3.2: Temporal envelope shapes used for the stimuli. All envelopes are made up of two 0.5s sections at the start and end, with a 1s middle section. (a) Ramp down (RD), (b) Ramp up (RU), (c) 5Hz modulation (MF), (d) 2Hz modulation (MS).

	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*

Table 3.1: Identifiers for the thirty–two double–feature stimuli. These are produced by combining each of the four pitch (P) envelopes with each of the loudness (L) and spectral centroid (C) envelopes.

for male and female participants. We therefore chose to use base pitches of 110Hz for males and 220Hz for females. These values are comfortably within the typical producible ranges presented in the above mentioned studies, and have the same musical note (A). The ranges for FFR are also different for male and female participants, therefore the pitch envelope ramps are based on 12ST deviations from the SFF. We opted to test a natural vibrato rate and slower than normal rate, at 2Hz and 5Hz. The depth is 3ST (centered around the mean F_0) for both modulation rates which equated to a pitch change of 30ST/s at 5Hz and 12ST/s at 2Hz. These parameters give the participants a realistic chance of accurately imitating the modulations. The range of the loudness ramp envelopes (LRU and LRD) is 24dB. The extent for LMF and

LMS envelopes is +/− 6dB (total min:max extent of 12dB). This gives a maximum rate of change of 120dB/sec at 5Hz and 48dB/sec at 2Hz.

As previously mentioned, the spectral shape is controlled by a variable cutoff frequency on a low pass resonant filter. Without knowing how people might imitate a varying cutoff frequency it is difficult to decide on realistically producible parameter values for these envelope shapes. It is also not clear how the simple filter model used to create the stimuli might map to the spectral shapes produced by the voice. However, the sound of a modulating cutoff frequency is somewhat similar to a 'wah–wah' sound. It is therefore conceivable that people might use the diphthong /ɑʊ/ (such as in the word 'bout') to create this effect. Using the 'wah-wah' example, a periodic modulation of spectral centroid can be achieved by periodically repeating /ɑʊ/. The results from Gay [1968] show that this diphthong glide can be voiced at moderate and fast speaking rates in a mean duration of 112ms and 98ms respectively. Ferragne and Pellegrino [2010] present mean values for males as: /ʊ/: 406Hz and 1358Hz; /ɑ/ 687Hz and 1477Hz, for F1 and F2 respectively. Lloyd [2005] gives F1 and F2 values for both males and females, as: /ʊ/ male: 286Hz and 1091Hz; /ʊ/ female: 364Hz and 1303Hz; /ɑ/ male: 731Hz and 1550Hz; /ɑ/ female: 951Hz and 1819Hz. The range used for both the ramp and modulation shapes in the stimuli is 300Hz to 1.3KHz, and is the same for both male and female participants. This corresponds to spectral centroid ranges of approximately 300Hz–900Hz for males and 400Hz–1kHz for females (note the difference between male and female is due to the SFF of the stimuli), which is comfortably within the producible ranges for speech [Přibil and Přibilová, 2012]. At a modulation rate of 5Hz the duration of a complete glide for the /ɑʊ/ diphthong is 100ms, and at 2Hz it is 250ms: both are suitably within the producible range.

3.2.3 Participants

Nineteen participants took part in the study. Of these, 16 were male and 3 were female. All of the participants had some experience in computer based music production (this was a stated prerequisite during recruitment), and over 5 years experience playing an instrument. The participant ages were 18–25 (n=2), 26–35 (n=13), and 36–45 (n=4).

3.2.4 Procedure

The study took place in an acoustically treated, sound deadened room. The recording chain was an AKG C414 microphone (cardioid polar pattern, low cut disabled, no pad engaged) and an Apogee Duet 2 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 monitor controller and PMC AML monitors. All audio was recorded at a sample rate of 44.1KHz 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 (Figure 3.3). 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.

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.

3.2.5 Feature extraction

The imitation files were manually edited to remove sections of silence (or more accurately, 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 barycentre of the spectrum, using the definition given by Peeters [2004]; loudness was calculated in sones, based on an implementation of the the loudness model by Moore et al. [1997], described by Peeters [2004]. All features were

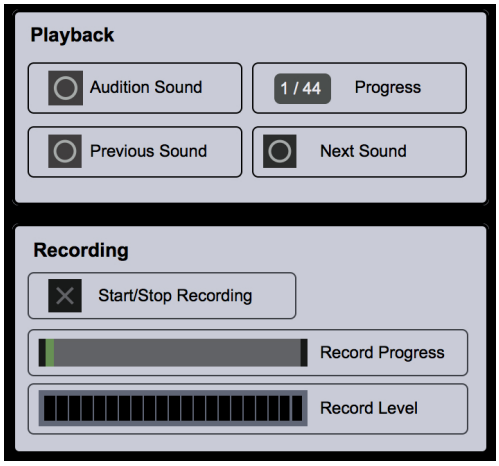


Figure 3.3: The participant–facing graphical user interface used for the vocal imitation study.

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.

3.2.6 Parameter extraction

Our goal is to test for the effect of single and double feature envelopes on imitation accuracy, therefore we require metrics 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 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.

3.2.6.1 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 6th order IIR filter with a cutoff of 10Hz and 5Hz for imitations of the *MF* and *MS* envelopes respectively.

2. Locate local maxima using a peak–picking algorithm.

3. Interpolate the maxima positions using quadratic interpolation to improve the rate calculation accuracy.

4. Remove any neighbouring maximum within the minimum period threshold (0.1 seconds for 5Hz and 0.2 seconds for 2Hz), keeping the greater maximum.

5. Find the minima between the maxima and (quadratically) interpolate the values.

6. Find the modulation area (first and last half–cycle with an extent $> 1/6$ of the extent in the stimulus, from the mean value). This is to remove any flat start and end sections in the imitation.

7. Calculate the per cycle rate (Figure 3.4): this is taken as the inverse of the distance between two maxima/minima [Dromey et al., 2003; Ferrante, 2011; Prame, 1994]. Note - whether minima or maxima are used will depend on whether the modulation area begins with a maxima or minima

8. Calculate the per cycle extent (Figure 3.4): 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 threshold, i.e. the participant had failed to vocalise a suitable modulation. This is a relatively small proportion of the imitations, however we note that they were mostly for the double–feature imitations of pitch (n=13) and loudness (n=9) envelopes. These cases were removed from the analysis. As an alternative approach to calculating modulation rate, we applied an FFT based method by picking the peak magnitude bin from a discrete Fourier transform

of the feature vector, however the periodicity was typically too noisy to give satisfactory results.

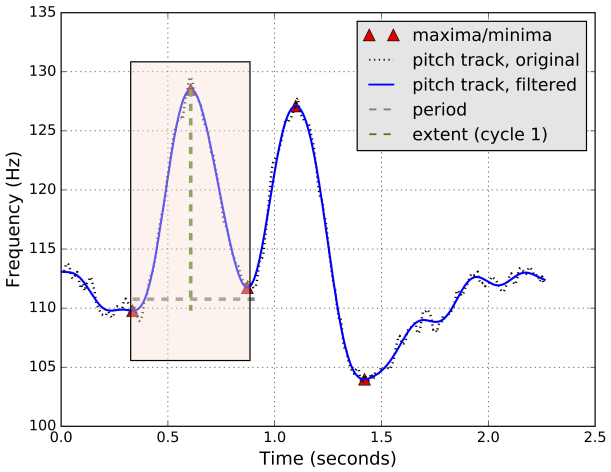


Figure 3.4: F_0 of one participant's imitation of the *PMS* envelope (2Hz modulation). The modulation rate is calculated as the inverse of the distance between two minima. Extent is the difference between highest and lowest values in a cycle (measured in ST). The shaded area highlights a single modulation cycle.

3.2.6.2 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 and taking the error, 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 them. 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 used to generate the stimuli are piecewise linear functions (see Figure 3.2). We therefore fit the frame–wise features of each imitation to such a function, to determine the range and slope parameters.

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 Figure 3.5). We then fit a continuous piecewise model that consists of 2 knots (k_1 and k_2), and where

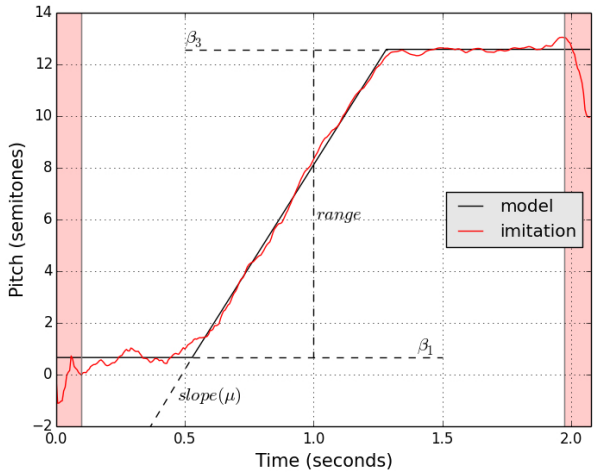


Figure 3.5: Pitch track (in ST) of one participant's imitation of the *PRU* envelope, overlaid with the fitted model. The shaded sections (first and last 5%) are ignored for the model fitting as they tend to have a large error due to variation as people settle on a pitch and end a vocalisation.

the slope for pieces 1 and 3 is 0. This model is given by:

$$y = \begin{cases} \beta_1 + \epsilon(\chi) & , \chi < k_1 \\ \beta_2 + \mu\chi + \epsilon(\chi) & , k_1 \leq \chi \leq k_2 \\ \beta_3 + \epsilon(\chi) & , \chi > k_2, \end{cases} \tag{3.1}$$

where $\beta_1, \beta_2, \beta_3$ 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 model is fitted by iterating through all possible integer values of χ for k_1, k_2, where $k_1 < k_2$ and each piece consists of at least 5 consecutive frames, minimising the sum of squared error (SSE), i.e. $\sum_{\chi=\lfloor N\times 0.05 \rfloor}^{\lfloor N\times 0.95 \rfloor} \epsilon(\chi)$, where N = number of feature frames for a given imitation. See Algorithm 1 for the operational process used to fit the model, and figure 3.5 for an example of such a model fitted to an imitated pitch envelope. 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 μ 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 $\frac{max(\beta_1,\beta_3)}{min(\beta_1,\beta_3)}$, and the 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 3–piece continuous linear functions.

Algorithm 1 Continuous linear plateau–ramp–plateau model fitting

procedure FITMODEL($x_{0...N}, y_{0...N}$)
  $start \leftarrow$ `truncate`$(N \times 0.05)$
  $end \leftarrow$ `truncate`$(N \times 0.95)$
  $best \leftarrow 2^{32} - 1$ ▷ larger than worst expected model fit
  for k_1 `from` $start + 5$ `to` $end - 10$ **do**
    for k_2 `from` $k_1 + 5$ `to` $end - 5$ **do**
      $\beta_1 \leftarrow \text{mean}(y_{start}...y_{k_1})$
      $\beta_2, \mu \leftarrow \text{linearRegression}(x_{k_1...k_2}, y_{k_1...k_2})$
      $\beta_3 \leftarrow \text{mean}(y_{k_2}...y_{end})$
      $error \leftarrow \text{calculateSSE}(x_{0...N}, y_{0...N}, \beta_{1...3}, \mu, k_1, k_2)$
      if $error < best$ **then**
        $best \leftarrow error$
        storeModelParameters$(\beta_{1...3}, \mu, k_1, k_2)$
function CALCULATESSE($x_{0...N}, y_{0...N}, \beta_{1...3}, \mu, k_1, k_2$)
  $error \leftarrow 0.0$
  for i `from` 0 `to` k_1 **do**
    $error = error + (y_i - \beta_1)^2$
  for i `from` k_1 `to` k_2 **do**
    $error = error + (y_i - (\beta_2 + (x_i \mu)))^2$
  for i `from` k_2 `to` N **do**
    $error = error + (y_i - \beta_3)^2$
  return *error*

This method is based on the assumption that participants did indeed imitate a linear function for the ramp portion of the envelope. To test this we first visually inspected each imitation feature vector plotted over its respective model (as shown in Figure 3.5). We then tested the linearity of the imitated ramps 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 < 0.2s, or the slope was in the opposite direction to that in the stimulus. These were mostly for the double–feature imitations of loudness (n=16) and spectral centroid (n=37) ramps when combined with pitch modulation envelopes. These cases were removed from the analysis because no meaningful parameters could be extracted for comparison. Two alternative means of modelling imitations of the ramp envelopes were also investigated: *i*) using the `breakpoints` function from the `strucchange` package [Zeileis et al., 2001] for `R` [R Core Team, 2016], and *ii*) an iterative mean squared error optimi-

sation method from Crawley [2012]. However, both these approaches do not constrain the piecewise function to be continuous, therefore the results were deemed unsuitable given the continuous nature of the extracted frame–wise features.

3.3 Statistical analysis

3.3.1 Single feature imitations

We tested for the effect of two factors on imitation accuracy: envelope and feature, using linear mixed effect regression (LMER). 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 effectively handle missing data (removal of the failed imitation cases resulted in an unbalanced dataset).

Separate LMER 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, 2016] and found no notable 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`. 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` (FDR = 5%).

3.3.1.1 Ramp envelopes

A full factorial ANOVA was conducted on the range and slope LMER 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 `feature` and `envelope` ($F(2, 93) = 5.1$, $p_{adj} = 0.024$). 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 interac-

tion 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 Figure 3.6a, 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 (LRD) and 0.83 (LRU). The imitation ranges are generally larger than the stimulus 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), however these differences are small in real terms, equating to a difference of only 36 cents. Imitations of pitch range are more accurate and have much lower variance than for loudness and spectral centroid.

In terms of ramp slope (Figure 3.6b), we found no significant interaction between `envelope` and `feature`, and no significant effect of `envelope` on imitation accuracy. There is however 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 stimulus slopes for all features and envelopes.

3.3.1.2 Modulation envelopes

As with the ramp envelopes, a full factorial ANOVA was conducted on the rate and extent LMER 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 (Figure 3.6c). 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 2Hz envelopes compared to 5Hz. This effect is observed for all features, but is largest for spectral centroid.

It is worth noting that an alternative, and perhaps 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 2Hz equates to an error of 1Hz, whereas a ratio of 1.5 at 5Hz equates to an error of 2.5Hz. Conceivably these errors are therefore not comparable 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 significant interaction between `feature` and `envelope`, and no significant effect of `feature`, but there is a significant effect of `envelope` ($F(1, 94) = 7.9, p_{adj} = 0.024$). We note that although the interaction is not significant, it is marginally so ($F(2, 94) = 2.4, p_{adj} = 0.093$), and there is clearly some effect of this, as can be seen in Figure 3.6d. This can be explained in the small difference between fast and slow modulation rates for loudness (where the extent for the imitations is consistently lower than for the stimuli), and the fact that a slower modulation rate appears to lead to a larger imitation extent for pitch and spectral centroid. The lack of significance for this effect is likely due to the large variance in both pitch and spectral centroid imitations. Overall, participants performed best when imitating the extent for *PMS*, indicating a positive effect of slower rate for pitch. This effect is not observed for loudness, or spectral centroid where there is an asymmetry in the direction of error fast and slow rates, but a similar absolute error.

3.3.2 Double–feature imitations

In this section we report how the accuracy of each single feature envelope (e.g. *PRU*) changes when it is combined with envelopes of another feature (e.g. each of the spectral centroid and loudness envelopes). We perform the analysis by modelling the imitation accuracy for each feature separately, again using LMER. This gives 8 LMER models for each feature: 4 for each set of controls by 2 parameters. Each model was fitted with a fixed effect of `stimulus type`, and a random intercept for each `participant`.

The factor of `stimulus type` has 3 levels for pitch (pitch, pitch+loudness, pitch+spectral centroid), and 2 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 of the LMER models to a one way type III ANOVA

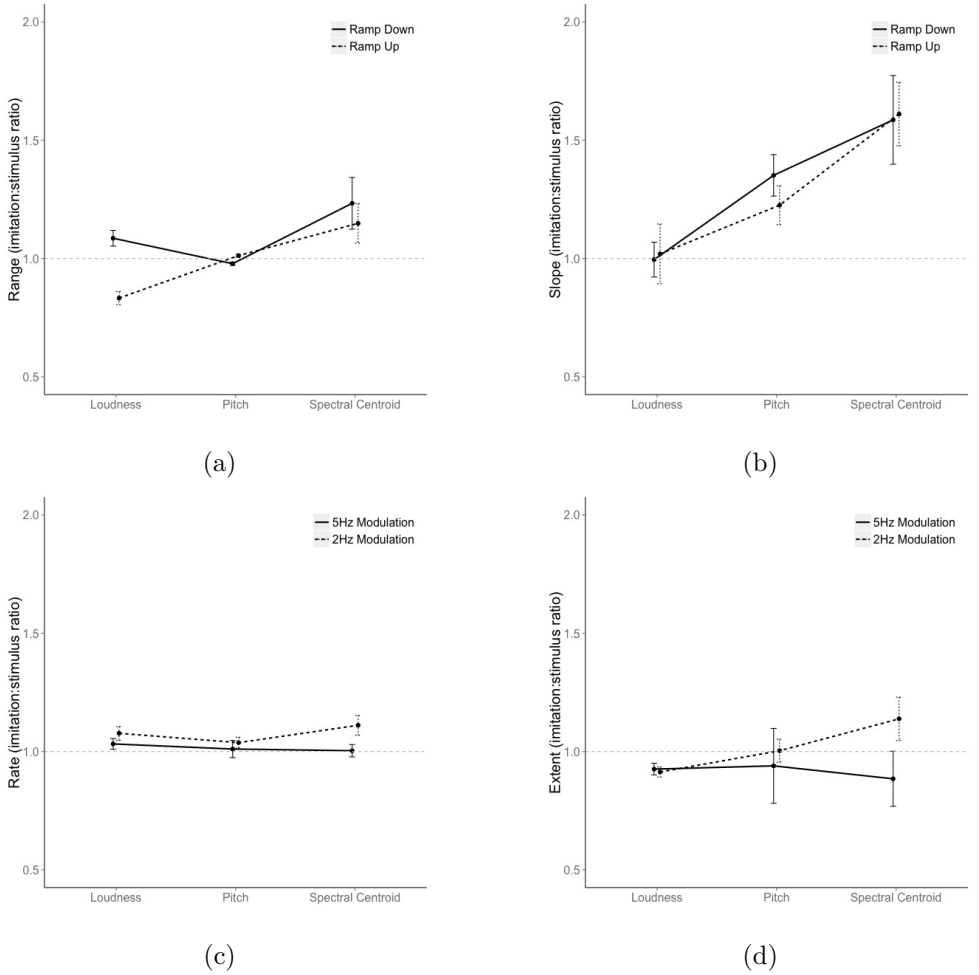


Figure 3.6: (a) Range, (b) slope, (c) rate, and (d) extent accuracy for imitations of the 12 control stimuli (2 ramp envelopes and 2 modulation envelopes for each of the 3 features), across all participants. Values are means across participants with standard error bars.

using Satterthwaite's approximation for denominator degrees of freedom, with a factor of `stimulus type`. As with the single–feature analysis, all p–values for each feature were adjusted using the Benjamini & Hochberg false discovery rate correction from the `p.adjust` function in `R` (FDR = 5%). The results for pitch, loudness, and spectral centroid are given in Tables 3.2, 3.4, and 3.6 respectively, and for completeness we also include the full results for every level of double–feature stimuli (not averaged over `stimulus type`) in Tables 3.3, 3.5, and 3.7.

3.3.2.1 Pitch

The results show no significant effect of the double-feature stimuli on accuracy of the range, rate, or extent parameters (Tables 3.2 and 3.3). 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 pitch+loudness and pitch+spectral centroid stimulus types.

Figures 3.7a and 3.7b 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 (Tables 3.5 and 3.7).

Although there are no significant effects of the double–feature stimulus types for the range, rate and extent 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 5Hz and 2Hz 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 5Hz pitch envelopes (*PMF*), and the double–feature stimuli appear to have a larger effect on extent accuracy for the *PMF* envelope than for *PMS*.

	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]

Table 3.2: 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.

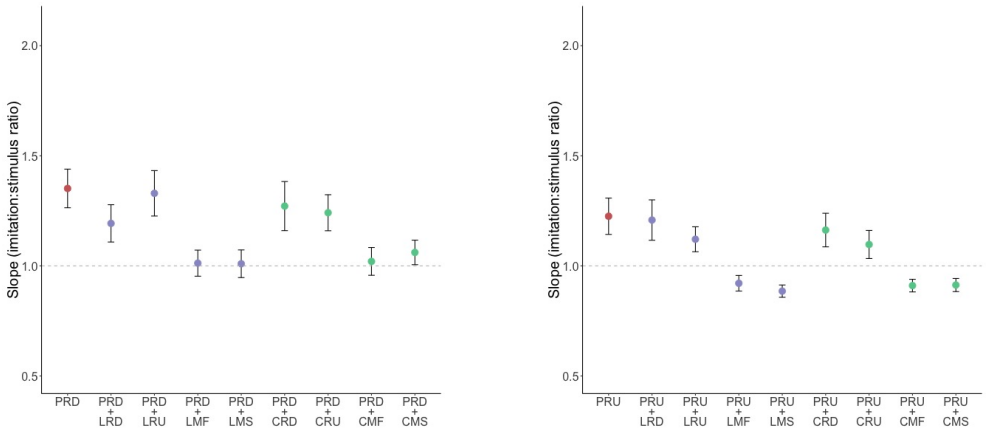
		Double-feature stimuli (combined with control)							
	Control	*LRD*	*LRU*	*LMF*	*LMS*	*CRD*	*CRU*	*CMF*	*CMS*
PRD									
Slope	1.35 [0.09]	1.19 [0.08]	1.33 [0.10]	1.01 [0.06]	1.01 [0.06]	1.27 [0.11]	1.24 [0.08]	1.02 [0.06]	1.06 [0.06]
Range	0.98 [0.01]	0.98 [0.01]	0.99 [0.01]	1.01 [0.01]	1.01 [0.01]	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 3.3: 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.

3.3.2.2 Loudness

The results for loudness envelopes show 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 3.4. Figure 3.8a 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 Figure 3.8b.

There are no statistically significant differences between stimulus types for



(a) (b)

Figure 3.7: 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 and envelopes with standard error bars.

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 (Table 3.5).

	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]**

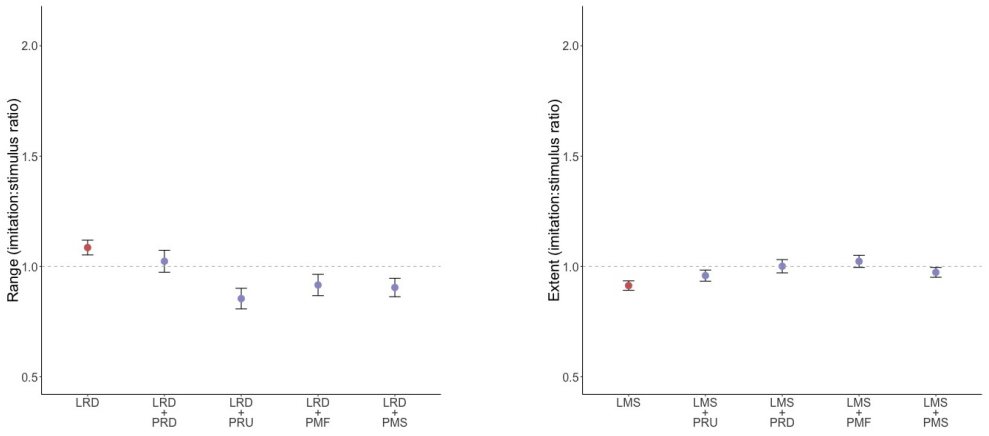
Table 3.4: Means (and standard errors) of loudness imitation accuracy for loudness vs. pitch+loudness stimulus types. Bold values indicate a significant effect of stimulus type (e.g. single vs. double–feature) on imitation accuracy.

3.3.2.3 Spectral centroid

In terms of spectral centroid envelopes, we found no significant effect of stimulus type on imitation accuracy for any of the parameters (Table 3.6). Interestingly, there is notably 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 amount of variance in the single–feature imitations.

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



(a) (b)

Figure 3.8: 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.

	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]

Table 3.6: Means (and standard errors) of spectral centroid imitation accuracy for spectral centroid vs. pitch+spectral centroid stimulus types.

		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]

Table 3.7: Results (mean and standard error) for imitations of the four spectral centroid envelopes, individually (*Control*) and when combined with each of the pitch envelopes.

3.4 Discussion

3.4.1 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 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. Interestingly, these results contrast those for 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 -24 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 2 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 that those in Mürbe et al. [2004], where professional singers exhibited a mean interval error of 19 cents when singing a slow, legato arpeggiated 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 ones 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 SPL 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, 2016] for the phoneme /ɑ/. Ectophonic loudness is the perceived loudness of sounds external to the body [Yadav, 2016], which also 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 Figure 3.6a).

The fact that spectral centroid and sharpness both correlate with brightness [Ilkowska and Miśkiewicz, 2006; Schubert and Wolfe, 2006] allows us to compare the spectral centroid imitations of our participants to those of Lemaitre et al. [2016b], 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. 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 300–900Hz for males and 400Hz–1kHz for females. These appear to fit 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 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 a shortening of the ramp envelopes (participants imitated the correct range over a shorter 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 modulation rate, with mean ratios for each stimulus ranging from 1.00 (*CMF*) to 1.11(*CMS*). As noted in Section 3.3.1.2, when measured as a ratio the modulation rate error is higher for the 2Hz stimuli than for 5Hz, across all features. This shows that for the two rates we have tested here, relative error appears to be inversely proportional to modulation rate, in contrast to previous findings on accuracy of singing vibrato at rates of 3 and 5Hz, where no such relationship was observed [King and Horii, 1993]. This is likely to be influenced by two factors: Firstly, the 5Hz 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; Prame, 1994; Sundberg, 1994b]; secondly, 2Hz is such a slow modulation rate that slight deviations in timing would cause a relatively large error compared to the 5Hz 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 2Hz and -18 cents at 5Hz (target extent for both rates was 3ST). The difference between modulation rates indicated that participants were more able to imitate the target range at 2Hz; 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 previously discussed.

3.4.2 What happens to imitation accuracy when people are asked to vocalise multiple feature envelopes simultaneously?

In general imitation accuracy was not significantly different between the single and double–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 [Alku et al., 2002; Gramming, 1991; Gramming et al., 1988]. This has also been shown to exist in singing [Sundberg et al., 1993] and for the sus-

tained vowel /ɑ/ [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 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 2Hz and 5Hz (for example the pitch rate accuracy ratio for both *PMF+LMS* and *PMF+CMS* is 0.73, which equates to 3.65Hz). This indicates an inability to accurately vocalise multiple feature envelopes with different modulation rates, as might be expected.

3.4.3 Effects of singing experience and sex

The effects of singing experience and sex are not within the scope of this study, 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 Section 3.2.2 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 years or more. We tested for the effect of both singing training and sex on the imitation accuracy of each parameter using LMER models with `participant` as a random effect and the following fixed effects: `feature`, `envelope`, `singing experience`, and `sex`. A full factorial ANOVA on the LMER models indicated no significant effects of either `singing experience` or `sex` on the imitation accuracy of any of the parameters. It is worth mentioning that this does not mean that singing experience 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.

3.4.4 Participant feedback

The participants completed a short feedback questionnaire following the study. A breakdown of the responses is shown in Table 3.8. 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 or 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 5Hz envelopes, however it is not the case for 2Hz 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.

	# 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"	3	3	13

Table 3.8: 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.

3.5 Summary and conclusions

The findings of this study complement previous work on vocal imitation 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. 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, even when they are not singers. This is an encouraging result that highlights the potential of using the voice as a medium for sound search in QBV applications. Most importantly though, the results of this study suggest that the participants were able to exercise control over 2 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 feature 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.
4. There is a significant effect of feature on slope accuracy: loudness is most accurate, followed by pitch and spectral centroid.
5. Participants generally imitated modulation rates of 2Hz and 5Hz with high accuracy for all features.
6. Modulation extent is more accurate for pitch than for loudness and spectral centroid.
7. There are clear effects of modulation rate (2Hz vs. 5Hz) on imitation accuracy: higher (and overestimated) imitation rates occur at the slower modulation rate.
8. A similar effect is observed for pitch and spectral centroid extent: higher extents are vocalised at the lower modulation rate (2Hz). This effect is not observed for loudness.
9. 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.
10. 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.

In this chapter we have investigated vocal imitation accuracy using computational methods. Whilst this serves to quantify the feature–level accuracy of the imitations, as was the aim of this experiment, it does not necessarily inform us about vocal imitation accuracy in perceptual terms. For example, a less accurate imitation in terms of the features and parameters used here may be perceptually more similar to the stimuli than a more accurate imitation. This scenario is conceivable if there are specificities or voice quality indicators beyond the low–level features tested here that contribute to vocal imitation accuracy. In addition, the stimuli used in this experiment were synthesised to vary only in the low–level features of interest. As such, they are not 'real–world' examples of sounds that might exist in sample libraries for music production. Now that we have established the potential for the voice as a medium for QBE, we turn to a more real–world application that was identified in Section 1.1: QBV for percussion sounds, and investigate the ability of musicians to vocalise sounds in terms of perceptual similarity.

Chapter 4

Vocal imitation of percussion sounds

In the previous chapter we investigated vocal control of elementary synthesised sounds. This allowed us to control the individual temporal envelopes for each of the acoustic features of interest, however, as noted in Section 3.5, these types of sounds do not necessarily represent the typical sounds that might be searched for in a music producers sound library. As discussed in Chapters 1 and 2, searching for drum sounds is a core part of the electronic music production work flow that might benefit from more intuitive and efficient search methods such as QBV [Andersen and Grote, 2015], therefore in this chapter we focus on the vocal imitation of such sounds. In particular, we are interested in the *perceptual* quality of vocal imitations, for example, whether musicians can vocally imitate percussion sounds such that listeners are able to identify the imitated sound from a set of same–category sounds (e.g. kicks, snares, toms etc.).

There has been a small but notable amount of research into vocalised percussion sounds such as beatboxing, which we discussed in Chapter 2, however much of this is concerned with either classifying vocalisations into drum categories [Kapur et al., 2004; Sinyor et al., 2005], understanding the linguistic–related mechanisms of producing such sounds [Blaylock et al., 2017; Guinn and Nazarov, 2018; Proctor et al., 2013], or understanding the relation between vocalised percussion and actual drum sounds based on acoustic features [Lederer, 2005; Stowell, 2010]. In addition, the perceptual relevance and communicative power of vocal imitations has been researched for non–percussion,

particularly environmental sounds [Lemaitre and Rocchesso, 2014; Lemaitre et al., 2011], however this type of analysis has not yet been conducted on vocalised percussion sounds. With the exception of Patel and Iversen [2003], we are not aware of any research that addresses the perceptual similarity between vocalised percussion sounds and their real–world counterparts (i.e. the imitated sounds). In this chapter we aim to address this knowledge gap by investigating the perceptual similarity between vocal imitations and same–category drum sounds, and how this might differ between sounds within and across drum categories.

The chapter is laid out as follows: the research questions and outline of the study are given in Section 4.1. The method and sounds used for the vocal imitation task are described in Section 4.2, and the method for collecting the perceptual similarity ratings is outlined in Section 4.3. Results are given and discussed in Section 4.4, in terms of how listeners were able to identify the imitated sounds from the imitations (Section 4.4.3) and the similarity ratings between imitations and same–category sounds (Section 4.4.4). Finally, summary conclusions are presented in Section 4.5.

4.1 Experiment outline and research questions

The work in this chapter is split into 2 parts: a vocal production task and a listening test. In the first part (vocal production), a group of 14 participants were tasked with imitating 30 drum sounds from 5 categories (cymbals, hats, kicks, snares, and toms), with 6 sounds in each category. The categories were chosen because they represent the 5 most common types of drum sounds used in Western popular music, and as we discuss in Section 4.2.1 the sounds were chosen to be a representative range of drums from each category. As with the previous chapter, we recruited musically trained participants to provide the imitations, as they are the target user group for the intended application of this work (QBV).

In the second part (listening test), listeners rated the similarity between the vocal imitations and 6 same–category drum sounds. We limited the scope of this study to same–category sounds, because whereas the problem of classifying vocalised percussion sounds into drum categories has been addressed in previous work [Kapur et al., 2004; Sinyor et al., 2005] (where relatively elementary methods have demonstrated very good results [classification accuracy

$> 90\%$], therefore we do not consider this a problem on which we should focus our efforts), the ability for algorithms or people to differentiate between same–category percussion sounds is unknown. If we are to apply the voice to search for such sounds then the potential for this ability must first be demonstrated. The core research questions are as follows:

1. Are musicians able to differentiate their vocal imitations of same–category percussion sounds such that listeners are able to identify the sounds being imitated?

2. How does this ability differ between categories of percussion sounds: are certain categories more *imitable* than others?

3. Will imitated sounds receive higher similarity ratings when rated against their respective imitations, compared to when they are rated against imitations of other sounds?

4.2 Vocal production task: recording the imitations

4.2.1 Selecting the drum sounds

We are specifically interested in drum samples that might be typically used by music producers. For this reason we selected the stimuli from a range of high quality drum sound recordings from the commercial FXpansion[1] *BFD3 Core* and *8BitKit* sample libraries. These contain recordings of classic acoustic drums (such as Mapleworks and Bosphorus) and popular electronic drum machines (such as the Roland TR808 and Roger Linn LM–1). The libraries contain recordings of each drum at up to 127 velocity levels, therefore we removed all duplicate velocity level samples, taking only the median velocity recording for each drum. We then selected only the drums for each of 5 categories: cymbals (n=99), hats (n=88), kicks (n=42); snares (n=60); toms (n=158). This gives 447 drum samples in total, which is too many sounds for each participant to imitate, so we selected a representative sample of 6 sounds from each category.

Our criteria for selecting the stimuli was to have a broad range of drum samples from each category to reflect the variety of sounds in the sample library.

[1] https://www.fxpansion.com

To achieve this we extracted samples from each category based on their similarity, using the *auditory image*[2] based drum similarity method from Pampalk et al. [2008]. This method has been shown to be highly correlated with perceptual ratings of similarity for same–category drum samples, so it is ideal for our task. In summary, similarity between two drum sounds is measured as the Euclidean distance between their vectorised auditory images. An auditory image is essentially a spectrogram image representation of a sound, with three dimensions: time, frequency and loudness. Pampalk et al. [2008] investigated a range of parameter settings for generating auditory images, and found the following settings gave similarity metrics that were most highly correlated with listener's perceptual ratings of drum sound similarity.

1. STFT size of 4096 samples (at 44.1kHz) and hop size of 512 samples (87.5% overlap).

2. Frequency bins grouped using 72 bins on the Bark scale (implemented using a triangular window).

3. Loudness in dB, scaled using Terhardt's model for the outer and middle ear [Terhardt, 1979].

Using these parameters, we constructed a similarity matrix of all pairwise similarities between the auditory images of samples for each category. The similarity between two auditory images was measured according to the method in Pampalk et al. [2008], described in two stages:

1. If the 2 images have a different number of columns (i.e. are of different duration), the shorter image is zero–padded to the length of the longer one.

2. The distance between 2 images is then calculated for every possible time alignment of the images (based on time shift steps of 4096 samples). This is calculated as the Euclidean distance between the vectorised auditory images. The lowest distance measure is taken as the similarity between the sounds (i.e. the similarity at the best time alignment).

We selected a subset of six samples from each similarity matrix. For each category, we first selected a random seed sample (S_1). We then selected the most and least similar samples to S_1 (S_2 and S_6 respectively). Finally, we selected samples S_3, S_4, S_5 such that they were equally spaced in distance

[2] we use the term *auditory image* in keeping with the authors' description

between S_2 and S_6, with respect to S_1. This gave 6 samples spanning the range from most to least similar to the seed sample. We note that this method does not guarantee the most diverse selection of samples will be selected from a category, and the selected samples are dependent on the seed. However, in practice we found that the selected samples spanned a broad range of the sounds from each category. The auditory images for the extracted samples are given in Figure 4.1 and the details of each sound are summarised in Table 4.1. It should be noted that in Figure 4.1 the sample S_1 (seed) does not always appear most dissimilar to S_6, however this is likely due to the difference in durations (for clarity the sounds are all represented by the same size image, however they differ in duration and therefore require zero–padding to be compared in the above terms).

4.2.2 Participants

The 14 participants (herein referred to as the imitators) were musicians and music producers, all of whom reported no hearing or speech impairments. The age range was 26–43 and the median age was 30.5. One female took part in the study, and the remaining imitators were male. We note that the severe sex imbalance means we cannot generalise our results across the sexes, however as we are not concerned with testing any effects of sex we decided not to exclude imitations from the female imitator in the proceeding listening study.

All imitators were experienced in making music or producing music containing drum samples on a computer, and reported having more than 5 years' experience playing an instrument. Three imitators reported experience in singing but stated the voice was not their sole or main instrument. One imitator was an experienced beatboxer ($>$ 5 years' experience). All imitators reported having more than 2 years' experience in making music using synthesisers and/or samplers, and eleven had more than 5 years' experience doing this.

Category	#	Drum	Articulation
cymbal	1	Paiste 2002 Power Bell Ride	bow
cymbal	2	Paiste Signature Dry Heavy Ride	edge
cymbal	3	Bosphorus China 20	mallet, edge
cymbal	4	Bosphorus Splash 10	mallet, bell
cymbal	5	Paiste Signature Full Crash	bell
cymbal	6	Paiste 2002 Power Bell Ride	edge
hat	1	Paiste Signature	closed
hat	2	Sequential Circuits DrumTraks	closed
hat	3	Zildjian New Beats-Mastersound	half open
hat	4	Linn LM-1	open
hat	5	Paiste 2002	open
hat	6	Bosphorus	brush, half open
kick	1	DW Mardi Gras Sparkle	kick in (mic)
kick	2	DW Mardi Gras Sparkle	kick out (mic)
kick	3	Gretsch Purple	kick out (mic)
kick	4	Linn LM-1	–
kick	5	Mapleworks Custom	kick in (mic)
kick	6	Roland 808	long duration
snare	1	Roland 909	medium duration
snare	2	Oberheim DX	medium duration
snare	3	Roland 909	short duration
snare	4	Ludwig Hammered Supraphonic	half edge
snare	5	Tama Bell Brass	full hit
snare	6	Tama Bell Brass	half edge
tom	1	Gretsch Purple High	rim
tom	2	Gretsch Purple Mid	rim
tom	3	Mapleworks Custom Floor	mallet, full hit
tom	4	Mapleworks Custom Mid	rim
tom	5	Sequential Circuits DrumTraks	–
tom	6	Mapleworks Custom Mid	brush, rim

Table 4.1: Selected drum samples used as stimuli for the vocal imitations. Descriptions and articulations are taken from the sample library documentation and are not exhaustive descriptions of the recording setup, strike style or drum machine settings etc. Unless specified otherwise the acoustic drums were struck with a stick.

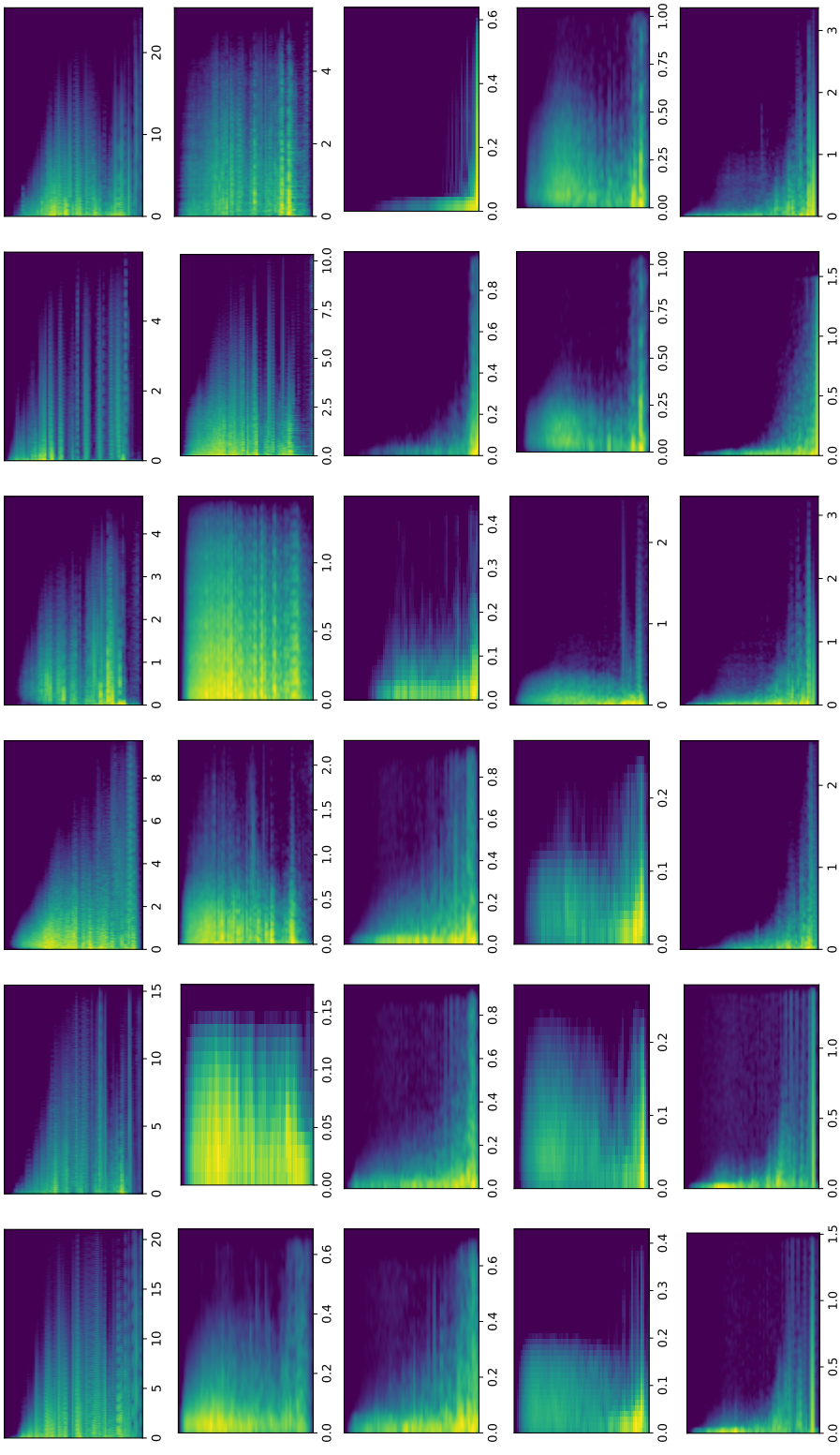


Figure 4.1: Auditory images of the drum stimuli. Rows 1–5 are for cymbals, hats, kicks, snares, and toms, respectively. Seed samples (S_1) are given in the first column. Remaining columns (S_2–S_6) are in descending order of similarity to the seed.

4.2.3 Procedure

The recordings took place in the same space and using the same equipment described in Chapter 3. The imitators were advised that the aim of the study was to listen to each drum sound and imitate it as accurately as possible using their voice. As in the previous vocal imitation study, the participants were seated at a computer and presented with a basic interface for auditioning the stimuli and recording their imitations (Figure 3.3). The lead researcher gave an overview of the interface and left the room for the duration of the study.

The order of the drum sounds was randomised. Each drum sound could be auditioned and rehearsed as many times as the imitator wanted. The imitators 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 before proceeding to the next drum sound. The imitators were advised that the final recording of each sound would be used for the analysis.

4.3 Listening study design

In this section we present the method for the listening study, where participants were asked to rate the perceptual similarity between the vocal imitations recorded in Section 4.2 and same–category drum sounds.

4.3.1 Participants

63 participants (herein referred to as the listeners) were recruited from professional and research networks. Of these, 46 were male and 16 were female (one chose to not disclose their sex). The age range was 18–66, with a median age of 30. All listeners reported no hearing impairments.

4.3.2 Stimuli and procedure

The 420 vocal imitations and 30 drum sounds from Section 4.2 were used as stimuli. For a given imitation, listeners rated the similarity between the imitation and the 6 same–category drum sounds, using a format based on the MUSHRA protocol described in Chapter 2. In the present study the known

reference was an imitation, and the 6 test sounds were the sounds from the same drum category as the imitation. Listeners rated the test sounds in terms of similarity to the imitation. The imitated sound can be considered a hidden reference, however we did not necessarily expect this to be rated as the most similar test sound, as it is possible that an imitation sounds more similar to one of the other drum sounds. This test format allowed us to collect individual ratings between an imitation and each of the test sounds, and additionally determine whether an imitation was similar enough to the imitated sound such that an independent listener could identify the imitated sound (by taking the top rated test sound).

The experiment was conducted remotely via a webpage built using the BeaqleJS framework [Kraft and Zölzer, 2014]. Each imitation was presented on a separate test page, made up of a reference (imitation) and 6 test items (same–category drum sounds). An example test page is given in Figure 4.2. Due to the large number of imitations, each listener only rated a random subset of 28 imitations plus 2 random duplicate imitations, giving 30 test pages each. The duplicate test pages were included to assess the intra–rater reliability of each participant. The listeners were instructed to rate the similarity of the test items with respect to the reference, using continuous unnumbered sliders from 'less similar' to 'more similar'. It was possible to navigate forward and backward through the test pages, adjust volume and loop the samples. The study took approximately 30 minutes to complete (approximately 1 minute per test page).

Each listener provided 180 similarity ratings (6 responses per test page, 30 test pages). There were 63 listeners in total, giving 1890 test pages and 11340 responses. We removed test pages if the listener gave the same rating to all 6 sounds: in every test page there was a notable difference between the most and least similar drum sounds, therefore this indicates that the listener did not follow the instructions properly. There were 80 invalid test pages (480 responses). The majority of these were from three listeners (25, 22 and 15 each). The web server was configured to maintain a balanced distribution of completed valid test pages per imitation: 408 imitations were rated four times and 12 were rated five times, excluding the duplicate test pages.

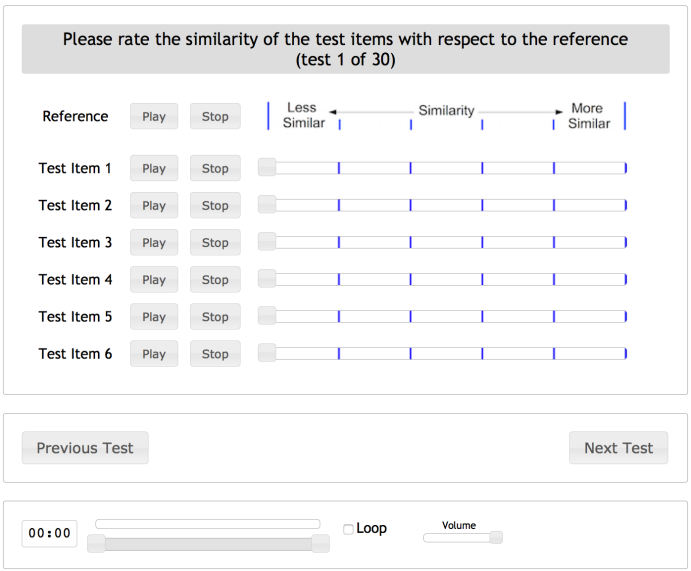


Figure 4.2: An example test page from the web based listening test.

4.4 Results and discussion

4.4.1 Intra–rater reliability

Listener reliability was assessed using the Spearman rank correlation between the two duplicate test pages for each listener. We used the ranks because we expected some variability due to the continuous response scale, and were mainly interested in whether participants could replicate the ordering of their responses, not the exact rating values. Reliable listeners were defined as those who were able to replicate their responses for at least one of the duplicates with $\rho >= 0.5$. We note that this value is somewhat arbitrary, but it indicates a large positive correlation [Cohen, 1988] hence was deemed suitable for the purposes of identifying unreliable listeners. There were 51 reliable listeners, for whom $\rho = 0.63/0.04$ (mean/standard error), giving 9126 responses from 1521 test pages.

4.4.2 Concordance of ratings (inter–rater agreement)

Concordant imitations are those for which there was agreement amongst the listeners regarding the similarity ratings. We computed Kendall's coefficient of concordance [Kendall and Smith, 1939] on the ranked ratings for each imitation, excluding those from unreliable listeners and duplicate test pages. The

mean coefficient for all imitations is 0.61 (standard error = 0.01), indicating strong to moderate agreement amongst the reliable listeners [Schmidt, 1997].

4.4.3 Identifying the imitated sounds

Here we assessed whether the listeners could correctly identify the imitated sound from the six same–category drum sounds. We therefore considered only the highest rated sound from each test page, and calculated the proportion of instances where this was also the imitated sound. Of 1419 completed test pages (excluding unreliable listeners and duplicates), there were 516 instances where the highest rated sound was the imitated sound, and 903 where the imitated sound did not receive the highest rating. This shows that listeners managed to identify the imitated sound with above chance accuracy, as per similar previous studies on vocalisations of everyday sounds [Lemaitre and Rocchesso, 2014] and text–based meanings [Perlman and Lupyan, 2017], however the overall identification accuracy was quite low, at 36.3%. Similar identification accuracy is observed when duplicates are included, at 36.5%. The method we used to select the drum sounds means that some drum sounds are more similar than others, and there should be high similarity between certain sounds, where we might expect some confusion in terms of identification accuracy. This appears to be the case: in 856 (60.3%) of the tests the imitated sound was rated first or second highest.

To investigate the effect of this confusion among the drum sounds, we constructed contingency tables for each drum class, which are given in Figure 4.3. This shows the proportion of times each sound was rated highest for each imitated sound. We conducted a one-way z–test for proportions on each matching imitated and rated sound pair, i.e. the diagonals in Figures 4.3a–4.3e. This tells us for which drum sounds listeners could identify the imitated sound. The tests within each contingency table were corrected for using the Benjamini and Hochberg false discovery rate correction (FDR = 5%). Of the 30 drum sounds, 16 were imitated such that listeners were able to identify the imitated sound with above chance accuracy. This is an encouraging finding considering that the imitations were only compared to same–category sounds, and remarkable given that the imitators were not specifically instructed to imitate the sounds such that they could be differentiated based on the imitations, or made aware that they would be used for the listening test. Identification accuracy ranged from 4% to 74%, which is considerably more variable that

the findings from Lemaitre and Rocchesso [2014] (75%–86% for imitations of identifiable everyday sounds and artificial sound effects respectively), which is likely due to there being greater similarity between our same–category stimuli. Imitations of all 6 hats, 4 cymbals and 3 kicks were correctly identified significantly above chance ($p_{adj} < 0.05$). Imitations of snares and toms were less well identified: Figures 4.3d and 4.3e show that certain sounds in the snare and tom categories were regularly rated highly, irrespective of the sound being imitated. This effect is largest for *snare6*, *tom4* and *tom6*. We will now discuss the imitation strategies and acoustic characteristics of the imitated sounds that may result in some sounds being more identifiable than others.

Listeners performed best at correctly identifying the imitated sounds for the hat category. Here the greatest confusion was between *hat1* – *hat2*, and *hat5* – *hat6*: *hat1* and *hat2* are both closed hats with very short decay times, whereas *hat5* and *hat6* are both open hats with relatively longer decay times (as shown in Figure 4.1). In addition, the interactions of the top and bottom plates are similar in *hat5* and *hat6*. In contrast, *hat3* is a half–open hat, with a unique and distinguishable amplitude envelope compared to the other hat sounds, and *hat4* has a clear, rhythmic repeating pattern in the decay, again giving it a unique temporal signature. It was apparent when listening to the imitations that the imitators tried to imitate these temporal cues, and listeners mostly struggled to identify the imitated sound when the temporal signatures for two sounds were similar. The differences in temporal features (such as duration and decay shape) are less extreme within the other drum categories, which may be why the hats were identified more successfully.

Identification accuracy was similar for the kicks and cymbals, with 3 and 4 sounds from each category correctly identified with above chance accuracy. There was notable confusion between *kick1*, *kick2* and *kick3*: these are all acoustic kick drums with similar resonance patterns and amplitude envelopes. In addition, they are notably brighter than the other kick sounds and have similar short, click–like attack characteristics (whereas the other kick sounds contain considerably less high frequency content). There was also confusion between *kick5* and *kick6*: these are lower in pitch compared to the other kick sounds, and are of a similar duration. Indeed, when listening to the imitations of *kick5* and *kick6* it was apparent that most imitators used similar techniques for both sounds, typically vocalising a stop consonant followed by a voiced decay.

Imitations of all three acoustic snare sounds (*snare4*, *snare5*, *snare6*) were (on average) rated as being most similar to sound *snare6*. This sound contains notably more (and a longer duration of) snare rattle compared to *snare4* and *snare5*, although they all contain some snare rattle. Many of the imitators used alveolar, post alveolar or retroflex fricatives to imitate the rattle, and in the imitations of *snare4* and *snare5* that were rated as being most similar to *snare6*, these fricatives were emphasised more compared to the other imitations. This indicates that imitators were not able to suitably differentiate their vocalisations with respect to the amount of snare rattle they were imitating, and instead any use of 'snare rattle–like' fricatives was generally associated with the sound containing the most amount of rattle. It has been previously demonstrated that imitators will emphasise the salient characteristics of a sound when imitating it [Lemaitre et al., 2011], therefore it may be that by over–emphasising the snare rattle on any rattling snare drum, the imitators inadvertently made their imitation sound most like the most 'rattly' snare sound.

As previously mentioned, when selecting the stimulus sounds we intentionally chose a range of sounds from each category, varying in the degree of similarity to a seed sample. We therefore expected some confusion in the identification of similar imitated sounds. This is evident for both the hats and kicks (and to some extent the cymbals), however the same effect is not observed for toms. Here we observe a kind of hubness [Flexer et al., 2012] in the perceptual space, where certain drum sounds appear to be closest to all imitations within a category. For example, *tom4* and *tom6* are consistently identified as being more similar to the imitations, even when they are not the sound being imitated. These sounds are from the same drum but played with different beaters (*tom4* = stick, *tom6* = brush – see Table 4.1). As such, the decay parts of the sounds are very similar, and it is mainly the attack portions that differ: *tom6* is slightly more noisy in the attack portion due to the use of a brush. When listening to the imitations of these sounds we found that some of the imitators adopted the same or very similar techniques to produce both sounds (7 of the 14 imitators used affricates to vocalise the attack for both sounds), making it difficult to identify which of the sounds were being imitated.

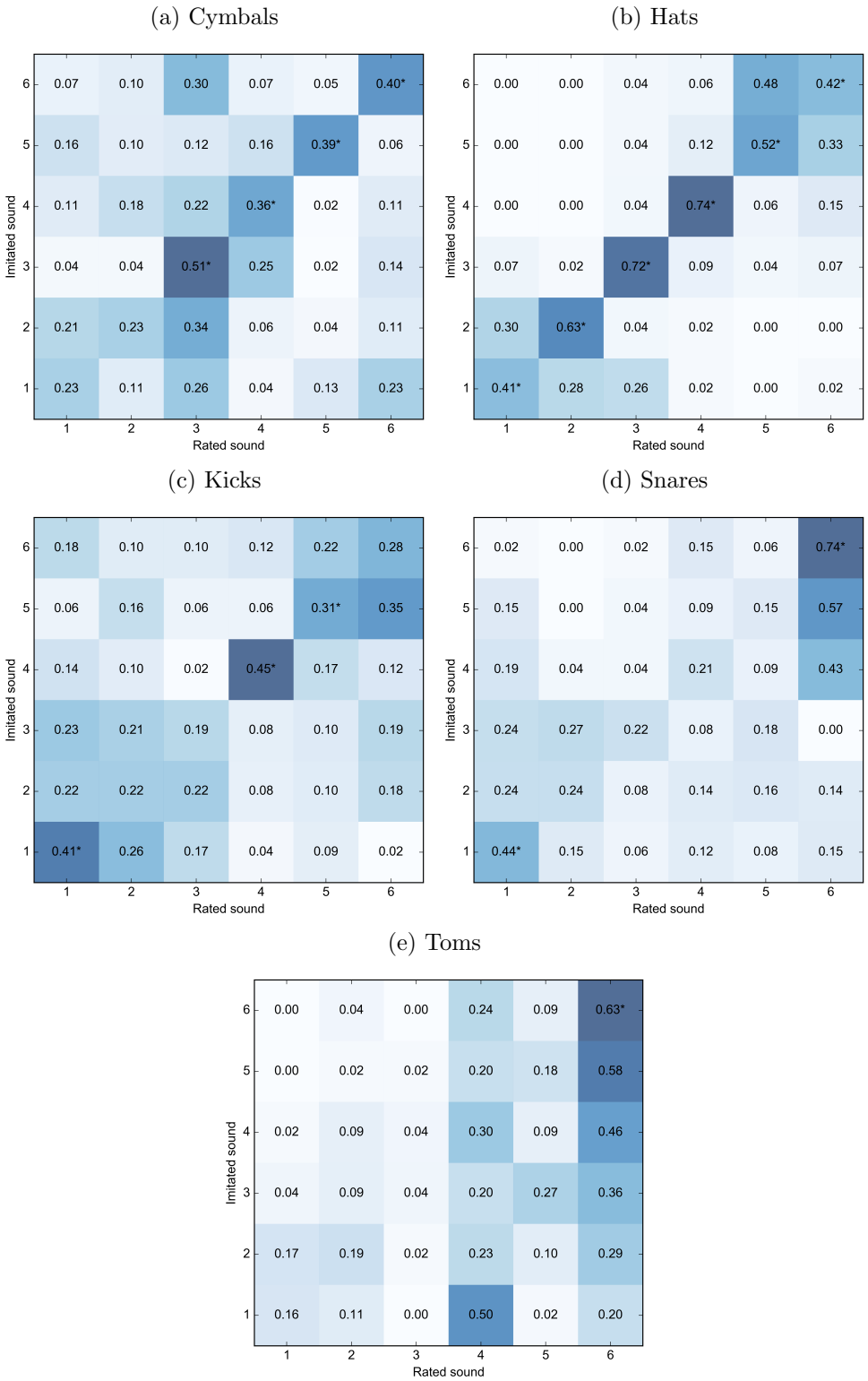


Figure 4.3: Contingency tables of the highest rated sound for each imitated sound, by drum category. Cell values and shading indicate the proportion (0–1) of tests for a given imitated sound where the rated sound received the highest rating. Asterisks in the diagonals indicate imitated sounds that were correctly identified with significantly above chance accuracy ($p_{adj} < 0.05$).

The sound of the stick striking the rim is particularly audible in *tom1* and *tom2*. It is difficult to vocalise this rim sound, and many of the imitators (12/14) seemingly did not attempt to do so: they used the same affricates or stop consonants for the attack portion as they did for the other tom sounds. Listeners may have failed to identify imitations of these toms because this distinguishable attack sound was generally not imitated. Indeed, there are two imitations for each of *tom1* and *tom2* where the rim sound was imitated using non–pulmonic consonant clicks: for these imitations, the imitated sounds were rated as being most similar when averaging across listeners. This highlights the importance of imitating the attack characteristics for a sound to be identified, at least for the types of tom sounds used in the present study. Attack time is a well known salient timbral descriptor for instrument or sound discrimination tasks [McAdams et al., 1995; Siedenburg et al., 2016], however it has been shown that people tend to not differentiate between subtle differences in attack times when imitating sounds [Lemaitre et al., 2016b]. This presents an interesting problem for modelling similarity between vocalisations and percussive sounds: if imitators apply the same imitation technique to vocalise perceptually different attack times then this descriptor may be of little practical use.

Finally, because the toms are pitched we expected pitch–accurate imitations to be correctly identified, assuming pitch as a salient feature. The F_0 for toms 1–6 is 84, 66, 79, 99, 66 and 102 Hz respectively. Most of the imitations for *tom1* and *tom2* were closest in pitch to the imitated sounds (10/14 for both). Therefore if the listeners based the similarity ratings mainly on pitch, we would expect these imitations to be correctly identified, or else confused with sounds of a similar pitch (the pitch differences between *tom1* and *tom3*, and *tom2* and *tom5* are very small). However, the majority of these imitations were rated as being most similar to *tom4* and *tom6*, which have a considerably different pitch. This indicates that the listeners did not necessarily use pitch as a cue for similarity between imitations and pitched percussion sounds.

4.4.4 Analysis of the similarity ratings

The results presented in the previous section suggest that vocalisations of certain sounds were more representative of the imitated sound than others, in terms of listeners being able to identify the imitated sound. In this section we investigate whether similarity ratings between imitations and imitated

(i.e. 'target') sounds are higher than for non–target sounds, and if this varies between drum categories. The listener similarity ratings were modelled with LMER using the `lme4` package [Bates et al., 2015] for `R`. LMER is well suited to this task given that all listeners did not provide ratings for all imitations but only a randomly–selected set of 28 imitations (giving an unbalanced dataset). In addition, it allows us to model the dependencies between ratings for each listener, drum category and imitator.

The full model was specified with `rating` as the response variable, fixed effects of `drum category` and `target` (a dummy variable indicating whether the rated sound was the imitated sound) with an interaction term, and random intercepts for each `listener` and `imitator`. Parameter estimates were then extracted for each combination of the fixed effect levels, and 95% Wald confidence intervals (CIs) were calculated. The results are given in Figure 4.4. For each drum category there is a statistically significant difference ($\alpha < 0.05$) between the ratings of target vs. non–target sounds, as indicated by the CIs. The effect of `target` is largest for hats, and similar for all the other categories. This shows that for all categories the target sounds were (at least on average) rated higher than the non–target sounds, regardless of whether they received the highest rating.

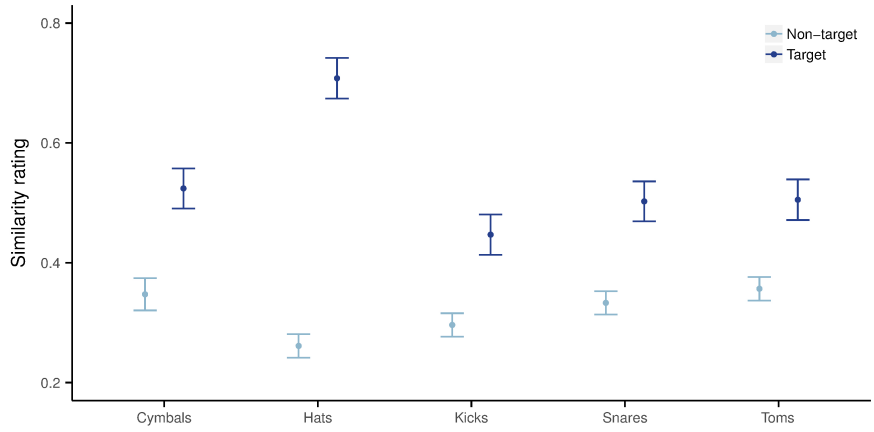


Figure 4.4: Comparison of similarity ratings between imitations and target vs. non–target sounds, by drum category. Values are mean rating parameter estimates with 95% Wald confidence intervals.

To establish whether this effect differed between drum sounds within each category we modified the above model, replacing the fixed effect of `drum`

`category` with `rated sound`. All ratings were first normalised for each listener to remove any bias from a listener not using the full range of the rating scale (in the MUSHRA standard, use of both hidden reference and anchor serves to encourage listeners to use the full range). The same parameter estimates were extracted, and are given in Table 4.2. For 22 of the 30 sounds the rated sound received a significantly higher ($\alpha < 0.05$) similarity rating when being rated against its respective imitation, compared to when rated against an imitation of another sound. For the 8 drum sounds where the upper and lower CIs overlap, the target sound was still rated higher on average than the non–target sounds.

Cymbals	1	2	3	4	5	6
target	0.54 (0.46, 0.62)	**0.47 (0.39, 0.55)**	0.60 (0.52, 0.68)	0.52 (0.43, 0.60)	0.49 (0.42, 0.57)	**0.53 (0.44, 0.61)**
non–target	0.34 (0.30, 0.38)	**0.38 (0.34, 0.43)**	0.44 (0.39, 0.48)	0.32 (0.27, 0.36)	0.19 (0.14, 0.23)	**0.42 (0.37, 0.46)**

Hats	1	2	3	4	5	6
target	0.68 (0.58, 0.75)	0.77 (0.69, 0.84)	0.69 (0.61, 0.77)	0.76 (0.68, 0.84)	0.72 (0.65, 0.80)	0.64 (0.56, 0.72)
non–target	0.22 (0.17, 0.26)	0.16 (0.11, 0.20)	0.28 (0.23, 0.32)	0.30 (0.25, 0.34)	0.33 (0.28, 0.37)	0.29 (0.25, 0.34)

Kicks	1	2	3	4	5	6
target	0.52 (0.44, 0.60)	**0.41 (0.34, 0.49)**	**0.34 (0.26, 0.42)**	0.48 (0.40, 0.57)	0.49 (0.41, 0.56)	0.45 (0.37, 0.53)
non–target	0.33 (0.28, 0.37)	**0.37 (0.32, 0.41)**	**0.30 (0.25, 0.34)**	0.20 (0.15, 0.24)	0.31 (0.27, 0.36)	0.28 (0.24, 0.33)

Snares	1	2	3	4	5	6
target	0.62 (0.55, 0.70)	0.48 (0.40, 0.56)	0.46 (0.38, 0.54)	**0.39 (0.31, 0.47)**	**0.42 (0.34, 0.50)**	0.63 (0.56, 0.71)
non–target	0.35 (0.31, 0.40)	0.32 (0.28, 0.37)	0.22 (0.18, 0.27)	**0.34 (0.29, 0.38)**	**0.38 (0.33, 0.42)**	0.38 (0.34, 0.43)

Toms	1	2	3	4	5	6
target	0.48 (0.40, 0.56)	0.40 (0.32, 0.48)	**0.33 (0.25, 0.41)**	**0.57 (0.49, 0.65)**	0.50 (0.42, 0.57)	0.75 (0.67, 0.83)
non–target	0.26 (0.22, 0.31)	0.25 (0.20, 0.29)	**0.26 (0.21, 0.30)**	**0.50 (0.45, 0.54)**	0.32 (0.28, 0.37)	0.55 (0.51, 0.60)

Table 4.2: Comparison of similarity ratings between imitations and target vs. non–target sounds, by drum sound. Values are mean rating parameter estimates with 95% CIs. Cases where the CIs overlap between conditions for each drum sound are given in bold.

This analysis compares the ratings for the imitated sound to the *average* ratings for the five same–category sounds that were not imitated. Therefore, even if one or two of the 'non–target' sounds were perceptually similar to the imitated sound, the less similar sounds bring down the average rating. The results in Table 4.2 illustrate that imitators were able to inadvertently identify and imitate the distinguishing characteristics of different sounds within each category, to an extent that enabled listeners to discard at least *some* of the sounds as less similar to the imitated sound than others, i.e. even when the 'target' sound was not rated highest, it was rarely rated low relative to the other sounds. This effect can be quantified in terms of the mean reciprocal rank (MRR) for the ratings of the 'target' sounds. A random ranking would give an MRR of 0.41, whereas we found the MRR of all target sounds to be

0.58, indicating that on average, the target sound was rated first or second. It is interesting to note that this is also the case for the sounds that were regularly rated as being most similar to any imitation (e.g. *snare6*, *tom4* and *tom6*), as mentioned in Section 4.4.3. For example, the non–target ratings for *tom6* are higher than the target ratings for most of the other tom sounds, although all the toms received their highest rating when they were compared against a target imitation).

The rank order of the similarity ratings was not the same for all tests from a single imitated sound, as might be expected if listeners based the similarity ratings entirely on similarity between the actual drum sounds. Instead, different imitations of the same sounds elicited different rankings. For example, when averaging across listeners, *tom6* was rated as most similar to 10/14 of the imitations of *tom5*. However, in 6/10 of these imitations the second most similar rated sound was *tom5*, whereas it was *tom4* for 4/10. This highlights that the ordering of the drum sounds in terms of similarity ranking with respect to the imitation changed (sometimes considerably) between imitations. It is apparent from listening to the imitations that for a given drum sound the imitation techniques differed between imitators, and also each imitator employed different techniques to imitate the different drum sounds: highlighting specific characteristics of the imitated sounds (such as attack time, pitch, and amplitude envelope). The difference in rankings for the same imitated sound may therefore be due to these characteristics being more or less perceptually relevant to the listeners. For example, the attack time may be most important for one imitation, but for another it may be the amplitude envelope, even when two imitations are of the same sound. A similar effect was previously identified by Lemaitre et al. [2016a], where similarity between imitations and imitated sound classes appeared to be based on specificities in each of the imitations. In addition, listeners may have focussed on different characteristics of the sounds depending on their critical listening experience [Lemaitre et al., 2010], however, as previously noted, overall there was moderate to strong concordance between listeners' rankings of the same sounds.

The similarity ratings for 'target' sounds were notably higher for the hats than for the other drum categories (all of which had similar rating values), indicating that the hats may be more imitable than the other drum sounds. It is conceivable that pitched sounds would lend themselves more to vocal imitation than non pitched sounds, however this is probably not the case when the sounds contain resonances and inharmonicities that are difficult or impossible

to reproduce with the voice, which is the case for many pitched percussion sounds. We found that many of the imitators used speech–related vocal techniques (as used when vocalising fricatives and affricates) to imitate the noisy spectral shape of the hats, and this, combined with the distinctive temporal shapes for each hat sound resulted in higher similarity ratings and identification accuracy (we note that there was less variance in temporal shape between the cymbals, which is likely the cause of the lower identification accuracy of cymbals compared to hats). There is also more contrast in the temporal shapes of the hat sounds compared to all the other drum categories, due to the inclusion of both open and closed hats (indeed, one might argue that open and closed hats are 'different' instruments: they have been treated as such in previous work on classifying beatboxed sounds [Sinyor et al., 2005]), and this was reflected in the way listeners used the scale, with generally greater contrast between the ratings for 'target' vs. 'non–target' sounds.

4.5 Summary and conclusions

In this chapter we have presented a 2 part study investigating the imitation of percussion sounds. In the first part, 14 participants imitated 30 drum sounds (6 from each of 5 categories - cymbals, hats, kicks, snares, and toms). In the second part, 63 listeners rated the similarity between the imitations and 6 same–category sounds. In addition to the individual similarity ratings between imitations and imitated sounds, the experimental design meant that for each imitation, listeners identified the same–category sound that was most similar to the imitation, and a rank ordering of similarity between the imitation and each of the same–category sounds. The main purpose of the work in this chapter was to establish whether musicians could imitate percussion sounds such that listeners could identify the sound being imitated, from a set of same–category sounds. In particular, we were interested in what types of sounds imitators were not able to differentiate, or where listeners exhibited confusion between similar percussion sounds.

In general we found that the identification of imitations varied considerably based on *i*) the imitated sound, and *ii*) the similarity between same–category sounds. In the worst case (*tom3*), the imitated sound was almost never identified from its imitations (4% of instances), and in the best case (*hat4*) it was correctly identified 74% of the time. This highlights that for

certain sounds both the imitation and listening tasks were difficult, and there was considerable confusion when the imitated sounds were similar to other same–category sounds or imitators adopted similar vocal techniques to imitate different sounds. The difficulty of the task is indicated by the fact that 12/63 listeners did not manage to reproduce their own results to a reasonable degree, when given duplicate tasks (based on a rank correlation threshold of 0.5). Nonetheless, there was moderate to strong concordance amongst the (ranked) ratings for the remaining 51 listeners, and they were able to identify the imitated sounds with above-chance accuracy for 16 of the 30 sounds. Importantly, when listeners did confuse percussion sounds, on average they still rated imitations as being more similar to their respective imitated sounds than the other same–category sounds, with a mean reciprocal rank of 0.58 across all tests. This further demonstrates that the diversity of same-category sounds was such that imitations were, on average, not confused with the sounds 'most different' from the imitated sound.

We limited the scope of this work to comparing imitations to same–category sounds, many of which have similar spectral distributions. In cases where the spectral distributions differ, the imitators did not necessarily emphasise these differences, and in some cases were not able to due to the limitations of the vocal apparatus (for example complex harmonic patterns in tom drum resonances). As such it appears that listeners relied on temporal shape and specificities: sounds with distinctive temporal signatures (such as the hats) were generally identified more successfully than those with similar temporal envelopes (such as the toms). Spectral features are typically used for measuring timbral similarity between percussion sounds, and this highlights an important issue with QBV for such sounds: the types of features used in drum category classification tasks may not be suitable to measure same–category similarity, particularly when the spectral differences are not easily differentiated using the voice. In the next chapter we investigate acoustic features for predicting the similarity ratings presented here. In particular, we seek to address the relative importance of temporal vs. spectral features for QBV of percussion sounds, and compare heuristic vs. learned features for this task.

Chapter 5

Audio features for query by vocalisation

In this chapter we investigate the suitability of different types of audio features for QBV of percussion sounds. As we discussed in Chapter 2, existing research on this topic has previously only considered the performance of QBV methods (in particular different types of audio features) based on the assumption that the imitated sound is the sound that should be retrieved, or ranked highest in a list of retrieved sounds. However, as we have shown in Chapter 4, although in general listeners considered the vocal imitations of a given drum sound to be more similar on average to the 'target' (i.e. imitated) sound, compared to other same–category sounds, the imitated sound may not always be the most *perceptually* similar sound to the imitation. As such, we approach this topic from the perspective of perceptual similarity, using the sounds and similarity ratings presented in Chapter 4. Instead of considering the suitability of audio features in terms of classification accuracy (i.e. identifying the imitated sound from the imitation), we will evaluate their suitability as predictors for the perceptual similarity ratings.

With regards to the audio features, we consider both heuristic and learned features. We conduct a comprehensive meta–review of the heuristic audio features used in previous studies on vocal imitation of both percussion–specific and non–percussion sounds. Whilst heuristic features have been the mainstay of many MIR tasks for the last 2 decades, there has been a notable shift towards using neural networks for learning audio features, particularly using deep learning methods. As such we compare the comprehensive set

of heuristic features, and suitable subsets thereof, to features derived using deep learning, namely convolutional–auto–encoders (CAEs). In addition to the heuristic vs. learned feature comparison, we are also interested in what elements of the vocal imitations are most salient for listeners when judging the similarity between imitations and percussion sounds. To address this we investigate the importance of spectral vs. temporal information for predicting the ratings, by manually selecting relevant subsets of the heuristic features and adapting the model architecture of the CAEs.

This chapter is laid out as follows: we outline the research questions and scope in Section 5.1. In Section 5.2 we present the feature sets that will be evaluated, for both heuristic (Section 5.2.1) and learned features (Section 5.2.2). In Section 5.2.2 we also discuss the model architecture, training data, and training procedure for learning the features. The method used to evaluate the different types of features is explained in Section 5.3, and the results are given in Section 5.4. Finally, conclusions are presented in Section 5.5.

5.1 Research questions and scope

We limit the heuristically derived audio features to those from the relevant literature on vocal imitations, excluding those that are not relevant to percussion sounds (see Section 5.2.1 for details). These include many of the standard audio features commonly used in both speech analysis and MIR tasks. The features are evaluated using the vocal imitation and similarity rating datasets from Chapter 4. As such, we limit the scope of this work to the same set of 30 drum sounds: 6 from each of 5 classes (kicks, snares, cymbals, hats, and toms), and consider only the similarity between the imitations and same–class sounds (e.g. similarity between the imitation of a kick drum and 6 kick drum sounds). In terms of the similarity ratings, we only consider those from the 51 reliable listeners (as defined in Chapter 4), which gives 9126 similarity ratings between imitations and same–category sounds. Our research questions are as follows:

1. What types of audio features best represent the perceptual similarity between vocalisations of percussion sounds and actual percussion sounds? In particular, do learned features outperform heuristically derived features?

2. How does this differ between categories of percussion sounds: e.g. do some features work best for kicks, and others for snares?

3. What can we learn from the performance of different types of audio features about how listeners consider the similarity between vocalisations and actual sounds?

   (a) Do listeners rely on a small subset of particularly salient audio features?

   (b) What is the relative importance of spectral vs. temporal information?

5.2 Feature sets

5.2.1 Heuristic features

The sets of heuristic features used in the experiments of this chapter are detailed below. Unless stated otherwise, all spectral features were computed from the power spectrum. Prior to any features being extracted the 420 imitations and 30 stimuli were manually edited (to remove sections of silence) and peak normalised as part of the listening study design from Chapter 4.

Set 1: Full feature set

The full feature set is taken from the review of literature on vocal imitation of both percussion and non–percussion sounds from Chapter 2, which were listed in Table 2.1. We exclude morphological features and wavelet coefficients for the reasons discussed in Section 2.4.4.1, namely the poor performance of wavelet coefficients for classifying vocalised percussion sounds (compared to MFCCs and LPC coefficients), and the unsuitability of morphological features for discriminating between vocalisations with a similar global morphological profile. The full set of features is given in Table 5.1. After being extracted, each feature was standardised to have zero mean and unit variance (across all sounds). The frame–wise features were extracted using a window size of 4096 samples and 87.5% overlap. These parameters were selected according to the findings in Pampalk et al. [2008], as discussed in Chapter 4. For frame-wise

features we take summary statistics to capture the temporal variation over the entire sound, in the form of either the median and IQR, or mean and variance. Median and IQR are more robust to sections of silence and spurious feature values at start and end frames [Peeters et al., 2011]. Therefore these statistics are more suitable for features where we expect the sound to settle on a value, and are not interested in largely different values around the start and end frames. This is the case for pitch and other pitch–related features (clarity, noisiness, roughness, and inharmonicity), where is it reasonable to expect no useful harmonic information in the attack portion of the sounds. For this reason these features are only extracted from frames proceeding the end attack time (see below for definition), ignoring very quiet tails from the sounds (defined as the part of the tail where the root mean squared (RMS) energy is below 1% of the maximum RMS in the signal). For the other features, we expect a skewed distribution of feature values over time, according to the typical profile of percussion sounds, consisting of only attack and decay portions. Here we are interested in the comparative weighting of the higher and lower valued features over time, therefore mean and variance are used. Many of the features listed in Table 5.1 have been extensively described and investigated in the MIR literature (see for example Bullock [2008]; Peeters [2004]; Peeters et al. [2011]; Stowell [2010]), however for completeness the specifications for features used in the present experiments are given in Appendix A.

The feature set consists of 155 features. It is conceivable that some of these features will not be independent, and closely related features will exhibit covariance, sharing mutual information. Stowell [2010] investigated the mutual information for many of these features and found high dependence between: spectral centroid and rolloffs; spectral rolloffs and band powers; and spectral spread and flatness, amongst others. As we will discuss in Section 5.3, for our evaluation we treat the feature space derived using these features as a Euclidean space, and are interested in the distance between vocal imitations and the percussion sounds. As such it is prudent to remove any correlated dimensions in the feature space, to reduce the computational complexity required to measure distance between samples and reduce the effect of the curse of dimensionality (where the sparsity of the data in a sampled space increases with the number of dimensions used to represent the data [Chávez et al., 2001]). We therefore apply PCA to the standardised features extracted on all samples. The number of components used for the final feature set is based on a threshold for explained variance, i.e. how many components are required to capture $n\%$ of

Feature	**Global**	**Frame**	**Med/IQR**	**Mean/Var**	**# Feats.**
Log attack time	✓				1
Temporal crest factor	✓				1
Duration	✓				1
Zero crossing rate	✓				1
Decay time	✓				1
Pitch		✓	✓		2
Pitch clarity		✓	✓		2
Noisiness		✓	✓		2
Roughness		✓	✓		2
Inharmonicity		✓	✓		2
Spectral centroid		✓		✓	2
Spectral rolloffs (×4)		✓		✓	8
Spectral crest factor		✓		✓	2
Spectral slope		✓		✓	2
Spectral spread		✓		✓	2
Spectral kurtosis		✓		✓	2
Spectral flatness		✓		✓	2
Spectral skewness		✓		✓	2
Spectral entropy		✓		✓	2
Spectral compactness		✓		✓	2
Strongest frequency		✓		✓	2
Spectral flux		✓		✓	2
Overall power		✓		✓	2
Band–specific powers (×5)		✓		✓	10
LPC coefficients (×10)		✓		✓	20
MFCCs (×13)		✓		✓	26
ΔMFCCs (×13)		✓		✓	26
ΔΔMFCCs (×13)		✓		✓	26

Table 5.1: The full set of global and frame–wise heuristic features extracted from the imitations and imitated sounds. Spectral rolloff features were calculated for the 95^{th}, 75^{th}, 50^{th}, and 25^{th} percentiles. Band–specific powers were calculated for 5 log–spaced bands from 50Hz–6.4kHz. Detailed feature specifications are given in Appendix A.

the variance in the dataset. We tested thresholds between 40–100%, in steps of 10%, and found that a threshold of 60% (consisting of the first 14 principle components) showed the best performance based on the evaluation method in Section 5.3. Consequently, this PCA–reduced feature space was used for the full feature set.

Sets 2 and 3: MFCCs and Δ/ΔΔMFCCs

MFCCs are commonly used descriptors of timbre, and have been shown to work well for many sound classification tasks, including measuring the perceptual similarity between sounds [Terasawa et al., 2005], genre classification

[Tzanetakis and Cook, 2002], speech recognition [O'Shaughnessy, 2003], and have been previously used as a baseline for comparing to learned features for QBV tasks [Zhang and Duan, 2015, 2016a]. In addition, first or second order ΔMFCCs are also often used to capture the variation of MFCCs over time [O'Shaughnessy, 2003; Stowell, 2010; Zhang and Duan, 2015]. We therefore include a subset of features containing only MFCCs (set 2), and to investigate whether the temporal variation in MFCCs is useful for our task, we also include a subset of ΔMFCCs and ΔΔMFCCs (set 3).

Set 4: Temporal

As noted in the research questions set out in Section 5.1, we are interested in the relative importance of spectral vs. temporal information for our task. Given the physical constraints of the vocal tract and transposition of spectral features discussed in Chapters 2 and 3 respectively, one might not expect the spectral range of the imitations to map directly to that of the percussion sounds. However, with the exception of attack time (which as previously discussed and identified by Lemaitre et al. [2016b], people may not be able to differentiate), there is no such constraint for accurately imitating the temporal features such as the overall energy envelope and duration. For this reason we include for comparison a subset of temporal features, namely the log attack time, temporal centroid, duration, temporal crest factor, zero crossing rate, decay time, and overall power.

Set 5: Auditory images (PHG)

We also include for comparison the similarity measure from Pampalk et al. [2008] that were used to select the drum samples in Chapter 4. This method (herein referred to as PHG) has been shown to be a good predictor of perceptual similarity between same–category drum sounds, therefore we are interested in whether it is also suitable for predicting the similarity between vocal imitations and real drum sounds. To recap, the distance between two sounds is measured as the Euclidean distance between their vectorised spectrograms, constructed with a 4096 sample window; 512 sample hop size, Bark scale (72 bins), loudness in dB and scaled using Terhardt's model for the outer and middle ear Terhardt [1979]. The lengths of the 2 spectrograms are equalised

by zero padding the shorter one prior to the vectorisation, in order to compare equal–sized feature vectors.

Set 6: MPEG–7

The final feature set (set 6) is taken from the MPEG–7[1] standard for a percussive timbre space, described by Peeters et al. [2000]. This is motivated by the findings of Lakatos [2000] who investigated the acoustic correlates of MDS–reduced timbre spaces derived from perceptual similarity ratings between percussion sounds. Lakatos found that for a 2D space, the first dimension is best correlated with a combination of log attack time (LAT) and temporal centroid (TC), and the second dimension is best explained by spectral centroid (SC). This feature set was also tested against the PHG method in Pampalk et al. [2008], although it did not perform as well for predicting the perceptual similarity between same–category sounds. Nonetheless, as it is relativity simple to implement, we include it here for comparison. The distance, d, between 2 sounds (a and b) is calculated as per Equation 5.1. According to the MPEG standard the relative weighting of LAT, TC, and SC is dependent on the dataset. For this reason we conducted a grid search of all possible weightings between 0–1 (with steps of 0.1), and found that weightings of 0.8, 0.2, and 1.0 worked best for LAT, TC, and SC respectively, based on the evaluation method in Section 5.3.

$$d = \sqrt{\left((LAT_a - LAT_b)\frac{w_{lat}}{10} + (TC_a - TC_b)\frac{w_{tc}}{10}\right)^2 + \left((SC_a - SC_b)\frac{w_{sc}}{10^5}\right)^2} \quad (5.1)$$

5.2.2 Learned features: CAE networks

Having described the heuristic feature sets, we now move on to the learned features, providing details of the convolutional neural networks, training data and procedure. In particular, we describe the general model architecture along with the variants used to generate feature sets with different sizes (128–2048) and resolution of the spectral and temporal dimensions.

[1] SO/IEC 15938 Information technology - Multimedia content description interface - Part 4: Audio (2002)

Model Architecture

The basic architecture is a CAE with four 2D convolution layers in each of the encoder and decoder sections. Each convolutional layer is followed by batch normalisation and (ReLU) activation layers. To avoid checker board artefacts caused by deconvolution layers [Odena et al., 2016] we apply upsampling prior to each decoding convolutional layer. As such, each decoding deconvolution layer is an upsampling layer followed by a 2D convolution layer with $(1, 1)$ stride. We vary the kernel size of the first and last layers while using fixed $(10, 10)$ kernels for the other convolution layers. The kernel size is varied in order to compare the shape of the encoded representation (i.e. square, wide, tall) and how this interacts with the shape of the kernels over layers. The encoding layers have [8, 16, 24, 32] kernels (layers 1–4 respectively) which is mirrored in the decoder, i.e. [32, 24, 16, 8]. Finally, a single–channel convolution layer is used as an output layer. The activation of the last layer of the encoder is flattened into a 1D vector which is used as the feature vector. All 11 the variants of the above model are presented in Table 5.2.

Training Data

To train the network requires considerably more data than was collected from the experiment in Chapter 4 (420 imitations and 30 sounds). Therefore we compiled a dataset consisting of a wide range of vocal and percussion related sounds including *i*) short, percussive, non-percussive, pitched, and unpitched sounds, and *ii*) a broad range of non–verbal vocalisations. Specifically, there are 24,294 percussion sounds, 4,884 sound effects and 4,523 single note instrument samples. These samples were all taken from the author's private sound library, amassed over a 10+ year period producing music. In addition to the instrument and sound effect samples, we included 4,429 vocal imitations of instruments, synthesisers and everyday sounds from Cartwright and Pardo [2015], and vocal imitations of the short synthesised sounds from Chapter 3. The combined datasets comprise of ~39k sounds, of which ~6k are vocal imitations. We do not include the vocal imitations and drum sounds used for the evaluation, to ensure the trained network is generalisable beyond the sounds specific to this task.

Pre–processing

The audio files were all pre–processed to produce fixed size time–frequency representations for training the networks. As discussed in Chapter 2, magnitude spectrograms with linear, log, or Mel–based frequency scales are often used as inputs to CNNs, with the latter two used to scale the visual representations according to how we perceive frequencies. We are not concerned with reconstructing the audio from the decoded layer, therefore we are able to apply non–linear scaling (e.g. log, Mel, Bark) to the frequency axis of the spectrograms. All 3 scales have similar characteristics, and there is little difference between them in terms of how well they contribute to predicting perceptual similarity between same–category drum sounds using the auditory image method (PHG) for feature set 5 [Pampalk et al., 2008]. We therefore apply Bark scaling (as opposed to log or Mel), in line with the PHG method. For each sound in the training set, we compute the bark–grams from power spectrograms with a 4096 sample window and 512 sample overlap, using 128 Bark bins. As with the PHG baseline, the magnitudes are modified via decibel scaling and Terhardt's ear model curves [Terhardt, 1979]. To achieve a fixed size representation for all sounds, we either zero-pad or truncate the Bark spectrograms to 128 frames ($\approx$ 1.5 seconds).

Training Procedure

The network models were implemented using Keras [Chollet et al., 2015] and Tensorflow [Abadi et al., 2016]. The dataset described in Section 5.2.2 was split 70/30% for training and validation respectively, and batch learning was applied using a batch size of 128. As the training dataset contains 5.5 times more audio samples than vocal imitations, and we are equally interested in learning both sound types, a 50/50% split of audio samples/vocal imitations was specified for each batch. The models were all fitted using the Adaptive Moment estimation (Adam) optimiser [Kingma and Ba, 2014] with the suggested default learning rate of 0.001, and mean squared error loss function. We used the early–stopping scheme for no improvement in validation loss after 10 epochs, to avoid the model overfitting to the training data. The best model for each parameter setting is selected for the analysis (i.e. the model with the lowest validation loss).

5.3 Evaluation method

In this section we describe the method used to evaluate the different feature sets (heuristic feature sets 1–6 and 11 variants of the CAE network). The complete evaluation work flow is given in Figure 5.1. The 30 drum sounds, 420 vocal imitations, and 9126 similarity ratings from reliable listeners (as defined in Chapter 4) were used to evaluate the performance of each of the feature sets. For a given feature set and drum category, distance is measured between each of the 84 imitations and the 6 within-category sounds, giving 504 distance values per category, and 2520 in total. We use Euclidean distance in keeping with the PHG method and previous approaches to using acoustic features for predicting similarity between vocalisations and referent sounds [Lemaitre et al., 2016a] or text–based meanings [Perlman and Lupyan, 2017]. For each feature set the distances were normalised between 0–1 to make the model parameters (in particular the slopes) comparable by removing any influence of distance scale on the estimated parameters. A linear mixed effect regression (LMER) model was fitted for predicting the continuous ratings from the continuous predictor variable (distance). LMER is well suited to this task given that all listeners did not provide ratings for all imitations but only a randomly-selected set of 28 imitations (giving an unbalanced dataset). In addition, it allows us to include the dependencies between ratings for each listener, imitator, and imitated sound, which are inherent in the experimental design from Chapter 4.

Maximum likelihood parameters for the model were estimated using the lme4 package in R [Bates et al., 2015]. The general model was fitted with `rating` as the dependent variable for each response, fixed effects of `distance` and `imitated sound`, with an interaction term between the fixed effects, and random intercepts for each `listener` and `imitator`. The model (as specified in R) is given by:

```
lmer(rating~distance * imitated sound + (1|participant) + (1|imitator))
```

We then calculated the slope of `rating` over `distance` for a given `imitated sound` (i.e. the slopes of the interaction term), with 95% Wald CIs. For imitated sounds where the upper CI < 0 we can infer that the slope is significantly below 0 ($\alpha < 0.05$) [Gardner and Altman, 1986]. This indicates that the feature set is a good predictor for the imitated sound in question. We note that as with the modelling of the same data in Chapter 4, we observed

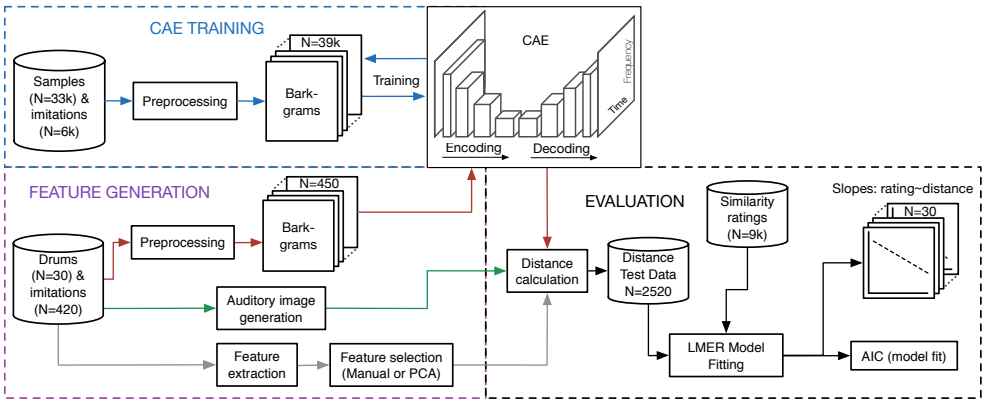


Figure 5.1: Overview of the complete evaluation work flow, for all 3 types of features (CAE, auditory image and heuristic). Audio features may be generated from the audio by taking any of the red, green or grey paths. Euclidean *distance* between each imitation and its imitated sound is then computed in the feature space, and fitted with the *rating* data to an LMER model. Performance of each feature set is measured by 1) AIC for model fit, and 2) the proportion of imitated sounds that have a significantly negative slopes for `rating` ~ `distance`.

heteroskedasticity in the residuals of the fitted models due to zero–inflation in the responses (rating data). Parameter estimates were therefore compared to those from robust models [Koller, 2016], and no major differences were found. As such the non–robust models were used for the analysis.

The performance of each feature set was evaluated using two metrics: the percentage of imitated sounds for which the slope of the interaction is significantly below 0 (accuracy); and Akaike's information criterion (AIC), which gives a measure of model fit (note: lower AIC = better model fit). An ideal feature set would have a significantly negative interaction slope for all 30 imitated sounds (perfect predictor = -1.0), and be a good fit to the rating data given the LMER model.

5.4 Results and discussion

The results for all 6 heuristic feature sets and the 11 CAE model variants are given in Table 5.2. The first notable finding is that all the learned feature sets outperform the heuristic features tested here, in terms of both evaluation metrics. This finding concurs with previous comparisons of heuristic vs. learned features for QBV tasks [Zhang and Duan, 2015, 2016a,b]. The best perform-

ing heuristic feature set in terms of AIC is PHG (2389), although both the MPEG–7 timbre space and the PHG method give the same result in terms of accuracy, with 53.3%, or 16/30 sounds for which the `rating` ~ `distance` slope is significantly less than 0 ($\alpha < 0.05$). The improved model fit with PHG over the MPEG–7 timbre space concurs with the findings of Pampalk et al. [2008], however we observe much lower correlation between similarity ratings than in the aforementioned study (where the similarity ratings were between drum sounds, not vocal imitations and drum sounds). The weightings of the 3 features for the MPEG–7 timbre space (LAT, TC, SC) were optimised via a grid search, in terms of minimising the AIC. Interestingly, Pampalk et al. [2008] also optimised the weightings, and identified TC (and to an extent SC) as the most important features for discriminating between drum sounds, whereas we found weightings of 0.8, 0.2, and 1.0 for LAT, TC, and SC respectively gave the best performance. This suggests that whilst LAT is less important than TC for discriminating between same–category drum sounds, the inverse applies when comparing vocal imitations. Overall these findings support the hypothesis that the auditory image similarity measure (and parameters) of Pampalk et al. [2008] are to some extent transferable for measuring similarity between vocalised and actual percussion sounds, at least compared to the other features tested here, including the MPEG–7 timbre space.

The other feature sets perform similarly in terms of accuracy (43.3–46.7%), however there is notable variance in terms of model fit. The accuracy measure does not consider the steepness of the slope, only that they are negative, therefore the AIC is a more informative evaluation metric. The AIC values are worst for MFCCs, although we observe a notable improvement when ΔMFCCs are included, indicating that the temporal evolution of these features is useful for predicting perceptual similarity between imitations and imitated sounds. This, and the fact that the temporal features also outperform MFCCs indicates that temporal information may be more relevant than the spectral envelope alone, in terms of how listeners discriminated between the same–category drum sounds tested here. This concurs with the findings of Lemaitre and Rocchesso [2014], where recognition of vocalised everyday sounds was more impaired by temporal inaccuracies than spectral ones, highlighting the importance of temporal information for judging the similarity between vocalisations and non–vocal sounds. That is not to say that there is no salient information in the spectral features of percussion sounds, evidenced in that the PHG method (which is essentially a comparison between time–frequency representations)

performs best, and by Tindale [2004], who demonstrated that a combination of temporal and spectral features outperform temporal features alone for the task of classifying different snare drums based on playing technique. However, imitators tend to transpose spectral features, as shown in Chapter 3 and by Lemaitre et al. [2016b], and listeners may not equate such transposition with decreased similarity.

To understand what is driving the results in Table 5.2 we may consider how the heuristic feature sets perform for each of the 30 drum sounds. The `rating` ~ `distance` slopes for all 30 sounds are given in Figure 5.2. This shows that the prediction performance for all feature sets varies considerably between the sounds. The PHG method performs better for snares and toms compared to kicks, which is similar to the findings of Pampalk et al. [2008], however, interestingly there does not appear to be any relationship between the identifiability of imitations of certain sounds and how well the acoustic distance measure performs: for example, *hat4*, *hat5* and *hat6* were all identified from their imitations with significantly above chance accuracy (see Chapter 4), yet for these sounds the upper CI of the slope estimates cross 0. Pampalk et al. [2008] did not include cymbal or hat sounds in their study, therefore it is not clear how suitable the measure is for comparing similarity for cymbal and hat sounds, let alone between imitations and imitated sounds for these drum types. Indeed, we might conclude from Figure 5.2e that this measure is not particularly good for these drum types unless the sounds are very short (as are *hat1* and *hat2*), and except for these cases, the MPEG–7 timbre space may be a better feature set for this task.

In terms of the full feature set, there are similar slope trends as for MFCCs+ ΔMFCCs (Figures 5.2a and 5.2c), with the exception of only the cymbals. This indicates that the MFCCs+ΔMFCCs are contributing more to the first 14 principle components of the PCA–reduced full feature set compared to the other features. We calculated the contribution (i.e. loadings) of each subset to the 14 principle components, and found that ΔMFCCs contribute 38% of the loadings, whereas the MFCCs alone contribute only 15%. This is somewhat expected because they make up the majority of the 155 features, and there are twice as many ΔMFCCs as MFCCs (52 vs. 26), making up 34% and 17% of the full feature set, respectively. However, this suggests that the ΔMFCCs are relatively more important than the MFCCs, at least in terms of how much of the variance they explain in the imitations and drum sounds tested here, although the difference is small.

Type	Feat. set	L1/8 kernel	Strides of conv./upsampling layers				L4 shape (×32)	# Dims	Results	
			L1/8	L2/7	L3/6	L4/5			AIC	Acc.
Full+PCA	1	–	–	–	–	–	–	14	2766	46.7
MFCC	2	–	–	–	–	–	–	26	3078	43.3
MFCC+ΔMFCC	3	–	–	–	–	–	–	78	2703	46.7
Temporal	4	–	–	–	–	–	–	9	2745	43.3
PHG	5	–	–	–	–	–	–	–	**2389**	**53.3**
MPEG–7	6	–	–	–	–	–	–	3	2807	**53.3**
CAE (Square)	7	(5, 5)	(2, 2)	(2, 2)	(2, 2)	(2, 2)	(8, 8)	2048	**1819**	**73.3**
	8	(5, 5)	(2, 2)	(2, 2)	(2, 2)	(4, 4)	(4, 4)	512	1924	66.7
	9	(5, 5)	(2, 2)	(2, 2)	(4, 4)	(4, 4)	(2, 2)	128	1953	66.7
CAE (Tall)	10	(5, 3)	(2, 2)	(2, 2)	(2, 2)	(2, 4)	(8, 4)	1024	**1609**	70.0
	11	(5, 3)	(2, 2)	(2, 2)	(2, 4)	(2, 4)	(8, 2)	512	1643	**73.3**
	12	(5, 3)	(2, 2)	(2, 4)	(2, 4)	(2, 4)	(8, 1)	256	2358	63.3
	13	(5, 3)	(2, 2)	(2, 4)	(2, 4)	(4, 4)	(4, 1)	128	2522	56.7
CAE (Wide)	14	(3, 5)	(2, 2)	(2, 2)	(2, 2)	(4, 2)	(4, 8)	1024	1918	66.7
	15	(3, 5)	(2, 2)	(2, 2)	(4, 2)	(4, 2)	(2, 8)	512	1864	73.3
	16	(3, 5)	(2, 2)	(4, 2)	(4, 2)	(4, 2)	(1, 8)	256	1390	**83.3**
	17	(3, 5)	(2, 2)	(4, 2)	(4, 2)	(4, 4)	(1, 4)	128	**1294**	**83.3**

Table 5.2: Results for all 17 feature sets, including details for the 11 CAEs and 6 heuristic feature sets. CAEs differ in the kernel shape of L1 and L8, and the shape of the encoded layer (determined by strides). Results are given in terms of *i*) the LMER model fit (AIC), and *ii*) the percentage of imitated drum sounds for which the `rating` ~ `distance` slope is significantly less than 0 ($\alpha < 0.05$). The learned feature sets are grouped according to the shape of the encoded layer, and the best performing feature sets for each group are given in bold.

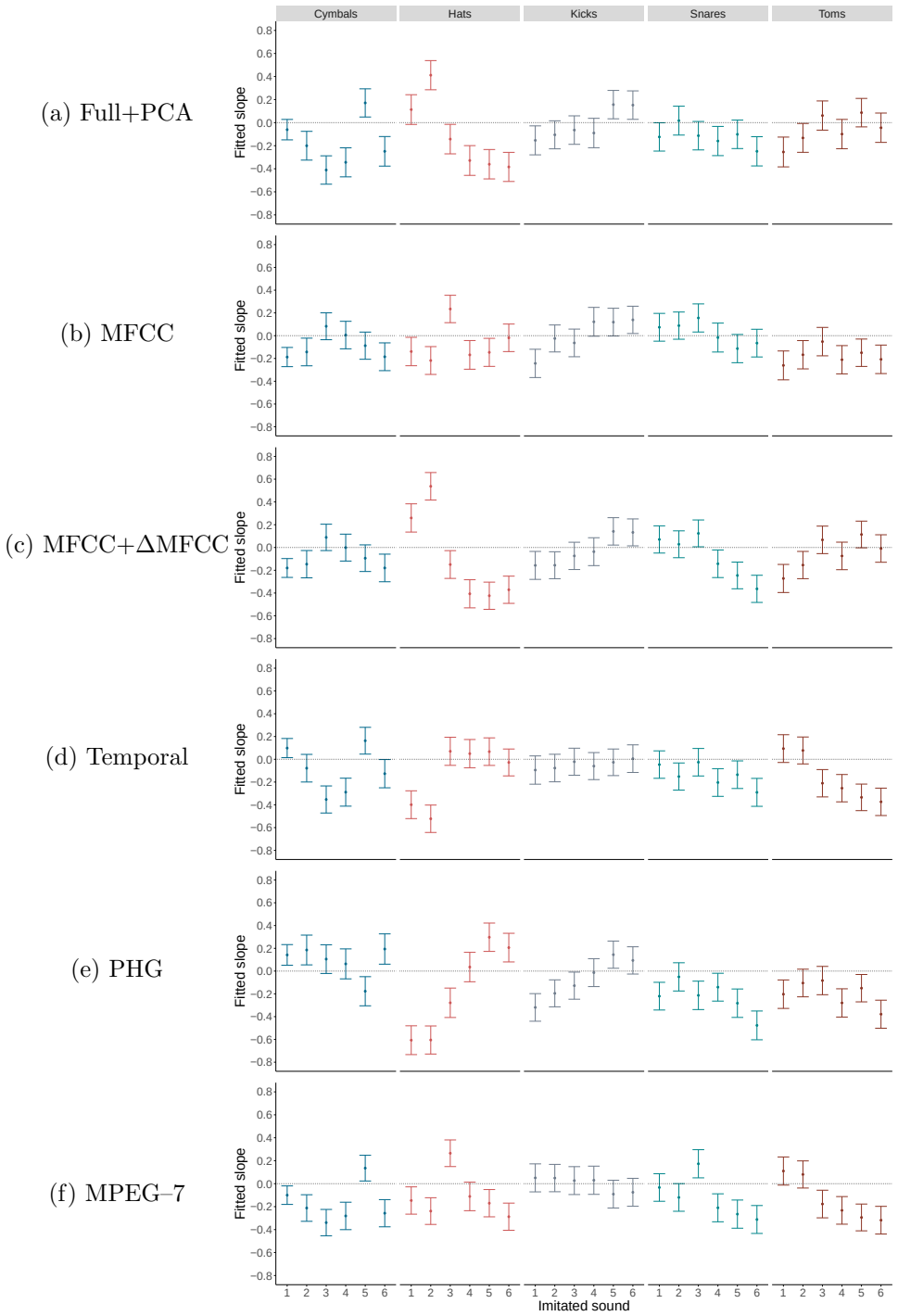
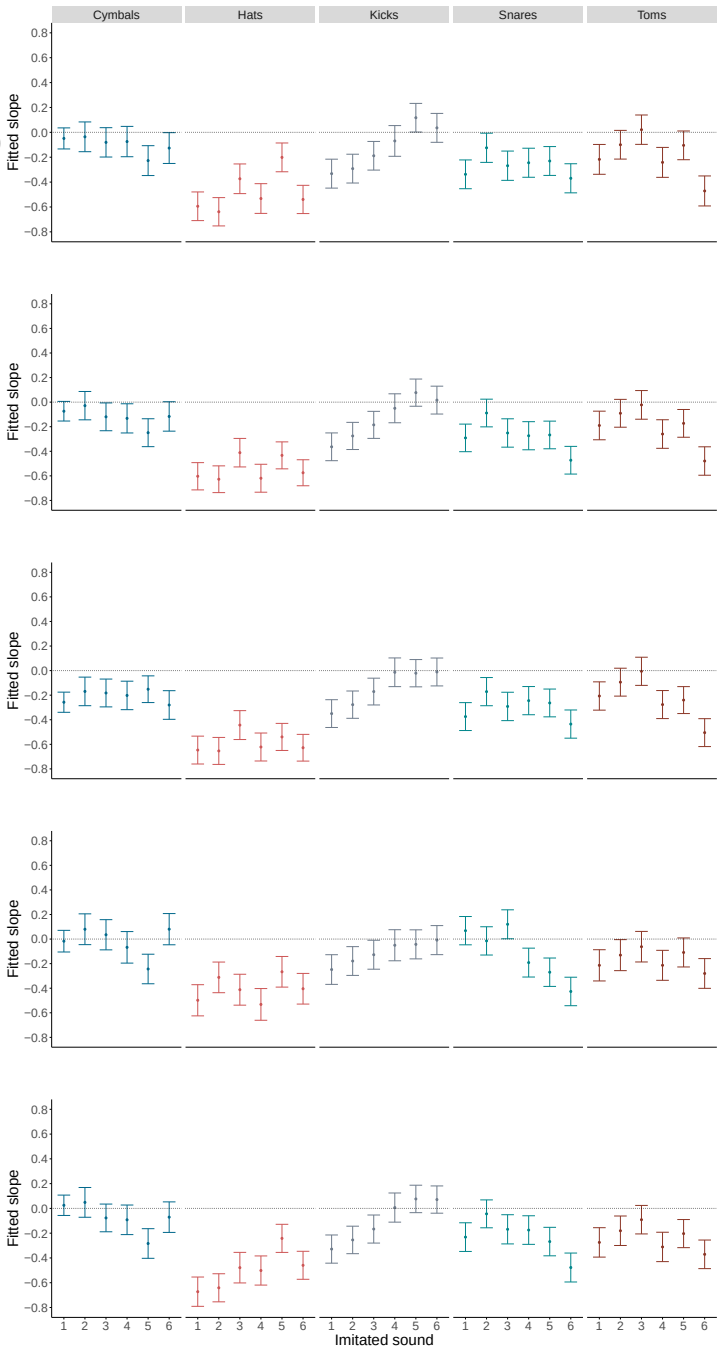


Figure 5.2: Slope estimates for the LMER models fitted using each of the heuristic feature sets. A negative slope indicates a decrease in perceptual similarity with an increase in distance, i.e. sounds for which the method performs well. Values are mean estimates across all imitations for each drum sound, with 95% Wald confidence intervals.



(a) CAEW 1024 (14)

(b) CAET 1024 (10)

(c) CAEW 128 (17)

(d) CAET 128 (13)

(e) CAES 128 (9)

Figure 5.3: Slope estimates for the LMER models fitted for the *wide* and *tall* CAE models with the largest encoded size: 14 (a), 10 (b); and the smallest encoded size for all 3 encoded shapes: 17 (c), 13 (d), and 9 (e). A negative slope indicates a decrease in perceptual similarity with an increase in distance, i.e. sounds for which the method performs well. Values are mean estimates across all imitations for each drum sound, with 95% Wald confidence intervals.

In terms of the learned features and different CAE architectures, the LMER model from the best performing feature set (17) gives fitted slopes for `rating` ~ `distance` that are significantly less than 0 ($\alpha < 0.05$) for 83.3% (25/30) of the imitated sounds, and has the lowest AIC. This feature set is generally a good predictor of perceptual similarity between the vocal imitations and imitated sounds tested here. Interestingly, as with the heuristic features, preservation of the temporal resolution appears to be more important than spectral resolution for our task. For CAEs with kernels wide in time and narrow in frequency (14–17), performance improves as the size of the encoded layer decreases. This indicates there is much redundancy in the spectral information: encoded shapes with spectral dimensions greater than 1 have an adverse effect on performance. Recall that the similarity ratings are only for sounds in the same class (e.g. kick, snare etc.), and we can expect high spectral similarity within each class, particularly for the attack portion of the sounds, which all contain broadband energy (as per Figure 4.1). As such, overall, energy differences in time may be more salient than the spectral distribution, providing the cues used by listeners when giving the ratings. This hypothesis is supported by comparing the square and tall CAEs: where reducing the temporal resolution generally decreases performance. However there is also some redundancy in the temporal information, as can be seen comparing feature sets 16 and 17. As a post–hoc analysis, we tested variants of CAE 17 using smaller encoded kernel shapes: $(1, 2)$ and $(1, 1)$, and found a decrease in performance below $(1, 4)$. This effect can also be seen in models 10–13, where performance decreases as width is reduced from 4 to 1.

Further analysis of the LMER models for the CAE feature sets is presented in Figure 5.3. Here we compare the slopes for each of the 30 sounds, for the *wide* and *tall* models with the largest encoded size (14 and 10 respectively), and the *wide*, *tall*, and *square* models with the smallest encoded size (17, 13, and 9 respectively). Firstly, we note that as with the heuristic features, there is considerable variation of performance between the sounds for all feature sets, however, generally the relative slopes within each feature set are similar across the different models. In other words, all the models presented in Figure 5.3 exhibit similar patterns, and the improvements seen in the best performing model (17, Figure 5.3(c)) appear to apply to most of the 30 sounds. We observe that prediction of ratings for cymbals suffered most from reductions in the temporal resolution, whereas the opposite effect was observed for toms. This indicates that although temporal resolution may be more important overall, it

is not the case for all drum categories. For the best performing model, there are 5 sounds for which the upper CI crosses 0 (3 kicks and 2 toms). These are all pitched sounds (although they are not the only pitched sounds in the dataset, indeed, all the toms are pitched), and as we showed in Chapter 4, pitch is not necessarily used by listeners as a salient cue for identifying imitated sounds from imitations. Nonetheless, this suggests that predictions for some pitched sounds may suffer from reducing the size of the encoded spectral shape to 1.

Finally, we note that the slopes, although generally below 0, do not approach -1. Listener rating data is inherently noisy, and the concordance amongst listeners varies across the sounds. As such, there will clearly be a glass ceiling for performance, and a perfect model fit would not be useful for a real world application of the LMER model. Indeed, a perfect model fit is not desirable if one is interested in generalisability of the fitted LMER model.

5.5 Summary and conclusions

In this chapter we investigated the performance of both heuristic and learned features for predicting the perceptual similarity between vocal imitations and actual percussion sounds. Seven heuristic feature sets were compared, including a set of 155 features from the literature on vocal imitation analysis and QBV, MFCCs, temporal features, spectrogram–based features, and the MPEG–7 percussion descriptors. In terms of learned features, we compared the encoded features from 11 variants of a convolutional auto–encoder (CAE) trained on over 39k percussion sounds, instruments and vocal imitations. Specifically, each of the networks varied in both the shape of the encoded layer in terms of the spectral and temporal dimensions, and the number of extracted features. Each of the feature sets was evaluated using the vocal imitations, drum sounds, and perceptual similarity ratings from the experiments of Chapter 4. For a given feature set, the distance between imitations and same–category drum sounds in the feature space was used as a predictor for the similarity ratings, in a regression model fitted using linear mixed–effect regression. The experiments in this chapter serve to identify what types of audio features best represent the perceptual similarity between vocal imitations and percussion sounds, in terms of *i*) whether learned features outperform heuristic features, *ii*) whether listeners rely on only a small subset of salient features, *iii*) the relative importance of spectral vs. temporal resolution in the CAE,

and *iv*) if, and how this differs across drum categories.

The results show that CAEs outperform all 7 sets of heuristic features by a considerable margin. In terms of the heuristic feature sets, we observed little difference in overall performance between the full set of 155 features (reduced to the 14 first principle components), MFCCs (with or without ΔMFCCs), and temporal features, and only marginal improvements for the spectrogram based features and MPEG–7 percussion descriptors. A category–level analysis showed that the MPEG–7 descriptors performed particularly well for cymbals and hats and reasonably well for snares and tom–toms, but poorly for kick drums. In contrast, the spectrogram based features performed best for snares, tom–toms, and (closed) hats, with particularly poor performance for cymbals and open hats. In terms of the CAE based features, the results show that (with the exception of some tom sounds) reducing the size of the encoded layer height (frequency) increases the predictive power of the learned features, yet reducing the width (time) has the opposite effect. This finding is partly unexpected given that the drum sounds generally have a similar overall temporal envelope (attack followed by a decay), however understandable given that we compare only same–category sounds, which also share similar spectral distributions.

These findings indicate that the suitability of different types of features for this task is dependent on the drum category. Human rating data is particularly noisy when the task requires discrimination between perceptually similar sounds, however there are instances where there are clear trends in the ratings that all of the feature sets failed to capture. As we saw in Chapter 4, *kick4* and *kick5* were identified with above chance accuracy from their imitations, and in general the target sounds were rated higher than non–target sounds for them, yet none of the feature sets (learned or heuristic) managed to sufficiently predict the similarity ratings. This suggests that there are distinguishing characteristics in the sounds that enabled listeners to identify the imitations, which are not being captured in any of the feature sets presented here, and highlights the challenge of finding a one–category–fits–all solution to predicting the perceptual similarity between vocalisations and percussion sounds. Nonetheless, overall, the best performing CAE feature set shows encouraging predictive power for QBV applications, with good performance for all the cymbals, hats, and snares, and most of the kicks and toms.

Chapter 6

Conclusions

The extraordinary ability of humans to effectively communicate sonic concepts and the practical issues associated with navigating large sound libraries give rise to the question of whether the voice might be a useful medium for sound search. In this thesis we explored this question by investigating the potential of musicians to accurately imitate musical sounds, and acoustic descriptors that might be useful for predicting the similarity between vocalisations and percussion sounds. In the following section we will outline the main findings of the work presented in this thesis and consider how these relate to the objectives set out in Chapter 1. We will then draw upon these contributions and some limitations of our work, to identify future directions and suggest potential areas for further research in Section 6.2.

6.1 Summary of contributions

In Chapter 2 we examined the existing literature that informed us to identify the three main threads of research set out in Chapters 3–5. This review highlighted the broad range of sounds that can be produced with the human vocal apparatus, beyond those used simply for speech and singing, and identified that the essence of many everyday, artificial, and environmental sounds can be effectively captured and communicated by means of vocal imitation. However it was not clear how well this ability applied to salient musical acoustic characteristics such as pitch, dynamics, and timbre, or musical sounds such as percussion instruments. The problem of navigating sound libraries was out-

lined, and the case for query by example (QBE) as a promising method for searching sounds was presented. This, combined with the potential for communicating sounds using the voice raised the question of whether the voice might be a useful means for querying sound libraries, i.e. query by vocalisation (QBV), and how the perceptual similarity between vocalisations and musical sounds might be modelled. We went on to identify many heuristic features that have been successfully applied to classification of vocalised sounds, and the emerging state–of–the–art features for many audio based tasks using deep learning methods. This raised the question of whether these types of features might be useful for QBV of musical sounds, and the merits of different types of audio representations for this task.

In Chapter 3 we embarked on the first line of enquiry: investigating vocal imitation of sounds varying in pitch, dynamics, and spectral shape. The focus of this work was to establish the potential for the voice as a medium for representing non–vocal sounds, and ascertain the level of control with which musicians were able to accurately imitate sounds with 1 or 2 features varying over time (controlled using ramp and modulation envelopes). In order to sufficiently control for each of these features and the interactions between them we synthesised sounds with parameters controlling for F_0 (pitch), loudness, and spectral centroid (determined by the cutoff frequency of a low pass filter). The results show that the extent and range of the envelopes were most accurately imitated for pitch, as one might expect given a well established relative scale (at least amongst musicians), and asymmetries in the imitations of pitch and loudness ramps, in agreement with previous findings from studies on sung pitch ramps [d'Alessandro et al., 1998] and loudness perception [Lane et al., 1970, 1961; Yadav, 2016]. Imitations of modulation envelopes showed that for all 3 features the musicians managed to accurately imitate the modulation rate, with negligible differences between features. Interestingly, on average, combining pitch with either loudness or spectral centroid envelopes did not have a significant effect on imitation accuracy for any of the features. To our knowledge this is the first such study on vocal imitation of these acoustic characteristics that includes combinations of features using ramp and modulation envelopes, the results of which highlight that musicians (including non–singers) are able to exercise a remarkable level of simultaneous control over pitch and loudness or pitch and spectral centroid.

In Chapter 4 we turned the focus to vocalised percussion sounds, with the aim of identifying whether musicians were able to imitate a set of drum sounds

such that third party listeners were able to identify the imitated sounds from the imitations, and establish the similarity between imitations and imitated sounds relative to other percussion sound from the same category (e.g. kick, snare etc.). The results showed that on average, the rated imitations were considered as more similar to their respective imitated sounds than the other same–category sounds, with a mean reciprocal rank of 0.58 (i.e. the mean ranking of the imitated sounds was between 1^{st} and 2^{nd} out of 6). The identification accuracy varied considerably between imitated sounds (ranging from 74% to 4%), which is lower than has been found for similar tasks using everyday sounds, in terms of classifying vocal imitations into groups [Lemaitre et al., 2011] and identification of individual referent sounds [Lemaitre and Rocchesso, 2014]. At first sight it therefore appears that percussion sounds may not be as imitable (in terms of communicative power) as everyday sounds, however we note that in the aforementioned studies the referent sounds were generally quite distinct, whereas we conducted a forced choice experiment where all options were percussion sounds from the same category. To our knowledge this is the first experiment of its kind, that specifically considers the ability of musicians to imitate percussion sounds such that same–category sounds can be differentiated based on the imitations. As such, the results of our experiment provide a contribution to understanding the effectiveness of vocal imitations for describing subtle differences between categorically similar sounds.

In Chapter 5 a set of heuristic and learned audio feature spaces were evaluated for their suitability for use in QBV of percussion sounds. The distance between sounds in each feature space was used in a linear mixed effect regression model to predict the similarity ratings from the listening experiment of Chapter 4. The results showed that the best performing learned features (extracted using a convolutional auto–encoder trained on percussion sounds and vocal imitations) outperformed all of the heuristic features for almost all sounds, and that temporal resolution of the learned features is more important then spectral resolution for this task. However, we found that for certain sounds, most notably kicks, none of the feature spaces were able to sufficiently predict the similarity ratings. These results support previous research indicating that learned features generally outperform heuristic features for many audio based tasks [Lee et al., 2009] including QBV [Zhang and Duan, 2015, 2016b], however there are a number of novel contributions in our work, namely: *i*) the range of heuristic features that was compared to the learned features. In the aforementioned studies only a handful of features are compared (namely

MFCCs), whereas we used a comprehensive set of features from the related literature specific to our application (vocal imitation analysis). *ii*) when comparing learned features we investigated the relative importance of spectral vs. temporal information for this task. For many audio tasks it is clear that a particular dimension will be more important (e.g. spectral for chord recognition [Humphrey and Bello, 2012], temporal for onset detection [Schlüter and Böck, 2013]), however for our task this was not the case, therefore we present an analysis of this factor using a range of different network architectures. *iii*) to our knowledge, this is the first such study that compares the perceptual relevance of different feature spaces including learned features, not only for QBV but across the field of MIR in general.

6.2 Future directions

The results presented in this thesis demonstrate the ability of musicians to control some (important) characteristics of the voice and vocalise percussion sounds, however the research field of vocalised musical sounds is still in its infancy, and despite a considerable amount of existing research on the singing voice there are many questions yet to be answered by research into non–verbal vocalisation of musical sounds, both in terms of what is possible (i.e. what people can vocalise and how they do it), and how we might model the similarity between vocalisations and non–vocal sounds. In this section we will highlight some potential areas for future research. Some of these suggestions came about as a result of the findings in the experiments of this thesis (not addressed due to inherent limitations of the presented experiments), and others identify work beyond the scope of this thesis that might guide future directions in the wider field of vocalised musical sounds.

Regarding control of vocal characteristics, we have only touched the surface in terms of *i*) what might be measured, and *ii*) the means of quantifying 'accuracy' or control. The work in Chapter 3 is based on synthesised sounds because this allowed for control over the parameters for each condition, whilst limiting any confounding factors from co-variance of characteristics that might exist in real–world sounds, for example in acoustic instruments. However, the resulting stimuli are quite simple both harmonically and in terms of the feature envelopes, permitting relatively simple methods for extracting the envelope parameters (slope, range, modulation rate and extent). Real–world sounds

will generally have more complex envelopes and potentially more interactions between the features. For these kinds of more complex shapes there may exist more suitable ways to quantify the difference between imitation envelopes and those of the imitated sounds, from the domains of time series analysis (such as dynamic time warping, or simply cross correlation), and curve fitting (such as smoothing spline ANOVA or additive modelling). In addition, the characteristic of 'timbre' was reduced to the audio feature of spectral centroid for the purposes of this experiment, and we know that the perceptual attributes of timbre cannot be fully quantified using this single measure (although it is an important timbral feature). As such, it is not clear how well our findings translate to real–world sounds. For example, one might consider whether people are able to imitate subtle fluctuations of features or whether more complex envelope shapes and combinations thereof can be imitated just as accurately as our stimuli. Therefore, a natural extension to this work is to apply a similar method (comparison of feature envelopes between imitations and imitated sounds) to a set of real musical instruments or synthesisers, extending the range of features tested in accordance with the types of sounds in the dataset. There is an existing dataset of such imitations [Cartwright and Pardo, 2015], making this an attractive next step for any further research in this area.

Furthermore, we might consider whether imitation accuracy should indeed be measured in terms of the accuracy of individual audio features, and if so, whether all features should be weighted equally when quantifying it. As we have seen in Chapter 4, human listeners appear to be good barometers for measuring imitation accuracy (based on the agreement, or concordance amongst listeners), yet it is not always possible to quantify their ratings by applying the same acoustic features (and weightings thereof) to every sound, as previously demonstrated by Lemaitre et al. [2016a]. This highlights a potential disjunction between the approach to analysing vocal imitations presented in Chapter 3 compared to that in Chapter 4. In Chapter 3 we quantitatively assessed imitation accuracy of synthesised sounds at a frame–based feature level. This informs us of the relative accuracy with which musicians might imitate specific feature envelopes, however it does not tell us anything about the perceptual accuracy of the imitations. Indeed, one might argue that some of the less accurate imitations in terms of frame–level features are more perceptually similar to the imitated sounds than the more accurate imitations. In contrast, in Chapter 4, where we turned the focus to percussion sounds, we did not consider the feature level accuracy, but focussed only on the perceptual

similarity between imitations and imitated sounds. The different approaches to quantifying imitation accuracy, whilst both valid, are complementary, and as such it would be insightful to 'complete the circle' of analysis by considering the perceptual accuracy of imitations in Chapter 3, and the feature–level accuracy of imitations in Chapter 4. Doing so would allow us to establish the relationship between and individual merits of these two approaches to evaluating vocal imitations.

In addition, it has been shown that people tend to agree on what salient acoustic features have been captured and vocalised from imitated sounds [Lemaitre et al., 2017], but the strategy of an imitator will depend on the sound that is to be imitated, and different imitators apply different strategies to imitate the same sounds (as discussed in Chapter 4 and by Lemaitre et al. [2016a]). For this reason, even if people are able to accurately imitate the acoustic features in a target sound, given the same task, 2 people may consider different features to be more or less important for vocalising the *essence* of a target sound. Moreover, they may both be correct, in terms of a listener being able to identify the imitated sound from the imitations! As such, there is clearly a perceptual bias that should be considered, or attempted to be controlled for, when quantifying vocal control of multiple acoustic features: it may not be that the imitator cannot exercise sufficient control over their vocal tract to imitate a particular sound, but rather that their interpretation of the sound does not require equally faithful reproduction of all features. In our experiments we have mostly ignored this perceptual bias (although the effect of imitator is included in all the statistical models used for inference), however it is clearly an important factor if one wishes to apply this type of method to real–world sounds, which are likely to contain variations of more than 2 features and with considerably more complex feature envelopes.

The objectives in Chapter 5 were focussed on exploring audio features by comparing distance between sounds in the feature spaces to perceptual similarity ratings, with the aim of finding a suitable Euclidean space that best represents the perceptual distance between vocalisations and percussion sounds. Some obvious 'low hanging fruits' for further research are to extend the presented methods, with variations on the network architectures, distance measures, feature learning methods, and data augmentation. In addition, alternative prediction models (such as logistic models and other supervised classifiers) may be equally, if not more effective than unsupervised models for the application of QBV, as demonstrated by Zhang and Duan [2017]. One

such approach would be to use deep neural networks to predict the similarity ratings *directly* from the features (similar to the approach by Zhang and Duan [2017], but using ratings instead of class labels). Such models are however subject to certain constraints. One major benefit of our approach is that searching a Euclidean space can be far more efficient than predicting and comparing the similarity between a query and all sounds in a library, as would be required in an end–to–end predictive model. Furthermore, to train such a model requires a large amount of labelled training data, and collecting similarity ratings can be resource–intensive. However, if the constraints of computational simplicity and data availability are lifted then it would be useful to investigate the performance of end–to–end models.

We may also consider whether the salient cues that listeners use to decide if an imitation is more or less similar to an imitated sound are the same for percussion and non–percussion sounds, and if not, how they differ. A first step in this direction could be to apply the models from Chapter 5 to new sounds, by refitting them without any fixed effects that are specific to the experimental design (such as the imitated sound), and using only the global parameters to predict the perceptual similarity for new vocalisations and non–vocal sounds. Both the CAE network and LMER models were trained on percussion sounds (although vocal imitations of non–percussion sounds and instruments were also used to train the CAE), therefore we might not expect this model to transfer well to vocalisations of non–percussion sounds. Nonetheless, applying the fitted models to different types of sounds will elucidate the generalisability of the estimated *similarity~distance* relationship. On a practical note, this type of experiment could be evaluated using (additional) similarity ratings for imitations of non–percussion sounds, or implemented as a complete QBV system and tested in a more qualitative manner.

Finally, we note that despite being motivated by the problem of search, the research presented in this thesis does not directly address it, but instead investigates a number of avenues that offer prerequisites for informing the design of QBV systems. There are many facets that are yet to be addressed, such as the user experience of QBV systems. To date there has been little (if any) research directly addressing how QBV systems compare to alternative search methods for music production and computer–based musicians. As noted by Stowell [2010], people behave quite differently when presented with an actual system that requires them to externalise their musical ideas in the form of vocalisations. They might feel inhibited or uncomfortable doing so, and the

way in which people interact with a music–sample based QBV system may, or indeed, probably does, differ from how they behave in a controlled vocal production experiment. This highlights some limitations about what we are able to infer from the results of the present experiments: the gap between our findings and real–world QBV systems is ripe for future research on user experience and sonic interaction design. Whilst further research is required to understand how the voice might be used to solve the problem of audio search, we have contributed to a number of areas that span far beyond the notion of QBV, including vocal analysis, communicability of vocal imitations, and audio feature learning.

References

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

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

K. Andersen and F. Grote. Giantsteps: Semi-structured conversations with musicians. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2295–2300, Seoul, Korea, 2015.

B. Andreeva, G. Demenko, M. Wolska, B. Möbius, F. Zimmerer, J. Jügler, M. Jastrzebska, and J. Trouvain. Comparison of pitch range and pitch variation in Slavic and Germanic languages. In *Proceedings of the 7th International Conference on Speech Prosody*, volume 7, pages 776–780, Dublin, Ireland, 2014.

M. Atherton. Rhythm-speak: Mnemonic, language play or song? In *Proceedings of the International Conference on Music Communication Science*, pages 15–18, Sydney, Australia, 2007.

J.-J. Aucouturier and F. Pachet. Music similarity measures: What's the use? In *Proceedings of the International Conference on Music Information Retrieval*, Paris, France, 2002.

J.-J. Aucouturier and F. Pachet. A scale-free distribution of false positives for a large class of audio similarity measures. *Pattern recognition*, 41(1): 272–284, 2008.

L. Bailly, N. Henrich, and X. Pelorson. Vocal fold and ventricular fold vibration in period–doubling phonation: Physiological description and aerodynamic modeling. *The Journal of the Acoustical Society of America*, 127 (5):3212–3222, 2010.

R. J. Baken and R. F. Orlikoff. *Clinical measurement of speech and voice*. Cengage Learning, 2nd edition, 2000.

S. Baldan, S. Delle Monache, D. Rocchesso, and H. Lachambre. Sketching sonic interactions by imitation-driven sound synthesis. In *Proceedings of the*

13th International Conference on Sound and Music Computing. Hamburg, Germany, 2016.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

Y. Bengio et al. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

A. Berenzweig, B. Logan, D. P. Ellis, and B. Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, 28(2):63–76, 2004.

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

R. Blaylock, N. Patil, T. Greer, and S. Narayanan. Sounds of the human vocal tract. In *Interspeech*, pages 2287–2291, Stockholm, Sweden, 2017.

D. Z. Borch and J. Sundberg. Some phonatory and resonatory characteristics of the rock, pop, soul, and Swedish dance band styles of singing. *Journal of Voice*, 25(5):532–537, 2011.

D. Z. Borch, J. Sundberg, P.-Å. Lindestad, and M. Thalen. Vocal fold vibration and voice source aperiodicity in 'dist' tones: A study of a timbral ornament in rock singing. *Logopedics Phoniatrics Vocology*, 29(4):147–153, 2004.

H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59(4):291–294, 1988.

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

J. Bullock. Libxtract: A lightweight library for audio feature extraction. In *Proceedings of the International Computer Music Conference*, volume 43, Copenhagen, Denmark, 2007.

J. Bullock. *Implementing audio feature extraction in live electronic music*. PhD thesis, Birmingham City University, 2008.

A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg. Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *The*

Journal of the Acoustical Society of America, 118(1):471–482, 2005.

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

M. Cartwright and B. Pardo. Synthassist: Querying an audio synthesizer by vocal imitation. In *Proceedings of the Conference on New Interfaces for Musical Expression*, London, England, 2014.

M. Cartwright and B. Pardo. Vocalsketch: Vocally imitating audio concepts. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul, Korea, 2015.

M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman. Fast and easy crowdsourced perceptual audio evaluation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 619–623, Shanghai, China, 2016.

M. Casey. General sound classification and similarity in MPEG-7. *Organised Sound*, 6(2):153–164, 2001.

E. Chávez, G. Navarro, R. Baeza-Yates, and J. L. Marroquín. Searching in metric spaces. *ACM Computing Surveys (CSUR)*, 33(3):273–321, 2001.

K. Choi, G. Fazekas, and M. Sandler. Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*, 2016a.

K. Choi, G. Fazekas, and M. Sandler. Explaining deep convolutional neural networks on music classification. *arXiv preprint arXiv:1607.02444*, 2016b.

K. Choi, G. Fazekas, M. Sandler, and K. Cho. Transfer learning for music classification and regression tasks. *arXiv preprint arXiv:1703.09179*, 2017.

F. Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

J. Cohen. *Statistical power analysis for the behavioral sciences*. Erlbaum, NJ, USA, 2nd edition, 1988.

G. Coleman. Mused: Navigating the personal sample library. In *Proceedings of the International Computer Music Conference*, Copenhagen, Denmark, 2007. Citeseer.

R. F. Coleman, J. H. Mabis, and J. K. Hinson. Fundamental frequency-sound pressure level profiles of adult male and female voices. *Journal of Speech, Language, and Hearing Research*, 20(2):197–204, 1977.

R. H. Colton. Vocal intensity in the modal and falsetto registers. *The Journal of the Acoustical Society of America*, 47(1A):105, 1970.

Y. M. Costa, L. S. Oliveira, and C. N. Silla. An evaluation of convolutional neural networks for music classification using spectrograms. *Applied Soft Computing*, 52:28–38, 2017.

M. J. Crawley. *The R book*. John Wiley & Sons, 2012.

C. d'Alessandro, S. Rosset, and J.-P. Rossi. The pitch of short-duration fundamental frequency glissandos. *The Journal of the Acoustical Society of America*, 104(4):2339–2348, 1998.

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

A. Del Piccolo and D. Rocchesso. Non-speech voice for sonic interaction: a catalogue. *Journal on Multimodal User Interfaces*, pages 1–17, 2016.

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

L. Deng, M. L. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton. Binary coding of speech spectrograms using a deep auto-encoder. In *Interspeech*, Makuhari, Japan, 2010.

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

S. Dieleman and B. Schrauwen. End-to-end learning for music audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6964–6968, Florence, Italy, 2014. IEEE.

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

M. Echternach and B. Richter. Vocal perfection in yodelling—pitch stabilities and transition times. *Logopedics Phoniatrics Vocology*, 35(1):6–12, 2010.

P. Edmiston, M. Perlman, and G. Lupyan. Creating words from iterated vocal imitation. In *39th Annual Conference of the Cognitive Science Society*, London, England, 2017.

T. M. Elliott, L. S. Hamilton, and F. E. Theunissen. Acoustic structure of

the five perceptual dimensions of timbre in orchestral instrument tones. *The Journal of the Acoustical Society of America*, 133(1):389–404, 2013.

D. P. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence. The quest for ground truth in musical artist similarity. In *Proceedings of the International Conference on Music Information Retrieval*, Paris, France, 2002.

P. Esling and C. Agon. Multiobjective time series matching for audio classification and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2057–2072, 2013.

H. Fastl and E. Zwicker. *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media, 2006.

E. Ferragne and F. Pellegrino. Formant frequencies of vowels in 13 accents of the British Isles. *Journal of the International Phonetic Association*, 40 (1):1–34, 2010.

I. Ferrante. Vibrato rate and extent in soprano voice: A survey on one century of singing. *The Journal of the Acoustical Society of America*, 130 (3):1683–1688, 2011.

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

J. L. Flanagan. *Speech analysis synthesis and perception.* Academic, New York, 1965.

A. Flexer, D. Schnitzer, and J. Schlüter. A MIREX meta-analysis of hubness in audio music similarity. In *Proceedings of the International Conference on Music Information Retrieval*, pages 175–180, Porto, Portugal, 2012.

F. Font and G. Bandiera. Freesound explorer: Make music while discovering freesound! In *Proceedings of the 3rd Web Audio Conference*, London, England, 2017.

D. J. Freed. Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. *The Journal of the Acoustical Society of America*, 87(1):311–322, 1990.

O. Fried, Z. Jin, R. Oda, and A. Finkelstein. Audioquilt: 2D arrangements of audio samples using metric learning and kernelized sorting. In *Proceedings of the Conference on New Interfaces for Musical Expression*, pages 281–286, London, England, 2014.

C. J. Frisbie. Anthropological and ethnomusicological implications of a comparative analysis of Bushmen and African Pygmy music. *Ethnology*, 10(3): 265–290, 1971.

H. Fujisaki. Dynamic characteristics of voice fundamental frequency in speech and singing. In *The production of speech*, pages 39–55. Springer, 1983.

M. Garcia. Observations on the human voice. *Proceedings of the Royal Society of London*, 7:399–410, 1854.

M. J. Gardner and D. G. Altman. Confidence intervals rather than p values: estimation rather than hypothesis testing. *British Medical Journal (Clinical Research Edition)*, 292(6522):746–750, 1986.

T. Gay. Effect of speaking rate on diphthong formant movements. *The Journal of the Acoustical Society of America*, 44(6):1570–1573, 1968.

M. P. Gelfer. The stability of total phonational frequency range. *The Journal of the Acoustical Society of America*, 79:83, 1986.

B. R. Gerratt and J. Kreiman. Toward a taxonomy of nonmodal phonation. *Journal of Phonetics*, 29(4):365–381, 2001.

X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

C. Gobl, A. Nı, et al. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1):189–212, 2003.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

M. Gordon and P. Ladefoged. Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29(4):383–406, 2001.

E. M. Grais and M. D. Plumbley. Single channel audio source separation using convolutional denoising autoencoders. *arXiv preprint arXiv:1703.08019*, 2017.

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

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

M. Gratier and E. Devouche. Imitation and repetition of prosodic contour in vocal interaction at 3 months. *Developmental Psychology*, 47(1):67, 2011.

J. M. Grey. Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, 61(5):1270–1277, 1977.

J. M. Grey and J. W. Gordon. Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America*, 63(5): 1493–1500, 1978.

D. Guinn and A. Nazarov. Evidence for phonological features and phonotactics in beatboxing vocal percussion. In *Old World Conference on Phonology*, London, England, January 2018. UCL.

B. Gygi, G. R. Kidd, and C. S. Watson. Similarity and categorization of environmental sounds. *Perception & Psychophysics*, 69(6):839–855, 2007.

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

M. A. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, University of Waikato, 1999.

P. Hamel, S. Lemieux, Y. Bengio, and D. Eck. Temporal pooling and multiscale learning for automatic annotation and ranking of music audio. In *Proceedings of the International Conference on Music Information Retrieval*, pages 729–734, 2011.

T. Harvey. *Keening as a cross-cultural phenomenon*. PhD thesis, The Ohio State University, 1993.

A. Hazan. Performing expressive rhythms with billaboop voice-driven drum generator. In *Proceedings of the 8th International Conference on Digital Audio Effects*, 2005.

S. Heise, M. Hlatky, and J. Loviscach. Aurally and visually enhanced audio search with soundtorch. In *Extended Abstracts on Human Factors in Computing Systems*, pages 3241–3246. ACM, 2009.

M. Helén and T. Lahti. Query by example methods for audio signals. In *Proceedings of the 7th Nordic Signal Processing Symposium*, pages 302–305. IEEE, 2006.

M. Helén and T. Lahti. Query by example in large databases using key-sample distance transformation and clustering. In *9th IEEE International*

Symposium on Multimedia Workshops, pages 303–308. IEEE, 2007.

M. Helén and T. Virtanen. Query by example of audio signals using Euclidean distance between Gaussian mixture models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages I–225, Honolulu, Hawaii, 2007. IEEE.

P. Helgason. Sound initiation and source types in human imitations of sounds. In *Proceedings of FONETIK*, pages 83–88, 2014.

P. Helgason, G. L. Salomão, and S. Ternström. A database of articulatory annotations of vocal imitations. Technical report, Royal Institute of Technology Sweden (KTH), 2016.

D. N. Henrich. Mirroring the voice from Garcia to the present day: Some insights into singing voice registers. *Logopedics Phoniatrics Vocology*, 31(1): 3–14, 2006.

N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo. Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. *The Journal of the Acoustical Society of America*, 117(3):1417–1430, 2005.

P. Herrera, A. Yeterian, and F. Gouyon. Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques. In *Proceedings of the International Conference on Music and Artificial Intelligence*, pages 69–80. Springer, 2002.

P. Herrera, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1):3–21, 2003.

N. Hewlett and J. M. Beck. *An introduction to the science of phonetics*. Routledge, 2006.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

H. Hollien. On vocal registers. *Journal of Phonetics*, 2:125–143, 1974.

H. Hollien, D. Dew, and P. Philips. Phonational frequency ranges of adults. *Journal of Speech, Language, and Hearing Research*, 14(4):755–760, 1971.

U. Hoppe, F. Rosanowski, M. Döllinger, J. Lohscheller, M. Schuster, and U. Eysholdt. Glissando: laryngeal motorics and acoustics. *Journal of Voice*, 17(3):370–376, 2003.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash, and K. Johnson. Formants of children, women, and men: The effects of vocal intensity variation. *The Journal of the Acoustical Society of America*, 106(3):1532–1542, 1999.

E. J. Humphrey and J. P. Bello. Rethinking automatic chord recognition with convolutional neural networks. In *11th International Conference on Machine Learning and Applications*, volume 2, pages 357–362. IEEE, 2012.

E. J. Humphrey, J. P. Bello, and Y. LeCun. Feature learning and deep architectures: New directions for music informatics. *Journal of Intelligent Information Systems*, 41(3):461–481, 2013.

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

International Telecommunication Union. ITU 1534-1: Method for the subjective assessment of intermediate quality level of coding systems. Technical report, International Telecommunication Union, 2003.

J. E. Jackson. *A user's guide to principal components*. John Wiley & Sons, 1991.

M. A. F. Jahn. *The singer's guide to complete health*. Oxford University Press, 2013.

J. Janer. *Singing-driven interfaces for sound synthesizers*. PhD thesis, UNIVERSITAT POMPEU FABRA, 2008.

A. Kapur, M. Benning, and G. Tzanetakis. Query-by-beat-boxing: Music retrieval for the DJ. In *Proceedings of the International Conference on Music Information Retrieval*, pages 170–177, Barcelona, Spain, 2004.

K. Kato and A. Ito. Acoustic features and auditory impressions of death growl and screaming voice. In *9th International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 460–463. IEEE, 2013.

P. Keating, M. Garellek, and J. Kreiman. Acoustic properties of different kinds of creaky voice. In *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, Scotland, 2015.

M. G. Kendall and B. B. Smith. The problem of m rankings. *The Annals of Mathematical Statistics*, 10(3):275–287, 1939.

R. D. Kent, J. F. Kent, and J. C. Rosenbek. Maximum performance tests of speech production. *Journal of Speech and Hearing Disorders*, 52(4):367–387, 1987.

J. B. King and Y. Horii. Vocal matching of frequency modulation in synthesized vowels. *Journal of Voice*, 7(2):151–159, 1993.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

M. Koller. An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, 75(6):1–24, 2016.

S. Kraft and U. Zölzer. BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality. In *Proceedings of the Linux Audio Conference*, Karlsruhe, Germany, 2014.

A. Krishnaswamy. Application of pitch tracking to South Indian classical music. In *Proceedings of the International Conference on Multimedia and Expo*, volume 3, pages III–389. IEEE, 2003.

J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

P. K. Kuhl and A. N. Meltzoff. Infant vocalizations in response to speech: Vocal imitation and developmental change. *The Journal of the Acoustical Society of America*, 100(4):2425–2438, 1996.

A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen. *lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package)*. R package version 2.0.33, 2016.

P. Ladefoged. *Preliminaries to linguistic phonetics*. University of Chicago Press, 1971.

P. Ladefoged. *A course in phonetics*. Hardcourt Brace Jovanovich Inc., New York, 3rd edition, 1993.

S. Lakatos. A common perceptual space for harmonic and percussive timbres. *Perception & psychophysics*, 62(7):1426–1439, 2000.

H. Lane, B. Tranel, and C. Sisson. Regulation of voice communication by sensory dynamics. *The Journal of the Acoustical Society of America*, 47(2B):

618–624, 1970.

H. L. Lane, A. C. Catania, and S. S. Stevens. Voice level: Autophonic scale, perceived loudness, and effects of sidetone. *The Journal of the Acoustical Society of America*, 33(2):160–167, 1961.

O. Lartillot and P. Toiviainen. A Matlab toolbox for musical feature extraction from audio. In *Proceedings of the International Conference on Digital Audio Effects*, pages 237–244, 2007.

K. Lederer. *The phonetics of beatboxing*. B.A. dissertation, Leeds University, UK, 2005.

H. Lee, P. Pham, Y. Largman, and A. Y. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in Neural Information Processing Systems*, pages 1096–1104, 2009.

G. Lemaitre and D. 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.

G. Lemaitre, O. Houix, N. Misdariis, and P. Susini. Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology: Applied*, 16(1):16–32, 2010.

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

G. Lemaitre, O. Houix, F. Voisin, N. Misdariis, and P. Susini. Vocal imitations of non-vocal sounds. *PloS one*, 11(12), 2016a.

G. Lemaitre, A. Jabbari, O. Houix, N. Misdariis, and P. Susini. Vocal imitations of basic auditory features. *The Journal of the Acoustical Society of America*, 137(4):2268–2268, 2016b.

G. Lemaitre, H. Scurto, J. Françoise, F. Bevilacqua, O. Houix, and P. Susini. Rising tones and rustling noises: Metaphors in gestural depictions of sounds. *PloS one*, 12(7):e0181786, 2017.

T. C. Levin and M. E. Edgerton. The throat singers of Tuva. *Scientific American*, 281(3):80–87, 1999.

P. Lloyd. Pitch (f0) and formant profiles of human vowels and vowel-like baboon grunts: the role of vocalizer body size and voice-acoustic allometry. *The Journal of the Acoustical Society of America*, 117:944, 2005.

A. Loscos and T. Aussenac. The wahwactor: a voice controlled wah-wah pedal. In *Proceedings of the Conference on New interfaces for Musical Expression*, pages 172–175. National University of Singapore, 2005.

E. Marchetto and G. Peeters. A set of audio features for the morphological description of vocal imitations. In *Proceedings of the 18th International Conference on Digital Audio Effects*, Trondheim, Norway, 2015.

W. L. Martens, A. Marui, and S. R. Area. Categories of perception for vibrato, flange, and stereo chorus: Mapping out the musically useful ranges of modulation rate and depth for delay based effects. In *Proceedings of the 9th International Conference on Digital Audio Effects*, 2006.

M. Mauch and S. Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663. IEEE, 2014.

D. Mauro and D. Rocchesso. Analyzing and organizing the sonic space of vocal imitations. In *Proceedings of the 10th Audio Mostly Conference on Interaction with Sound*. ACM, 2015.

S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3):177–192, 1995.

P. McLeod and G. Wyvill. A smarter way to find pitch. In *Proceedings of International Computer Music Conference, ICMC*, 2005.

E. Mercado III, J. T. Mantell, and P. Q. Pfordresher. Imitating sounds: A cognitive approach to understanding vocal imitation. *Comparative Cognition & Behavior Reviews*, 9, 2014.

D. Mitrović, M. Zeppelzauer, and C. Breiteneder. Features for content-based audio retrieval. *Advances in Computers*, 78:71–150, 2010.

B. C. Moore, B. R. Glasberg, and T. Baer. A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–240, 1997.

R. J. Morris, W. Brown, D. M. Hicks, and E. Howell. Phonational profiles of male trained singers and nonsingers. *Journal of Voice*, 9(2):142–148, 1995.

D. Mürbe, F. Pabst, G. Hofmann, and J. Sundberg. Effects of a professional

solo singer education on auditory and kinesthetic feedback—a longitudinal study of singers' pitch control. *Journal of Voice*, 18(2):236–241, 2004.

T. Nakano, J. Ogata, M. Goto, and Y. Hiraga. A drum pattern retrieval method by voice percussion. *Database (Musical Instrument Sound)*, 3:1, 2004.

J. Nelder. A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, pages 48–77, 1977.

J. G. Neuhoff. Perceptual bias for rising tones. *Nature*, 395(6698):123–124, 1998.

J. G. Neuhoff. An adaptive bias in the perception of looming auditory motion. *Ecological Psychology*, 13(2):87–110, 2001.

A. Novello, M. F. McKinney, and A. Kohlrausch. Perceptual evaluation of music similarity. In *Proceedings of the International Conference on Music Information Retrieval*, pages 246–249, 2006.

A. Odena, V. Dumoulin, and C. Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

J. J. Ohala and W. G. Ewan. Speed of pitch change. *The Journal of the Acoustical Society of America*, 53(1):345–345, 1973.

D. O'Shaughnessy. Interacting with computers by voice: automatic speech recognition and synthesis. *Proceedings of the IEEE*, 91(9):1272–1305, 2003.

E. Pampalk, P. Herrera, and M. Goto. Computational models of similarity for drum samples. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):408–423, 2008.

S. Parekh, F. Font, and X. Serra. Improving audio retrieval through loudness profile categorization. In *2016 IEEE International Symposium on Multimedia*, pages 565–568. IEEE, 2016.

E. Parizet and V. Koehl. Application of free sorting tasks to sound quality experiments. *Applied Acoustics*, 73(1):61–65, 2012.

A. D. Patel and J. R. Iversen. Acoustic and perceptual comparison of speech and drum sounds in the north Indian tabla tradition: An empirical study of sound symbolism. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 925–928, 2003.

G. Peeters. A large set of audio features for sound description (similarity and

description) in the CUIDADO project. *CUIDADO Project Report*, 2004.

G. Peeters, S. McAdams, and P. Herrera. Instrument sound description in the context of MPEG-7. In *Proceedings of the International Computer Music Conference*, pages 166–169, 2000.

G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams. The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916, 2011.

M. Perlman and G. Lupyan. The potential for iconicity in vocalization. *bioRxiv*, page 148841, 2017.

P. Q. Pfordresher and S. Brown. Poor-pitch singing in the absence of "tone deafness". *Music Perception: An Interdisciplinary Journal*, 25(2):95–115, 2007.

P. Q. Pfordresher, S. Brown, K. M. Meier, M. Belyk, and M. Liotti. Imprecise singing is widespread. *The Journal of the Acoustical Society of America*, 128 (4):2182–2190, 2010.

E. Prame. Measurements of the vibrato rate of ten singers. *The Journal of the Acoustical Society of America*, 96(4):1979–1984, 1994.

J. Přibil and A. Přibilová. Comparison of complementary spectral features of emotional speech for German, Czech, and Slovak. *Cognitive Behavioural Systems*, 7403:236–250, 2012.

M. Proctor, E. Bresch, D. Byrd, K. Nayak, and S. Narayanan. Paralinguistic mechanisms of production in human "beatboxing": A real-time magnetic resonance imaging study. *Journal of the Acoustical Society of America*, 133 (2):1043–1054, 2013.

P. Proutskova, C. Rhodes, T. Crawford, and G. Wiggins. Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing. *Journal of New Music Research*, 42(2):171–186, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

L. R. Rabiner and R. W. Schaffer. *Digital processing of speech signals*. Prentice Hall, 1978.

A. F. S. Ramires. Automatic transcription of vocalized percussion. Master's thesis, Universidade do Porto, 2017.

J. Reigado and H. Rodrigues. Vocalizations produced in the second year of life in response to speaking and singing. *Psychology of Music*, pages 1–12, 2017.

D. Rocchesso, D. A. Mauro, and C. Drioli. Organizing a sonic space through vocal imitations. *Journal of the Audio Engineering Society*, 64(7/8):474–483, 2016a.

D. Rocchesso, D. A. Mauro, and S. D. Monache. mimic: The microphone as a pencil. In *Proceedings of the Tenth International Conference on Tangible, Embedded, and Embodied Interaction*, pages 357–364. ACM, 2016b.

G. Roma and X. Serra. Querying freesound with a microphone. In *Proceedings of the First Web Audio Conference*, Paris, France, 2015.

T. D. Rossing, F. R. Moore, and P. A. Wheeler. *The science of sound*. Addison Wesley, San Francisco, 3rd edition, 2001.

B. Roubeau, N. Henrich, and M. Castellengo. Laryngeal vibratory mechanisms: The notion of vocal register revisited. *Journal of Voice*, 23(4): 425–438, 2009.

S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

J. L. Santacruz, L. J. Tardón, I. Barbancho, A. M. Barbancho, and E. Molina. Voice2tuba: transforming singing voice into a musical instrument. *Multimedia Tools and Applications*, 76(7):9855–9875, 2016.

G. Scavone, S. Lakatos, P. Cook, and C. Harbke. Perceptual spaces for sound effects obtained with an interactive similarity rating program. In *Proceedings of the International Symposium on Musical Acoustics*, pages 487–491, Perugia, Italy, 2001.

D. Scherer, A. Müller, and S. Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 92–101. Springer, 2010.

J. Schlüter and S. Böck. Musical onset detection with convolutional neural networks. In *Proceedings of the 6th International Workshop on Machine Learning and Music*, 2013.

J. Schlüter and T. Grill. Exploring data augmentation for improved singing voice detection with neural networks. In *Proceedings of the International*

Conference on Music Information Retrieval, pages 121–126, 2015.

R. C. Schmidt. Managing delphi surveys using nonparametric statistical techniques. *Decision Sciences*, 28(3):763–774, 1997.

E. Schubert and J. Wolfe. Does timbral brightness scale with frequency and spectral centroid? *Acta Acustica united with Acustica*, 92(5):820–825, 2006.

W. Sethares. *Tuning, timbre, spectrum, scale*. Springer, New York, 1998.

J. Sharpe. Jimmie riddle and the lost art of eephing. `https://www.npr.org/templates/story/story.php?storyId=5259589`, 2006.

S. Siddiq, C. Reuter, I. Czedik-Eysenberg, and D. Knauf. Towards the comparability and generality of timbre space studies. In *Proceedings of the Third Vienna Talk on Music Acoustics*, pages 237–240, 2015.

K. Siedenburg and S. McAdams. Four distinctions for the auditory wastebasket of timbre. *Frontiers in Psychology*, 8:1747, 2017.

K. Siedenburg, I. Fujinaga, and S. McAdams. A comparison of approaches to timbre descriptors in music information retrieval and music psychology. *Journal of New Music Research*, 45(1):27–41, 2016.

E. Sinyor, R. Fiebrink, C. McKay, D. McEnnis, and I. Fujinaga. Beatbox classification using ACE. In *Proceedings of the International Conference on Music Information Retrieval*, pages 672–675, London, England, 2005.

M. Slaney. Web-scale multimedia analysis: Does content matter? *IEEE MultiMedia*, 18(2):12–15, 2011.

C. Spevak and E. Favreau. Soundspotter-a prototype system for content-based audio retrieval. In *Proceedings of the 5th International Conference on Digital Audio Effects*, 2002.

T. Sporer, J. Liebetrau, and S. Schneider. Statistics of MUSHRA revisited. In *Proceedings of the 127th Audio Engineering Society Convention*, pages 323–331, New York, USA, 2009.

J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

D. Stowell. *Making music through real-time voice timbre analysis: machine learning and timbral control*. PhD thesis, Queen Mary University of London, 2010.

D. Stowell and M. D. Plumbley. Characteristics of the beatboxing vocal style: C4dmtr-08-01. Technical report, Queen Mary, University of London, UK, 2008.

A. M. Sulter, H. K. Schutte, and D. G. Miller. Differences in phonetogram features between male and female subjects with and without vocal training. *Journal of Voice*, 9(4):363–377, 1995.

J. Sundberg. Data on maximum speed of pitch changes. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 4:39–47, 1973.

J. Sundberg. Articulatory interpretation of the "singing formant". *The Journal of the Acoustical Society of America*, 55(4):838–844, 1974.

J. Sundberg. *The Science of the Singing Voice*. Northern Illinois University Press, Illinois, USA, 1989.

J. Sundberg. Vocal fold vibration patterns and phonatory modes. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 35:69–80, 1994a.

J. Sundberg. Acoustic and psychoacoustic aspects of vocal vibrato. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 35(2-3):045–068, 1994b.

J. Sundberg, I. Titze, and R. Scherer. Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source. *Journal of Voice*, 7(1):15–29, 1993.

H. Terasawa, M. Slaney, and J. Berger. The thirteen colors of timbre. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 323–326, New Paltz, New York, 2005.

E. Terhardt. Calculating virtual pitch. *Hearing Research*, 1(2):155–182, 1979.

A. Tindale. Classification of snare drum sounds using neural networks. Master's thesis, McGill University, 2004.

I. R. Titze. Comments on the myoelastic-aerodynamic theory of phonation. *Journal of Speech, Language, and Hearing Research*, 23(3):495–510, 1980.

C.-G. Tsai, L.-C. Wang, S.-F. Wang, Y.-W. Shau, T.-Y. Hsiao, and W. Auhagen. Aggressiveness of the growl-like timbre: acoustic characteristics, musical implications, and biomechanical mechanisms. *Music Perception: An Interdisciplinary Journal*, 27(3):209–222, 2010.

C. Turquois, M. Hermant, D. Gómez-Marín, and S. Jordà. Exploring the benefits of 2D visualizations for drum samples retrieval. In *Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval*, pages 329–332. ACM, 2016.

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

K. Ullrich, J. Schlüter, and T. Grill. Boundary detection in music structure analysis using convolutional neural networks. In *Proceedings of the International Conference on Music Information Retrieval*, pages 417–422, 2014.

B. White, A. Mehrabi, and M. B. Sandler. An archival echo: Recalling the public domain through real-time query by vocalisation. In *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, pages 42:1–42:4, New York, NY, USA, 2017. ACM. doi: 10.1145/3123514.3123546.

G. Wichern, J. Xue, H. Thornburg, and A. Spanias. Distortion-aware query-by-example for environmental sounds. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 335–338. IEEE, 2007.

F. Wickelmaier, N. Umbach, K. Sering, and S. Choisel. Comparing three methods for sound quality evaluation with respect to speed and accuracy. In *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

E. Wold, T. Blum, D. Keislar, and J. Wheaten. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 3(3):27–36, 1996.

D. Wolff and T. Weyde. Learning music similarity from relative user ratings. *Information Retrieval*, 17(2):109–136, 2014.

Y. Xu and X. Sun. How fast can we really change pitch? Maximum speed of pitch change revisited. In *Proceedings of the 6th Internation Conference on Spoken Language Processing*, Beijing, China, 2000.

Y. Xu and X. Sun. Maximum speed of pitch change and how it may relate to speech. *The Journal of the Acoustical Society of America*, 111(3):1399–1413, 2002.

J. Xue, G. Wichern, H. Thornburg, and A. Spanias. Fast query by example of environmental sounds via robust and efficient cluster-based indexing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5–8. IEEE, 2008.

M. Yadav. *Loudness of the singing voice: A room acoustics perspective.* PhD thesis, University of Sydney, 2016.

A. Young. The voice-index and digital voice interface. *The Leonardo Music Journal*, 2014.

A. Zeileis, F. Leisch, K. Hornik, and C. Kleiber. strucchange. An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7:1–38, 2001.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*, pages 818–833. Springer, 2014.

T. Zhang and C.-C. Kuo. Hierarchical classification of audio data for archiving and retrieving. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3001–3004. IEEE, 1999.

Y. Zhang and Z. Duan. Retrieving sounds by vocal imitation recognition. In *Proceedings of the 25th IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, Boston, USA, 2015.

Y. Zhang and Z. Duan. Imisound: An unsupervised system for sound query by vocal imitation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2269–2273, Shangai, China, 2016a.

Y. Zhang and Z. Duan. Supervised and unsupervised sound retrieval by vocal imitation. *Journal of the Audio Engineering Society*, 64(7/8):533, 2016b.

Y. Zhang and Z. Duan. IMINET: Convolutional semi-Siamese networks for sound search by vocal imitation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, 2017.

R. I. Zraick, J. L. Nelson, J. C. Montague, and P. K. Monoson. The effect of task on determination of maximum phonational frequency range. *Journal of Voice*, 14(2):154–160, 2000.

Appendix A

Heuristic feature specifications

This appendix contains further details of the heuristic features used for the experiments in Chapter 5. Note that unless stated otherwise, the implementations of these features are based on the definitions given by Peeters [2004].

Global features:

Log attack time is taken as per the fixed threshold method from Peeters [2004]. This is calculated from the energy envelope, which is the instantaneous RMS of the time–domain signal calculated per frame, using a frame size of 512 samples (11ms). This is considerably shorter than the 100ms window size suggested by Peeters [2004], however it is more suited to percussion sounds (where we are interested in relatively high temporal resolution), and matches the hop size used to extract the spectral features. As the audio files were manually edited to start at the beginning of the sound, the beginning of the file and the maximum value are selected as the start and end points of the attack, respectively.

Temporal crest factor, as defined in Stowell [2010] is the ratio of the maximum value over the mean, and is also computed on the energy envelope.

Duration is taken as the entire length of the edited audio file (as opposed to the effective duration defined in Peeters [2004]).

Zero crossing rate is taken as the number of times the time–domain signal

crosses zero, per second.

Decay time is taken as the time from the maximum value to 50% of the maximum value in the time–domain, as per Patel and Iversen [2003] (computed from the smoothed energy envelope).

Frame–wise features:

Pitch and (pitch) clarity are computed using the same approach as Stowell [2010], which is an auto–correlation based time–domain method from McLeod and Wyvill [2005]. This produces a *normalised square difference function*, giving the difference between 2 copies of the same signal at each lag position. The pitch is then calculated from the strongest peak (ignoring the value at a lag of 0), and clarity is taken as the strength (i.e. magnitude) of that peak.

Noisiness is the ratio of noise energy over total energy. The noise energy is computed as the difference between harmonic and total energy, where harmonic energy is taken as the sum of energy over all harmonics (odd and even) of F_0.

Roughness is calculated using the model from Sethares [1998], which is similar to the approach used by Lartillot and Toiviainen [2007]. Spectral peaks are extracted using the *contrast* method from Lartillot and Toiviainen [2007] and the sum of sensory dissonances is calculated for each partial pair (taken from the power spectrum), as per Sethares [1998], weighted by the product of the loudness for each partial pair (in Sones). The median and IQR roughness values are calculated using only frames where more than 1 contrasting peak is detected.

Inharmonicity is the energy weighted difference between the partials and integer multiples of F_0. The partials are calculated using the same peak picking method that is used for roughness.

Spectral centroid is defined as the amplitude weighted mean frequency (i.e. the barycentre of the spectrum).

Spectral rolloff is the frequency below which n% of the energy exists. A value of n between 85%–95% is commonly used. In line with Stowell [2010] we calculate this for a high percentile (95%), the median (50%) and quartiles

(25% and 75%).

Spectral crest factor is the frequency–domain equivalent to the temporal crest factor, taken as the maximum amplitude value over the mean [Stowell, 2010].

Spectral slope is the slope calculated from a linear regression of the spectrum.

Spectral spread is the variance of the spectral centroid.

Spectral kurtosis is the 4th order moment of the spectrum (i.e. the flatness or peakiness of the distribution).

Spectral flatness is the geometric mean of the spectrum over the arithmetic mean.

Spectral skewness is the 3rd order moment of the spectrum (i.e. the asymmetry of the spectral distribution).

Spectral entropy is the Shannon entropy, calculated from the probability density function of the spectrum.

Spectral compactness measures the sum of amplitude differences between adjacent bins. For a given bin, n, the difference, d_n is taken as the difference between the amplitude of n and the mean of the amplitudes for bins: $n-1$, n, and $n+1$, as per the definition by Sinyor et al. [2005].

Strongest frequency is simply the frequency of the highest amplitude bin [Sinyor et al., 2005].

Spectral flux describes how much the spectrum varies over time, measured as the difference between 2 successive frames as per Peeters et al. [2011]. For 2 given spectral frames, X_t, X_{t-1}, this is calculated as $1 - \frac{X_t \cdot X_{t-1}}{\|X_t\| \|X_{t-1}\|}$.

Band–specific power is the sum of power in a particular frequency band. There is no standard definition for which band values to select: Hazan [2005] uses bands 100Hz–2kHz, 2–6kHz, 6–10kHz; Stowell [2010] uses 5 log spaced bands: 50–400Hz, 400–800Hz, 800Hz–1.6kHz, 1.6–3.2kHz, and 3.2–6.4kHz; and Herrera et al. [2002] use the relative percentage of energy in each of 8 bands, which were chosen by experimentation and observation of the data. We use the 5 log spaced bands from Stowell [2010]. We also include a

measure for overall power, simply calculated as the sum of power in all bands.

LPC coefficients provide a compact estimation of the spectral envelope and are commonly used in speech coding. They exploit the source–filter nature of the voice by estimating the formant frequencies using all–pole filters [Rabiner and Schaffer, 1978]. This provides both a spectral envelope (i.e. the LPC coefficients) and a source signal (the residual from inverse filtering the coefficients with the original signal). We take the 10 coefficients from an 11th order LPC (ignoring the first coefficient), using the Python implementation of LPC from the scikits.talkbox[1] Python package.

MFCCs and ΔMFCCs. MFCCs are calculated from the discrete cosine transform (DCT) of the log–magnitude, Mel–scaled spectrum. We take the first 13 MFCC coefficients (ignoring the first coefficient) from the implementation in the scikits.talkbox Python package, and include both first and second order Δ, giving 39 features per frame.

[1] https://scikits.appspot.com/talkbox