DEPARTMENT OF LINGUISTICS AND ENGLISH LANGUAGE

Verbalising Timbre: A Neural Network Approach

Leo Barlow

A dissertation of 10,402 words, submitted to the University of Edinburgh in accordance with the requirements of the degree of Master of Arts with Honours Linguistics in the School of Philosophy, Psychology & Language Sciences.

Tuesday 27th July, 2021

Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA	:	Analysis of Variance
MDS	:	Multidimensional Scaling
DNN	:	Deep Neural Network
VAME	:	Verbal Attribute Magnitude Estimation
MFCC	:	Mel-Frequency Cepstral Coefficient
ReLU	:	Rectified Linear Unit

List of Symbols

ρ_T	:	Tau-equivalent reliability coefficient
r	:	Pearson product-moment correlation coefficient
ρ	:	Spearman's rank correlation coefficient

Acknowledgements

I'd like to thank my supervisors, Prof. Graeme Trousdale and Dr. Hannah Rohde, who guided this project from incoherent rambling to rhetorical-question-free results. The rest I owe to my family, for their support and endless patience.

Chapter 1

Introduction

Timbre is notoriously an elusive concept. On hearing a musical instrument, an unfamiliar voice, or even the piercing whistle of a kettle, we can easily pick out how they 'sound'. However, when it comes to describing them, our vocabulary seems hopelessly vague. What acoustic properties might *piercing* — a tactile metaphor — convey? Nonetheless, using metaphorical scales like *warmth* or *brightness* to verbalise timbre is pervasive, even in the music industry (Porcello, 2004). Attempts to pin down what these scales mean stretch back to Von Helmholtz (1877), but this is still not a solved problem: timbral metaphors just don't map neatly to acoustic properties of the sounds they describe. As a result, predicting judgements as to the metaphorical *warmth* or *brightness* of a timbre is difficult to automate.

In recent years, neural networks have exploded in popularity for their ability to solve this sort of 'classification problem'. Although there has been no shortage of neural models in the more objective art of instrument recognition, their potential for placing sounds on abstract, metaphorical scales hasn't been thoroughly explored. Automatically verbalising timbre this way is most obviously applicable to music production software, including categorising audio clips so they can be searched for using metaphors, and making synthesiser controls more intuitive. The ability to accurately predict a timbre's *warmth* or *brightness* would also add to our theoretical picture, by indicating that these metaphors are expressive, despite feeling vague. As proposed by theories of cross-modal metaphor, *warmth* might be used to describe a timbre that subconsciously feels warm (Marks et al., 1987; Marks, 1982).

This dissertation presents a neural network approach to verbalising timbre, by creating networks to predict metaphorical timbre judgements. In Chapter 2, we will see that these judgements ought to be reasonably predictable. However, attempts to do so statistically may be flawed, and existing neural approaches are underdeveloped. Chapter 3 covers how I created a dataset of metaphorical judgements for neural network modelling. We will then go over recent advances in deep learning, and how they were put into practice with two neural networks to predict those judgements. The network results and judgement data statistics are given in Chapter 4, showing that the best network's predictions were on average 87%

likely to fall within two points of the ten-point metaphorical judgements in a separate set of test data. These findings are then critically discussed in Chapter 5, which concludes by considering directions for future research. Finally, Chapter 6 summarises the dissertation as a whole.

Chapter 2

Background

Following the history of the word ‘timbre’ reveals it is actually a fairly modern concept. Widely considered its first circulated definition, Rousseau’s entry in a 1765 encyclopedia struggles to disentangle timbre from emotional impact (violins being ‘soft’ and therefore ‘beautiful’) (Saitis & Weinzierl, 2019). In a large-scale corpus of words from the nineteenth to twenty-first centuries, the English term only began to take off around 1860, before relinquishing its impressionistic meaning and reaching a peak in usage during the signal processing boom of the 1970s and 1980s (Wallmark & Kendall, 2018). However, it hasn’t become any easier to define: the Oxford English Dictionary gives it as the “character or quality of a musical or vocal sound” (OED online, 2020), where scientific work can only agree that “it is not loudness and it is not pitch” (Bregman, 1994, p. 93). Section 2.1 of this chapter discusses whether the abstract concept of timbre is something we can actually verbalise. Section 2.2 then reviews a number of ways it can be verbalised automatically, as well as the practical and theoretical implications of an accurate solution to this problem.

2.1 Timbral ineffability

The subjective is typically taken to be ineffable: when describing how an apple tastes, or how a trumpet sounds, words fail to fully capture those stimuli — there’s only so much that *sweet* or *loud* can convey about inherently multidimensional experiences. However, some of these dimensions seem harder to put into words than others. A trumpet can be *loud* or *quiet*, with a pitch *high* or *low*, but it isn’t easy to characterise its timbre. We can therefore ask whether timbre is fundamentally ineffable.

2.1.1 Lexical coding

A reason to think it might be is that it doesn’t seem to be lexically coded. Cross-linguistically, the sensory modalities aren’t all coded equally. For example, most languages have far fewer words to describe smell and taste than visual characteristics like colour and size (Majid et al., 2018). This effect is so pronounced that smell has been designated a 'muted' sense (Olofsson & Gottfried, 2015). It is speculated that these differences in coding reflect evolutionary preferences. For example, verbalising vision might have been the most important to our survival (Levinson & Majid, 2014). Indeed, we talk about some modalities more than others: in a large English subtitle corpus, references to vision were twice as common as references to audition, and more than ten times as frequent as references to smell (Winter et al., 2018).

Timbre seems to fall into the same category as smell or taste. Where colour can be described with domain-specific words like *red*, *white* and *blue*, when verbalising timbre we resort to onomatopoeia (*hissing*, *clicking*, *booming*), properties of the source instrument or object (*wooden*, *metallic*, *nasal*) or most commonly, cross-modal metaphors (*warm*, *bright*, *smooth*) (Wallmark & Kendall, 2018). This isn't a problem for pitch, where it's clear what the spatial metaphors *high* and *low* refer to. However, for timbre it isn't obvious whether the different cross-modal metaphors are actually expressive. A metaphor is expressive according to how well it characterises perceptual effects, and how consistently it is used within and between individuals.

2.1.2 Cross-modal metaphor

Unlike stimuli that span modalities, such as foods (which have both a smell and a taste), or surfaces (which have both colour and texture), sounds are highly exclusive (Lynott & Connell, 2013). Any objectively cross-modal effects are likely to arise only from the lowest frequencies, for example when feeling the throb of a powerful sound system, or seeing the strings vibrate on a double bass.

That said, cross-modal metaphors could be less vague than they seem, according to the theory that modalities like colour, taste or timbre are unified. There is debate as to whether this unity is partially innate (a violin's sound is *thin* because it is processed similarly to the feel of thin objects) or entirely acquired (violins are thin, and thin objects usually have a particular sound) (Marks et al., 1987; Marks, 1982). Nevertheless, it is accepted that these metaphors aren't completely arbitrary. For instance, the metaphorical use of *piercing* reflects the mapping between auditory and tactile perception, which is substantiated by neurophysiological research (for a review, see Eitan & Rothschild, 2011). Furthermore, in the spirit of 'neurons that fire together, wire together', the Neural Theory of Metaphor argues that because metaphors trigger neural responses in the modality they reference (Lakoff, 2008), this type of metaphor can also be translated back into its original modality. This way, sunlight poetically described to be *roaring* is interpreted as having brightness equal to the loudness of a *roaring* sound (Marks et al., 1987).

Of course, links between the modalities aren't the only reason timbral metaphors like *thin* might be expressive — they could also have acquired conventional meanings over time. Groups with more need to verbalise timbre develop common interpretations of these metaphors, including saxophonists (Nykänen et al., 2009), violinists (Stepánek, 2006; Fritz et al., 2008), guitarists (Traube, 2004) and pianists (Cheminée, 2006; Bellemare & Traube, 2006).

2.1.2.1 Consistency in timbral metaphor

If timbral metaphors are expressive, either through cross-modality or convention, we would expect them to refer to specific acoustic patterns. As a result, we could say that timbre is not ineffable if different people produce similar judgements for how well a metaphor describes a particular sound.

To test the consistency of metaphorical timbre judgements, Faure et al. (1996) played 32 participants synthesized instrument sounds, and collected numeric judgements for the appropriateness of French metaphors, including *soft* (*doux*) and *wide* (*large*). They found that 75% of the correlations between participants' judgements were significant ($p < .01$). Similarly, Darke (2005) played 22 participants recordings of orchestral instruments and collected judgements from 0 to 5 for how well they were described by metaphors like *bright* and *thin*. The judgements were highly consistent between the participants for ten of the twelve scales tested ($\rho_T > .85$), although a one-way ANOVA revealed that some instruments evoked significantly inconsistent judgements. Disley et al. (2006) and Disley & Howard (2004) provide few details, but also reach the conclusion that metaphorical timbre judgements are mostly consistent between participants. Where participants are not limited in their choice of metaphors ('free' verbalisation), results are less conclusive: Brookes & Williams (2010) and Zacharakis & Reiss (2011) found that some sounds evoked particular metaphors, where others prompted a wide variety of them.

If timbral metaphors are expressive, we would also expect the same sound to evoke the same metaphorical judgement in repeated trials. This is demonstrated in Wallmark (2019b), where 46 participants showed "modest" consistency over the same sounds. Again, free verbalisations weren't as clear, with participants in Zacharakis & Reiss (2011) usually opting for a timbral synonym in place of their original wording.

2.2 Verbalising timbre

Overall, the studies reported in Section 2.1.2.1 indicate that timbre is not ineffable, as judgements for metaphorical properties like timbral *softness* seem to be consistent between and within participants. As a result, we should be able to verbalise timbre automatically, by predicting the metaphorical judgements a listener would make.

2.2.1 Statistical approaches

Much of the work on verbalising timbre has accompanied the effort to narrow down its perceptual dimensions — the axes over which a timbre can vary independently. Comparatively little has been done on the acoustic correlates of timbre's semantic dimensions, knowing which would allow us to verbalise it from the audio signal. The number and complexity of these correlates means this is still very much an open field of study (Zacharakis & Reiss, 2011).

2.2.1.1 Semantic correlates of perceptual dimensions

Typically, the perceptual dimensions of timbre have been derived using Multidimensional Scaling (MDS). This entails participants rating the similarity of pairs of timbres to produce a timbre 'space', in which more similar timbres are closer together (for example, Kendall & Carterette, 1993a). Timbre spaces balance the accuracy of these distances with the resulting number of dimensions: while a twenty-dimensional solution might be the most accurate given the data, condensing timbre to two might capture most of the detail with only slight distortion. Figure 2.1 shows one example of a low-dimensional timbre space.

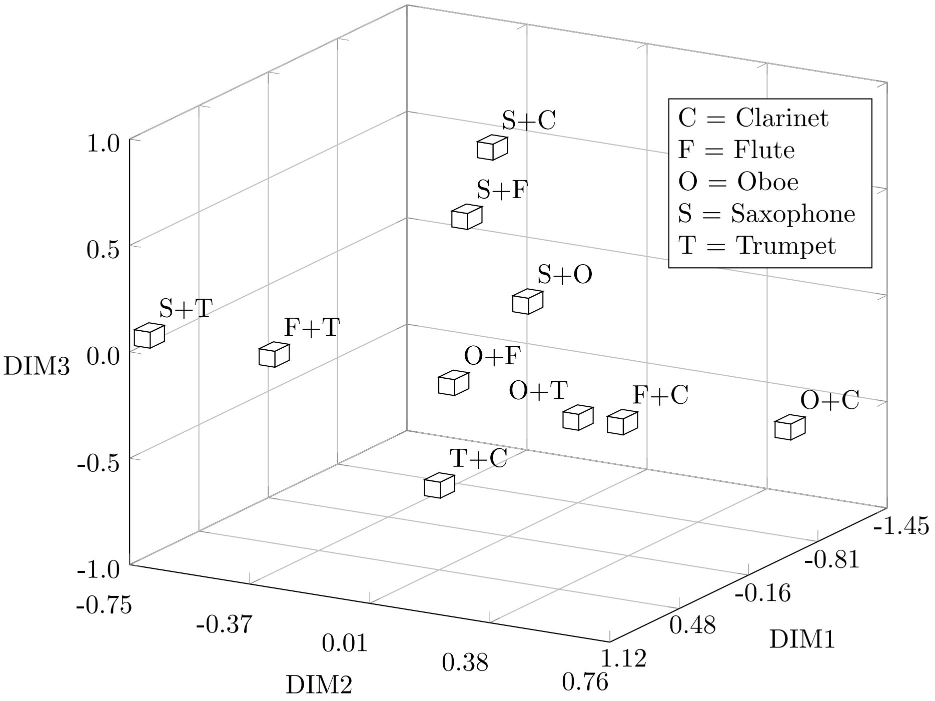

Figure 2.1: A three-dimensional MDS solution for timbres created by blending wind instruments, reproduced from Kendall & Carterette (1993a). Distances, for example F+T to O+C, reflect how similar participants thought those two timbres were.

Because MDS doesn't require similarity judgements to refer to anything specific, it doesn't presuppose any of the resulting dimensions (McAdams et al., 1995). However, not defining these dimensions also makes the space highly abstract — to interpret the position of a timbre verbally, we first need to know what the dimensions mean in terms of timbral metaphors. To this end, Kendall & Carterette (1993a) aimed to understand the dimensions of the timbre space in Figure 2.1, by correlating them with the dimensions of a 'semantic space'. They played participants orchestral timbres, collecting judgements for how well each could be described by eight timbral metaphors, such as *hard* and *sharp*. These judgements were then also interpreted with MDS, by replacing the perceived similarity of a pair of timbres with their verbal similarity. The resulting semantic space, like Figure 2.1, represented each timbre as a point in abstract dimensions, but with distances between them calculated from metaphorical judgements. Mapping this semantic space to the perceptual space was only partially successful, as correlations between the perceptual and semantic dimensions weren't particularly strong. Kendall & Carterette (1993b) replaced the metaphors in Kendall & Carterette (1993a) — originally used to describe synthetic timbres — with those used in orchestration manuals, to match the orchestral stimuli participants heard. This change produced a semantic space with stronger perceptual correlations, suggesting similarity judgements could be converted into timbral metaphors. Samoylenko et al. (1996) performed a similar experiment, using free verbalisations instead of predetermined scales. While their results were harder to interpret, they also reported metaphorical judgements to be representative of timbre perception.

However, while ubiquitous in the field, MDS can only be used to draw theoretical conclusions. Because it works from similarity judgements, it can't be used to verbalise timbre automatically (Cosi et al., 1994). Verbalising a new timbre would involve manually judging its similarity to each timbre already in the space, giving it a position in the space, then reading off its semantic coordinates. What is more, these coordinates aren't guaranteed to be meaningful. It isn't safe to assume, for example, that perceptually similar timbres will evoke similar metaphorical judgements. By imposing semantic spaces onto perceptual spaces, the result is "neither musical nor verbal" (Kendall & Carterette, 1991, p. 391), and, as we will see in Section 2.2.1.3, the true mapping between them may be much more complex.

2.2.1.2 Acoustic correlates of metaphorical scales

Another approach to verbalising timbre is to identify acoustic properties, or 'descriptors', of the sounds a particular metaphor correlates with. How well a timbre is described by that metaphor can then be predicted from the descriptor measurement and a linear regression model. This process is undoubtedly practical, and can be run directly on raw audio data. However, identifying the descriptors behind common metaphors like *warm* is easier said than done.

It's worth noting that research in this area tends to look for acoustic correlates of several metaphorical scales at once, as many appear to be synonymous. In their seminal paper, von Bismarck (1974) used factor analysis to reduce thirty scales to just four orthogonal groups of synonyms that encoded 90% of the same information. Similarly, Disley et al. (2006) found four factors encoding 91.7% of fifteen scales, and Zacharakis et al. (2014) a three-factor solution for 82% of 30 scales: a *brilliance* group, a *harshness* group and a *thickness* group.

One fairly well established correlate exists for the group often referred to as *brightness*, which is thought to include *clarity*, *purity* and *sharpness* (Disley et al., 2006; von Bismarck, 1974; Stepánek, 2006). Lichte (1941) found that when asked to judge which of two synthesized sounds was *brighter*, participants easily and consistently chose the sound with the higher mean frequency, a descriptor that has since been dubbed 'spectral centroid'. Illustrated in Figure 2.2, spectral centroid repeatedly appears as the strongest correlate of this group (Disley & Howard, 2004; Schubert et al., 2004; Almeida et al., 2017), predicting as much as 93% of the variance in similarity judgements referring to *brightness* (Marozeau & de Cheveigné 2007; see also Wessel 1979).

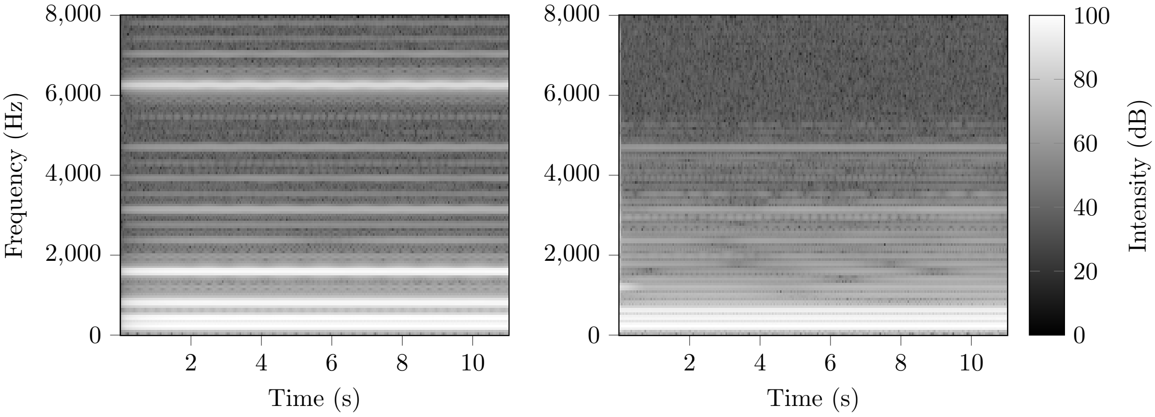

Figure 2.2: On the left: the spectrogram of a sound with a high spectral centroid; on the right: a sound with a low spectral centroid.

Where *brightness* might then be considered 'solved', finding descriptors for the other scales is an "open problem" (Cosi et al., 1994, p. 71). There has been some success in predicting the *roughness* group, which may include *harshness* (Stepánek, 2006; Zacharakis et al., 2014). Nykänen et al. (2009), for example, correlated *roughness* judgements for saxophone timbres with two descriptors, which accounted for 81% of the variance. However, it is debatable how distinct *roughness* is from *brightness* — the strongest descriptor was again similar to spectral centroid, and Wallmark (2019b) found sufficient correlation to justify one underlying group ($r(91) = .84,\ p < .0001$).

Scales not referring to spectral centroid are harder to predict, as their acoustic correlates aren't so clear cut: von Bismarck (1974), for example, did not find any evidence to link *hollowness* or *thinness* with a more striated frequency distribution, as proposed in Lichte (1941). Similarly, Alluri & Toiviainen (2010) reported moderate correlations between how much neighboring frequencies vary in strength, or 'spectral flux', and scales such as *strength*, *softness* and *fullness*. However, their results directly contradict Caclin et al. (2005), which did not find spectral flux to be a salient descriptor. Alluri & Toiviainen (2010) even call the established descriptor for *brightness* into question, identifying only a weak relationship to spectral centroid ($r(98) = .27,\ p < .1$). Likewise, Zacharakis & Reiss (2011) found 55% of participants would describe a timbral change in terms of *brightness* where there was no spectral centroid variation at all, suggesting the problem of its correlates might not be 'solved' after all.

Some metaphorical scales seem to resist correlation altogether — *warmth*, for example, has not yet found any reliable descriptors, and attempts to link it to the intensity of a timbre's low frequency components have been unsuccessful (Zacharakis & Reiss, 2011). Two of the three groups in Alluri &

Toiviainen (2010) also showed no strong correlations to seven possible descriptors, producing models with only 22% and 36% accuracy. As a result, while Section 2.1.2.1 found most scales to be used consistently, in this section we have failed to find clear acoustic correlates for some scales.

2.2.1.3 Acoustic correlate complexity

As we've seen, judgements on some metaphorical scales have been predicted from the strength of just one descriptor, like spectral centroid for *brightness*. Others may require two, like the pair of descriptors modelling *roughness* in Nykänen et al. (2009), or the seven in Alluri & Toiviainen (2010). However, it could be that many metaphors are actually the product of innumerous, more subtle descriptors. If so, these descriptors would appear insignificant in statistical analysis when taken individually, and might even be too specific to consider trying — one component of *softness*, for example, could be the ratio of spectral flux between three particular frequency bands. As a result, predicting most scales with simple descriptors has been described as "doomed to failure" (Disley et al., 2006, p. 62). This could explain why one model in Jiang et al. (2020) managed a reasonable 66.5% average accuracy, predicting judgements over a number of scales: it used correlations to not seven but fifty-four descriptors.

We can also question the assumption that these relationships are linear. Analysis of MDS results in Esling et al. (2018) revealed that common descriptors like spectral centroid aren't perceived the same way throughout timbre spaces. If descriptors for other metaphorical scales behave the same way, it doesn't make sense to predict them with linear regression. This is essentially the same problem we've seen above: we can't make accurate predictions from a descriptor without knowing the nature of its nonlinearity, just as we can't know which descriptors to try in the first place (Lemaitre & Susini, 2019).

Aucouturier (2006) offers a convincing hypothesis for these complex acoustic correlates: timbral metaphors also express higher-level cognitive reasoning, such as contextual biases and expectations. This way, a steel drum's timbre might be described as very *bright* for little other than the tropical contexts it is associated with. Another such example is the branding of digital synthesisers as *cold* in comparison to their analogue predecessors — a decade before, it was the analogue synthesisers that were *cold*, with *warmth* being reserved for acoustic instruments (Goodwin, 1988). People with these sorts of cognitive biases might therefore arrive at the same judgements, for example of a timbre's *warmth*, just as we've seen participants generally agree in Section 2.1.2.1. However, using *warmth* to mean 'acoustic instrument-like' would produce a highly complex set of descriptors, poorly captured by simpler alternatives like spectral centroid or flux. In Disley et al. (2006), the participants themselves reported one such bias, saying they often used metaphors like *metallic* and *wooden* according to the material of the instruments they were hearing, regardless of how *metallic* or *wooden* they actually sounded.

A stronger version of this idea is the 'semantic coding hypothesis': we might convey timbre entirely through these acquired biases, for example, describing the low register on a clarinet as *warm* only because other people do. The cross-modal links discussed in Section 2.1.2 — our other explanation for consistent timbral metaphors — then follow the semantic association, causing the register to subconsciously feel *warm* (Martino & Marks, 1999). That said, it's more likely these acoustically complex biases are only part of how we express timbre, as we've seen scales like *brightness* are largely used to refer to clear descriptors. Wallmark (2019b) also rejects the strong hypothesis: participants in their experiment often made judgements deviating from stereotypical descriptions of orchestral instrument timbres, such as the *nasality* of an oboe.

2.2.2 Neural network approaches

Where simple acoustic correlates fail to accurately predict metaphorical timbre judgements, neural networks might succeed. They have long been speculated to be better at connecting metaphorical scales with their descriptors (see Miranda, 1995), as they can produce complex, nonlinear mappings of their input data. However, only a handful of neural models have actually been tested.

2.2.2.1 An introduction to neural networks

Artificial neural networks are one form of machine learning: they develop a mapping linking inputs to outputs, and the more input-output pairs they are given, the more accurate that mapping becomes. Each 'neuron' in a neural network is responsible for replicating part of this mapping, by applying a nonlinear function of its inputs. When arranged in a sequence of 'layers', the outputs of a set of neurons can become the inputs to another set, as in Figure 2.3. Successively applying nonlinear functions this way means that networks with mediating (or 'hidden') layers can transform an input into practically any output. These 'deep' neural networks (DNNs) can therefore accurately model extremely complex mappings.

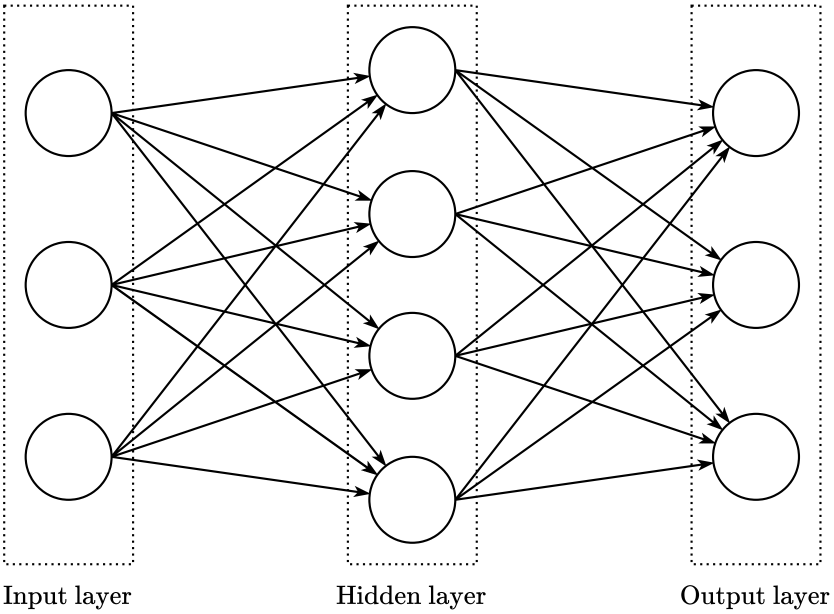

Figure 2.3: A network of 'fully connected' layers, where the neurons in each layer feed into every neuron of the next layer.

To understand how DNNs actually 'learn' anything, we first have to consider the workings of individual neurons. Figure 2.4 shows how their connections are weighted: all a neural network learns is what these weights should be. When 'trained' on input-output pairs, a network passes the inputs through its neurons, and compares the results of the final layer against the 'correct' outputs of the pairs. This comparison (the 'loss') is then used to adjust the weight of each connection using the backpropagation algorithm (Rumelhart et al., 1986), such that weights which didn't transform the input in the right way are modified to reduce the loss. Training on more pairs sets the weights closer and closer to any underlying patterns in the data, progressively increasing the network's accuracy.

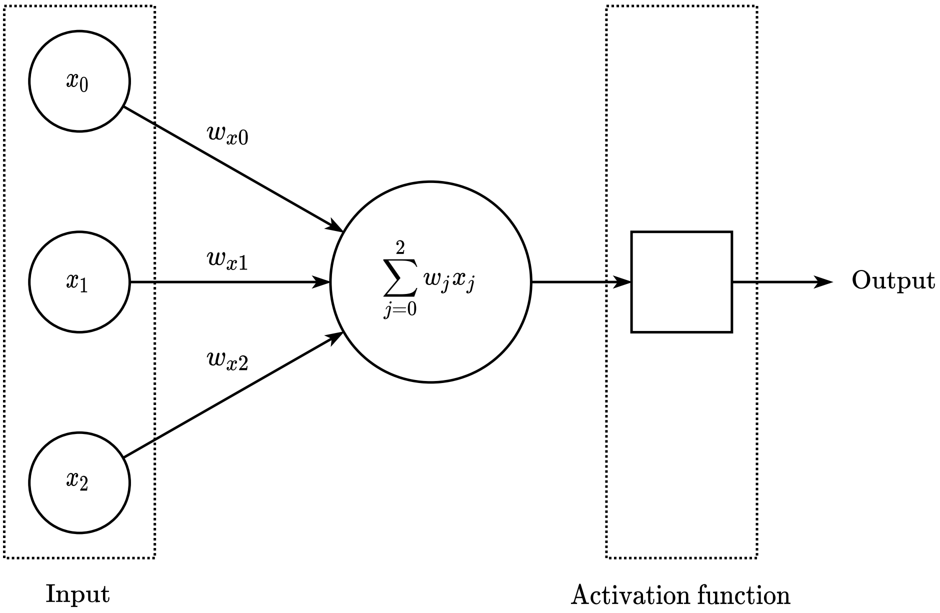

Figure 2.4: One artificial neuron from the hidden layer of Figure 2.3. Its inputs are the three input neurons x, which have weights w_x. The weighted sum of these inputs is sent through a nonlinear 'activation' function, producing an output to the next layer.

While they can be accurate, neural networks also introduce a series of problems. For one, their accuracy comes with a price — data. To successfully learn nontrivial patterns, DNNs usually need to be trained on large amounts of it, and labelling each 'correct' output becomes very time consuming. Training can also be slow, as increasing the number of neurons in a network exponentially increases the number of connections between them, each of which has a weight to adjust. This brings us to another problem: natural stimuli are enormously detailed. For example, a four-second audio file might consist of 64000 samples, which capture the audio signal value at regular time intervals. Training a network with an input neuron for each of these samples would entail slowly adjusting millions of weights during training, not to mention the amount of data needed to make those adjustments accurate. As a result, using a DNN typically involves extracting 'features' that simplify the input, which risks discarding relevant information. It can also be hard to see which patterns have been learned, as traditional neural networks are 'black boxes': where the relationship between descriptor and scale is clear in linear regression, it isn't easy to interpret the thousands of weights defining how a network makes predictions.

Despite these problems, the potential for better accuracy makes DNNs an attractive way to predict metaphorical timbre judgements, particularly if we aren't interested in how those predictions are made. Like similarity judgements in MDS, neural methods make no assumptions beyond the feature extraction stage. Instead of choosing descriptors, checking how well they correlate with a metaphorical scale, and making predictions from those correlations, neural networks can learn the most relevant descriptors themselves. These descriptors can therefore be much more sophisticated than we might think to test for — the kind needed to capture any complex cognitive biases encoded in our metaphorical scales.

2.2.2.2 Instrument recognition

Because of their accuracy, DNNs have already seen extensive use in automatic instrument recognition (Pons et al., 2017; Han et al., 2016). Although not in the way we've considered, these networks are also verbalising timbre, and the scales they use are still intuitive — in place of *brightness* or *warmth* on the controls of a synthesiser, we could equally use *violin-like*. However, description in terms of another sound (or 'mimesis') isn't very flexible. For example, where would a trumpet's timbre fall on the scale *violin-like*? It is perhaps for this reason that musicians avoid it: when asked to group timbres by similarity, listeners with extensive musical training used acoustic differences to justify their groups twice as often as mimesis, whereas non-musicians used mimesis almost three times more (Lemaitre et al., 2010). Similarly, in a corpus of eleven orchestration treatises, only 10.9% of the references to timbre were mimetic (Wallmark, 2019a). As a result, we can't rely on instrument recognition models when verbalising timbre — they aren't appropriate beyond the sounds of specific instruments.

2.2.2.3 Existing neural network approaches

Feiten & Ungvary (1991) were the first to predict metaphorical timbre judgements using neural networks, with the goal of searching through large sound databases verbally. They labelled 120 sounds with four binary oppositions, such as *sparse* vs. *rich*. Each of these sounds was then reduced to 32 features, capturing the intensity of a series of frequency bands. The four oppositions were predicted by separate networks, all trained on 100 of the labelled sounds. During testing on the remaining 20 sounds, the networks achieved a reasonable average accuracy of 71%. However, the predictions they could make were very coarse: to be more useful, they'd need to predict exactly how *rich* a sound is. The network accuracies were also likely to have been limited by the amount of training data.

Training DNNs on large datasets has become much more feasible since this early study, thanks to improvements in computing hardware. In the last 10 years, DNN research has also found various ways to improve their performance and data efficiency, such as 'convolutions', 'dropout' and 'batch normalisation'. However, neural network approaches to predicting metaphorical timbre judgements have yet to make use of them. One approach formed part of an intuitive sound synthesis technique (Gounaropoulos & Johnson, 2006). The network design was straightforward: fully connected, with two hidden layers of 100 neurons each. For its outputs, sounds were labelled with values from one to ten on nine metaphorical scales, including *warmth*, *brightness* and *harshness*. As inputs, twenty features were extracted from each sound, with fifteen to describe the overall frequency distribution, one to measure a detuning effect, and another four to capture frequency variation over time. The system was tested on 30 synthesised sounds. Although they discuss some preliminary results, Gounaropoulos & Johnson (2006) don't report an accuracy, nor a dataset size. It is therefore hard to evaluate their network, although we can speculate that the large number of neurons made training difficult.

A more recent neural network approach also reports few details (Jiang et al., 2020). 34 participants scored the timbres of 72 instruments on five nine-point metaphorical scales, such as *dark-bright* and *coarse-pure*. These scores were then averaged, and used as labels for the network to predict. For its inputs, 166 features were extracted from each sound, capturing a variety of spectral and temporal characteristics. The network — only described as having a hidden layer — managed an average of 91.4% accuracy, although it isn't clear what it was actually tested on.

Overall, these networks have produced promising results. However, there aren't any networks to reflect advances in the field, and nothing that could be called comprehensive. As a result, it remains to be seen just how well DNNs can predict metaphorical timbre judgements.

2.2.3 Practical applications

One application for predicting these judgements is in music production software. Synthesisers have improved substantially since their genesis in the 1960s, but they still force new users to learn a barrage of confusing controls before they can make the timbres they want (Ethington & Punch, 1994). The reason is simple — we don't fully understand what intuitive scales like *warmth* mean, so the synthesiser controls must instead modify acoustic concepts like envelope attack, filter cutoff and resonance. To solve this problem, programs for generating sound with intuitive, metaphorical controls can use metaphorical judgement predictions in repeatedly evaluating and tweaking the synthesised output, to approximate the desired timbre (Gounaropoulos & Johnson, 2006; Brookes & Williams, 2007). A similar idea could be applied to alert sounds for mobile devices. Alarms might be more effective with a *bright* timbre, where gentle reminders could be *soft*. An objective measure of these qualities could therefore help designers choose more effective alert sounds, although more research would be needed to determine which metaphorical qualities evoke particular responses.

Another musical application is the organising of audio clips, or 'samples'. Electronic musicians may have tens of thousands of samples, which become tedious to search through (Lemaitre et al., 2010). The neural network boom has produced software to arrange them by genre and perceptual similarity, such as

Algonaut's 'Atlas' or 'Sononym'. However, as of yet there is no commercial software that allows the user to search for samples verbally, using timbral metaphors.

Lastly, the ability to predict metaphorical timbre judgements could be used to create a standardised way to code timbre. Just as colour can be coded with values for red, green and blue, timbre could be coded with numeric judgements for scales like *brightness*, *softness* and *warmth*. A common frustration in research on verbalising and perceiving timbre is that existing experimental results are hard to compare, as most studies use different sound stimuli (see Esling et al., 2018). If the stimuli were all coded in a standardised metaphorical way, these results would become much easier to analyse in future research.

2.2.4 Theoretical implications

While the studies discussed reported in Section 2.1.2.1 do indicate that timbre is not ineffable, they aren't conclusive. In addition to its practical value, a way to accurately predict metaphorical timbre judgements would therefore support the idea that timbre is something we can verbalise, by showing that timbral metaphors express specific acoustic patterns.

2.3 Summary

In this chapter, we first saw that while timbre isn't lexically coded, we can verbalise it through cross-modal metaphors. Although these metaphors feel vague, experimental research indicates they are used quite consistently, suggesting judgements on metaphorical scales like *warmth* could be predictable. However, using similarity judgements and MDS to position timbres on semantic axes can't be automated, and measuring descriptors in the audio signal and mapping them to metaphorical judgements with linear regression has produced mixed results. We discussed how the acoustic correlates of timbral metaphors might be far more complex than simple descriptors, and why neural networks, which can capture these complex mappings, might be more successful. Nonetheless, existing neural approaches are neither thorough, nor up to date with modern deep learning techniques.

We can therefore hypothesise that DNNs are a good way to predict metaphorical timbre judgements. In addition to a number of practical applications, the ability to predict these judgements accurately would support the case against timbral ineffability.

Chapter 3

Method

This chapter describes the development of two neural networks to verbalise timbre. As we've seen, there is no comprehensive work on how well neural networks can predict judgements on metaphorical scales like *brightness*. This approach therefore tests the hypothesis that DNNs are a good way of predicting these judgements, acting as a proof-of-concept for the practical applications in Section 2.2.3, and adding to the body of evidence that timbral metaphors are mostly expressive and predictable.

Timbral metaphors don't necessarily refer to the same descriptors in different types of sound — what makes a voice *soft* might be different to what makes a piano *soft*. As a result, the classification problem was limited to musical sounds. Section 3.1 describes the creation of a DNN-friendly dataset of metaphorical timbre judgements for these sounds. Section 3.2 then explains a number of modern deep learning techniques, and how they were applied to the two network designs.

3.1 Data collection

The data collection process can be broken down into four main steps: selecting a set of sounds, labelling them with metaphorical judgements, simplifying them into features, and some final data processing. We will cover each of these steps in turn, eventually reaching the final inputs and outputs used to train and test the networks.

3.1.1 Sound data

For the initial sound data, I used the NSynth dataset (Engel et al., 2017), which contains a mix of synthesized and acoustic monophonic musical sounds. It is also large, having been designed for machine learning. I excluded the most percussive sounds (labelled as 'percussive' in the dataset), as they often appear distinct from other sounds in MDS perceptual spaces (for example, Iverson & Krumhansl, 1993). They therefore might have had different descriptors to the rest of the sounds for the same metaphors.

As well as describing timbre, judgements on metaphorical scales like *brightness* have been shown to be related to pitch (Iverson & Krumhansl, 1993; Eitan & Rothschild, 2011; Marozeau & de Cheveigné, 2007), suggesting that a piccolo, for example, might only be described as *bright* for the high pitches it plays. To avoid having to collect judgements on sounds at multiple pitches for the networks to learn these pitch effects, I limited the problem further to predicting timbre judgements for sounds at a single pitch. I selected sounds with the most frequent pitch in the dataset (G3, or roughly 196Hz), and discarded the rest, yielding a total of 905 sounds, which I reduced to 900 for easier division.

Loudness has also been shown to affect metaphorical timbre judgements (Eitan et al., 2010; Eitan & Rothschild, 2011). To again limit the classification problem, I set each sound to a similar loudness using root mean square normalisation, taking care to avoid 'clipping' artefacts. The NSynth dataset also provides variants of each sound for different dynamics (staccato through legato). However, many sound almost identical, so only one instance of each sound was kept in the final dataset, but with a random dynamic to cover the widest possible range of musical timbres.

Two seconds or less is often considered long enough to get an impression of a sound's timbre in perception studies (Lichte, 1941; Zacharakis & Reiss, 2011; Disley et al., 2006). Sounds in the NSynth dataset are four seconds long, but many decay back to silence within two seconds. The 900 sounds were therefore cropped to two seconds each, reducing the number of features they would produce and thus requiring fewer weights in the networks.

3.1.2 Output labels

The sound data was then labelled with timbre judgements on metaphorical scales, as outputs for the networks to predict.

3.1.2.1 Collecting judgements

While I would have ideally collected metaphorical judgements on all 900 sounds from multiple participants, it would have been impossible to recruit them for such a long task. However, having multiple participants judge different subsets of the data wouldn't have worked well either. Differences in their judgement criteria wouldn't have been averaged out, which might have created an inconsistent set of labels. As a result, I labelled the timbres myself.

A valid concern is that my judgements might not reflect how most people use timbral metaphors. I have some musical training, and we've already seen one way this can affect the description of timbre. Musicians also tend to be more consistent in their timbre judgements (von Bismarck, 1974; McAdams et al., 1995). Conclusions based on mine will therefore need to be tested for generality in future work.

To check that I was making reliable, repeatable judgements, I made a second set of judgements for 10% of the sounds selected at random. I expected fairly high consistency between this sample and the fully labelled dataset — at least as consistent as the between and within-participants results covered in Section 2.1.2.1. However, we can't put too much weight on my consistency within sounds, as there was likely to have been an element of repetition priming: I may have remembered my previous responses (Darke, 2005).

3.1.2.2 Choosing metaphorical scales

Early work on timbral metaphors collected judgements on 'semantic differential' scales (for example, von Bismarck, 1974; Pratt & Doak, 1976), which consist of a metaphor and its antonym, such as *bright-dark*. However, these metaphors aren't necessarily two ends of the same scale — *bright* could refer to a completely different set of descriptors to *dark*. As a result, Kendall & Carterette (1993a) introduced Verbal Attribute Magnitude Estimation (VAME) scales, which range between an metaphor's negation and its assertion, for example *not bright* to *bright*. I chose to make my judgements along these VAME scales, to avoid the risk of capturing two sets of descriptors with one scale. Originally, VAME used graphical sliders, but I settled on integers between 0 (*not bright*) and 9 (*bright*) for simplicity.

To collect judgements covering as much of the timbral variation in the sound data as possible, the scales themselves were chosen based on the factor analyses mentioned in Section 2.2.1.2. As we saw, these generally find three or four mostly orthogonal scale groups, so I took scales from the first three. However, these factor analysis groupings are often relaxed (for example, Zacharakis et al., 2014). I therefore expected some degree of correlation between the three scales.

The first of these groups is almost always made up of luminance metaphors (Pratt & Doak, 1976; Stepanek & Moravec, 2005; Disley et al., 2006; Zacharakis et al., 2014). Of these, *brightness* is generally considered the most common timbral metaphor (Alluri & Toiviainen, 2010), which made it an obvious choice for the first scale. The second group usually refers to texture, and often includes *softness* (Nykänen et al., 2009; Alluri & Toiviainen, 2010; Elliott et al., 2013), which I used as my second scale. The third group is more ambiguous between studies, but is commonly described as *fullness* or *richness* (Lichte, 1941; von Bismarck, 1974; Pratt & Doak, 1976; Kendall & Carterette, 1993b; Alluri & Toiviainen, 2010; Zacharakis et al., 2014). I chose *fullness*, as the metaphor I felt more confident using.

3.1.2.3 Making metaphorical judgements

To keep the scales independent, I judged the sounds on each of them separately. I wrote a program to play them in a random order and accept a VAME score between zero and nine. This way, I made three passes over the 900 sounds, listening through headphones, first for *brightness*, then *softness* and finally *fullness*. These passes were spread over three full days to avoid fatigue. The second set of judgements for the 10% sample were made a day after the final pass.

When making *brightness* judgements, I generally considered guitar, brass and bell-like timbres to be *bright*, and the more sinister-sounding vocals and gongs *not bright.* For *softness*, I found myself largely describing the loudness profile of the timbres: bowed strings and horns were *soft* where plucked strings and staccato piano timbres were *not soft.* I was also influence by overtones, with simpler, flute-like timbres falling on the *soft* side. My *fullness* judgements were based on a brassy quality — trombones were *full* but most woodwind instruments *not full.*

3.1.3 Input features

The sounds were then converted into features, as explained in Section 2.2.2.1.

3.1.3.1 Choosing features

The features had to balance how well they would represent the sounds they were taken from with the number of weights needed in the networks — more network inputs create more weights. They also needed to capture not just the frequencies of a sound, but their development over time, as MDS studies widely agree on a temporal dimension to timbre perception (Grey, 1977; Iverson & Krumhansl, 1993; Elliott et al., 2013). A popular approach in automatic speech recognition is to use vectors of Mel-frequency cepstral coefficients (MFCCs) that capture the frequency distributions of slices of audio (time 'bins'). MFCCs are also a common choice to measure timbral similarity (for example, Aucouturier & Pachet, 2003). By using the Mel scale, these coefficients model the fact that human frequency perception isn't linear: differences between low frequencies sound larger than differences between high frequencies. Human loudness perception also isn't linear, so the amplitudes of MFCCs are typically put on a logarithmic scale. However, MFCCs were designed to represent the formant trajectories of speech, where I instead needed to capture musical timbres. I therefore used spectral rather than cepstral log Mel coefficients. I considered 20 of these coefficients (covering frequencies from 0 to 8kHz) for every 100ms time bin in a sound to be enough to capture its timbre.

3.1.3.2 Extracting features

To extract log Mel spectral features from a sound, I used the library `python_speech_features`, which performs Fourier transforms on segments of audio and allocates the outputs to Mel scale frequency bins. I made those segments twice the length of a time bin and applied a 'Hamming' window, to avoid windowing artefacts. Finally, I took the log of each feature, and scaled them between 0 and 255 like the pixel values of an image. I wrote a program to extract and store features this way for all the sounds I'd made judgements for. The resulting feature vectors can be thought of as low-resolution images, as shown in Figure 3.1.

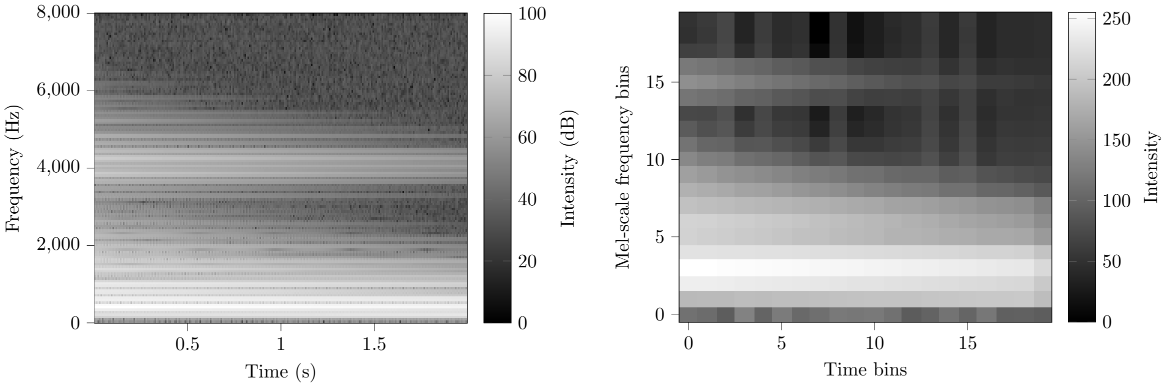

Figure 3.1: On the left: the spectrogram of a sound in the dataset; on the right: the log Mel spectral features of that sound. The Mel scale compresses high frequencies, so the band around 4kHz only appears in a few Mel bins around bin 15.

3.1.4 Data processing

3.1.4.1 Splitting the data

Standard practice in machine learning is to create three subsets of the input and output data. The first of these is the ‘training’ set, which is used to adjust the weights in a neural network during the training phase. The second is the ‘validation’ set, which is used during the design process to check how well a model performs. The last is the ‘test’ set, which is kept separate and only used to evaluate a model once. This ensures test set accuracy reflects how a model would perform on new data, as it can’t be tuned to perform better specifically on the test data.

There are no standard proportions for the training, validation and test sets. I considered 90 sounds enough to reflect the performance of the networks on new data, so I maximised the size of the training set by using an 80% training, 10% validation and 10% test split, selecting data for each set at random.

3.1.4.2 Augmenting the data

A common technique in deep learning is to augment the training data by creating variants of each sound, but giving them the same labels. In image recognition, this usually takes the form of rotating and mirroring images (for example, Krizhevsky et al., 2012). While a dog is still the same dog regardless of its orientation, to a neural network it looks like new data, which can be used to adjust its weights. For audio data, this can be any number of modifications, as long as they aren’t large enough to make the labels inaccurate (see Salamon & Bello, 2017). I chose noise addition, time shifting, time stretching and time squashing. For noise addition, I added uniform random noise as a percentage of the maximum possible amplitude over the whole sound. To shift a sound in time, I added silence to the start, then deleted the same amount of audio from the end to keep it two seconds long. I didn’t do the same in reverse, to avoid losing the initial attack of a sound, which is perceptually significant (Grey, 1977; McAdams et al., 1995; Elliott et al., 2013). To stretch and squash sounds, I used the Python library `pyrubberband`, again adding silence or deleting audio to keep the sounds two seconds long. Exaggerated examples of these techniques are shown in Figure 3.2.

I wrote a program to generate variants of each sound for three levels of the four techniques: variants with 0.33%, 0.66% and 1.00% noise, variants with 1600, 3200 and 4800 samples of opening silence, and variants with 0.33%, 0.66% and 1.00% stretching and squashing. This increased the size of the training set from 720 to 9360 training items.

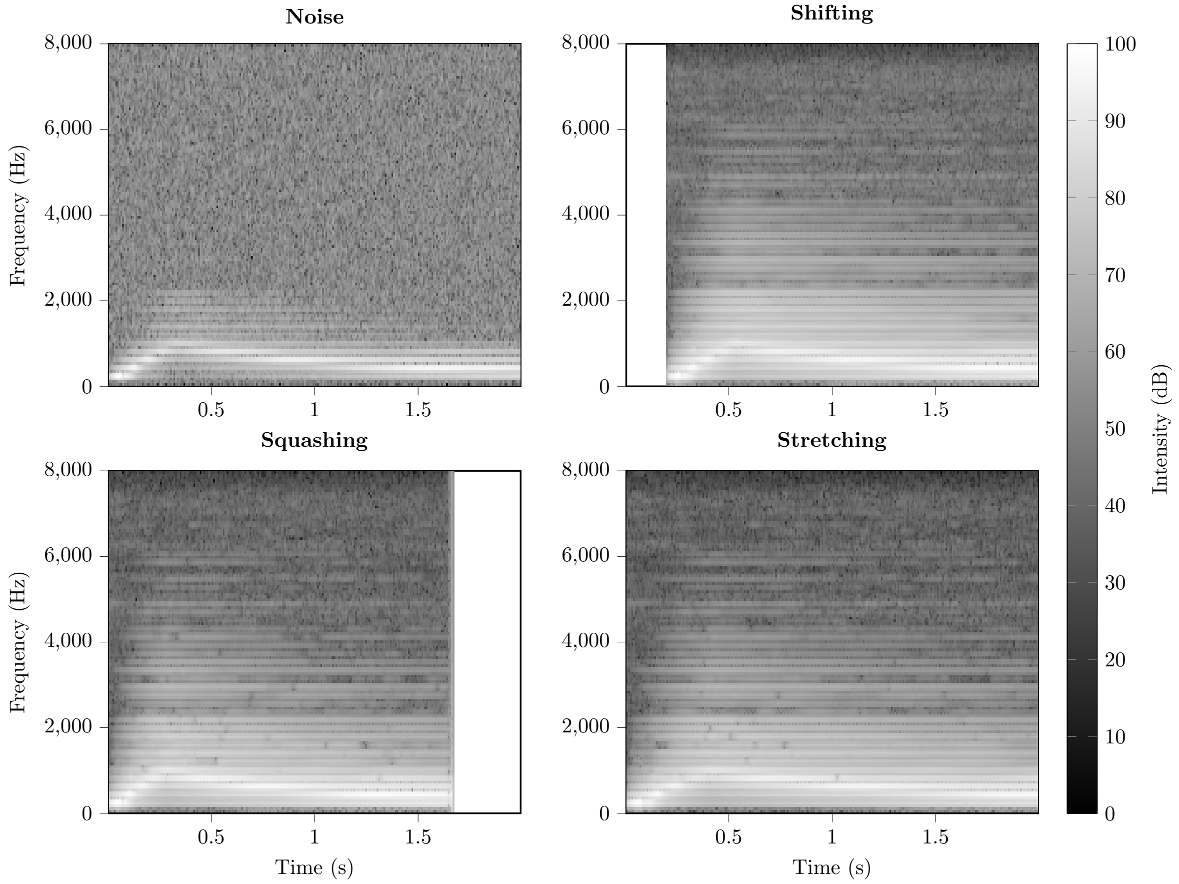

Figure 3.2: Exaggerated examples of the four augmentation techniques.

This concluded the data collection stage. Figure 3.3 shows the full pipeline, from the original audio files to the final training, validation and test sets.

3.2 Neural network modelling

We've seen that the feature vectors can be thought of as low-resolution images, which formed the network inputs, making the problem akin to an image classification task.

Neural networks tend to work better when they output categorical distributions (most famously, Krizhevsky et al., 2012), so the outputs were defined as probability distributions over the ten possible values for *brightness*, *softness* and *fullness*. This section describes the design and implementation of two networks to produce these distributions.

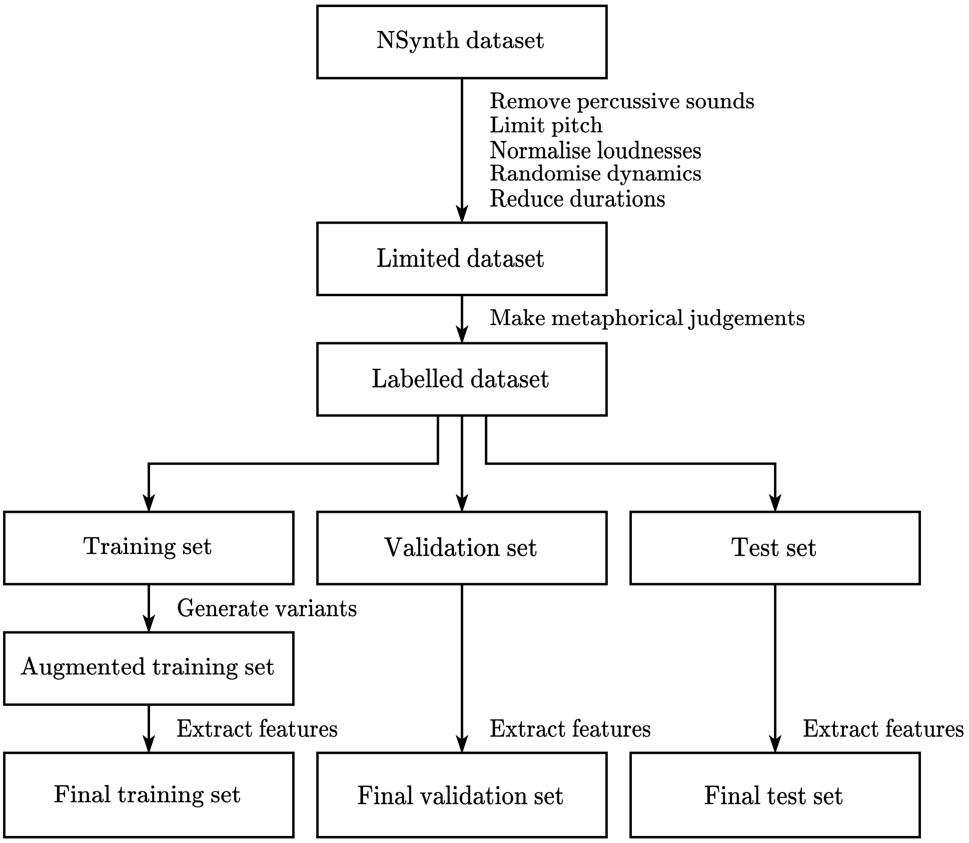

Figure 3.3: The data collection pipeline.

3.2.1 Modern neural networks

Even after augmentation, there wasn't much training data by deep learning standards. Fully connected layers like those shown in Figure 2.3 create lots of weights to adjust, so the networks needed to avoid using many of these layers — a common intuition is that the number of weights in a network should be of the same order of magnitude as the number of training items. They also needed to predict the same judgement for a timbre regardless of where it started in an audio file: we don't expect the *brightness* of a timbre to change based on how much silence is before or after it, so this position invariance is essential in making a timbral classifier robust to real-world data. This is another reason why fully connected layers aren't ideal: to learn position invariance, they need to be trained on data covering every possible position (Chollet et al., 2018). As a result, the networks were designed around a number of more modern deep learning concepts, explained below.

3.2.1.1 Convolutions

'Convolutional' layers are a popular way to both reduce the number of weights in a network, and to build in position invariance, meaning less training data is needed (LeCun et al., 2015). Illustrated in Figure 3.4, convolution is based around the idea of stepping a template set of values (or 'kernel') over the layer input, producing an output at each possible position reflecting how well the kernel matched the input values underneath it. The weights in a convolutional layer are just the values in that kernel, which the network can adjust during training like weights on the connections in a fully connected layer. This way, the kernel learns to match the most relevant patterns in its input, and the network doesn't need a weight for every input value. This also makes the network inherently position invariant: a kernel will match a pattern regardless of where it is in the input.

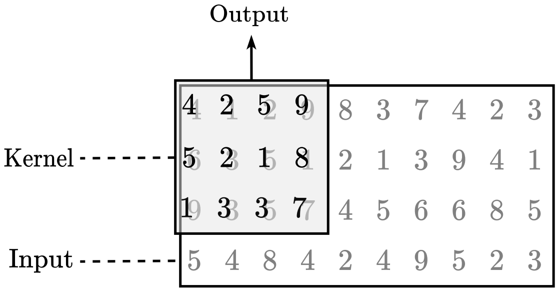

Figure 3.4: The concept of convolution: a kernel outputs the dot product of its values and the values underneath it, then steps to the next position.

Convolutional layers typically step several of these kernels over their input, each of which learns to match a different pattern and produces a different set of outputs — these sets are the output 'channels'. A convolutional layer can also take in several sets of inputs (input channels), which in image recognition are usually the red, green and blue components of an image. The kernels then become multidimensional to match patterns across the input channels, although they still only output one value. Convolutional layers can therefore be stacked on top of each other: output channels from one become input channels to another. Because a kernel usually can't be stepped past the edges of the layer input, a convolutional layer produces fewer signals than it takes in, so stacking them this way can progressively condense the information passing through a network. This means we can use several convolutional layers to learn the most relevant patterns in the network input, before a final fully connected layer that learns to connect those patterns to the right network outputs: the convolutional layers reduce the number of inputs to the fully connected layer, keeping the number of weights in the network low. This also creates a bottleneck in the network, which limits the detail the network can represent at that point. As a result, it can't make predictions based on anything too specific, which helps it to generalise to new data (see He et al., 2016).

3.2.1.2 Batch normalisation

'Batch normalisation' is an operation to keep the signals in a network in a reasonable numeric range, by setting the mean and standard deviation of a layer's outputs over a batch of training data to zero and one respectively. Avoiding large numeric fluctuations also helps a network to train faster (Ioffe & Szegedy, 2015).

3.2.1.3 Dropout

Another modern concept is 'dropout', which is as simple as ignoring some of the outputs of a layer at random (Srivastava et al., 2014). Adding dropout means a network is forced to capture a pattern in fewer weights, making those weights less specific to the nuances of the training data. Once again, this helps it to generalise.

3.2.1.4 Max pooling

A 'max pooling' operation is typically applied to the output of a convolutional layer (Chollet et al., 2018). This boils down to taking a set of neighbouring outputs, and only keeping the largest of them — it doesn't matter whether a pattern was found by a kernel in one position or in the next position, as long as that pattern was found. Only keeping the largest is therefore an easy way to reduce the number of inputs to the next layer, keeping the number of weights in the network low.

3.2.1.5 Softmax

Lastly, the output of a neuron is unbounded, so when using categorical outputs the final fully connected layer is usually followed by a 'softmax' function. This takes a log and scales the outputs to sum to one, meaning they can be interpreted as probabilities (Chollet et al., 2018).

3.2.2 Designing the networks

I built networks applying these concepts using the fairly accessible deep learning Python library `Keras`, taking the 'Simple MNIST convnet' image classification example as a starting point (Chollet et al., 2015). While there are a few design guidelines for DNNs, what works largely depends on the data, so I tuned the networks through experimentation.

3.2.2.1 A convolutional style network

The first network design applied the idea of multiple convolutional layers with a final fully connected layer. I experimented with various additional operations, arriving at the final structure in Figure 3.5. The number of output channels in each convolutional layer was set to ten, as a compromise between more (to capture more patterns in the input) and fewer (to keep the number of weights in the network down). While kernels in a convolutional layer are often stepped over two dimensions, keeping pitch constant in the sound data meant they could be the full height of their input and step along one dimension: the timbres were all at the same 'height' in the input images, so the kernels didn't need to move up and down to match them. All the kernels were set to be five time bins wide, creating 16 unique positions over the 20 input time bins in the first convolutional layer. This reduced the 400 input features to just 160 output signals.

As I developed the network, I found adding batch normalisation and dropout after each convolutional layer lowered the validation loss (here, the mean squared error between the validation set labels and the network's predictions). The final design also used max pooling after the second layer, which was set to take the largest of pairs of neighbouring kernel outputs, halving the number of output signals. I added one last convolutional layer before the fully connected layer, meaning it only needed 20 input neurons and creating a signal bottleneck. The 30 outputs of the fully connected layer were scaled to be probability distributions using three softmax functions — one for each scale. The other layers used the popular 'ReLU' activation function.

I tried a series of standard ways to reduce the final validation loss of the network. One was changing the 'learning rate', which dictates how much a network can adjust its weights by after seeing a batch of training data. Large weight adjustments can overshoot, resulting in a higher loss. Dropping the learning rate from 0.0005 to 0.0001 reduced the validation loss, and made the network only slightly slower to train. The final validation loss was also improved when adding dropout of 0.1, which meant it was ignoring 10% of the outputs from each convolutional layer. With these settings, I trained the network from scratch ten times, and kept the version with the lowest final validation loss.

3.2.2.2 A fully connected style network

I also built a network with a focus on fully connected layers, closer to the design described by Gounaropoulos & Johnson (2006). As shown in Figure 3.6, I was forced to use an initial convolutional layer (again, with ten channels and max pooling over the output pairs) to keep the number of weights down. Nonetheless, it still had over five times as many weights as the convolutional style network, and only limited position invariance. Like the convolutional style network, the medial layers used ReLU activation functions, and dropout of 0.1. However, the learning rate had to be reduced to 0.00001, so it took much longer to train than the first design.

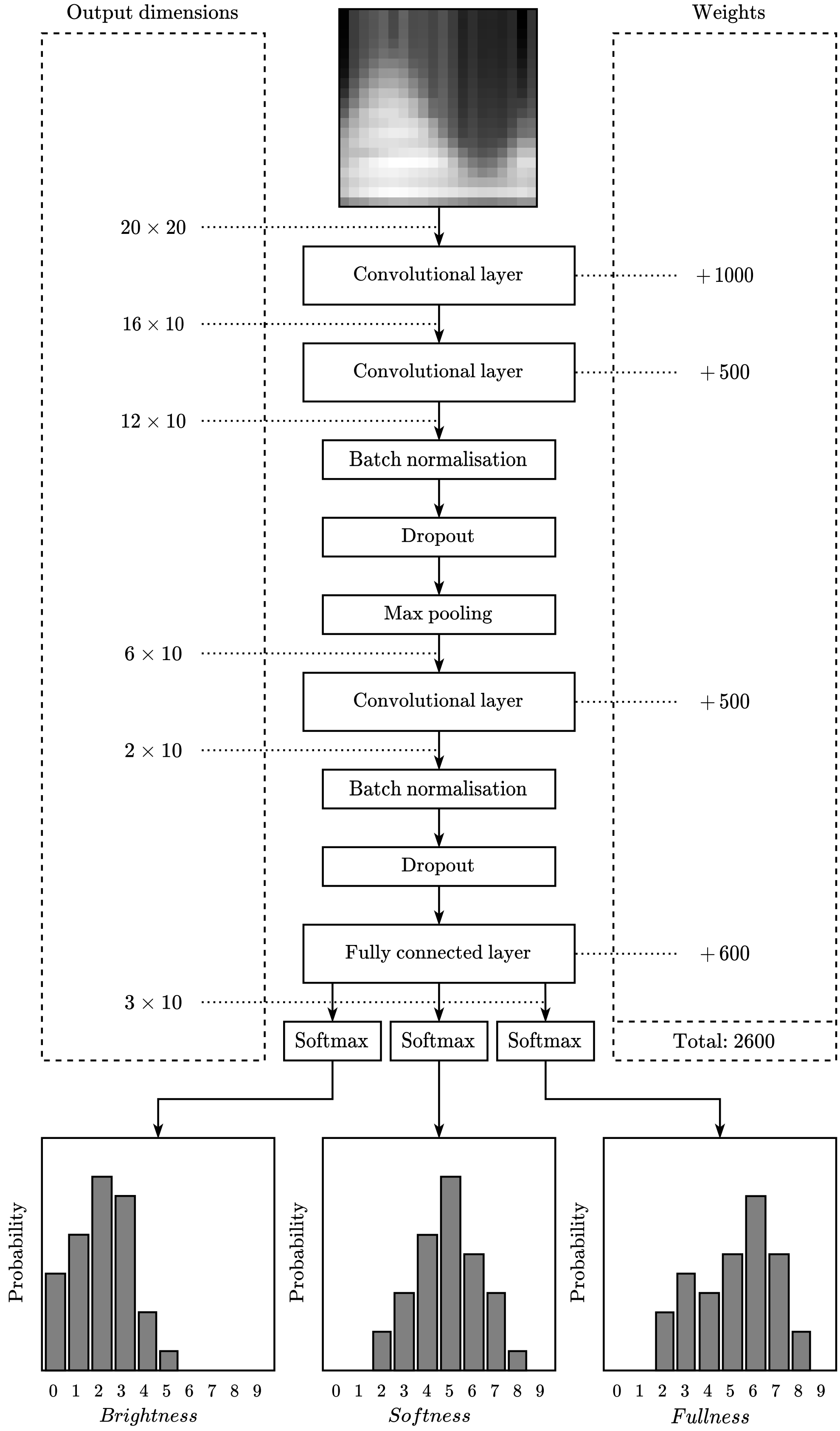

Figure 3.5: The convolutional style network. Note how the convolutional layers and max pooling operations reduce the number of inputs to the fully connected layer.

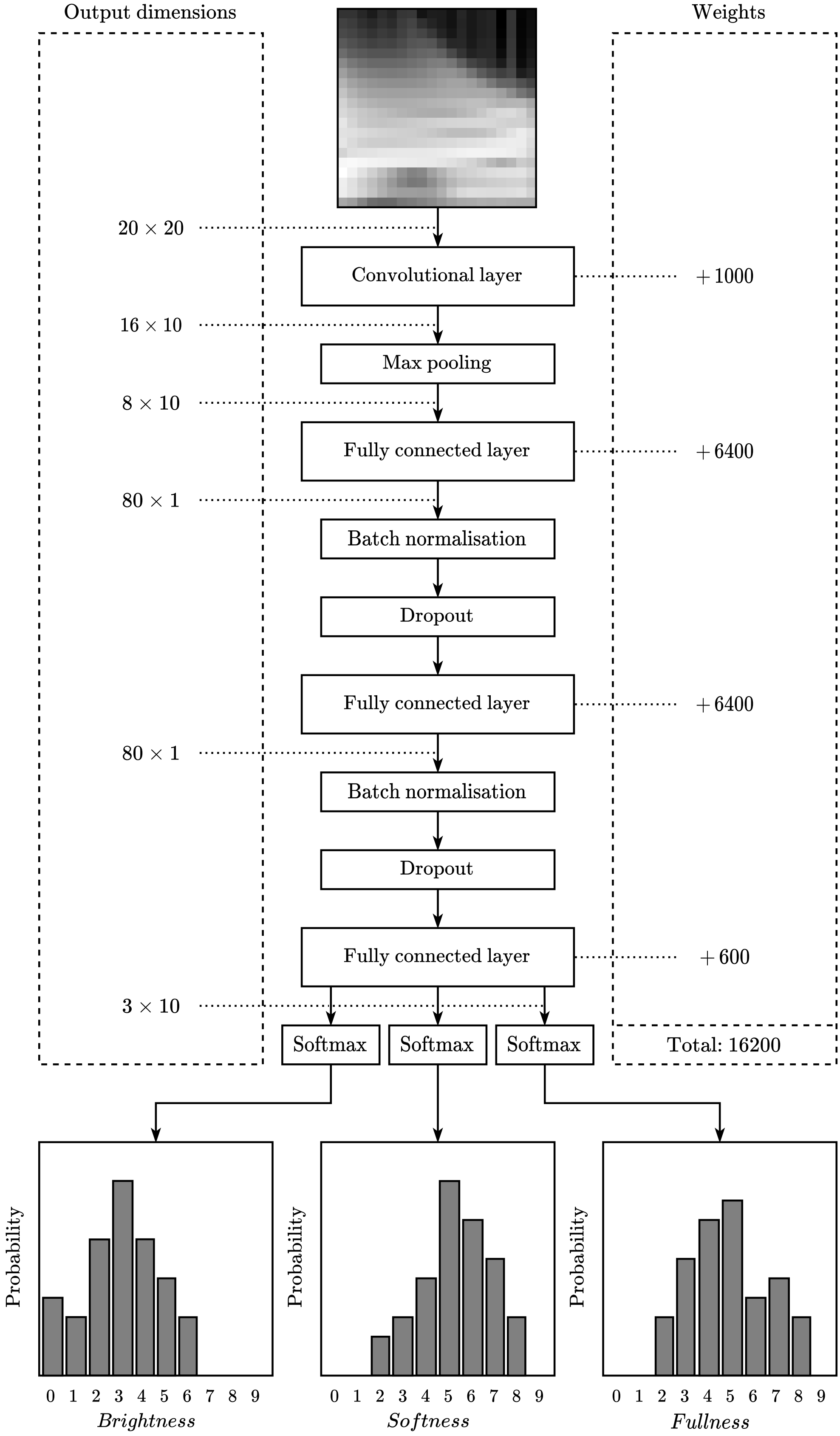

Figure 3.6: The fully connected style network. Note the large total number of total weights.

Chapter 4

Results

In Chapter 2, we hypothesised that DNNs are a good way to predict metaphorical timbre judgements, due to the complexity of their acoustic correlates. When creating a dataset of these judgements, I introduced two more hypotheses. Firstly, my judgements along the three scales would be mostly orthogonal. Secondly, my judgements for the dataset would be highly consistent with a sample of repeated judgements. Section 4.1 gives an overview of the metaphorical judgement data, which confirms these hypotheses. Results from the networks are then presented in Section 4.2, showing that both were accurate, but the convolutional style network with fewer weights performed better.

4.1 Data analysis

Figure 4.1 gives the distributions of my metaphorical judgements over each of the three VAME scales: *brightness* judgements were mostly neutral, whereas *softness* judgements had the most spread.

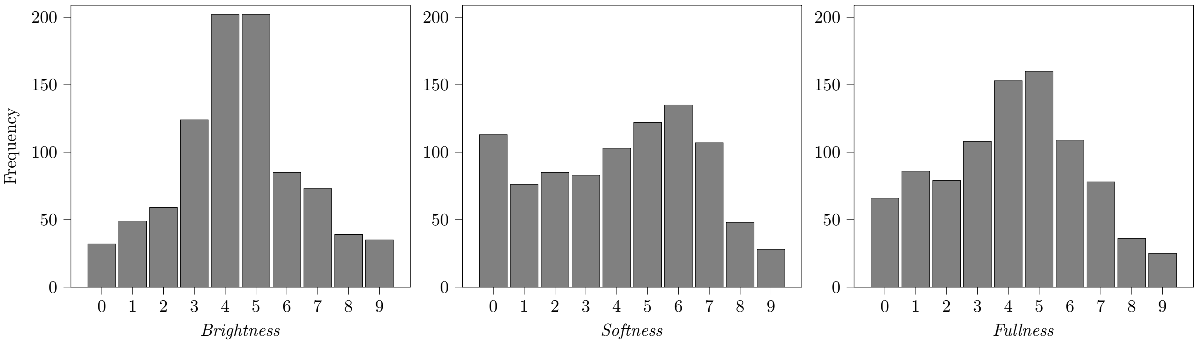

Figure 4.1: Scale judgements by frequency in the total (pre-splitting) dataset.

4.1.1 Scale orthogonality

As discussed in Section 3.1.2.2, I tried to choose scales that covered as much of the timbral variation in the sound data as possible, but I didn't expect them to be perfectly orthogonal. I measured correlations between them using the non-parametric Spearman's rank correlation coefficient (ρ), which is appropriate for ordinal data. There was almost no correlation between *brightness* and *softness* ($\rho(900) = -.02,\ p < .0001$), but *brightness* and *fullness* showed a weak positive correlation ($\rho(900) = .28,\ p < .0001$) and *softness* and *fullness* a weak negative correlation ($\rho(900) = -.29,\ p < .0001$). Figure 4.2 visualises these correlations as heatmaps.

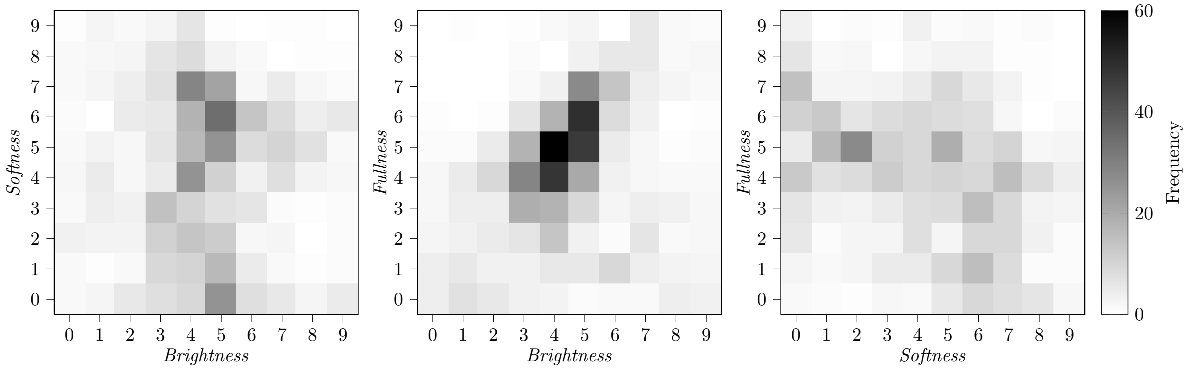

Figure 4.2: Heatmaps showing the correlations between my scale judgements in the total dataset: *brightness* and *fullness* were weakly positively correlated, where *fullness* and *softness* were weakly negatively correlated.

4.1.2 Consistency within sounds

To see how consistently I used the scales in repeated judgements, I made a second set of judgements for 10% of the sounds, as outlined in Section 3.1.2.1. To compare this set to my judgements on those sounds for the dataset, I calculated the Tau-equivalent reliability coefficient (ρ_T) for each scale — a common consistency metric for Likert scale-like data. As expected, my judgements were very similar between the two, with values greater than 0.9 for *brightness* ($\rho_T = 0.91$), *softness* ($\rho_T = 0.96$) and *fullness* ($\rho_T = 0.92$).

4.2 Neural network performance

4.2.1 The convolutional style network

The convolutional style network performed well on the validation set, and stopped improving after 150 passes over the training data (or 'epochs'). Figure 4.3 shows how the network continued to fit to the training data after 150 epochs, but this didn't affect how well it generalised to the validation data. Training took 92.5 seconds overall.

For each sound in the test set, I recorded the VAME score predicted to be most likely by the network, and compared it to the actual label to get a prediction error. For example, a *brightness* prediction of 4 for a sound with *brightness* labelled 6 produced an error of -2. Figure 4.4 plots these prediction errors by frequency, showing that the network's predictions were overwhelmingly identical or close to the labelled scores.

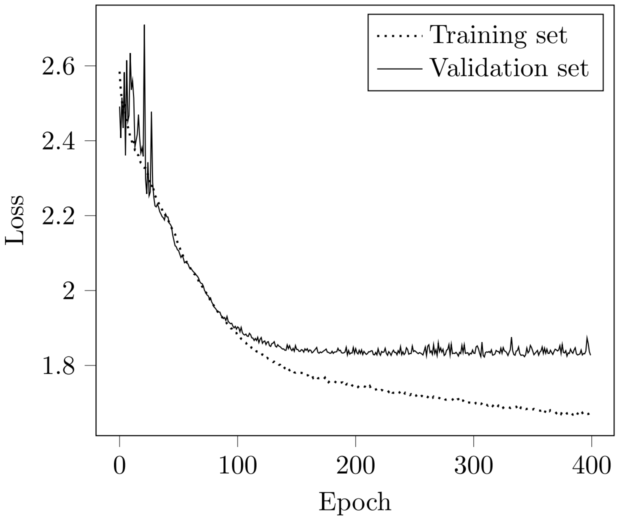

Figure 4.3: The performance of the convolutional style network during training. Note how the training set loss continues to fall while the validation set loss doesn't.

However, because the distributions of my judgements were unbalanced (Figure 4.1), these results didn't necessarily mean the network was making sophisticated predictions — it could have produced low errors just by guessing the most frequent scores for each scale. I therefore weighted the test set predictions inversely to the frequency of each scale's VAME scores in the test set: a prediction for a sound labelled as having *brightness* 5 contributed less than a prediction for a sound with *brightness* 9. I then normalised the total frequencies of each error to sum to one, converting them into probabilities. The weighted errors are plotted in Figure 4.5, which shows that the network errors were still mostly low when accounting for the unbalanced judgement distributions. By simply adding the bars between a positive and negative weighted error size, we can assign a probability to the network's predictions falling within a given range of the labelled score for a sound in the test set. These probabilities are given in Table 4.1.

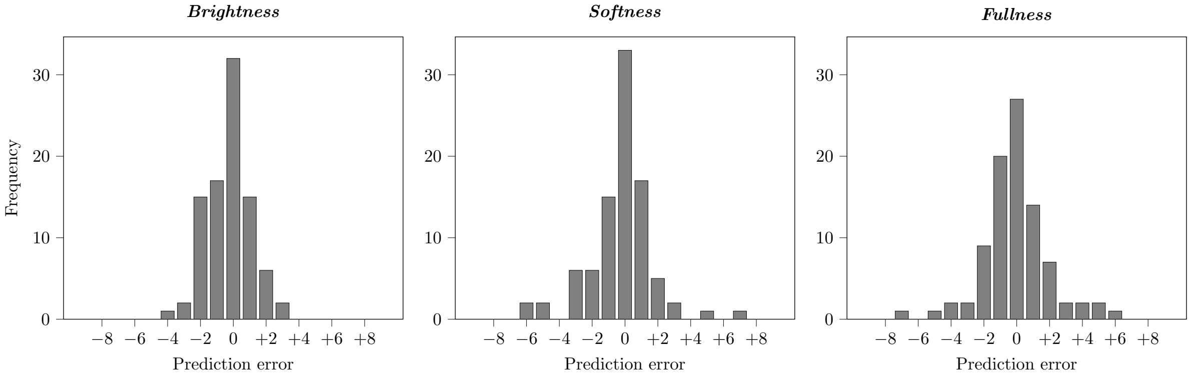

Figure 4.4: Test set error by frequency for the convolutional style network.

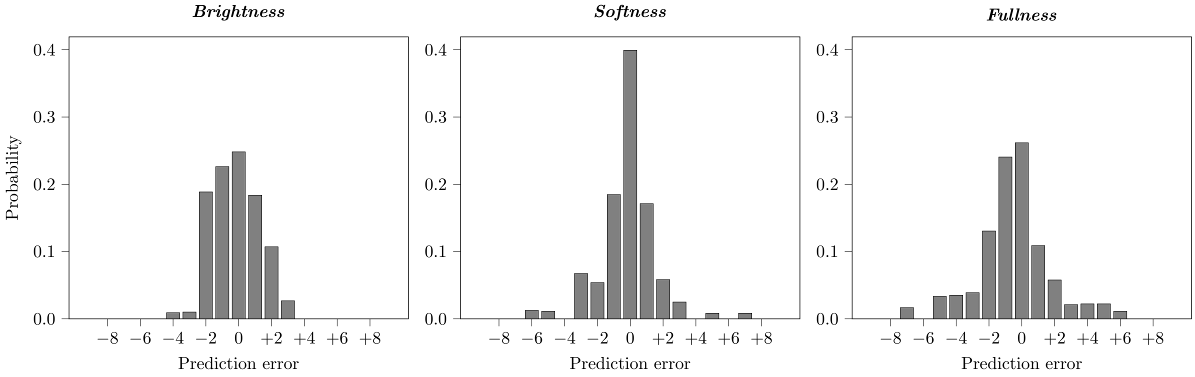

Figure 4.5: Weighted test set error for the convolutional style network. Probabilities of zero arise where the network didn't make any errors of that size in the test set.

Prediction error	***Brightness***	***Softness***	***Fullness***	Mean
±0	0.25	0.40	0.26	0.30
±1	0.66	0.76	0.61	0.68
±2	0.95	0.87	0.80	0.87

Table 4.1: Weighted test set prediction error size probabilities for the convolutional style network. For example, the network's best guess for the *brightness* of a sound in the test set was 95% likely to fall within 2 points of the labelled score. This corresponds to the total height of the middle five *brightness* bars in Figure 4.5.

4.2.2 The fully connected style network

Lastly, the fully connected style network needed to be trained for 1000 epochs before it stopped improving, because its learning rate had to be set quite low. Training therefore took 228.3 seconds — more than twice as long as for the convolutional style network. However, it actually performed slightly worse overall: the probabilities of its predictions falling close to the labelled scores in the test set, given in Table 4.2, were generally lower than for the first design. As a result, the discussion in Chapter 5 focuses on the less computationally intensive convolutional style network.

Prediction error	***Brightness***	***Softness***	***Fullness***	Mean
±0	0.22	0.37	0.37	0.32
±1	0.62	0.68	0.71	0.67
±2	0.87	0.80	0.80	0.82

Table 4.2: Weighted test set prediction error size probabilities for the fully connected style network.

Chapter 5

Discussion

The modelling in Chapter 3 tested the main hypothesis that neural networks are a good way to predict metaphorical timbre judgements. From the large judgement dataset I created, I first found my *brightness*, *softness* and *fullness* judgements to be mostly orthogonal. Repeating these judgements also revealed that they were highly consistent over the same sounds. Most significantly, a modern neural network was able to predict them with quite high accuracy. Section 5.1 discusses these results in relation to the arguments in Chapter 2, and highlights the limitations of my approach. Section 5.2 then concludes the chapter by considering directions for future research.

5.1 The neural network approach

The predictions by the convolutional style network (or 'convnet' for short) were on average 87% likely to fall within 2 points of the labelled score for a sound in the test set (Table 4.1). This accuracy substantiates the promising results from the early neural networks discussed in Section 2.2.2. What is more, it seems to show the superiority of neural approaches to linear regression: while the convnet can't be directly compared to models predicting judgements by different participants on different sounds, it was accurate across all three scales, where we've seen linear regression struggles to make predictions on scales outside the *brightness* group. This accuracy was on a distinct test set, meaning the network generalised well to new sounds. In addition, the *brightness*, *softness* and *fullness* judgements it predicted weren't strongly correlated, so the fact that it was accurate across all three shows it learned to verbalise a wide range of possible timbres. Overall, these results strongly support the hypothesis that DNNs are a good way of predicting timbre judgements on metaphorical scales.

The network accuracies also indicate this approach has practical value. The convnet in particular presents several advantages over previous neural networks: its predictions on ten-point VAME scales are much more detailed than the binary outputs of Feiten & Ungvary (1991). It is also very quick to train, where the very large network in Gounaropoulos & Johnson (2006) likely wasn't. Finally, its convolutional layers make it much more position invariant than any of the fully connected designs.

As for the broader question of whether timbre is something we can verbalise, the convnet's ability to predict my metaphorical judgements, as well as the high consistency of those judgements over the same sounds, suggests these metaphors do refer to specific acoustic patterns. This supports the studies we saw in Section 2.1.2.1 showing that people mostly agree on which metaphors best describe a timbre, and the conclusion that timbre is not ineffable. A strength of my approach is that it may better reflect how consistently timbral metaphors are used than studies comparing judgements from multiple participants — individuals might use a metaphor consistently, but if they disagree on what it refers to, that metaphor will appear inconsistent. Using a neural network means we can measure consistency within participants, but without the risk of repetition priming that comes with asking them to judge a sound multiple times.

However, the convnet wasn't perfectly accurate. There are a number of possible reasons why. For one, my judgements probably weren't perfectly consistent, adding an element of randomness to the training and test data. Another explanation is that it failed to completely learn the acoustic patterns behind *brightness*, *softness* and *fullness* — its design might have been too simple to capture them, or there might not have been enough training data. However, neither of these seem very likely. The convnet performed similarly to the fully connected network, which had far more trainable weights, and the number of weights in the convnet (16200) was of the same order of magnitude as the number of items in the training set (9360). A bottleneck in the feature extraction stage is more plausible, as the log Mel spectral features were fairly coarse. The fact that the networks were quite accurate indicates these features captured most of the descriptors relevant to my metaphorical judgements. However, they couldn't have captured micro-spectral descriptors like the spectral flux mentioned in Section 2.2.1.2, which might have been relevant too. This may be why *fullness* was slightly less predictable — I didn't extract a feature for the detuning of a sound as Gounaropoulos & Johnson (2006) did, which is often associated with a 'broadening' effect (Ethington & Punch, 1994). Similarly, Figure 5.1 shows how the extracted features poorly represented vibrato smaller than a single frequency bin. A final explanation is that some of my judgements reflected specific cognitive associations, like the example of a steel drum's *brightness* given in Section 2.2.1.3, but they weren't common enough in the training set for the networks to learn.

5.1.1 Limitations

There are several important limitations to these results, most obviously that the judgement data was collected from only one person. The reliability coefficients for how consistent I was within sounds were a little higher than the consistency between participants coefficients reported by Darke (2005), as expected. This suggests I might have been exceptionally consistent in my judgements for the dataset too, either

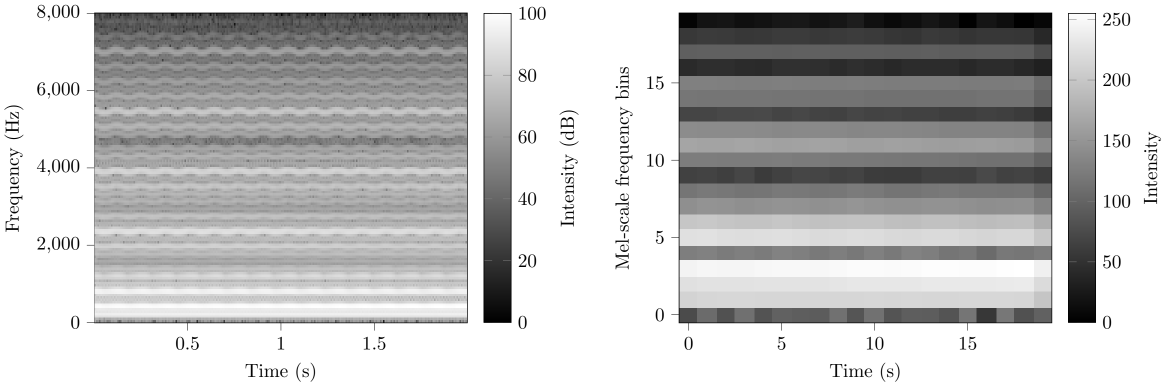

Figure 5.1: On the left: a sound in the dataset with vibrato; on the right: log Mel spectral features failing to capture that vibrato.

through musicality, or just from being too analytical. As a result, they might have been easier to predict than metaphorical timbre judgements normally are. This also applies to the orthogonality of the three scales: I might have subconsciously aimed to keep my judgements distinct.

Secondly, limiting the non-timbral variation in the sound data may have excessively simplified the problem. Like most studies on timbral metaphor, I only collected judgements for sounds at a single pitch and loudness. However, for any practical purposes a network would need to learn how these factors affect metaphorical judgements, which would likely require much more training data. Agus et al. (2019) also raises the point that most timbral classifiers are trained and tested on cleanly recorded sounds in isolation, even though this is rarely how we are exposed to sounds in real life. Training the networks to learn how metaphorical judgements are affected by the context of a timbre would again require more data.

There may also have been too much timbral variation in the sound data — I collected judgements for a wide range of synthetic and acoustic sounds, where most studies have limited their stimuli much more, to orchestral (Elliott et al., 2013), synthetic (von Bismarck, 1974), or even just organ timbres (Disley & Howard, 2004). My judgements might therefore have been describing the sources of the different sounds rather than their timbres, making them more "classificatory than continuous" (McAdams et al., 1995, p. 191). This would have essentially turned the classification problem into instrument recognition, which we already know DNNs are capable of.

5.2 Future research

This dissertation set out to show that DNNs are a good way to predict metaphorical timbre judgements. Based on my judgements, this seems to be true. The results also support research to show timbre is not ineffable. That said, there are some clear directions for future work. Firstly, investigating whether a DNN can accurately predict judgements by other people, who might be less consistent. It would also be interesting to see how well the networks generalise to judgements from multiple participants. They might associate the three scales with different descriptors, and their judgements would likely reflect different cognitive biases. Similarly, we could ask whether the predictions for *brightness*, *softness* and *fullness* are still accurate in other languages. There is some evidence for language-independent interpretations of timbral metaphors (see Zacharakis et al., 2014), but they don't always translate well — Kendall & Carterette (1993a) couldn't replicate German speakers' use of *sharpness* with English speakers, and Disley & Howard (2004) even found differences between the metaphorical judgements of British and American English speakers.

Future research could also work on the practical applications discussed in Section 2.2.3, such as searching sample libraries. This would involve experimenting with different features, trying more network designs, and incorporating pitch and loudness variation. As it stands, the data collection pipeline could be made more efficient by augmenting the training set after feature extraction. The networks could also be improved by using more detailed inputs: a trend in machine learning is that given enough training data, models are more accurate when their inputs aren't simplified into features. For example, the state-of-the-art speech synthesiser WaveNet uses raw audio data (Oord et al., 2016).

Lastly, the networks could be used to try and make predictions on scales that linear regression struggles with, such as *warmth*. Most research on timbral metaphor treats all scales the same way, but it's possible that some are more expressive than others. Lichte (1941) proposed that *fullness* is less fundamental than *brightness*, which does seem consistent with the results given here, as both networks were slightly more likely to predict *brightness* judgements within two points of a label than *fullness* judgements. If *warmth* judgements aren't predictable even with a DNN, then perhaps it isn't an expressive metaphor.

Chapter 6

Conclusion

Timbre is something that feels ineffable: we intuitively describe it with metaphors like *brightness*, but it's hard to explain what these metaphors actually refer to. In Chapter 2, we saw that people generally use these metaphors consistently, and that a way to make accurate predictions along metaphorical scales would be useful. Nonetheless, statistical methods have had mixed success, likely because timbral metaphors have very complex acoustic meanings. This suggests that neural networks might do better. However, while existing neural approaches have produced promising results, they are outdated, and all very light on details. This dissertation therefore aimed to show that neural networks are a good way to predict metaphorical timbre judgements. Chapter 3 described how I created a dataset, which involved manually labelling 900 sounds with 10-point scale judgements for the *brightness*, *softness* and *fullness* of their timbres. I then designed and built two neural networks to predict those judgements, using a number of modern deep learning techniques. The results, given in Chapter 4, showed that the better of the two achieved an average of 87% accuracy within two points of the judgements in a separate set of test data. As discussed in Chapter 5, this indicates that neural networks are more than capable of predicting metaphorical timbre judgements, with a number of potential practical applications. It also supports the conclusion that timbre is not ineffable, by showing that timbral metaphors are predictable from the sounds they describe. However, more research is needed to show that these findings extend beyond my judgements alone.

References

Agus, T. R., Suied, C., & Pressnitzer, D. (2019). Timbre recognition and sound source identification. In *Timbre: Acoustics, perception, and cognition* (pp. 59–85). Springer.

Alluri, V., & Toiviainen, P. (2010). Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception*, *27*(3), 223–242.

Almeida, A., Schubert, E., Smith, J., & Wolfe, J. (2017). Brightness scaling of periodic tones. *Attention, Perception, & Psychophysics*, *79*(7), 1892–1896.

Aucouturier, J.-J. (2006). *Ten experiments on the modeling of polyphonic timbre* (Unpublished doctoral dissertation). Université Pierre et Marie Curie (Paris 6).

Aucouturier, J.-J., & Pachet, F. (2003). Representing musical genre: A state of the art. *Journal of new music research*, *32*(1), 83–93.

Bellemare, M., & Traube, C. (2006). Investigating piano timbre: Relating verbal description and vocal imitation to gesture, register, dynamics and articulation. In *Proceedings of the 9th international conference on music perception and cognition.*

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound.* MIT press.

Brookes, T., & Williams, D. (2007). Perceptually-motivated audio morphing: Brightness. In *112th audio engineering society convention.*

Brookes, T., & Williams, D. (2010). Perceptually-motivated audio morphing: Warmth. In *128th audio engineering society convention.*

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *The Journal of the Acoustical Society of America*, *118*(1), 471–482.

Cheminée, P. (2006). "vous avez dit 'clair'?" le lexique des pianistes, entre sens commun et terminologie. *Cahiers du LCPE: Dénomination, désignation et catégories*, *7*, 39–54.

Chollet, F., et al. (2015). *Keras.* https://keras.io.

Chollet, F., et al. (2018). *Deep learning with Python.* Manning Publications.

Cosi, P., De Poli, G., & Lauzzana, G. (1994). Auditory modelling and self-organizing neural networks for timbre classification. *Journal of New Music Research*, *23*(1), 71–98.

Darke, G. (2005). Assessment of timbre using verbal attributes. In *Conference on interdisciplinary musicology.*

Disley, A. C., & Howard, D. M. (2004). Spectral correlates of timbral semantics relating to the pipe organ. *Speech, Music and Hearing*, *46*, 25–39.

Disley, A. C., Howard, D. M., & Hunt, A. D. (2006). Timbral description of musical instruments. In *International conference on music perception and cognition* (pp. 61–68).

Eitan, Z., Katz, A., & Shen, Y. (2010). Effects of pitch register, loudness and tempo on children's use of metaphors for music. In *11th international conference on music perception and cognition.*

Eitan, Z., & Rothschild, I. (2011). How music touches: Musical parameters and listeners' audio-tactile metaphorical mappings. *Psychology of Music*, *39*(4), 449–467.

Elliott, T. M., Hamilton, L. S., & Theunissen, F. E. (2013). 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.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Eck, D., Simonyan, K., & Norouzi, M. (2017). *Neural audio synthesis of musical notes with wavenet autoencoders.*

Esling, P., Bitton, A., & Chemla-Romeu-Santos, A. (2018). Generative timbre spaces: regularizing variational auto-encoders with perceptual metrics. In *Proceedings of the 21st international conference on digital audio effects.*

Ethington, R., & Punch, B. (1994). Seawave: A system for musical timbre description. *Computer Music Journal*, *18*(1), 30–39.

Faure, A., Mcadams, S., & Nosulenko, V. (1996). Verbal correlates of perceptual dimensions of timbre. In *4th international conference on music perception and cognition* (pp. 79–84).

Feiten, B., & Ungvary, T. (1991). Organisation of sounds with neural nets. In *Proceedings of the international computer music conference* (pp. 441–444).

Fritz, C., Blackwell, A., Cross, I., Moore, B., & Woodhouse, J. (2008). Investigating English violin timbre descriptors. In *Proceedings of the 10th international conference on music perception and cognition* (pp. 638–639).

Goodwin, A. (1988). Sample and hold: pop music in the digital age of reproduction. *Critical Quarterly*, *30*(3), 34–49.

Gounaropoulos, A., & Johnson, C. (2006). Synthesising timbres and timbre-changes from adjectives/adverbs. In *Workshops on applications of evolutionary computation* (pp. 664–675).

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

Han, Y., Kim, J., & Lee, K. (2016). Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(1), 208–221.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–456).

Iverson, P., & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbrea. *The Journal of the Acoustical Society of America*, *94*(5), 2595–2603.

Jiang, W., Liu, J., Zhang, X., Wang, S., & Jiang, Y. (2020). Analysis and modeling of timbre perception features in musical sounds. *Applied Sciences*, *10*(3), 789.

Kendall, R. A., & Carterette, E. C. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Perception*, *8*(4), 369–404.

Kendall, R. A., & Carterette, E. C. (1993a). verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives. *Music Perception*, *10*(4), 445–467.

Kendall, R. A., & Carterette, E. C. (1993b). verbal attributes of simultaneous wind instrument timbres: II. Adjectives induced from Piston's "Orchestration". *Music Perception*, *10*(4), 469–501.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097–1105.

Lakoff, G. (2008). The neural theory of metaphor. In R. W. Gibbs Jr. (Ed.), *The Cambridge handbook of metaphor and thought* (p. 17–38). Cambridge University Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

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

Lemaitre, G., & Susini, P. (2019). Timbre, sound quality, and sound design. In *Timbre: Acoustics, perception, and cognition* (pp. 245–272). Springer.

Levinson, S. C., & Majid, A. (2014). Differential ineffability and the senses. *Mind & Language*, *29*(4), 407–427.

Lichte, W. H. (1941). Attributes of complex tones. *Journal of Experimental Psychology*, *28*(6), 455–480.

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior Research Methods*, *45*(2), 516–526.

Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., O'grady, L., ... others (2018). Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences*, *115*(45), 11369–11376.

Marks, L. E. (1982). Synesthetic perception and poetic metaphor. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(1), 15.

Marks, L. E., Hammeal, R. J., Bornstein, M. H., & Smith, L. B. (1987). Perceiving similarity and comprehending metaphor. *Monographs of the Society for Research in Child Development*, i–100.

Marozeau, J., & de Cheveigné, A. (2007). The effect of fundamental frequency on the brightness dimension of timbre. *The Journal of the Acoustical Society of America*, *121*(1), 383–387.

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception*, *28*(7), 903–923.

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

Miranda, E. R. (1995). An artificial intelligence approach to sound design. *Computer Music Journal*, *19*(2), 59–75.

Nykänen, A., Johansson, Ö., Lundberg, J., & Berg, J. (2009). Modelling perceptual dimensions of saxophone sounds. *Acta Acustica United with Acustica*, *95*(3), 539–549.

Olofsson, J. K., & Gottfried, J. A. (2015). The muted sense: neurocognitive limitations of olfactory language. *Trends in Cognitive Sciences*, *19*(6), 314–321.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Pons, J., Slizovskaia, O., Gong, R., Gómez, E., & Serra, X. (2017). *Timbre analysis of music audio signals with convolutional neural networks.*

Porcello, T. (2004). Speaking of sound: Language and the professionalization of sound-recording engineers. *Social Studies of Science*, *34*(5), 733–758.

Pratt, R., & Doak, P. (1976). A subjective rating scale for timbre. *Journal of Sound and Vibration*, *45*(3), 317–328.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Saitis, C., & Weinzierl, S. (2019). The semantics of timbre. In *Timbre: Acoustics, perception, and cognition* (pp. 119–149). Springer.

Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, *24*(3), 279–283.

Samoylenko, E., McAdams, S., & Nosulenko, V. (1996). Systematic analysis of verbalizations produced in comparing musical timbres. *International journal of psychology*, *31*(6), 255–278.

Schubert, E., Wolfe, J., & Tarnopolsky, A. (2004). Spectral centroid and timbre in complex, multiple instrumental textures. In *Proceedings of the 8th international conference on music perception and cognition* (pp. 112–116).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Stepánek, J. (2006). Musical sound timbre: Verbal description and dimensions. In *Proceedings of the 9th international conference on digital audio effects* (pp. 121–126).

Stepanek, J., & Moravec, O. (2005). Verbal description of musical sound timbre in Czech language and its relation to musicians profession and performance quality. In *Conference on interdisciplinary musicology.*

timbre, n.3. (2020). In *OED online.* Oxford University Press. Retrieved 2021-04-08, from http://www.oed.com/view/Entry/202089

Traube, C. (2004). *An interdisciplinary study of the timbre of the classical guitar* (Unpublished doctoral dissertation). McGill University.

von Bismarck, G. (1974). Timbre of steady sounds: A factorial investigation of its verbal attributes. *Acta Acustica united with Acustica*, *30*(3), 146–159.

Von Helmholtz, H. (1877). *On the sensations of tone as a physiological basis for the theory of music* (A. J. Ellis, Trans.). Longmans, Green.

Wallmark, Z. (2019a). A corpus analysis of timbre semantics in orchestration treatises. *Psychology of Music*, *47*(4), 585–605.

Wallmark, Z. (2019b). Semantic crosstalk in timbre perception. *Music & Science*, *2*, 1–18.

Wallmark, Z., & Kendall, R. A. (2018). Describing sound: The cognitive linguistics of timbre. In *The Oxford handbook of timbre.*

Wessel, D. L. (1979). Timbre space as a musical control structure. *Computer Music Journal*, 45–52.

Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition*, *179*, 213–220.

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2014). An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. *Music Perception: An Interdisciplinary Journal*, *31*(4), 339–358.

Zacharakis, A., & Reiss, J. (2011). An additive synthesis technique for independent modification of the auditory perceptions of brightness and warmth. In *130th audio engineering society convention.*

Appendix A

Source code

The data and source code for this project are available online at github.com/notefive/verbalising-timbre.