University of Amsterdam

Masters Thesis

The Music Emerges: A Computational Approach to the Speech-to-Song Illusion

Author:
Arran Lyon

Supervisor:
Dr. Makiko Sadakata

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Computational Science

in the

Computational Science Lab
Informatics Institute

August 2020

Declaration of Authorship

I, Arran LYON, declare that this thesis, entitled 'The Music Emerges: A Computational Approach to the Speech-to-Song Illusion' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at the University of Amsterdam.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: 10 August 2020

UNIVERSITY OF AMSTERDAM

Abstract

Faculty of Science
Informatics Institute

Master of Science in Computational Science

The Music Emerges: A Computational Approach to the Speech-to-Song Illusion

by Arran Lyon

The Speech-to-Song Illusion is an auditory effect whereby a perceptual transformation occurs during an exact repetitions of some short, naturally spoken phrase, where the listener begins to hear as if the speaker is singing, when initially it sounded exactly like speech. Understanding the cognitive mechanisms underlying the illusion could reveal the inner workings of speech, song and music perception. We use a collection of audio stimuli (300) related to this illusion along with a parallel set of perceptual rating data obtained from a previous study. We describe an algorithm that automatically transcribes the melody as a sequence of tones within the natural voice that improves upon the original formulation. We define a set of 33 melodic, rhythmic, audio and dissonance related features measured from the raw audio and the extracted tone sequence, including features previously connected to the occurrence of the illusion. New features include a metric that measures the distance of the melody to typically composed melodies (according to a Bayesian model), along with a novel set of features that captures the dynamic change of tension and release of the musical phrase based on a measure of sensory dissonance. We then propose a suitable method to best assign a binary label (transforming or non-transforming) to the stimuli based on the continuous range of scores provided by participants of an experiment. After under going a feature selection procedure, several data classifiers (linear and non-linear support vector machines and logistic model, along with two ensembles) that predict if the stimulus will transform from the features all obtain balanced accuracy scores between 0.66 and 0.71, where each model used between 7 and 10 features. We confirm that stability plays an important part in the illusion, along with the closeness of the melody to typical melodies and the consonance of the ending of the melody, suggesting transforming stimuli are those with characteristics of ordinary musical phrases. Finally, 55 participants took part in a follow up validation experiment with 98 new stimuli to test if the models generalised well, and found that one model in particular (a linear support vector machine) maintained a score above baseline on the new vocal stimuli (balanced accuracy 0.65 on the new stimuli).

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Makiko Sadakata for introducing me to this curious illusion, and for all her continuous and encouraging support in this research.

Next, to Bas Cornelissen for his valuable and foundational contribution for which much of this work is based on, and without this would not have been possible, and also to Gerben Groenveld for his feedback and audio materials from his own research into the illusion.

A huge thanks to Prof. Henkjan Honing who has kindly contributed his time to evaluating the work, and for the valued early feedback in the process.

I would like to offer a my sincerest gratitude to my friends Valentin Vogelmann, for his helpful advice on theoretic concepts, and to Mattia Capelletti for his cultural perspective.

A warm thank you goes to my good friend Tanja, for her continued support over many long walks.

Finally, my gratitude extends to friends Charlie, Charlotte, Efi, Gerben, David and Alice for their advice, feedback and inspiration, to Black Gold Amsterdam for the music, and to all my friends and family across the continent.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Repetition is not repetition…
the same action makes you feel something
completely different by the end.
— Pina Bausch

1.1 The Mystery of Repetition

In her book *On Repeat*, Margulis (2014) recounts the importance and power of repetition in music, and some of the paradoxes that it brings about. The simple act of repeating short segments of otherwise wavering, atonal contemporary music increased the listeners response to the music to the point that they were rated as more enjoyable and artistic then the unaltered material, despite being original work from well celebrated composers (Margulis, 2013b). Somehow, reiterating a sound introduces a new perspective on the segment that could not be appreciated when played only once, and transforms it into something beyond its original form to such a degree that the listener feels very different towards it.

Repetition itself materialises on all timescales of the music listening experience — from short immediate recurrence of rhythms and melodies, to the reiterated choruses in pop music, to the repeat listening of entire songs or albums over hours, days or years. In all these levels it seems that there is pleasure in listening to the same sound, ad verbatim, despite already knowing exactly how it sounds. In fact this effect is rather strong, such that repeated exposure of a piece of music can increase the listeners preference towards it simply through the familiarity of the piece, a phenomenon known as the *mere exposure effect* (Zajonc, 1968).

It would appear then that recognition of the sound and the expectancy of musical events are important factors to the listeners satisfaction of it, however this idea is contradicted in another study by the same author (Margulis, 2010). She found that listeners unfamiliar with Beethoven String Quartets enjoyed the performance of the music significantly less when they were presented with a description of the piece beforehand, suggesting that prior knowledge of either the dramatic or structural aspects is actively detrimental to the listeners experience. The paradoxical nature of these results suggests that both the expected and unexpected are somehow necessary for the enjoyment of music. Perhaps then the pleasure arises from discovering the music in the sound by oneself, and through repetition this search for musical information is facilitated by the multiple opportunities on each loop to find this.

Repetition is also a signifier of intention — in his own development of *musique concrète*, Pierre Schaeffer states that exact reproduction of a sound is unnatural, and so to hear a repeated sound implies a synthetic and human process, that the original sound was not by accident and not that of some random chance (Schaeffer, 1952). In music in general, this could be an unusual melodic or rhythmic structure that on first listen appears 'wrong', but when this phrase is repeated multiple times the perception shifts — what originally 'broke the rules' of music turns out to be a compositional choice by the creator, and the musical idea is cemented and legitimised. This has been demonstrated by Margulis and Simchy-Gross (2016) with a set of experiments that show that randomly generated sequences of notes are rated highly as sounding musical when it is repeated several times, even though it does not conform to any rules of musical theory. In the authors own experience, while attempting to create an artificial neural network that can compose original music, it was found the listener responses to the models output was greatly improved by repeating short phrases produced by the computer. Untouched, the models output sounded uncanny and unlike that of any human composition, almost random even, until repetition seemed to legitimise the music to the point that it sounded like 'real' music.

This technique has of course been utilised by composers throughout history and musical cultures. Repetition is an integral (and almost defining) characteristic in many forms of music, from ritualistic rhythms, to the cannons of classical, to contemporary compositions, electronic dance music, and the sample culture and beat-loops of hip-hop and pop music of the modern era. Most notably, Steve Reich is perhaps the most well known composer of the twentieth century to embrace the power of repetition, where his music often used tape loops, repeated musical ideas and reiterated vocals to form complete works. Despite the exact reproduction of the sounds, there is still somehow a dynamic shift throughout the piece, as each loop has a slight change in its context, thus giving an ever changing perspective to the sounds.

Whilst being so prevalent in music, repetition appears to be less fundamental in other forms of art or expression, its use of which is reserved only as a specific device when the idea calls for it. For example, in Andy Warhol's mass produced prints of Marilyn Monroe, the repetition itself *is* the art and makes the intended narrative, rather than the individual prints themselves. In spoken word and poetry, repetition of a phrase serves to add emphasis to the statement, and to add an extra layer of communication beyond the literal meaning of the words. In music however the opposite is true — it is rare to question the use of a composers use of repetition, we think nothing of it for the most part, and it has even become a point to actively *avoid* the use repetition at all in some contemporary musical pieces. To understand repetition then is to gain deep insight into the workings of music itself.

There appears to be a deep connection between music and natural language in that both of these are built by an ordered, hierarchical structure of smaller units — these could be words in a sentence or notes in a melody. While there is a lot of discussion around the nature of a musical grammar (e.g. Lerdahl and Jackendoff, 1983), what is clear is that music has some form of long-term dependencies and a complex organisation between events within it, and that the surrounding context of the events are important. Nonetheless, Patel (2003) proposes and provides evidence of a deep connection of music and language in his shared syntactic integration resource hypothesis (SSIRH), suggesting that there indeed appears to be a common baseline between both. Despite this connection however, natural language does not require repetition in its production, as reiterating words ad verbatim does not carry new information, and so any theory of musical grammar must include a provision for repetition. Understanding repetition in music then could reveal more on how such a grammar could function[1]. This is important, as all sound is simply a collection of pressure waves within a medium, but it is a uniquely human decision to distinguish a particular assortment of these waves that we experience as being musical, or not. Therefore, to understand what aspects of the signal makes us call it *music* reveals some deep inner mechanisms of the human mind, and the mystery of repetition could lead us in this direction of understanding.

1.2 The Transformation of Sound

1.2.1 Auditory Illusions

Illusions offer a unique opportunity to study the cognitive mechanism surrounding perception — a stimulus of the senses that somehow 'breaks' the normal perception of the

[1] A formal musical grammar system should allow for its own rules to be broken if the rule violation is repeated enough times.

source highlights the differences between the human sensor, the brain, and an ideal receptor. Such effects are of interest to researchers as they allow a chance to probe these shortcomings in an attempt to understand these discrepancies, and from there construct models which capture the dynamics of the percept. For example, in the striking *Café-wall* optical illusion, perfectly horizontal lines surrounded by offset rows of tiles of black and white squares appear to tilt the lines sideways. Ultimately, this illusion motivated theories and a model of how the human retinal system detects and perceives the edges in an image (Nematzadeh and Powers, 2016).

Illusions extend beyond the optical domain — auditory illusions appear when the listener hears something that does not exist in the audio signal, or if the sound they hear shifts from one thing to another. For example, in the missing fundamental illusion, the brain can hear a fundamental frequency in an audio signal made up of only its harmonics frequencies (Licklider, 1951), and through this phenomenon neural models of pitch perception have been proposed that account for this illusion (e.g. Chialvo, 2003). Cognitive phenomenon such as the McGurk effect relate both hearing and vision in speech recognition (McGurk and MacDonald, 1976), such that mismatched ques from both sensors leads to the experience of a third, different sound, and this has led to further research in the multi-modal aspects of speech recognition. Diana Deutsch, a predominant researcher in the psychology of music, has discovered several music related illusions, such as the *Octave Illusion* (Deutsch, 1974a), *Scale Illusion* (Deutsch, 1974b), and the *Tritone Paradox* (Deutsch, 1986) that reveal inaccuracies in the human pitch perception of sound. Through these anomalies, we can begin to understand more of the auditory system, and how the brain processes and categorises what it hears.

1.2.2 Repetition Based Illusions

Several auditory illusions occur when a sound source is repeated multiple times in a row and a perceptual shift occurs, where subsequent reiterations of the sound are perceived in a very different manner than on the original play-through. In the *verbal transformation effect* (VTE, Warren and Gregory, 1958), when a short word is repeated quickly without pause or modification, the listener begins to hear other words from the sound. For example, the word *ripe* transforms and flips between similar sounding words such as *right*, *white*, *bright* and even *bright-light*. This effect is quite unstable as the alternate words tend to switch around on each loop, or the original word returns to the listener. Another related effect is *semantic satiation*, first studied by Bassett and Warne (1919). Here, the word is repeated in a similar fashion but instead of transforming into other words, the sound decomposes into almost nonsensical, incoherent sounds such that the original word sounds alien and unfamiliar, and even loses all meaning. Unlike the VTE, this perceptual shift is quite strong and lasts for some time after the priming phase before the word begins to sound normal again.

In both effects, the act of repetition recontextualises the sound — without the support of the surrounding context of other words and utterances, the sound is free to take on different meanings, or even lose all definition completely. Semantic satiation occurs when the brain no longer focuses on the meaning of the word due to a 'fatigue' of the neural pathways (Jakobvits, 1962), and instead shifts the attention to the component sounds of the word itself (Margulis, 2013a). While these two illusions pertain to language they do not carry across to music. A short melody does not descend into musical nonsense on repetition, nor flip between alternate melodies in the same manner, but if anything it becomes stronger and perhaps even more 'musical' than at first (Margulis and Simchy-Gross, 2016).

1.2.3 The *Speech-to-Song* Illusion

While editing her audiobook on musical illusions, Diana Deutsch accidentally looped a short sentence of her voice only to discover that it began to sound as if she was singing the phrase "*sometimes behave so strangely*", when at first she was merely talking naturally. This led her to the unearth an effect that she later named as the *Speech-to-Song* illusion, and was first presented on her CD (Deutsch, 2003). The effect is simple — a short stimulus of a few spoken words intended to be heard as speech is repeated, and after around three or four loops sometimes the phrase is heard as if the speaker is singing. In relation to the other repetition based illusions, this seems to be closer to semantic satiation rather than the VTE, as the effect is stable and lasts a long time[2], but instead of losing meaning, something is actually gained — namely, music. Crucially however, this does not happen for all stimuli, and later studies find that this transformation is stronger for some sounds, and for different participants.

In their first study, Deutsch et al. (2011) asked participants to rate how song-like they perceive the phrase before and after multiple repetitions, and found that most participants agreed that a transformation into song occurs, and this effect is quite dramatic. A few different manipulations to the audio recording was tried — by either altering the transposition of the speech or jumbling the syllables after each repetition destroyed the effect, leading to the conclusion that the illusion requires that each loop must remain intact. Participants were even asked to sing back the song that they heard, and it was found that not only did everyone hear the same melody, but when reproducing the sung phrase they were much more accurate in faithfully imitating the original pitches of

[2] i.e. once the melody is heard, it persists for a while and cannot be 'unheard' (Groenveld et al., 2019)

the recording, compared to a control group who only heard the segment once without repetition.

After the first documented account of the effect by Deutsch a long line of enquiry into details of the nature of the illusion began. An influential study by Tierney et al. (2012) confirmed through fMRI scans that there is heightened activity in areas of the brain associated with pitch processing and song production when participants experience the illusion, showing that speech can really be heard as song. They also found that the set of transforming stimuli tended to have more stable, non-fluctuating F0 pitch contours than the speech that did not illicit the illusion. This result was confirmed in a later set of experiments from the lead author, where stimuli were digitally manipulated to flatten the pitch contours during the syllables, and found these vocals were rated higher on a scale of song-likeness (Tierney et al., 2018a). In this same study, they also found that the melodies contained in the transforming stimuli tended to be similar to those found in Western music, according to some probabilistic model of melody.

It was reported soon after the discovery of the illusion that it exists in multiple languages, such as is German (Falk and Rathcke, 2010) or in tonal languages, such as Mandarin (Zhang, 2010). These results were followed up by Margulis et al. (2015) who hypothesised the effect will be reduced if the listener does not understand the language, or finds it hard to pronounce such that they cannot 'sing-along' in their head. However, they found that this only boosted the effect, leading to the conclusion that by not understanding the semantic meaning of the words the listener focuses on other aspects of the sounds themselves, such as pitch, timbral or rhythmic qualities, and thus can find the music sooner. An experiment by Leung and Zhou (2018) saw that the semantic and emotional content of the spoken words had no bearing on the illsion, and this was taken to the extreme in the experiments of Tierney et al. (2018b), who created sounds that were simple tones which recreated the pitch contours of transforming stimuli were also rated as transforming into music. The effect can be experienced by people of different musical experience (Vanden Bosch der Nederlanden et al., 2015).

While there is ample evidence of pitch playing an important role in eliciting the illusion, the influence of rhythm and meter is less clear. Falk et al. (2014) saw mixed effects — a regular accent distributions only seemed to effect the time before the illusion is experienced, and not the probability of its occurrence. However, in a second experiment they found evidence that durational contrasts of accented and unaccented events is indeed connected to the illusion, suggesting rhythmic meter facilitates a perception of music. On the other hand, Tierney et al. (2018a) reported no change in ratings between control stimuli and those that were manipulated to have more isochronous timings of syllables, perhaps indicating that the role of rhythmic aspects is not as straightforward effect as pitch and other musical ques. It seems clear that the illusion reveals something about how the brain distinguishes speech from song, and although it takes a few repetitions for a listener to perceive it as song, there are some similarities in features between transforming stimuli and recordings of singing, most notably stable notes and meter. As the perception of song is not immediate, it seems as if speech and song are not distinct categories, and that the stimuli which transform lie on the blurred boundary between them.

1.2.4 Beyond Speech and Song

The natural followup of Speech-to-Song is to ask if the effect extends beyond spoken word, and if non-vocal sounds can become musical through repetition as well. It seems reasonable that this can also happen, as incorporating recorded samples of sounds and noises has been utilised by musicians as a musical technique for decades, however such usage has attracted little academic interest. Nonetheless, Simchy-Gross and Margulis (2018) ran the same experiment setup as Deutsch et al's original experiment but replaced the vocal stimuli with environmental sounds instead. These were typically sounds that one would not consider to be musical, however after several reiterations of the sound participants rated them higher on a scale of music-likeness. Recently, Rowland et al. (2019) also confirmed the illusion extends to water dripping sounds. Contrasting with the rest of the speech to song results however, they found that randomly ordering segments of the sound did not break the illusion, suggesting that environmental sounds do not have to played back exactly to illicit the perception of music. This implies the Speech-to-Song illusion a subset of a broader phenomenon of Sound-to-Music, which could even be contained in a larger space of *Sound-to-Something* of sound transformation illusions.

1.3 Thesis Overview

The main goal of this research is to analyse the Speech-to-Song phenomenon in greater detail by taking advantage of the largest collection of audio stimuli and data on the topic to date. Through the use of computational techniques, we analyse the data on a large scale to distil which characteristics seem to quantitatively impact the perception of musicality in a repeated sound, and distinguish how these features differ between those that transform and those that do not. Considering the importance of repetition in music, the illusion offers a unique opportunity to study the role of it from a cognitive perspective, and work towards narrowing down the aspects of audio that get teased out by repetition that causes the perceptual switch can lead to further clues to what makes music *music*.

To begin, in Chapter 2 we outline a method to automatically extract the melody that a listener could perceive with the audio signal that we will use to compute some characteristics on this melody. The incentive to automate the transcription of the song is so that we can analyse many more stimuli in a large scale survey, enabling the use of techniques from data science. Previous studies that looked at music theoretic traits used either hand annotated stimuli, or a rudimentary algorithm to measure the notes of the melody — our automatic method offers a slight improvement over more naive approaches to this task. We then extend this method with a Bayesian approach to search for an similar, more 'musical' melody. With this new melody, we can compare it to the one that was extracted to measure directly how likely it is to be found in a musical composition. This feature, along with many others are outlined in Chapter 3. We describes a series of algorithms to make measurements that we take directly from the audio itself and the note sequences to produce a feature vector that represents the audio stimulus. We include such features that have been discovered in past work, along with a set of new and novel measurements to test other aspects of the melody. It will be from these traits of the sound that we will further analyse and attempt to predict the behaviour of the participants and the probability of the illusion occuring.

Chapter 4 collects the stimuli along with human rating data from a past experiment and prepares it for analysis. This involves aggregating the final scores given by listeners on how strong the illusion materialised, filtering the data to obtain higher quality results, and devising a scheme to apply a binary label (*transforming* or *non-transforming*) for classification models to predict. We then test if there exists any direct correlations between feature and transformation scores, and measure if there exists any significant differences in features of the transforming and non-transforming stimuli. We fit several models to the data in Chapter 5, ranging from simple linear models, to non-linear kernel based methods, and then to ensembles of models, and evaluate their performance at predicting the labels. If the models have success, then we know that there is some information contained in the feature vector that facilitates the prediction that gives us a clue in what contributes to the effect.

In Chapter 6 we report an experimental setup that we conducted with a whole new set of stimuli to collect fresh empirical data, and evaluate how effective the models predict these new sounds. The stimuli is a diverse collection of spoken word compiled from a range of speaking styles, and in three different languages. We also include non-speech sounds to assess how well the models generalised to these novel sounds, despite the algorithms having been optimised for vocals. This data is analysed and compared to that of the previous experiment to confirm if the trends hold in this new data. Finally, a summary and discussion of all the methods, results and main conclusions can be found in Chapter 7.

Chapter 2

Melody Extraction

I haven't understood a bar of music in my life,
but I have felt it.
— Igor Stravinsky

2.1 Audio Analysis Methods

For the automatic and computational approach to music analysis we turn to the field of music information retrieval (MIR). This area of research combines methods and theories from signal processing, informatics, psychoacoustics, musicology, and machine learning to develop algorithms to accomplish many tasks that are of interest to researchers, commercial entities and consumers of music. Typical uses for such algorithms include music classification (Fu et al., 2011), harmonic and tonal analysis (Ni et al., 2012), genre detection(Li et al., 2003), track identification (Mohri et al., 2010), and recommendation systems (Rosa et al., 2015) to name just a few. We are interested here in the task of automatic music transcription (AMT), whereby an annotated musical score is generated from a raw audio signal that represents the melody of the music.

This is a vary active area of research with many applications, made apparent by the sheer number of articles published on this, and the availability of software and services on the market that accomplish this task (Chordify, Melody Scanner, ScoreCloud and Tony are just some of many). They utilise a range of algorithms and techniques, but mainly rely on first extracting fundamental frequencies of the notes and sounds, then some timing information of musical 'events', and finally some post processing to produce the final transcription (for an overview of AMT see Benetos et al., 2013) Most algorithms are optimised for instrumented music rather than the human voice, by assuming pitch is strongly present in the audio signal and that note pitches and timing fall on some grid (in a piano roll style representation). Even in the cases of transcribing sung melodies by inexperienced singers (for example in 'query-by-humming', see Ghias et al., 1995, Haus and Pollastri, 2001), it is assumed that the singer is actively attempting to produce a salient, reasonably accurate and stable melody. This is not the case in natural speech where there is no intention from the speaker to follow a melodic line and so pitch loosely fluctuates (which in the case of tonal languages provides additional semantic information). Therefore, to extract a melody from these audio clips requires adapting the methods and assumptions, and designing heuristics to identify a possible melody that could be perceived by a listener.

The field of phonetic research provides useful tool kits for analysing speech computationally. As we are interested in the melody within vocal stimuli in our study into the Speech-to-Song illusion, identifying the syllables that make up the rhythm and notes is a well researched area with established methods to accomplish these tasks in vocal analysis. Praat (Boersma and Weenink, 2012) is an open-source software package the contains a vast array of algorithms and measurements to analyse (and manipulate) all facets of speech, and has become an industry standard within the community. Particularly, we make use of two algorithms from this tool kit — the first to extract a pitch contour of the fundamental frequency (also called F0) over time, and the second to get the intensity of the audio, also over time.

2.2 The Melody Extraction Algorithm

In order to make correlations and inferences on the melody contained within the audio sample it is necessary to extract from the source an accurate set of notes, with their start times, lengths, and pitch values. With this information, measurements about the harmonic qualities of the melody can be made, alongside rhythmic information and pitch salience. However, while the data set on hand is modest in size, manually transcribing the melody information for each audio sample would be unfeasible. It is important to obtain an objective measurement of the melody, as not all listeners necessary perceive the same notes, so an automatic method is less subjective. Moreover, as the goal is to produce an algorithm that can identify new material from a large set of potential stimuli, an automatic process to accomplish this task is desirable. As the material of the study is entirely on the Speech-to-Song illusion, we will assume that the source sound is that of human speech, and not the more general goal of identifying melody from any sound source. This means taking vocal features as indicators of melody and rhythm, and optimising parameters to best fit these, and to use tool kits optimised for vocals.

The foundation of the process described here is based on the thorough (unpublished) work of Cornelissen (2015), who attempted to solve the problem of melody extraction in normal speech for studying the Speech-to-Song illusion. The algorithm first detects when the notes occur, then finds the melody that best fits given the pitch contour. The following presented here builds upon his method with some modifications and improvements.

2.2.1 Problem Statement and Definitions

An acoustic stimulus is divided into n discrete time steps, such that the time between steps is sufficiently short. Let $\boldsymbol{p} = \{p_1, p_2, \ldots, p_n\}$ be the F0 pitch contour in Hertz obtained from Praat[1] of the sample at each time step, where $p_i = 0$ when there is no pitch information (e.g. during moments of silence or noise where there is no F0 frequency). Let $\boldsymbol{I} = \{I_1, I_2, \ldots, I_n\}$ denote the measured intensity values of the signal (measures in dB relative to $2 \cdot 10^5$ Pascal) over the same time steps. An example of $\boldsymbol{p}$ and $\boldsymbol{I}$ are plotted for a stimulus in the central plot in Figure 2.1. It is the task of the algorithm to compute the melody from these two vectors of information alone.

Let $\boldsymbol{t} = \{t_1, t_2, \ldots, t_N\}$ be a sequence of N individual note values (as its fundamental frequency in Hertz) that make up the complete melody that we are attempting to extract. The start (onset) times of these notes forms the set $\boldsymbol{o} = \{o_1, o_2, \ldots, o_N\}$, and their last time steps $\boldsymbol{l} = \{l_1, l_2, \ldots, l_N\}$. Therefore, note t_i starts at time step o_i and ends at time step l_i. Grouping these sets into the tuple $\boldsymbol{M} = \langle \boldsymbol{o}, \boldsymbol{l}, \boldsymbol{t} \rangle$ forms a representation of the extracted melody. Furthermore, the set $\boldsymbol{s}_i = \{p_{o_i}, p_{o_i+1}, \ldots, p_{l_i}\} \subseteq \boldsymbol{p}$ contains all the pitches that make up the note t_i (i.e. the F0 contour during note i), and these are collected of every tone to form the set $\boldsymbol{S}$.

Therefore there exists two unknown functions — $\Lambda : \boldsymbol{I}, \boldsymbol{p} \mapsto \boldsymbol{o}, \boldsymbol{l}$ that segments and identifies when notes are detected in the audio sample, and $\Gamma : \boldsymbol{s}_i \mapsto t_i$ that gives the perceived tone t_i of note $\boldsymbol{s}_i$ from the pitch contour within it. The goal of this section is to estimate these functions.

2.2.2 Note Segmentation

The first step to extracting the melody is determining where the notes occur within the audio signal. While note boundaries are often easily distinguished in human listening, automatically identifying these from the audio signal alone is not trivial — for example simply segmenting the notes by unpitched time steps (i.e. where $\boldsymbol{p} = 0$) is not sufficient,

[1] We use the default settings for this algorithm, and for all other Praat measurements.

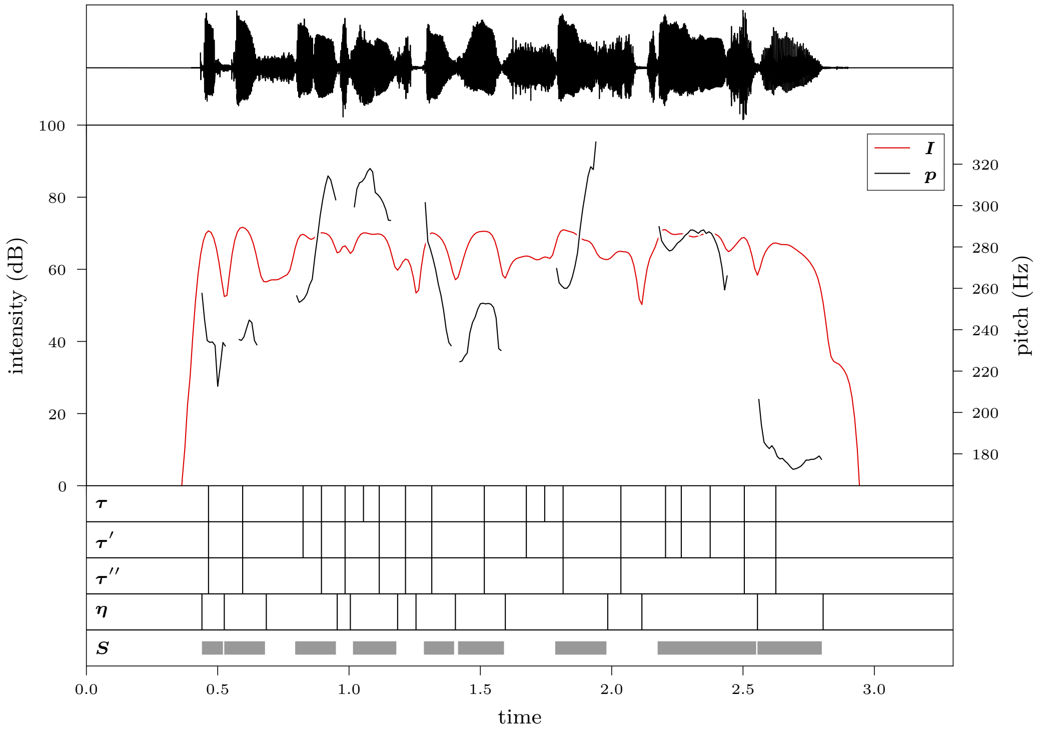


Figure 2.1: The raw audio signal (top) is reduced to intensity and pitch information (middle plot). The lower section shows the various steps of the note segmentation algorithm Λ': from identifying the peaks, two stages of filtering, deducing the note boundaries and finally the segmented note regions. These blocks correspond to the nine syllables of the phrase "*but they some-times be-have so strange-ly*"

as there may be multiple notes over a continuous pitch contour (such as glide from one stable note to another).

Typically in practice this task is accomplished by one of two methods: either segmenting by amplitude information or by pitch information (McNab et al., 1995). The former is often simpler to implement, but can fail if the notes are not acoustically isolated (i.e. when there are no brief dips in intensity between note events), while the later works well if the pitch contour is stable and not fluctuating during the note duration (e.g. in the presence of vibrato). More recent transcription algorithms for singing use hidden Markov models (e.g. Ryynänen and Klapuri, 2006), where note events are identified from multiple features, including dynamics of the F0 pitch curve, onset strengths, the detected vocal accents and salience (the prominence of the F0 pitch). As our data deals mostly with speech rather than singing or instrumentation, we take an approach that is less susceptible to unstable pitch contours.

To determine when notes are voiced within speech, it is important to recognise which aspect of the vocals align to the notes perceived such that these points can be identified automatically. The most salient part of a word is the *vocalic* part, specifically where the nucleus of the syllable occurs. This is typically (but not always) where the vowel of the syllable lies, and hence to segment the notes is to segment the nuclei of the speech source. De Jong and Wempe (2009) outline a method to make this segmentation, however Cornelissen improves on this algorithm and reports success on the identification of note boundaries. Therefore we use this, as outlined below.

The algorithm will estimate the unknown function Λ, that we denote as $\hat{\Lambda}$, and has four hyper-parameters: `maxDip`, `minDipBefore`, `minDipAfter`, and `threshold`. It is these hyper-parameters that are optimised in the parameterisation stage in Section 2.2.4. First, the time steps $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_m\}$ of all local maxima of $\boldsymbol{I}$ are found. Then, each peak at time step τ_i are removed from $\boldsymbol{\tau}$ to form $\boldsymbol{\tau}'$ if any of the following conditions are met:

$$\begin{aligned} I_{\tau_i} &< \texttt{threshold} \\ \forall \tau : \tau_i - \varepsilon \leq \tau < \tau_i + \varepsilon, \quad p_\tau &= 0 \\ |I_{\tau_i} - \min(\{I_\tau : \tau_i \leq \tau < \tau_{i+1}\})| &< \texttt{maxDip} \end{aligned}$$

The first condition checks that the peak is not too quiet, and the second condition ensures the pitch is voiced by checking if there is some pitch information available around the peak for some small margin ε[2]. The third condition checks that the peak is prominent enough relative with the next peak[3], with this accounting for slight fluctuations in the intensity during the nuclei.

This reduced set $\boldsymbol{\tau}'$ of candidate peaks is filtered once more ($\boldsymbol{\tau}''$) to find the most significant peaks by the following two conditions:

$$\begin{aligned} |I_{\tau_i} - \min(\{I_\tau : \tau_{i-1} \leq \tau < \tau_i\})| &< \texttt{minDipBefore} \\ |I_{\tau_i} - \min(\{I_\tau : \tau_i \leq \tau < \tau_{i+1}\})| &< \texttt{minDipAfter}. \end{aligned}$$

These conditions are similar to before, where peaks of low prominence compared with neighbouring peaks are removed[4]. Two stages of filtering avoids the algorithm being too 'greedy' that might cause some of the peaks to be discarded too early in the procedure.

Since $\boldsymbol{\tau}''$ are the time steps to the peaks of the syllables, it remains to compute the actual starting (and end) points of the syllable, which will become the final note boundaries. This is a simple process of finding the local minimum of $\boldsymbol{I}$ between two sequential peaks in $\boldsymbol{\tau}''$. For the very first peak, it is assumed the syllable starts when $\boldsymbol{I}$ is first greater

[2] If this pitch information actually belongs to a neighbouring syllable, this will be realised in later steps of the algorithm

[3] In the edge case where $i = |\boldsymbol{\tau}'|$ and thus there is no next peak, this condition is not tested.

[4] Again, in the case of no previous or subsequent peak (when $i = 1$ or $i = |\boldsymbol{\tau}''|$), then only the valid condition is checked.

than 50, and similarly for the final peak the note is assumed to end when $\boldsymbol{I}$ is less than 50. With this, we have a set of note boundaries $\boldsymbol{\eta} = \{\eta_1, \eta_2, \ldots, \eta_m\}$.

Finally, $\boldsymbol{S}$ is formed by collecting $\boldsymbol{s}_i$ for each pair of note boundaries η_i, η_{i+1} and taking all the values of $\boldsymbol{p}$ that fall between these time step boundaries. If $\boldsymbol{s}_i = \emptyset$ then it is excluded from $\boldsymbol{S}$. This concludes the note segmentation algorithm.

2.2.3 Note Extraction

The exact nature of function Γ that maps the frequencies to tones is unknown, as pitch perception is an ongoing endeavour of psycho-acoustic research (e.g. Jacoby et al., 2019). While spectral decomposition to find the fundamental frequency of a sound source is standard, the actual perceived tone is far from trivial: harmonics, overtones, timbre, salience and loudness are just some of the aspects that can affect the identified pitch. There are even individual bias at play, and that two different listeners may even disagree on which exact pitch they hear. Nonetheless, we attempt to estimate this process with a rule-based system that outperforms a naive baseline, and improves on previous work of Cornelissen (2015).

For each note $\boldsymbol{s}_i$ we wish to extract a candidate tone t_i that will be perceived from the pitches. As these pitches are almost never completely stable, the function $\hat{\Gamma}$ must handle all possible pitch changes and fluctuations.

A reasonable baseline, and the function used by Cornelissen, is to simply take the mean of the pitch during the note, i.e. $\hat{t}_i = \bar{\boldsymbol{s}}_i$. This has the advantage of mitigating the effects of vibrato or other fluctuations around the main pitch, however, it is susceptible to pitch transitions when the pitch contour changes from one note to the next which pulls the mean pitch away from the perceived value. An example of this can be seen in Figure 2.1, where the second last note starts with a stable pitch but and drops quickly at the end as it transitions to the final note, and so the mean would under-estimate the pitch. In some cases, the pitch of the note is constantly climbing up or down throughout the time of the note (see the seventh note in the same example), but the transcription of such a note is typically not in the centre where the mean would estimate it to be. This can be improved somewhat by simply removing the end points of $\boldsymbol{s}_i$, or by taking the mean of the the most stable regions by estimating the first derivative. This later strategy is still far from ideal — in cases where there are two stable regions with a large jump between them, this would still guess a note in between the two correct pitches. In such a case, it would be prudent to distinguish this as two separate notes. Figure 2.2 illustrates a typical stimulus with its transcription, along with the notes obtained by taking the mean of the stable regions of $\boldsymbol{s}_i$, and shows that a more robust method is required to obtain a satisfactory representation of the melody.

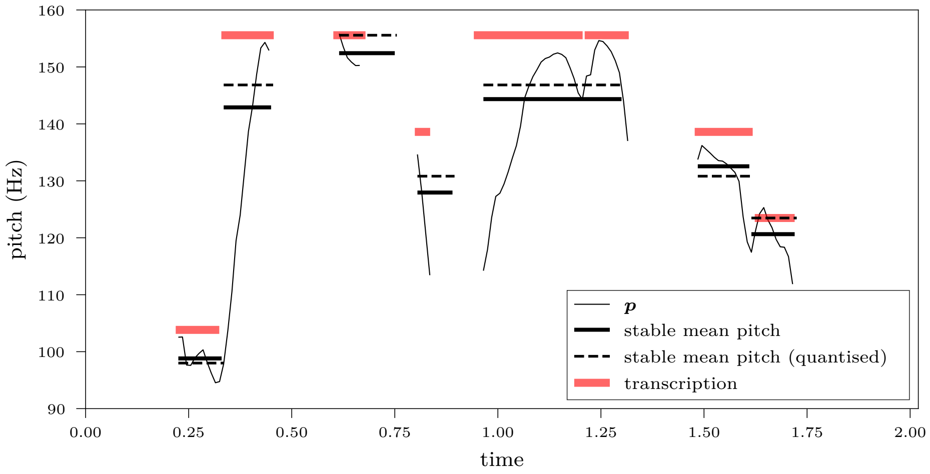


FIGURE 2.2: Comparison of the transcription (red lines) versus the stable mean pitch note extraction method (black lines). The note at 0.35s shows how this approach fails in cases where the pitch contour transitions from one note to another.

The proposed algorithm requires three hyperparameters: `minLength`, `unstable` and `maxUnstable` that are also parameterised in the process described in Section 2.2.4. For each $\boldsymbol{s}_i$ in $\boldsymbol{S}$, the following process is carried out to find $\hat{t}_i$:

Let $\boldsymbol{s}'_i = \{|p_j - p_{j+1}| : p_j, p_{j+1} \in \boldsymbol{s}_i\}$, i.e. the absolute first difference of the pitch values of the note $\boldsymbol{s}_i$.

1. if $\ell(\boldsymbol{s}_i) <$ `minLength`, (where $\ell(\cdot)$ is the length of the note in seconds), then no note is extracted
2. if $\bar{\boldsymbol{s}}'_i >$ `maxUnstable`, then $\hat{t}_i$ is the mean of the final half of $\boldsymbol{s}_i$
3. otherwise $\hat{t}_i$ is the mean of the set of pitches during the most stable portion of $\boldsymbol{s}_i$, which is the largest continuous subset of $\boldsymbol{s}'$ such that each element of this subset is less than `unstable`.

In other words, if the note is very unstable, take the mean of the final portion of the note's pitches, otherwise take the mean of the largest, flattest part of the note's pitch contour. The second case will estimate the pitch from the final part of the note's pitch contour, as it seems the case that transcriptions of glissando notes tend to put the note value at where the contour ends.

This algorithm $\hat{\Gamma} : s_i \mapsto \hat{t}_i$ in combination with $\hat{\Lambda}$ completes the melody extraction process. We continue by parameterising these two processes to get the most reliable algorithms.

2.2.4 Parameterisation and Evaluation

Across the two algorithms, there are seven free variables that need to be found such that the algorithms match as close as possible with human perception. Optimising the parameters involves minimising a loss function that measures a notion of 'distance' of the extracted melody of a stimulus from the ground truth transcription that was labelled by a human listener.

For evaluating the extraction algorithm, we have a modest collection of transcriptions of some number of stimuli. These transcriptions were made by Cornelissen in his essay to evaluate the segmentation and note values extracted with his implementation, and so will be used here too. There are a total of 50 annotated stimuli where the emergent melody is notated by start and end times of perceived notes, along with the corresponding pitch value. These stimuli are from a larger collection of 300 speech samples used in a previous experiment on the Speech-to-Song illusion (described later in Section 4.1), and most of them are speech samples that transform. Melodies were transcribed to be played on a keyboard in 12 tone equal temperament tuning (with reference pitch A_4=440Hz), so for a fair evaluation the notes extracted from the algorithms should be quantised to their nearest note value on such a tuning system. Quantisation is also required in Section 2.3, so we detail the conversion now.

To quantise a tone t (measured in Hertz) to its nearest note, first it is converted to its MIDI value m with (2.1), then m is rounded to the nearest integer, then converted back to frequency f with (2.2):

$$m = 69 + 12 \cdot \log_2 \frac{t}{440\text{Hz}} \tag{2.1}$$

$$t = 2^{(m-69)/12} \cdot 440\text{Hz}. \tag{2.2}$$

It should be made clear this quantisation step only affects the note value, and not the segmentation (timing) of the notes. Temporal quantisation where onset times and lengths are snapped to a discrete grid is a technique used in automatic music transcription to iron out expressive timing in a musical performance — as the transcriptions are not assumed to align to a musical metre, this is not an issue for our purposes.

As acknowledged earlier, a weakness of the pitch measurement can introduce octave errors within the melody, and where possible we try to avoid any measurements that are sensitive to these inaccuracies. This is the case for the loss function — the algorithm is not punished for being a whole octave away, since if the pitch extraction had been accurate to the human transcription it would have been correct.

Fortunately, a rather simple transformation of the note values can project them into a space where this is not a problem. Since any octave of a note with frequency f (in Hertz) is of the family of frequencies $f \cdot 2^k, k \in \mathbb{Z}$, by taking the logarithm with base 2 then octaves of the original frequency can be identified as being an integer distance away in this transformed space[5]. That is, frequencies f_1 and f_2 are octaves of each other if and only if $\log_2(f_1) = \log_2(f_2) + k$ for any $k \in \mathbb{Z}$. By taking modulo 1 on both sides simplifies the condition further: $\log_2(f_1) \bmod 1 = \log_2(f_2) \bmod 1$ if and only if f_1 and f_2 are octaves of each other.

In this space, distance can be formulated by the following. Let

$$\delta = \Big| \log_2(f_1) \bmod 1 - \log_2(f_2) \bmod 1 \Big|,$$

then:

$$d(f_1, f_2) = \begin{cases} \delta & \text{if } \delta \leq 0.5 \\ 1 - \delta & \text{if } \delta > 0.5. \end{cases} \tag{2.3}$$

Although quite technical in its formulation, this distance metric[6] has a simpler, geometric interpretation: the values of $(\log_2(f_1) \bmod 1)$ and $(\log_2(f_2) \bmod 1)$ are points around a circle with unit circumference, and so $d(f_1, f_2)$ is the shortest distance along the circumference between the points. Therefore, the furthest two points can be is 0.5, (i.e. half an octave).

Let $\boldsymbol{M} = \langle \boldsymbol{o}, \boldsymbol{l}, \boldsymbol{t} \rangle$ be the transcribed melody (with onset time, notes end times and note value), and $\hat{\boldsymbol{M}} = \langle \hat{\boldsymbol{o}}, \hat{\boldsymbol{l}}, \hat{\boldsymbol{t}} \rangle$ be the estimated melody (the number of notes in both might not be the same), then the loss function $\mathcal{L}(\boldsymbol{M}, \hat{\boldsymbol{M}})$ that evaluates the extraction be as follows: Initiate `loss` as zero, then for each time step $1 \leq i \leq n$ where the transcription $\boldsymbol{M}$ indicates there is a note t_j, if $\hat{\boldsymbol{M}}$ also indicates a note $\hat{t}_k$, then `loss` is incremented by their distance $2 \times d(t_j, \hat{t}_k)$. If $\hat{\boldsymbol{M}}$ does not indicate a note, then `loss` is incremented by 1. Finally, `loss` is normalised by dividing it by total number of time steps where $\boldsymbol{M}$ indicates a note as being sounded, such that `loss` $\in [0, 1]$

It should be noted that the loss function is not increased when the algorithm makes a false positive prediction at a time step, although this can certainly be implemented

[5] This makes the unison interval (i.e. when two notes the same) to be the ‘zero-th’ octave.

[6] It is easy to see that d satisfies the conditions to be a proper distance metric: $d(x, y) \geq 0$, $d(x, y) = 0$ if and only if $x = y$, $d(x, y) = d(y, x)$, and $d(x, z) \leq d(x, y) + d(y, z)$ (the triangle inequality)

to make a stricter loss. However, from observations of the predictions it was noticed that the algorithm tended to over-estimate the length of a note compared with the transcriptions (this is also demonstrated by the third and fourth note in Figure 2.2). This could be for a couple of reasons, such as the quality of the transcription (more focus on getting the note onset time accurate and less emphasis on its length), and the inherent weakness of the segmentation algorithm.

With a loss function defined, the procedure to evaluate the parameters is straight forward: for any set of parameter values predict the melodies of all 50 stimuli for which there exists a transcription, compute the loss for each one and average the losses to obtain a final score from 0 to 1, with 0 being the perfect loss. Using this score we can compare different parameter sets and thus find the optimal values that minimises the loss.

Since there are only seven parameters and evaluation is quick, a simple grid search that steps through all the combinations of values over their ranges is performed and the performance is evaluated with this loss function. Initially the grid search was quite coarse over a large range of values to estimate a ballpark set of values, then progressively finer grained searches narrows in on the best possible minimum. This procedure yielded the parameters is Table 2.1. For notes segmentation, it appears that imposing a minimum intensity threshold reduces the effectiveness of the algorithm (indicated by `threshold` $=$ 0), and that the small value for `maxdip` suggests only the weakest peaks should be dropped.

parameter	value	description
`maxDip`	0.5	intensity peak local prominence
`minDipBefore`	2.1	
`minDipAfter`	0.5	
`threshold`	0.0	minimum intensity peak
`minLength`	0.05	minimum length (in seconds) of a note
`unstable`	4.5	stability thresholds on deciding which part of the note's pitch contour to use when taking mean
`maxUnstable`	8.8	

Table 2.1: Optimal values found for the parameters used by the note extraction algorithms $\hat{\Lambda}$ and $\hat{\Gamma}$.

Table 2.2 summarises the losses for a naive baseline that predicts a single tone equal to $\bar{\boldsymbol{p}}$ for the complete length of the stimuli, using the mean pitch (where $t_i = \bar{\boldsymbol{s}}_i$), the stable mean method (where the mean of only the most stable portion of the note's pitch contour is used), and the extracted melody using the full algorithm $\hat{\Gamma}$, both before and after quantising the notes to MIDI values.

method	loss	
	normal	quantised
naive baseline	0.3876	—
mean (all of $\boldsymbol{s}_i$)	0.2464	0.2385
mean (stable regions of $\boldsymbol{s}_i$)	0.2382	0.2306
rule based	0.2250	0.2145

Table 2.2: Evaluation of note extraction methods, before and after quantisation. Lower loss indicates better performance.

First, it can be seen that the stable mean pitch method is already an improvement over the basic mean method. The current method performs significantly better than the naive baseline, and modestly outperforms either of the mean pitch methods. While at first it seems unsurprising that quantising the notes matches the transcriptions better, this does suggest also that the unquantised notes values were initially sufficiently close to the ground-truth values — a weaker algorithm would have a 50% chance that the rounding would shift the note closer or further from the true note.

We have outlined and parameterised the two algorithms which together recover a melody from the spoken word. Cornelissen then takes a next step to use a Bayesian approach to find the most likely *musical* note sequence, given melody we have just extracted. However, we separate this extra step and use this alternate, Bayesian melody as a way of objectively comparing how musical the 'raw' melody we extracted is, and use this later (Section 3.2) as a possible feature in identifying illusionary stimuli.

2.3 Bayesian Melody Search

Once we have a sequence of notes, the natural question to arise in a study of music perception is simply 'how musical is this melody?' Such a question is perhaps ill-defined in an objective investigation — definitions of musicality are very subjective and prone to critiques ranging from cultural and historical factors to more individualistic and personal aspects. This has long been a challenge in musicology, where any assumption of the universalities of music are often met with criticism, so researchers must tread very carefully when making any such claim.

One perspective to take the musicality question is in a comparative sense, where we can ask a similar, more quantative question of 'how is this melody *like* other typical melodies?' This calls for a statistical approach that involves developing a model that capture aspects of musical phrases that we can apply probability theory to, such that we can measure how likely a sequence of notes might appear in some music theoretic framework. Rather, we compare the extracted melody to one that is more typical to obtain a distance of how close the melody is to a musical one. This method is the one described by Cornelissen (2015), however we contribute a set of parameters for the model and outline a stronger search method.

2.3.1 Method Outline

The tone sequence $\boldsymbol{t}$ given the set of syllable frequencies $\boldsymbol{S}$ can be modelled naturally by taking a Bayesian approach:

$$P(\boldsymbol{t}|\boldsymbol{S}) = \frac{P(\boldsymbol{S}|\boldsymbol{t}) \cdot P(\boldsymbol{t})}{P(\boldsymbol{S})} \propto P(\boldsymbol{S}|\boldsymbol{t}) \cdot P(\boldsymbol{t}). \tag{2.4}$$

As we are trying to find the most likely tone sequence $\boldsymbol{t}$ given the syllable information $\boldsymbol{S}$, we maximise (2.4), in other words the maximum a posteriori estimate $\hat{\boldsymbol{t}}_{\text{MAP}}$ is given by

$$\hat{\boldsymbol{t}}_{\text{MAP}} := \arg\max_{\boldsymbol{t}} \Big(P(\boldsymbol{S}|\boldsymbol{t}) \cdot P(\boldsymbol{t}) \Big). \tag{2.5}$$

The first part of the right hand side of (2.4) is the *likelihood* — the probability of observing the noisy syllables $\boldsymbol{S}$ for a given tone sequence. A simple model assumes each frequency $p_j \in \boldsymbol{s}_i$ is (independently) drawn randomly from a normal distribution centred around t_i with some precision β. Therefore, $P(\boldsymbol{S}|\boldsymbol{t})$ can be formulated as

$$P(\boldsymbol{S}|\boldsymbol{t}) = \prod_{i=1}^{N} \prod_{f \in \boldsymbol{s}_i} \mathcal{N}(f|t_i, \beta^{-1}) \tag{2.6}$$

where $\mathcal{N}(x|\mu, \sigma^2)$ is the probability density function of the normal distribution with mean μ and standard deviation σ at x. The independence assumption in the likelihood is not ideal — clearly the frequency sequence is highly dependent on the previous time steps, and so could better be modelled with some Markov Chain or Random Walk. However, for this purpose, the simplification is sufficient.

2.3.2 Distribution of Melodies

The second part of (2.4) denotes the *prior* probability of $\boldsymbol{t}$, that is the probability of observing $\boldsymbol{t}$ from the distribution of *all* tone sequences. Of course, such a distribution over all melodies is unknown, however there exists several statistical models of music that attempt to capture this. For starters and to set a naive baseline, we could assume that $P(\boldsymbol{t})$ is uniform, and so the posterior simplifies to $P(\boldsymbol{t}|\boldsymbol{S}) \propto P(\boldsymbol{S}|\boldsymbol{t})$ and thus $\hat{\boldsymbol{t}}_{\text{MAP}}$

would involve maximising the likelihood function in (2.6). This results in

$$\hat{\boldsymbol{t}}_{\text{MAP}} = \{\bar{\boldsymbol{s}}_i : 1 \leq i \leq N\}, \tag{2.7}$$

which estimates the tones as the mean of the pitches during the syllable.

A more sophisticated probabilistic model of music is proposed in Temperley (2008) in his study of melody perception. The model makes basic assumptions on how notes are distributed given some previous context and key signature, and is able to capture much of the structure of the music that he fitted the model to, being able to predict the notes based on previous contexts and identify key signatures. In his paper, he applied the model to music from the Essen Folk Song collection (Schaffrath, 1995) to analyse typical melodic patterns in Western folk music. It works as follows:

1. First a central pitch c is drawn from a Normal distribution $\mathcal{N}(\mu_c, \sigma_c^2)$. This is somewhat like the 'tonic', but is generally the mean pitch over the entire phrase.
2. Each note is drawn from a range centred on c, with shorter intervals being more probable. This is modelled with a *range* distribution $\mathcal{N}(c, \sigma_r^2)$.
3. Each note t_i (for $i > 1$) are also constrained by it's *proximity* to the previous note t_{i-1}, again with larger intervals being less probable. This is another normal distribution $\mathcal{N}(t_{i-1}, \sigma_p^2)$.
4. Finally, the probability of a note is given by one of the 24 key profiles k, where notes outside the key will be less probable.

Formally, these conditions can be combined as follows:

$$t_1 \sim \text{RK}(c, k) \propto \mathcal{N}(c, \sigma_r^2) \cdot \mathcal{K}(k) \tag{2.8}$$

$$t_i \sim \text{RPK}(t_{i-1}, c, k) \propto \mathcal{N}(c, \sigma_r^2) \cdot \mathcal{N}(t_{i-1}, \sigma_p^2) \cdot \mathcal{K}(k) \tag{2.9}$$

where $\mathcal{K}(k)$ is the probability distribution of key profiles. Therefore, the joint probability distribution of the tone sequence $\boldsymbol{t}$ is obtained by marginalising over keys k and tonal centers c:

$$P(\boldsymbol{t}|c, k) = P(t_1|c, k) \cdot \prod_{i=2}^{N} P(t_i|t_{i-1}, c, k) \tag{2.10}$$

$$\Rightarrow \quad P(\boldsymbol{t}) = \sum_k \int_c P(c) \cdot P(k) \cdot P(\boldsymbol{t}|c, k).dc. \tag{2.11}$$

While this model is rather simple, it remains computationally intractable, mostly due to the itergral over c. However, since (2.11) is used in the maximisation of (2.5), we do not need to search over all c — it would be sufficient to fix c to some sensible value based on $\boldsymbol{S}$. The natural choice is to approximate μ_c by taking the mean of all (non-zero) pitches in $\boldsymbol{p}$, i.e.

$$\hat{c} = \bar{\boldsymbol{p}}, \tag{2.12}$$

as the maximising value would be very close to this mean. Therefore, (2.11) reduces to

$$\begin{aligned} P(\boldsymbol{t}) &= \sum_k P(k) \cdot P(\boldsymbol{t}|k) && (2.13) \\ &= \sum_k P(k) \cdot \prod_{i=1}^{N} P(t_i|k) \cdot \prod_{i=1}^{N} \mathcal{N}(t_i|\hat{c}, \sigma_r^2) \cdot \prod_{i=2}^{N} \mathcal{N}(t_i|t_{i-1}, \sigma_p^2). && (2.14) \end{aligned}$$

All that remains is to estimate the variances σ_r^2 and σ_p^2, and the choice of key probabilities and the distribution of notes under this key. The variances can be computed rather easily from the transcriptions themselves using an unbiased estimator. We find these values to be $\hat{\sigma}_r^2 = 20.61$ and $\hat{\sigma}_p^2 = 21.15$. These differ from the values that Temperley himself found in his corpus, although generally they agree (he estimated range variance between 17.0 and 29.0 and proximity variance between 7.2 and 70.0, depending on which dataset and estimatation technique he used).

2.3.3 Note and Key Calculation

Here we define the probabilities a note falls in a certain key, and the probability of such a key occurring. Formally, a key is a discrete collection of pitch classes that forms the basis of a composition, where choice of chords and melodies are typically restricted to the notes that are within the key, with some notes having more 'importance' than others — essentially how well the notes fit in the context of the key. In theory a key can be any collection of notes, but we restrict ourselves to the main two keys in Western music, namely the major and minor keys.

The most typical formulations of $P(t_i|k)$ in statistical models of music are defined using key profiles collected by Krumhansl and Kessler (1982), a pioneering study that discerned how well all 12 of the chromatic notes fit within a key, based on human listening trials. Temperley however recomputes these profiles from the Essen Folk Song collection, thus yielding slightly different probabilities, and this is what we use here. We must map the tones $\boldsymbol{t}$ to the 12 chromatic notes by quantising them using the method described in Section 2.2.4. This limits the possible tone sequences to those that can be played on a piano keyboard.

We denote a key k by the tuple $\langle q, r \rangle$, where q is the quality of the key (either major or minor), and r be the root of the key as a pitch class from 0 to 11, where C $\mapsto$ 0,

p_c	0	1	2	3	4	5
name	C	C♯/D♭	D	D♯/E♭	E	F
$P(p_c \mid \langle\text{major}, 0\rangle)$	**0.184**	0.001	**0.155**	0.003	**0.191**	**0.109**
$P(p_c \mid \langle\text{minor}, 0\rangle)$	**0.192**	0.005	**0.149**	**0.179**	0.002	**0.144**
p_c	6	7	8	9	10	11
name	F♯/G♭	G	G♯/A♭	A	A♯/B♭	B
$P(p_c \mid \langle\text{major}, 0\rangle)$	0.005	**0.214**	0.001	**0.078**	0.004	**0.055**
$P(p_c \mid \langle\text{minor}, 0\rangle)$	0.002	**0.201**	**0.038**	0.012	**0.053**	0.022

Table 2.3: Key profiles of the twelve pitch classes (and their names when in C tonic, i.e. $r = 0$) for major and natural minor keys. Bold face where pitch class belongs to the key.

C♯/D♭ $\mapsto 1$, D $\mapsto 2$ etc[7]. Therefore the set of all keys is then

$$\boldsymbol{k} = \{\langle q, r\rangle \;:\; q \in \{\text{major}, \text{minor}\},\; r \in \{0, 1, \ldots, 11\}\}. \tag{2.15}$$

To label a note and compute its position within a key, it is helpful to convert the tone t (with frequency in Hertz) of the sound into a *pitch space*, where changes of one semitone correspond to a difference of 1 in the pitch space (and thus the octave is repeated every 12 steps): This corresponds to the standard MIDI labelling of notes, which assume twelve tone equal temperament tuning with A = 440Hz, and with middle C centred on the value $n = 60$. The conversion of frequency to MIDI note was defined in (2.1). *Pitch class space* contain values in the interval $[0, 12)$, with 0 being the tonic of the key. To convert the MIDI note m to pitch class p_c of key $\langle q, r\rangle$:

$$p_c \equiv (m - r) \bmod 12 \tag{2.16}$$

A similar conversion was made in Section 2.2.4 to get the note distance, however in that case a step of 1 corresponds to a whole octave jump, whereas here a step of 1 corresponds to a semi-tone (of which there are 12 in an octave).

From here, the probability of a pitch class given a key is then a simple lookup of the probabilities derived from the Krumhansl-Kessler values, given in Table 2.3. With $P(t_i|k)$ defined, all that remains is the probability of a particular key $P(k)$. In his model, Temperley estimated

$$P(\langle q, r\rangle) = \begin{cases} \frac{0.88}{12} & \text{if } q = \text{Maj} \\ \frac{0.12}{12} & \text{if } q = \text{Min.} \end{cases}$$

This concludes Temperley's model of melody.

[7] We assume for simplicity that pitches with two possible name (e.g. G♯ and A♭) are the same, even though in music theory they have different functions.

2.3.4 Search Space

With the components that make the posterior $P(\boldsymbol{t})$ and $P(\boldsymbol{t}|\boldsymbol{S})$ defined, we can begin the search of the optimal tone sequence $\hat{\boldsymbol{t}}_{\text{MAP}}$. Of course, checking all possible sequences is intractable, but by initialising the search at a reasonable guess and searching the space in the local neighbourhood of similar melodies will likely find the global maximum. One option is to start by maximising the likelihood using (2.7), however we make the natural choice of initialising the search with the tone sequence already extracted by the algorithm:

$$\boldsymbol{t}_{\text{INIT}} = \{\hat{\Gamma}(\boldsymbol{s}_i) : 1 \leq i \leq N\}. \tag{2.17}$$

The search for similar sequences involves altering the melody note by note and computing the posterior for each transformation, and taking the sequence that maximised (2.4). This is done by exploring all possible combinations of transposing notes up or down by a number of semitones and computing the posterior of this new melody. Limiting the number of semitones we transpose by to c, for each note there are $2 \cdot c + 1$ possible values (including no transposition), and given a sequence length of N we have a total of $(2 \cdot c + 1)^N$ potential melodies that need to be checked. It is easy to see however that even for a modest length sequence the total number of sequences can grow very fast due to the exponentiation — for example, with an 8 note sequence and transposing notes up to a distance of 3 semitones results in nearly 5.8 million combinations that need to be checked. Even on reasonable hardware this is too many computations to be practical.

There are some exploration strategies that can be adopted to limit the search, such as only shifting the longest or most salient notes, or dynamically adjusting the choice of c based on the length N such that the total is feasible. As we are interested in how the melodic phrase resolves at the end, we choose to limit the search to the final notes of the sequence. In practice, we check for $c = 2$, and if a sequence is longer than 6 notes, we only iterate over the transpositions of the final 6 notes — this results in a maximum of 15,625 melodies, which takes around 10 seconds to evaluate on an 2.4GHz Intel CPU.

Chapter 3

Feature Engineering

There are no wrong notes in jazz:
only notes in the wrong places.
— Miles Davis

In this chapter, we outline the methods and algorithms to measure information of the raw audio signal and the melody we extracted from Chapter 2. By making particular measurements (*features*) we aim to capture certain characteristics of the sound such that we can determine which of these are common to the stimuli that transform into song, and those that do not. This would reveal which qualities of the sound and melody the brain focuses in on such that the perceptual shift occurs, and ideally uncover some of the differences between speech and song. Several of the features we describe here are formulated directly from previous studies, or developed in a way to capture known results. We also include some new, novel features in the hope our investigations can expand on the growing set of characteristics that facilitate the Speech-to-Song illusion[1].

3.1 Audio Features

The most significant characteristic that previous researchers have found that correlates with the transformation rating is the stability of the fundamental frequency (F0) pitch contour. Stable pitches implies that there is some structure or intention behind the sound — musical instruments typically produce constant pitches (aside from stylistic fluctuations such as tremolo or glissando), and trained singers hold there target notes with stability (e.g. Thompson, 2014, found that voice with small variations in F0 indicate a singing vocals). The method we use is a slight variant of the measurement outlined in

[1] A full summary of all the features described here is presented in Appendix A.

Tierney et al. (2012), where we take the sum of the average absolute first differences of the pitch contour of each note[2]. The measurement is then given by:

$$\texttt{stability} = \sum_{i=1}^{N} \frac{1}{|\boldsymbol{s}_i|} \sum_{j=o_i}^{l_i-1} |p_j - p_{j+1}|. \tag{3.1}$$

Contrary to its name, larger values indicate an *unstable* pitch track, and a perfectly stable pitch contour scores zero. If there is a gap in the pitch track (where $p_i = 0$) during a detected note, then this gap is interpolated to estimate the missing values. Ignoring the breaks and including the jump across the empty gap will cause spikes in the absolute difference yielding spurious results[3]. It should be made clear that these measurements are made only during the time steps that a note has been identified, and so any large steps between two distinct notes are not included, as is desired (measurements that pertain to the jumps between notes are accounted for in later features).

As the pitch contour seems a very important aspect of the illusion (as evident by the pure tone stimuli of Tierney et al., 2012), measuring the percentage of the sound source that actually has a detectable pitch might prove indicative — a sound that is very noisy and does not contain a comprehensible frequency might not transform at all, whereas one that has a well defined pitch perhaps does. `percent_pitched` quantifies this by calculating the percentage of time steps where a pitch can be detected:

$$\texttt{percent\_pitched} = \frac{1}{n} \Big| \{p_i : p_i > 0\} \Big| \tag{3.2}$$

It seems reasonable to assume that the higher proportion of the sound has a pitch, then it is easy to conclude the audio is music, so we include this here to test if this is the case. Of course, this measurement is rather sensitive to the algorithm (and its parameters) that produces the pitch contour. However, this will only largely affect the scale of this feature, rather than the quality of the measurement. If we make sure that the algorithm we use remains consistent, this should not pose any major problems.

Another simple measurement is the length of the stimuli — presumably if it is too long then the illusion will not materialise. As we only want to measure the time when the sound is audible, `length` is the time (in seconds) from the first moment $\boldsymbol{I}$ is greater than 50 until the last moment $\boldsymbol{I}$ drops below this threshold. This way, any leading or trailing silence in the audio file is not counted towards its length.

[2] Tierney et al. scale this value by multiplying by 100.

[3] It is unclear how the original authors dealt with breaks in the contour.

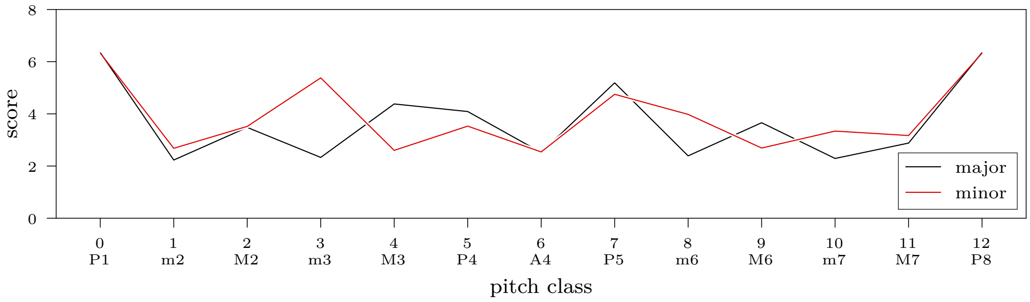


FIGURE 3.1: Krumhansl-Kessler key profiles for each pitch class (with their interval name). Non-integer interval values are linearly interpolated between points. Note that in this model the octave (12, P8) has the same score as the unison (0, P1).

3.2 Melodic Features

Some simple counts and statistics of the notes and the intervals (jumps) between them have been used in previous studies with some success (such as in the unpublished report by Graber, 2015). Music is characterised by jumping between notes to form a melodic contour, a feature less important in adult-directed speech (Corbeil et al., 2013), so they are included here to assess if these are distinguishing traits. These are `num_jumps` which counts the number of notes, `max_jump` and `mean_jump` compute basic statistics on the interval sizes in semitones, and `range` is the difference between the highest and lowest note, also in semitones.

Some features also look at characteristics of the final note — `last_jump` is the final interval size, `last_note_length` measures the final note length as a percentage of the stimulus total length (as defined above), and `last_note_lowest` is an indicator variable the equals 1 if the last note is the lowest note of the whole melody, 0 otherwise. Huron (1996) found in his analysis of musical phrases that typically the melody is arched-shaped, with the last notes tending to be the lowest of the phrase. These aim to test if the melody is characteristic of the resolution of a musical phrase, and where a longer final note could suggest a completion of the melodic line.

We also make some more music theoretic measurements, namely the fit of the notes to some key. Several results from previous studies found that notes which fit well to a musical key are more likely to transform (e.g. Groenveld et al., 2019, Tierney et al., 2018a). We measure `key_fit` in a similar way as described by Tierney et al., however we modify the score such that it is invariant to tuning, and that note lengths contribute weights in the calculation. The procedure is outlined in Algorithm 3.1, where the Krumhansl-Kessler values are the original found by their study (illustrated in Figure 3.1), and not the same as those found in Table 2.3. This measure falls in the range $[0, 6.35]$, with higher values indicating the melody fits well with some key profile[4].

Algorithm 3.1: `key_fit` score calculation

```
Compute weights $\boldsymbol{w}$ from lengths of notes in $\boldsymbol{t}$ (such that $\Sigma \boldsymbol{w} = 1$);
key_fit $\mapsto 0$;
foreach key quality $q \in \{major, minor\}$ do
    foreach tonic note $t_i \in \boldsymbol{t}$ do
        score $\mapsto 0$;
        foreach note $t_j \in \boldsymbol{t}$ do
            $c_j = 12 \times \log_2(f_j / f_i) \mod 12$;
            Lookup key fit score $f_j$ of $c_j$ from Krumhansl-Kessler profiles for key $q$.
              For non-integer $c_j$, linearly interpolate between the surrounding integer
              values;
            score $\mapsto$ score $+ f_j \times w_j$;
        key_fit $\mapsto \max($key_fit, score$)$
return key_fit
```

Similarly, we look at the intervals of the note jumps and compute a measure on how 'ideal' they are, assuming that intervals should be integer valued. Essentially, this scores the melody on it's fit to the 12 Tone Equal Temperament tuning system, the standard of Western music, and how much contesting is required to fit it to this system[5]. A similar feature was tested for in the control stimuli of Falk et al. (2014, Experiment 1), where stimuli were adjusted in pitch to include intervals sizes of 5.5 which does not appear in any Western musical scale. They found there was a marginally significant effect, so we generalise and form a measure that quantifies this for any stimuli. For each of the intervals between each sequential note (in semitones), it is scored by its rounding error to the nearest interval — from zero if it is ±0.5 semitones from a whole number to 1 if t is exactly integer. Formally, this function is:

$$\mathcal{I}(c_i) = 1 - 2 \times \min\left(\Big|\lceil c_i \rfloor - c_i\Big|, 0.5\right) \tag{3.3}$$

where the operation $\lceil x \rfloor$ rounds x to the nearest integer. `scalar_interval` is then the mean of all the interval scores, and so lies in the range $[0, 1]$, where higher scores indicate closer fit. As this is only computed over the note intervals, this is invariant to a reference pitch (such that pitching the whole melody away from standard concert pitch will not affect this score).

[4] `key_fit` can only be zero if the sequence is empty, and a sequence of only one note trivially scores maximum fit (since it is tonic of all keys.)

[5] Just Intonation, an alternate tuning system, might offer a more natural choice here. However, some of the measures in Section 3.4 will capture the same relationships between notes that this system is based on (e.g. simple fractions of frequencies between notes).

Next, we score the melody on the presence of certain melodic intervals that are prevalent in musical compositions, notably the Perfect Fifth, and the Major and Minor Thirds. As these are ubiquitous in Western music, their existence within a stimulus could support the impression of a complete musical phrase, and thus more likely to transform. To quantify this, a matrix of all note pairs and their intervals is constructed to form an $n \times n$ matrix (for a sequence of length n), where entries i, j is the interval (in semitones) between notes t_i and t_j modulo 12. Therefore, if two notes are exactly a Perfect Fifth apart, this matrix would contain a 7 in one of its entries. From this, `i_p5`, `i_3` and `i_m3` can be directly calculated by finding the maximum of (3.4) over all intervals $c_{i,j}$, where c is the interval size we are testing for.

$$\mathcal{I}(c, c_{i,j}) = 1 - 2 \times \min\left(\left|c - c_{i,j}\right|, 0.5\right) \tag{3.4}$$

In the case of the Perfect Fifth $c = 7$, and the Major and Minor Thirds have interval sizes 4 and 3 respectively. Simply put, `i_p5` scores 1 if there is a seven in the matrix, 0 if no entry rounds to a 7, or a value in between that equals 1 minus the smallest rounding error of an entry to 7, such that a maximum score is obtained when there is a strong presence of the Perfect Fifth.

Finally, we use the melody found from the Bayesian melody search in Section 2.3 to compute a distance of the extracted notes to that of a more likely musical phrase. Given a note sequence $\boldsymbol{t}$, we compute the Bayesian maximum a prior sequence $\hat{\boldsymbol{t}}_{\text{MAP}}$ and compare them note for note, by summing up their distances using (2.3) weighed by the notes length to give `bayesian_distance`. This score has a value zero if the extracted melody exactly matches $\hat{\boldsymbol{t}}_{\text{MAP}}$, and higher values indicate the notes are further from the more probable sequence, and thus are less likely to be a note sequence found in a musical composition. Tierney et al. (2018a) used the same model of melody as we have used in the Bayesian melody search to test if the likelihood of the melody has any correlation with the probability of transformation, and found that there is some effect.

3.3 Rhythmic Features

Timing and metre are significant characteristics of music (Cooper and Meyer, 1960) and so any attempt to find the musicality of sound sources should also consider rhythmic components of the audio. For example, meter invokes the listener with some form of expectancy (Large and Kolen, 1994), which is argued to be an important trait of the perception of music (Meyer, 1956, Rohrmeier and Koelsch, 2012), and so evidence of metre within the audio stream could suggest the illusion will materialise. Of course, the act of repetition itself will invoke some artificial metric structure, since there becomes an inherent and obvious grouping of regular recurrent rhythmic events (namely, the entire sample itself), however we are interested in the finer detailed structure. There are two facets of the consistency of rhythm that are of interest, specifically temporal regularity (the structure of the meter), and the steadiness of strength of the rhythmic cues that cement a rhythmic idea. As a basis to our rhythmic measurements, we identify the onset times the rhythmic cues (such as notes or sonic pulses) occur at. There are two main ways of identifying these cues — by observing energy spikes of the audio, or using the times when the notes occur.

The onset envelope, which returns for each time frame the amount of increasing spectral energy in the audio signal, is a very important measurement used in MIR research[6]. From this envelope, heuristics (such as Böck et al., 2012, Ellis, 2007) exist to pick out the prominent peaks that in turn identify the potential beat and tempo within the audio, even when onsets are soft or weak. An example of an onset envelope is illustrated in Figure 3.2 (black solid line). The other option of which rhythmic events to analyse is to use the information already in the extracted melody. In voice, the prosody and rhythm of speech is induced by predominant patterns of stressed and accentuation of syllables (Cutler, 1991), for which we have already identified the time steps they occur at. The disadvantage of this approach of course is that it is only suited for the human voice — a non-vocal sound could have unpitched rhythmic indicators that the onset envelope would detect.

Figure 3.2 illustrates the difference between the peaks of the onset envelope (black dots) and the note start times we extracted. It can be seen that peaks do not necessarily align to when the notes start, possibly due to plosives in the speech that cause loud bursts in the signal but that are not indicative of note boundaries. Initial testing of both these methods found that using the note onset times for temporal specific rhythm features produced better results, whereas the onset envelope was needed for measuring the salience of the cues.

The first rhythmic measure attempts to quantify how balanced the onset times are by looking at the ratios of the inter-onset intervals (IOIs), as proposed by Scott et al. (1986):

$$\texttt{onset\_variability} = \frac{1}{m} \sum_{i \neq j} \left| \log(\frac{d_i}{d_j}) \right| \tag{3.5}$$

where $d_i, d_2, \ldots, d_m$ are the times between onsets (see Figure 3.2). This measure has some nice properties — evenly spaced intervals will score zero (since $\log(1) = 0$) with higher values indicating greater irregularities, the measure is symmetric (as the absolute

[6] In fact, it is so useful that an annual competition is held by MIREX to advance the effectiveness of this measurement (Downie et al., 2005).

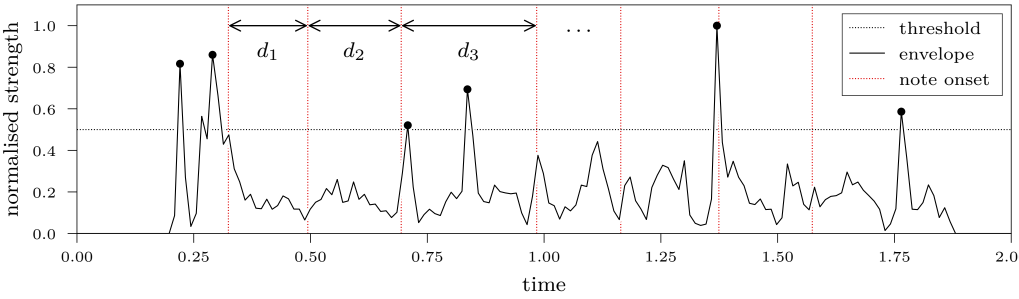


FIGURE 3.2: Example of a normalised onset envelope, along with the note onset times. The strongest peaks above the threshold are indicated.

value of the logarithm ensures that flipping the ratio will not change the measure), the ratio is invariant to tempo (as $kd_i/kd_j = d_i/d_j$), and can allow comparisons across different sequence lengths. One could also simply use the standard deviation of the IOIs and have many of the same properties, but would be sensitive to the tempo of events. Regular musical rhythms are certainly not limited to those aligning to an equally spaced temporal grids — metres with some swing (e.g. a short-long short-long rhythm) are very common within music across the globe. The measure above is somewhat unaffected by such rhythms if they are 'regularly irregular', especially if they are repeated periodically. This is certainly not the case if we used the standard deviation, which will only increase as the irregularities are repeated, even though it can be argued that the sequence is *more* regular because of this.

The (normalised) Pairwise Variability Index (nPVI, Gibbon and Gut, 2001, Low and Grabe, 1995) is another more recent measure of rhythm in speech and was developed to make comparative measures of timing of vowels in spoken word across languages:

$$\texttt{npvi} = \frac{100}{m-1} \sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| . \tag{3.6}$$

Lower values indicate more regular onset times, and to our knowledge has not been used in the study of Speech-to-Song. A recent paper by Condit-Schultz (2019) warns of some of the dangers of using this measure alone in quantifying rhythms qualities and styles of music — rhythm is too complex to be distilled into a single value. However, we are using this as an objective measure of variability of the onset times and not as a measure of complexity of the rhythm.

Next, we look at the consistency of the onset strengths. Steady and unchanging peaks of energy in the audio could facilitate the illusion by suggesting an underlying rhythmic process. For this, we use the onset envelope mentioned above, as this yields an indication of the energy spikes within the audio signal. We use the Python module Librosa (McFee et al., 2015) to obtain this which is then normalised in the range $[0, 1]$, and a standard peak finding algorithm to identify the strengths of the most prominent spikes above a certain threshold. Finally, the standard deviation of these peaks are then calculated to produce `onset_strength`, such that smaller values signify more regular strength patterns. As we are only concerned with the most significant peaks, the threshold level is set to 0.5.

3.4 Dissonance Features

We are interested in characteristics of the musical phrase that may or may not be present within the audio sample, and the structure of the harmony within it. We analyse this by making measurements of the harmonic relationships between individual notes that make up the sequence, and attempt to capture the dynamics of *tension* and *release* that composers typically create in their musical phrasing. This is usually achieved through use of harmony, but other techniques (such as rhythmic changes or developing timbre) can invoke the same effect. While it is tricky to quantify the level of tension and release, we present a measurement that could prove useful for our task. Here we move away from classic music theory and use ideas from psychoacoustics as a basis for these measurements.

First, to make comparisons of two notes, a notion of 'harmonic distance' between their pitches is defined. That is, we wish to measure how close two different pitches sound in a harmonious way, where two notes are near if they sound harmonious together, and distant if they sound unrelated and mismatched. This is a perceptual distance, rather than something absolute or numerical (for example, the linear distance of the notes on a piano keyboard, or difference in frequencies), and has a complex, non-linear relationship (as evident in Figure 3.1). Previous attempts of defining a harmony space often involve setting pitches on a simple two dimensional lattice (named a *Tonnetz,* Euler, 1739), where steps along a specific dimension from one pitch are pitches that are a perfect fifth or major third away. This notion has been developed to the extreme to form more complex spaces, including a highly intricate five-dimensional double helix wrapped around a helical cylinder (Shepard, 1982) that attempts to capture all possible musical intervals. While these formulations are useful for music theorists and composers, they are limited to a discrete space, where pitches are distinct and fall within a framework (such as 12 tone diatonic system), and that frequencies that map somewhere in between these points are hard to interpret or are ill-defined. We turn to psychoacoustics to formulate a model of harmony that exists in a continuous domain, that can compare any two frequencies.

3.4.1 Quantifying Dissonance

Dissonance is a key concept from psychoacoustics that happens when two sounds are perceived as unpleasant, jarring or harsh, whereas consonant sounds are those that are harmonious, pleasant and warm. There is a cultural component to the definition of what notes are consonant or dissonant, and that throughout history these labels have shifted within musical styles and practises. However, a more pragmatic study of this phenomenon was first introduced in Von Helmholtz (1875). He reasoned that if two sine waves similar in frequency can sound 'rough' if their constructive interference causes a beating sound, and that if any sound can be deconstructive into component sine waves, then dissonance is caused when this rough beating occurs with the partials of these sounds. Numerous studies on this that aim to quantify the degree on how this effects sensory dissonance levels along with experimental data (e.g. Hutchinson and Knopoff, 1978, Kameoka and Kuriyagawa, 1969, Plomp and Levelt, 1965, to name a few).

Vassilakis (2001) reviews such models and proposes the following to address some of the concerns: the roughness R of two pure sine tones with frequencies f_1, f_2 and amplitudes A_1, A_2 is given by:

$$R = X^{0.1} \cdot \frac{1}{2} Y^{3.11} \cdot Z \tag{3.7}$$

where:

$$X = A_1 \cdot A_2, \qquad Y = \frac{2 \cdot \min(A_1, A_2)}{A_1 + A_2},$$

$$Z = e^{-3.5 \cdot s |f_1 - f_2|} - e^{-5.75 \cdot s |f_1 - f_2|}, \qquad s = \frac{0.24}{0.0207 \cdot \min(f_1, f_2) + 18.96}.$$

Finally, the roughness of a signal made up of more than two component sine waves is estimated by summing together R of all pairs of sine waves that make up the signal.

To measure the dissonance of two notes, we compute the roughness of the signal generated by the two notes playing at the same time, a method used by Sethares (2005). A note has a fundamental frequency f_0 and n harmonics at frequencies $f_k = kf_0$ with amplitudes $A_k = p^k$, for some factor $p \in [0, 1]$. In Sethares' study and the construction of the tone stimuli of Tierney et al. (2018b) set $n = 6$, however Sethares uses $p = 0.88$, whereas Tierney et al. have $p = 1$. For consistency with regard to the Speech-to-Song illusion, we use the parameters of Tierney et al.

Figure 3.3 illustrates how the dissonance of two notes behaves for a range of interval sizes. This plot is generated by fixing one note, and computing the dissonance with a second note over a range of frequency ratios. For example, the octave interval occurs

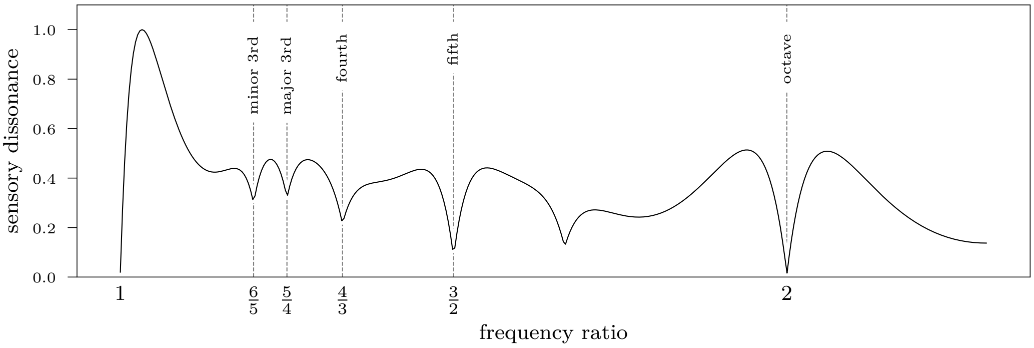
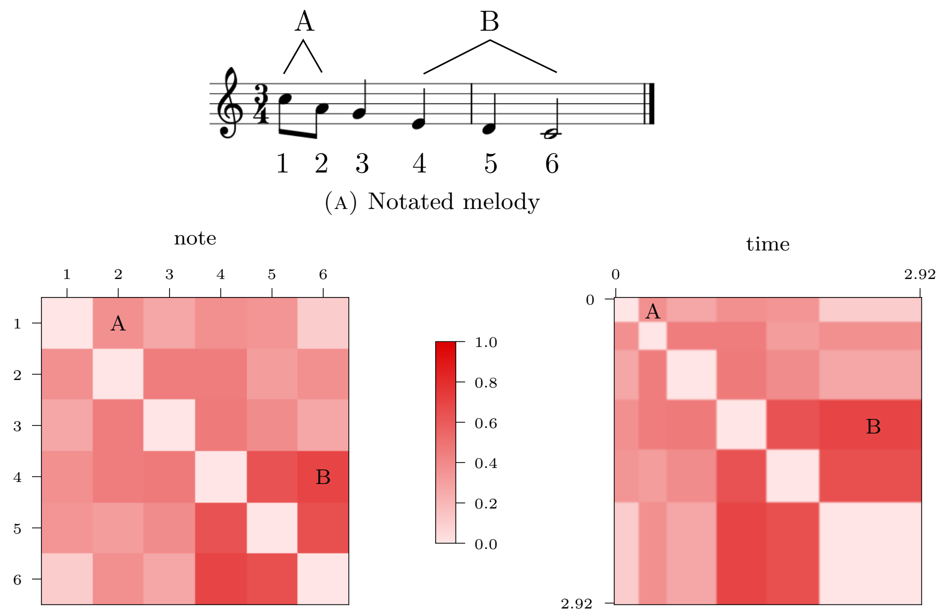


FIGURE 3.3: The dissonance measure for two notes for a range of ratios of their fundamental frequency. Local minima align well for common intervals tuned with just intonation.

when the fundamental frequency of the second note is double that of the first, and this is represented on the plot at the ratio 2. Common intervals in Western music emerge as local minima of this curve, such as the perfect fifth and the major and minor thirds, whereas intervals that just off the ideal ratio have comparatively higher dissonance. The choice of parameter values in the number of partials in the note and their amplitude fall-off effect the shape of this curve, however typically the main characteristics remain — increasing n adds more valleys at more 'complex' interval ratios, and increases the overall height of the curve, and p influences the relative depth of the dips. So long as the values we chose remain consistent, the comparisons we make remain valid. Armed with a method of assessing the relative fit of two notes, we can make measurements of the relationships of the note sequence in the audio sample.

3.4.2 The Self-Similarity Matrix

Self-similarity matrices are used extensively in MIR to identify hidden structure and patterns over the length of the sequence (Foote, 1999, Klapuri et al., 2010), and are used in a diverse range of tasks, including segmentation (Foote, 2000), automatic summarisation (Cooper and Foote, 2002), chorus detection (Goto, 2006), and rhythmic analysis (Foote and Uchihashi, 2001). These representations of the audio are created with some measure of similarity (such as homogeneity or distance) of some audio feature (e.g. chroma, spectral envelope, MFCCs) and compares every time step to every other time step. This yields an $n \times n$ square matrix, with time on both axis such that the element at (i, j) is the similarity of time steps i and j. Depending on the metric used, this matrix is often symmetrical along the main diagonal. Points near to the main diagonal refer to comparisons of the feature that are local temporally, and further points are those of longer term relationships, and thus a full picture of the internal structure is exposed.



(A) Notated melody

(B) Matrix **I**

(C) Matrix **II**

Figure 3.4: Self-similarity plots using the dissonance measure of a six note piano melody. Darker squares correspond to higher dissonance between the notes.

Analysis over these matrices, such as using auto-correlation, then reveal characteristics of the structure, such as repeated sections or parts that have the same instrumentation.

We create a self-similarity matrix using the dissonance measure, such that each cell represents the dissonance comparison of each note pair. To demonstrate this, Figure 3.4b shows the matrix for a short melodic sequence of notes played on a piano, where the both axes represent the note numbers. For example, the cell marked *A* compares the first and second note (Figure 3.4a) using the measurement outlined above. Cell *B* is computed of the fourth and sixth note pair which have a longer distance relation, and all possible note pairs are represented as a cell in the matrix, with comparisons over longer temporal distance further away the main diagonal. All entries along this main diagonal are zero (as the unison interval is perfectly consonant by this method), and the matrix is symmetric since our dissonance measure is symmetric. Alternatively, we can construct another matrix where axes denotes time instead of note numbers, such that the row heights and column widths are proportional to note lengths[7]. This is demonstrated in Figure 3.4c where it can be seen that the information of the first matrix is contained, but with the addition that the note lengths in the melody are preserved. In this matrix, longer notes contribute a greater 'weight' to any statistics taken over the representation, which will prove desirable for some of the measurements we make. We denote these two different matrices **I** and **II** to differentiate between the two types.

[7] Silent or unpitched gaps in between notes are ignored.

From these representations we can infer some features of the melody, such that the last few notes have some dissonance between them but are consonant with the rest of the phrase, which could suggest there is a build up of tension that resolves on the conclusion of the phrase with the last note being very consonant to the first note. By taking some measurements of particular regions of the matrix we can design various features that capture such structure.

3.4.3 Feature Measurements

First, we can make some basic statistical measurements over the entire matrix **II** — `max_dissonance`, `mean_dissonance` and `sd_dissonance` computes the maximum, mean and standard deviation of the time dissonance matrix. The next two measurements, `last_mean_diss` and `last_max_diss`, infer some information about the final note with the idea that a resolving note would be harmonious to much of the rest of the phrase. These are computed over the final column of **I** to detect if there the last note 'fits' well with the rest of the melody. Next, we take statistics over different quadrants of the matrix **II** to expose some broader aspects of the structure of the note sequence. `mean_diss_1` is the average dissonance of the top left quadrant which represents the first half of the melody, `mean_diss_3` that of the bottom right quadrant (final half), and `mean_diss_2` is the mean of the top right quadrant, which measures the average dissonance *between* the two halves. Finally, we take measurements of some of the diagonals of the matrix **I**. The first diagonal directly above the main (i.e. from cell (2, 1) to (6, 5) in Figure 3.4b) are the relation between each successive note, and so measurements over these cells can offer information about the finer structure of the melody. The second diagonal above the main then refers to relations of every note pair separated by two, (i.e. pairs 1 and 3, 2 and 4, 3 and 5 etc), and finally the third diagonal for notes separated by three. Taking the mean and standard deviation of the cells in the first diagonal returns `mean_d_order_1`, `sd_d_order_1`, and similarly the same is done for diagonals 2 and 3.

Each of these features have varying degrees of sensitivity to the order on which the notes occur in the melody, depending on which subset of the matrix they measure over. The first set that take basic statistics over the whole matrix are indifferent to where each note appears, the last note features of course are not affected by the order of the rest of the notes, the quadrants are invariant to the note order within each half, and finally the diagonal measurements are completely dependant of the arrangement. This way, the structure of the melody is quantified, and that two stimuli with the same notes but rearranged in different ordering yield different statistics.

Chapter 4

Data Analysis

Music. . . can name the unnameable and communicate the unknowable.

— Leonard Bernstein

4.1 Audio Stimuli

For this study we have a mixed selection of audio recordings and a parallel set of human ratings data from previous experiments to use for our analysis. The data is aggregated from various studies and are used in to fit and validate our models. With this data, we analyse the distributions and propose a labelling strategy for our binary classifier models to predict.

Most experiments into Speech-to-Song use their own set of materials in their experiments using a small number of recordings, and often with some manipulation of certain features of the stimuli. The original study by Deutsch et al. (2011) used only one recording and various alterations of pitches and word order, and this same stimulus is used in Vanden Bosch der Nederlanden et al. (2015) in their test on how musical experience influences the illusion. Falk et al. (2014) used only two German sentences to test rhythmic differences, and in their studies into the effect of tonal languages, Jaisin et al. (2016) use 6 stimuli (each in a different language), whereas Leung and Zhou (2018) use 6 English and 6 in Mandarin/Cantonese. Few experiments use larger collections of samples, most notably the assortment of 48 English language stimuli used by Tierney et al. (2012) that are used for several follow up studies (Graber et al., 2017, Tierney et al., 2018a,b). Most experiments so far are modest in size and scope, and for a data driven approach on the illusion we require a broader and richer set of stimuli along with a large number of ratings.

4.1.1 Materials

The first set of material are the 300 recordings used in the web experiment of Cornelissen et al. (2016), along the human ratings collected on the transformation level of each stimuli. This is the largest study on the Speech-to-Song to date and so we can utilise data driven techniques. Contained in this set are the original 48 stimuli of Tierney et al. (2012), along with 252 new samples also taken from audiobooks in multiple languages. As the material used by Tierney et al. have been used in several studies that established some of the main results, it is useful to denote these such that we can relate previous results from these studies to our measurements on the same stimuli, and check if the same observations hold on this larger dataset. We denote the 48 stimuli by AT for reference, the other 252 by UVA and the union of these two compilations by MCG.

Also obtained for this study are the material used by Groenveld et al. (2019), where 15 of the least transforming stimuli from MCG are altered in such a way that participants rated the modified stimuli as more transforming than the unmodified audio. For each of the 15 stimuli, there are three levels of alterations (plus the originals) which provides 60 stimuli in total. As the ratings for these stimuli are not a part of the MCG study but obtained through a different experimental setup, it would not be suitable to include them in the mix when fitting the models. Instead, we use these as validation data — if a model can accurately predict whether the speech transforms into song or not then it should mimic the results of this study and rate the modified stimuli as more likely to invoke the illusion. Finally, for validation, we also use the very original Diana Deutsch recording that is of course transforming, along with two control stimuli (white noise and complete silence) that we know should not transform to further test the robustness of the models.

4.1.2 Human Ratings

During the experiment of Cornelissen et al. (2016), participants were asked to rate how musical they perceived the speech as song-like using a continuous slider, with the position of the slider being recorded at short frequent time steps such that information about how the slider moves over time is included. This provides a wealth of data on not only the extent in which a the speech transformed, but at what point the transformation happened. With this information, it would be possible to model how the slider changes over time during the repetitions, or as a method of filtering out the stimuli that already sound musical to begin with. However, we are not interested in modelling the temporal aspect of the illusion (see Rowland et al., 2019), but rather the simpler task of making the classification itself of whether the speech will transform or not, and for this we need binary labels on the stimulus. Therefore, we take the final position of the slider as the final rating of the stimulus by the participant. To make the labels, an aggregation scheme that collects the multiple ratings for each stimulus into an overall score and the threshold value that categorises them based on this score needs to be decided.

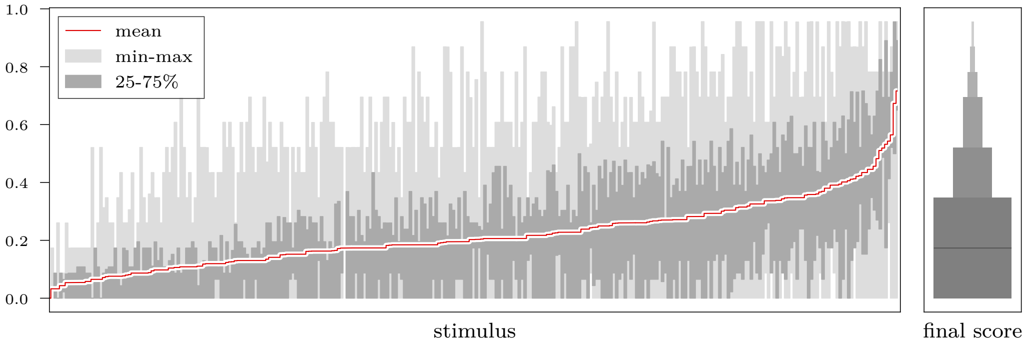


Figure 4.1: Distributions of the final scores for each stimulus in the MCG experiment, ranked by the mean final score. Enhanced boxplot of all the final ratings on the right.

For each stimulus, we have a number of ratings by the participants of the level they perceive the illusion in the range $[0, 1]$, where a score 0 corresponds to *exactly like speech* and 1 with *exactly like song*. Each stimuli has between one and 15 ratings (mean 8, standard deviation 1.15), with a total of 2404 individual data points. The final rating has a very skewed distribution — 49% of ratings are below 0.1 and only 17% above 0.5 (as shown by the enhanced boxplot[1] on the right of Figure 4.1) with an average score of 0.22. Within each stimulus, participants only agree to a certain extent: on average a stimulus has a standard deviation of 0.22, with the maximum being 0.39. Figure 4.1 summarises the final ratings for every stimulus, illustrating the mean final score along with the 25–75 percentile range and the minimum and maximum score.

This also shows a positive trend between mean final score and variance of the score ($r = 0.73$) that suggests higher rated stimuli generally have more disagreement between participants. This trend can be explained for three reasons — first, not every listener experiences the illusion to the same degree so this already accounts for a large amount of the variance. Second, it can be seen that nearly every stimulus has at least one rating of exactly zero, even for the highly transforming stimuli, either because the participant really did not perceive the illusion at all, or there are trials where the user did not complete the task properly. As this was an online experiment, the participants' attention and environment could not be controlled, compromising on the reliability of the results.

[1] An extension to the standard boxplot, useful for large data (Hofmann et al., 2011).

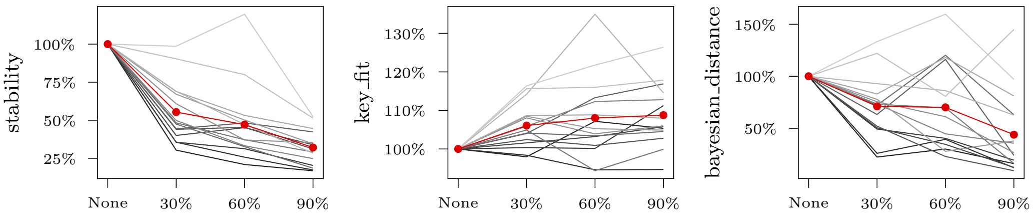


FIGURE 4.2: Illustrations of how three levels of manipulation (30%, 60% and 90%) are reflected by a steady increase or decrease of the measurements of three of the features, as a percentage of the measurement for the original stimulus. Marked line indicates the mean change of the feature.

Finally, presumably participants have a different sense of scale in their rating strategy on how far up they position the slider for transforming stimuli. A rating of *exactly like speech* is easier to align with than *exactly like song*, so there is less controversy with low scoring stimuli. Being a subjective study that requires people to scale the extent to which an illusion is experienced, high variance is to be expected, as Rowland et al. (2019) point out after observing similar behaviour from the ratings of their experiment. Nonetheless, we outline a method in Section 4.2 to interpret the data as reliably as possible.

4.1.3 Manipulated Stimuli

While the MCG data will be used to fit the models, we have several other audio samples that were not a part of that particular experiment, but where the behaviour of the illusion is known (or expected). The manipulated stimuli of Groenveld et al. (2019) are "auto-tuned" in two ways — first the pitches of the syllables are shifted to the nearest diatonic scale, and the F0 contours are stabilised to produce stimuli with more song-like characteristics. This has a direct consequence on two of our features that we measure, such that the `stability` feature should decrease and `key_fit` score increases. These are shown in the first two plots of Figure 4.2, where it can be seen that on average the there is quite a marked change from the unaltered stimulus measurements through the different levels of manipulation. Manipulation of the key fit score has indirect interaction with other features that are sensitive to musical key, such as `scalar_interval`, and the strength score of certain intervals such as `i_p5`. As the key profiles are integral to the posterior computed in the `bayesian_distance` measurement, this feature also decreases (as illustrated), suggesting that the manipulated melodies are also more typical of Western music according to this model.

These manipulated stimuli offer two modes of validating our process. First, as just mentioned, they confirm that the methods to measure F0 stability, key fit and related features are justified and correctly recognise the nature of the manipulations, albeit with some noise due to the imperfect automatic extraction of melody. Secondly, we verify the models by comparing the judgement of the manipulated stimuli to their originals to test if they accurately behave the same as the human trials and agree that the manipulated speech are more likely to transform. This is outlined in the validation procedure in Section 5.2.

4.2 Data Preparation

Before we can fit models to the data, we must prepare the raw data obtained by extracting all the features of Chapter 3 from the audio stimuli described above, and aligning them with the ratings given by the participants of the previous experiment.

4.2.1 Aggregation Schemes

We reduce the multiple ratings per stimulus into an overall score on how likely the fragment of speech will transform. There are several options for this, the most obvious of which is to simply take the mean final rating by each participant. Unfortunately, as evident from the high variance of ratings, and the fact that different people experience the illusion to different degrees, this approach can 'wash-out' some of the higher rated stimuli if there are some low rated scores. Taking the median offers a natural choice as this is less sensitive to outliers, but this brings the disadvantage of reducing the level of the highest rated stimuli again.

An alternate approach is to consider only the highest ratings of a given stimulus. It should be noted that during the experiment, the slider position is initialised at the far left corresponding to the rating *Exactly Like Speech*. This imparts a natural bias towards the lower end of the rating scale, and so if the slider ends on the far right then this implies the participant felt compelled enough to rate the stimulus highly transforming. With this reasoning, higher ratings would be more reliable and indicative of whether the stimulus transforms. Therefore, we take the mean of the top three highest ratings for a stimulus as the final score. This way, a sample that transforms strongly will indeed receive a high rating, whereas if the top three scores are low then we can be sure that transformation does not occur (or is very weak).

In their dataset, Tierney et al. organised their stimuli so the first 24 are transforming, and the other 24 are non-transforming. We can see from Figure 4.3 that taking the mean of the top three highest ratings preserves this distinction such that the transforming stimuli

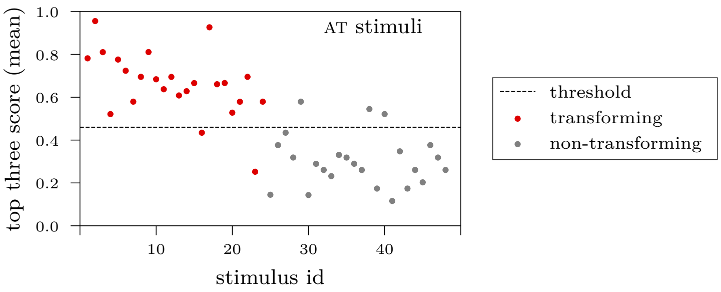


FIGURE 4.3: Distribution of data points of Tierney et al. (AT) stimuli from the ratings obtained from the MCG experiment. Transforming stimuli have IDs from 1 to 24.

receive a higher rating than almost all the non-transforming stimuli. Additionally, the distribution seems to suggest they numbered their stimuli by the rank of the ratings they obtained from their own experiment, as evident by a Spearman rank-correlation coefficient of 0.82 of MCG ratings vs stimulus ID. This also shows that the participants of the MCG study agreed strongly by the classification of transforming and non-transforming stimuli from the original dataset, and the aggregation scheme works well in interpreting their ratings.

While Tierney et al. provided labelled stimuli as transforming or not, unfortunately we do not have such a straight forward distinction for the UVA data. We devise a method to label the data according to some threshold level, where a stimulus is classed as transforming if the mean top three score is above this threshold. By observing the distribution of ratings of AT we can find this threshold that makes a split as close to the original distinction as possible, and use this to label the rest of the MCG data. There is a slight discrepancy here in that some non-transforming stimuli are rated higher than transforming stimuli, and vice versa, so a threshold split cannot segment the data according to Tierney et al's classifications perfectly. Nonetheless, as shown in Figure 4.3, a threshold of 0.46 makes a sufficient split which has an 0.87 accuracy score against the original labels for the stimuli, where only 5 of the 48 are mislabelled[2]. From this, we can label the UVA stimuli.

4.2.2 Data filtering

Figure 4.4 shows the distribution of UVA ratings (grey shaded area), along with the bimodal distribution of the AT ratings (black line) for comparison. Unfortunately, it has a very different distribution to the classic AT stimuli, and is closer to a unimodal

[2] By this methodology, different aggregation schemes yield different threshold values — using the average score as the metric to make the split suggests a threshold of 0.36, while using the median score a level of 0.35 is most suitable.

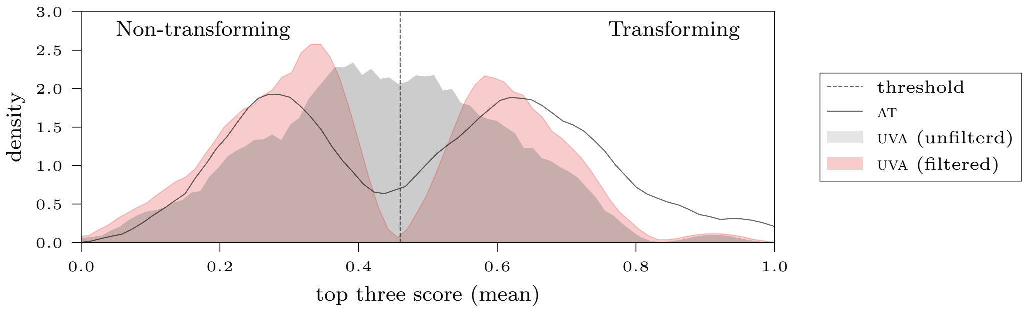


FIGURE 4.4: Distribution of MCG data subsets, and the effects of filtering the data.

distribution centred around the threshold value. Therefore, many of these stimuli are ambiguous — it is unclear how they should be labelled with any confidence. It also suggests that Tierney et al. were sure of selecting fragments that firmly invoke the illusion or not, and that such care was not taken when sourcing the stimuli of UVA. To give the classifier the best chance of success, we filter out the ambiguous data to leave only the stimuli that we can confidently label, and to attempt to form a stronger bimodal distribution. This way, when the models are fitted they are fed positive and negative stimuli from the more extreme ends of the rating scale, and fewer from the inconclusive region. As our models make classifications by segmenting boundaries in the feature space, having a broader distribution in this space only aids the fitting process and adds confidence to model.

Filtering the data is a simple process of removing some percentage of the data closest to the threshold value to squash the peak and separate the dataset more distinctly. First, we drop any stimulus that has less than three ratings to ensure reliable values, which loses 6 data points. Next, we opted to remove 30% of the stimuli closest to the threshold value decided from the AT set. This struck a good balance between forming a bimodal distribution while not discarding too much data, given that 252 is quite a modest size for a machine learning approach and that as much data as possible should be conserved. The final distribution is shown in Figure 4.4 (red shaded area), and it can be seen that the two peaks have formed, both lined up somewhat to AT. This yields 173 stimuli of UVA, with a total of 218 in the set MCG[3], and is well balanced (56% negative samples, 44% positive), ideal for a binary classification task. Finally, the models we choose to fit this data ideally require the data to be have a sample mean of 0 and a standard deviation of 1, so we scale and shift the data of each feature to have this property. This is a standard step in most machine learning pipelines, and for consistency we save the scaling and shifting parameters and apply these to any future data.

[3] Three of the audio files from AT were corrupted (numbers 25, 30 and 46), so in practise we only worked with 45 stimuli from this dataset. However, the scores for these stimuli were still included in the above analysis for setting the threshold level etc.

4.3 Feature Distributions and Correlations

We look at the distribution of the ratings obtained in Section 4.2 against the stimulus features computed in Chapter 3 to measure if there is any direct predictive power from the features alone. Ideally, there should be evidence of correlation that corresponds to established theories on the features that facilitate the transformation, although considering the complex nature of the illusion and the noisiness of the perception ratings a single feature is not expected to be sufficient in predicting the transformation alone.

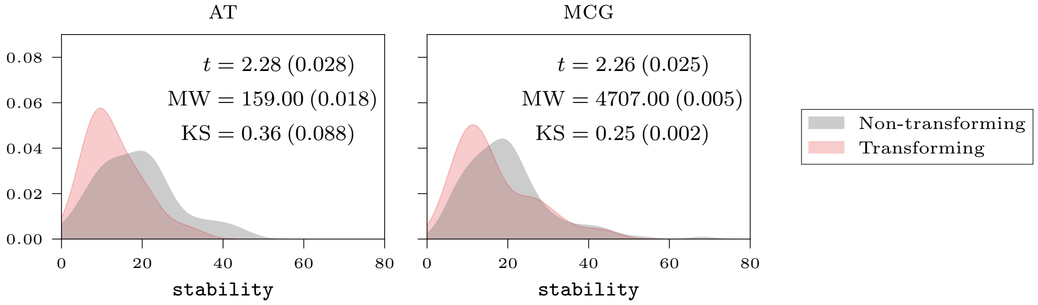


FIGURE 4.5: Distribution of stability for transforming and non-transforming stimuli, along with statistical tests and their p-values testing if the distributions are the same.

It has long been hypothesised that `stability` is a key factor in the illusion, with Tierney et al. (2012) finding a significant effect of this measure on the transformation. To make this claim, they compared the distribution of stability scores for transforming and non-transforming stimuli then conducted a Student's t-test to confirm that the sample means of these distributions are significantly different (reported $p < 0.0001$). We conduct the same test on their stimuli with agreeable results, although we obtain a somewhat weaker p-value ($p = 0.028$)[4], as shown on the left plot in Figure 4.5. This measure remains significant on the full MCG dataset (right plot) and a so appears to be consistent, although it should be noted the support of both distributions are the same. However, this statistical test only measures the significance of the difference of the sample means with the underlying assumption that the distributions are normal and with the same variance. According to the Shapiro-Wilk test for normality (Shapiro and Wilk, 1965), the `stability` measure fails this test, and so the t-test is inappropriate. The non-parametric Mann-Whitney U-test (MW, Mann and Whitney, 1947) checks if two distributions have different locations without the need for the assumption of normality, and is the first resort when this condition fails. Under this test, `stability` remains significantly different, in both Tierney et al. stimuli and the complete MCG set.

[4] There are several possible reasons for the discrepancy: they annotated the note boundaries by hand (whereas we used automatic methods), we are missing three of the audio files, and there could be slight differences in the parameters of the algorithm that extracted the F0 contour.

The two-sample Kolmogorov-Smirnov (KS, Hodges, 1958) test offers a stronger, alternate choice. Unlike Student's t and MW tests which only checks differences in the location (mean and median, respectively) of the distributions, KS is sensitive to the shape (such as skewness, dispersion), and so is a more powerful assessment of the dissimilarity between the two samples. If the KS test statistic is small (high p-value), then we cannot reject the hypothesis that the distributions are similar. For the `stability` score across the AT stimuli, this test yields KS $= 0.36$ ($p = 0.088 > 0.05$), so we cannot conclude that they are different, contrasting with the conclusion of Tierney et al. on their own dataset. Nonetheless, for the larger MCG data $p = 0.002$ and we can confidently make the claim they are indeed different. This demonstrates that larger data sizes are important for making claims with any certainty, and so we could be cautious of the claims made with Tierney et al's sample size.

Continuing this line of inquiry, we test if any of the remaining features have significantly different distributions. With the Kolmogorov-Smirnov test, we find that only 10 of the features hold significance (for $\alpha = 0.05$), most notably `key_fit`, `length`, `i_M3` and `mean_diss_3`, along with the mean, maximum and count of intervals. According to this test, `stability` has the most significantly different distributions. Of all the features, `mean_diss_1` is the only feature that passes the Shapiro-Wilk test of normality and has statistically significant divergent means across transforming and non-transforming stimuli according to the Student t-test. Nonetheless, this feature also fails the KS test, countering this result. Under the weaker Mann-Whitney test, 16 of the 33 features are significantly different, including all those claimed by the KS test, along with `percent_pitched`, `scalar_interval`, `i_p5`, and `bayesian_distance`. No rhythmic features can make this claim.

We consider the values of the features on the final rating of the stimulus itself, and test if there is any correlation (positive or negative) that directly relates the measurement to the rating. While correlation does not imply causation, it is still useful to understand which relationships exist within the data, and how significant they are. For this, we use Pearson's r statistic and compute an associated p-value that reflects the probability that an uncorrelated set produces an r as least as strong as the observed value, and so small values indicate significance. However, the computation of the p-value assumes normality, and as shown this does not hold for most features (only eight features do), so any conclusions drawn from this value should be intrepeted with caution.

Returning to `stability`, we only find a very weak correlation within the Tierney et al. dataset ($r_{\text{AT}} = -0.25$) which is not significant ($p > 0.05$). Nonetheless, we do find that there is a significant correlation of this feature over the entire MCG dataset ($r_{\text{MCG}} = -0.21$). As expected, the negative correlation is consistent with the theory

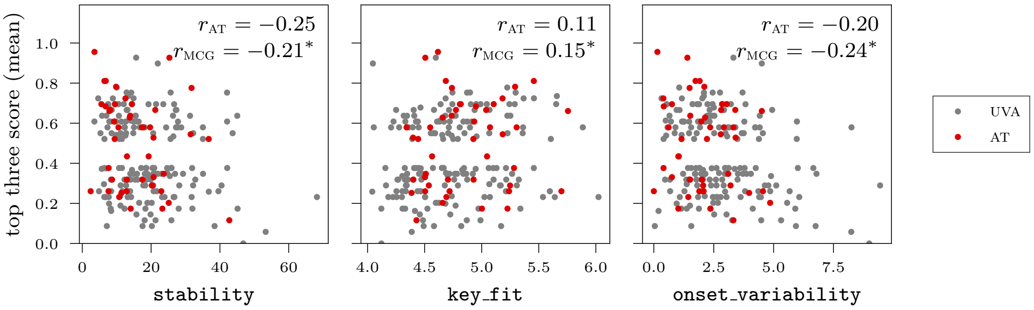


FIGURE 4.6: Distribution of stimuli for three features against their ratings, with Pearson's correlation coefficient for the Tierney et al. subset, and the full MCG set. Asterisk indicates $p < 0.05$

that stable pitch contours invoke the illusion, however from the distribution this does not appear to be a sufficient requirement — many non-transforming stimuli also have stable contours. Figure 4.6 illustrates this distribution along with two other significant examples and their correlation coefficients for the original AT dataset, and the complete MCG dataset[5]. Again, `key_fit` has a weak but significant effect on the transformation only over the full dataset, and `onset_variability` is the only rhythmic feature to have any strong direct influence on the illusion, with the negative correlation implying that more even spaced onsets are more likely to transform the speech.

This point regarding significance seems to be trend in the datasets — if the rest of the UVA stimuli where not included in the experiment, then several of the results from previous studies would fail to pass significance. A complete table (Table B.1) is presented in Appendix B of all the features and their correlations to the scores of both AT and UVA individually and combined. It can be seen that no one feature is consistently significant across both subsets, and that a handful are only significant on the full dataset and not on either subset. None of them are particularly strong, with the largest effect over the entire dataset being the `length` of the stimulus with $r = -0.27$, followed by the variability of onsets ($r = -0.24$) and the strength of a major third interval (`i_M3`, $r = -0.23$).

The discrepancy between these correlations highlights that the illusion is not a trivial effect of any one of the features, and that a given measurement of a speech clip is not enough to confidently tell how it will be perceived on repetition. Rather, it appears to manifest due to a dynamic interaction of these characteristics, or perhaps those beyond the ones measured here. We turn to machine learning techniques to attempt to capture these interaction and determin if there really is any predictive power in using these features to make the classification of a stimulus.

[5] The appearance of clusters in this plot is a result of the filtering which removed central points.

Chapter 5

Classification Models

Intelligibility in music seems to be impossible without repetition.

— Arnold Schoenberg

5.1 Statistical Modelling

When modelling statistical data there are two main approaches — either classification where data is categorised with discrete labels, or regression, where a dependent variable is modelled based on a set of predictors (i.e. features). In the case of Speech-to-Song, we could either predict the binary labels of the audio sample (that is, will it transform or not), or use a regression model to predict the mean final slider position. Considering how noisy and inconsistent the slider ratings are, classification is a relatively easier task. Nonetheless, classification models often output more than just its prediction but also a probability ('confidence') value that can be interpreted as the probability that the given input is labelled transforming or not. Although less precise than a regression task, this confidence level is somewhat analogous to the slider position if we interpret the participants rating as the certainty level that the illusion materialises.

Most studies into the Speech-to-Song use linear models that are fitted to their data to measure the effect of a small number of features (e.g. the Pearsons r of stability or musical key in Tierney et al., 2018a) but do not go beyond this level of statistical analysis. Groenveld et al. (2019) use a Probit model (with interaction terms) to model the stimuli scores from the manipulations they made, whereas Falk et al. (2014) used a binomial generalised linear model to perform a logistic regression. Graber (2015) in her attempt to model the AT dataset used decision trees to predict the classifications, which divides the data by a nested set of conditions on the features until a label can be applied. This type of model is effective in that they are not only non-linear but also provides a clear set of rules about how the model makes a prediction, however their performance suffers on noisy data. We use a range of more powerful models, some still linear, which are fitted to the features and validated to check how well they can describe the data. We then attempt to analyse the models to understand how the classification is made and which features support the decision.

All models are implemented in the machine learning Python library Scikit-learn (Pedregosa et al., 2011), and unless stated are parameterised by the defaults provided by the library. Let θ be a model, and $P_\theta(\hat{y} = 1 \,|\, \boldsymbol{x})$ be the probability the model assigns the label 'Transforming' given the feature vector $\boldsymbol{x} = \{x_1, \ldots, x_k\}$. We class the stimulus as transforming if this probability is greater than 0.5, otherwise as 'Non-Transforming'.

5.1.1 The Logistic Model

The first model we use is the standard binary logistic model as a classifier (Menard, 2002). Given a set of explanatory variables (features), the model outputs a probability of belonging to a class by mapping the result of a linear regression to the interval $[0, 1]$ through the logistic function:

$$P_\theta(\hat{y} = 1 \,|\, \boldsymbol{x}) = \frac{1}{1 + \exp(-\beta_0 - \sum_{i=1}^{k} \beta_i x_i)} \tag{5.1}$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the coefficients of the regression. Fitting this model to the data is a standard process of computing the maximum likelihood estimation using some iterative process until the solution converges.

By its construction, the logistic model naturally outputs a probability of the class, rather than just its prediction, so no further work is needed. Analysing the weights is straightforward too, since the term inside the exponential is essentially a linear regression so the relative magnitude of the coefficients reveals the feature importance, and is simple to interpret (Molnar, 2019). This model does not capture interaction effects explicitly — this would require engineering new features by including additional terms $\beta_{i,j} x_i x_j$ for $i \neq j$ within the exponential term of (5.1). We do not expect this to perform too well because of this, however this weaker model is included to serve for comparison with the more capable models.

5.1.2 Support Vector Machines

Support vector machines (SVMs, Vapnik, 1998) are a rather robust and flexible mathematical model that classify inputs by segmenting the multi-dimensional feature space into regions and categorising all points within a region one of two labels. This is achieved by slicing the space with hyper-planes, called decision boundaries, such the size of the margin between the nearest points either side of the boundary is maximised. In other words, the division of the space attempts to make the 'cleanest' partitioning such that all points are as far from the boundaries as possible, and would require new data points to deviate greatly from the training data to cause a missclassification. During the fitting procedure, only the data points that are closest to boundaries (the support vectors) have an influence on the orientation of the hyper-plane, and outliers that are far away do not contribute to its position. In this way, SVMs are robust to outliers. If the data is not linearly separable, then the hyper-plane cannot divide the space perfectly but a best-fit plane is found, such that a loss that measures the trade off between margin width and missclassification is minimised.

The hyper-planes follow a form $\boldsymbol{w} \cdot \boldsymbol{x} - b = 0$ to segment the space, so is linear, and like the logistic function does not model interaction effect implicitly. From Section 4.3 however, the illusion appears to manifest as a non-linear interaction of the features, and thus the linear SVM could be insufficient at capturing a relationship of features to classification. Fortunately, SVMs can segment non-linear space using the 'kernel trick' (Vert et al., 2004, page 34), where the data is transformed in a non-linear fashion with a kernel function ϕ, such the transformed space is linearly separable. This also has the added benefit that interaction is now captured in the space, as the dot product is replaced by this kernel function that can incorporate interaction terms. However, the cost of gaining more fitting power with non-linear SVMs is we that we lose intuitive interpretation of the space and ultimately how the model makes the classification.

We use three non-linear kernels — the radial bias function (RBF) that can capture clustering, a polynomial function which has some degree of flex in curving the space, and the hyperbolic tangent sigmoid function. In their implementation, the parameters of these are set to sensible defaults, and the polynomial degree is set to three. Platt (2000) outlines a method to map the decision of an SVM to a probability by fitting another sigmoid function to the distance of the point to the decision boundary. This requires an extra training step to calibrate the parameters of the sigmoid function to produce a proper probability, which could be considered an expensive operation when dealing with larger datasets, though for our use this is acceptable.

5.1.3 Ensemble Methods

Finally, we also combine the outputs of the different models to produce an ensemble classifier. Typically, collecting the outputs of several, independent models improves the performace of any single one classifier as the variance in outputs are reduced. Several strategies exist in combining the outputs, for example in a voting classifier the final decision is a simple majority vote of the individual classifiers. We opt for taking the average probability as the final classification statistic. This way, one very confident model can sway the decision if the other classifiers are indecisive (i.e. whose confidence is around the threshold of 0.5). Let $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_n\}$ be a set of individual models, then the ensemble classification is given by

$$P_{\boldsymbol{\Theta}}(\hat{y} = 1 \,|\, \boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} P_{\theta_i}(\hat{y}_i = 1 \,|\, \boldsymbol{x}_i),$$

where $\boldsymbol{x}_i$ is the feature vector required for model θ_i, which may be different between models, depending on which features they operate best with. We construct two ensemble classifiers — one that collects the outputs of just the SVMs, and another that includes both the SVMs and the logistic. The models in the ensembles are not retrained, but instead are assembled from the best fitting model of each type, and we measure there performance together as is.

5.2 Feature Selection

There are two main approaches to feature selection in machine learning — filter methods and wrapper methods. The former utilises information from the data itself to select the most important subset of features independent from the induction algorithm, whereas wrapper methods optimise the performance of the learning model itself through selection of the feature subsets (Kohavi, 1994). We opt for the later method, as the filter method typically does not test for interactions, as each features importance is considered in isolation.

Typically, a feature subset is found in a top-down approach, where the features are ranked by some criterion and the weakest features are removed one-by-one, such as in Guyon et al. (2002) where the square of the coefficients are used to sort the most relevant weights. Ranking features however is not without some unintuitive pitfalls, for example some features can appear to have small weights but contributed in a significant and dynamic way (Guyon and Elisseeff, 2003). Likewise, the addition of more features can even be detrimental to the performance of the model, so narrowing down the feature space to the optimal set is not trivial.

Fortunately, our feature space is modest in the number of dimensions (33), and with a fast fitting procedure (due to the moderate data size), a broad and deep search is attainable in reasonable time. We conduct a combinatorial search with a basic heuristic over many subsets of features (up to some limit) and evaluate each model on that feature set such that the best subset can be found rather quickly.

5.2.1 Evaluation

For a given subset of features, $\hat{\boldsymbol{f}}$, we evaluate a model θ by a series of scores that test certain attributes we wish the model to have. These scores are weighted and summed to produce a final evaluation score of that particular model and feature set combination, such that the complete evaluation procedure attempts to maximise this score.

First, we test how well the model can describe the data by means of K-fold cross validation. This entails first splitting the data into K equal subsets, then the model is fitted on a training set consisting of $K-1$ of the subsets, and tested on the remaining subset for some chosen evaluation metric. This fitting and testing is done for each combination of training and test sets, and the overall score is then the mean evaluation score over each of these 'folds'. This method is ideal for small and noisy datasets, where producing a fixed representative training and test set is not feasible. In our procedure, we choose $K = 5$, and evaluate the model using the balanced accuracy score. The data is not shuffled between evaluations so that the mean and variance are comparable across models. Ideally, if the features are truly representative of the data and help the model, then the standard deviation of scores across the folds should be low, indicating that the model is consistently fitting the different folds well. On the other hand, if this deviation is high then it is likely a sign the model is over-fitting on some folds and under-fitting on others. We denote the mean accuracy score μ_{acc} and the standard deviation σ_{acc}, and ideally the mean should be maximised and standard deviation minimised.

For the remaining scores, the model is first fitted to the full dataset before the next set evaluations. As mentioned in Section 4.1, we have a set of stimuli that have been manipulated such that there is a significant increase in the probability that they transform in the illusion, and so we take advantage of this to measure if the model also acknowledges this difference. First, the probabilities of the unaltered and highly altered (90% manipulation) stimuli are collected from the current model, then we calculate the mean change of probability and divide by the standard deviation of the altered stimuli probabilities to produce a score δ_{alt}. This way, a score of $\delta_{\text{alt}} \geq 1$ indicates a one sigma confidence the manipulated stimuli have an increased probability of transforming according to the model, and therefore we wish for δ_{alt} to be maximised too.

Next, we assess the robustness of the model by computing the area under the curve (AUC) of the receiver-operator characteristic (ROC) curve. The ROC curve plots the false positive rate against the true positive rate over the full range of threshold values, and measuring the area under this curve essentially indicates how well the model separates the false positives and false negatives. This is a standard metric in evaluating classifiers, as a score of 1 indicates a perfect model (with no false positives or false negatives), and a random model scores 0.5.

We also check how the model appraises two control stimuli — Diana Deutsch's *"sometimes behaves so strangely"* excerpt, and an audio sample that consists of pure white noise. These stimuli should be classed as transforming and not transforming respectively by θ, so let

$$\begin{aligned}\gamma_{\text{d}} &= P_\theta(\hat{y} = 1 \,|\, \boldsymbol{x}_{\text{d}}),\\ \gamma_{\text{wn}} &= 1 - P_\theta(\hat{y} = 1 \,|\, \boldsymbol{x}_{\text{wn}})\end{aligned}$$

where $\boldsymbol{x}_{\text{d}}$ and $\boldsymbol{x}_{\text{wn}}$ are the feature vectors of the two stimuli. These scores are highest when the model correctly assigns the appropriate labels to both of them.

Given some feature set $\hat{\boldsymbol{f}}$, the final evaluation score then becomes

$$\mathcal{S}(\theta, \hat{\boldsymbol{f}}) = \frac{3}{4}(\mu_{\text{acc}} - \sigma_{\text{acc}}) + \frac{1}{16}\Big(\min(\delta_{\text{alt}}, 1) + \text{AUC} + \gamma_{\text{d}} + \gamma_{\text{wn}}\Big), \tag{5.2}$$

where a perfect score is 1, a random classifier has an expected score 0.42, and a naive model (that always predicts 'transforming') achieves 0.46. Most of the weight is assigned to the models performance in the cross-validation so that we obtain a model which is not only accurate, but also precise and is not prone to over/under fitting. The score also rewards confident models — the stronger the predictions are (i.e. assigning probabilities further from 0.5 and closer to 0 or 1), the higher the scores for both the control stimuli and the altered stimuli. A confident model has more effectively segmented the feature space, such that the points are far from the decision boundaries.

5.2.2 Search Procedure

Searching the entire space of feature combinations in a brute-force manner involves checking 8.6×10^9 possible subsets of features, so instead we use a semi-greedy, bottom-up search to grow the list of effective features, capped to some maximum number. A simple, fully greedy algorithm would start with finding the one feature that scores the best on its own, then iteratively extending this list by the next feature that improves the score the most and continues until the score can no longer be increased. Unfortunately this succumbs to the typical pitfalls of greedy algorithms — while simple and quick to implement, they often reject large regions of the search space too soon and converge quickly to local maxima, and do not explore a broader area of the space. Therefore, we construct an algorithm that is essentially a number of these simple greedy algorithms searching in parallel over a wider region.

Let $\boldsymbol{\mathcal{F}} = \{f_1, f_2, \ldots, f_k\}$ be the full set of all the features (hence $\hat{\boldsymbol{f}} \subseteq \boldsymbol{\mathcal{F}}$), `FEATURES_LIMIT` be the maximum length of a feature set we search up to, and `TOP_NUMBER` which limits how broad the search is.

Algorithm 5.1: Search for the best feature set $\hat{\boldsymbol{f}}$ for model θ.

```
// initiate results with tuple containing empty feature set and score
0
$\boldsymbol{r} \leftarrow [(0, \{\})]$ ;
$\boldsymbol{c} \leftarrow \{\}$ ;                                              // set up cache
for $i \leftarrow 1$ to FEATURES_LIMIT do
    $p \leftarrow [\,]$ ;                  // pool of candidate feature sets to be evaluated
    foreach $(s, \boldsymbol{f}) \in \boldsymbol{r}[:$ TOP_NUMBER$]$ do
        foreach $f \in \boldsymbol{\mathcal{F}}$ do
            $\boldsymbol{g} \leftarrow \boldsymbol{f} \cup \{f\}$ ;        // extend every feature set by each feature
            if $\boldsymbol{g} \notin \boldsymbol{c}$ then
                $p \leftarrow [p, \boldsymbol{g}]$ ;               // append new feature set to pool
                $\boldsymbol{c} \leftarrow \boldsymbol{c} \cup \{\boldsymbol{g}\}$ ;                    // add feature set to cache
    foreach $\boldsymbol{f} \in p$ do
        $q \leftarrow (\mathcal{S}(\theta, \boldsymbol{f}), \boldsymbol{f})$ ;                          // evaluate feature set
        $\boldsymbol{r} \leftarrow [\boldsymbol{r}, q]$ ;                      // and append to the results list
    $\boldsymbol{r} \leftarrow$ sort$(\boldsymbol{r})$ ;      // sort the results by score in descending order
return $\boldsymbol{r}[0]$ ;                     // returns the highest ranked feature set
```

The procedure starts as the greedy algorithm does by evaluating each feature alone, then sorting by the top scores. However, instead of continuing with the top rated and extending this feature set, the `TOP_NUMBER` of feature sets are extended by each remaining feature and then evaluated. It then continues iteratively by sorting all the previously evaluated features sets by their score, taking the `TOP_NUMBER` highest rated, extending each of these sets, and this process is repeated `FEATURES_LIMIT` number of times[1], where finally the highest rated set is returned. The algorithm is described in pseudocode in Algorithm 3.1. The maximum number of evaluations carried out by this procedure is upper bounded by `FEATURES_LIMIT` × `TOP_NUMBER` × $|\boldsymbol{\mathcal{F}}|$, and in practise

[1] This number limits the maximum size a feature set can grow to.

much less, since it is likely many feature sets remain in the top rated sets (the cache avoids repeat evaluations). This marks a huge speed-up compared to the $2^{|\mathcal{F}|}$ evaluations of a brute force search.

Ideally we prefer a model that uses only a handful of features, as too many could be susceptible to overfitting, so we set `FEATURES_LIMIT` to 10, and by setting `TOP_NUMBER` to 40 we still have quite a broad search space. These values were found to strike a good balance between efficiency and space exploration.

5.3 Results

The results of running this evaluation over the five models (logistic and four SVMs) and two ensemble models along with two baselines are presented in Table 5.1. Listed alongside the highest score is the average (balanced) accuracy score over the K-folds along with their standard deviation, and the number of features the model uses (where applicable). The ensembles are constructed from the highest scoring models of each type, and are not subjected to the feature selection procedure above.

model	$\mathcal{S}$	$\mu_{\text{acc}}(\sigma_{\text{acc}})$	# features
random	0.42	0.50 (0.06)	—
naive	0.46	0.50 (0.00)	—
logistic	0.67	0.66 (0.01)	9
SVM - linear	0.68	0.66 (0.01)	9
SVM - RBF	0.69	0.70 (0.03)	7
SVM - poly-3	0.66	0.64 (0.03)	9
SVM - sigmoid	0.68	0.68 (0.02)	10
ensemble (SVMs + logistic)	0.67	0.72 (0.09)	—
ensemble (SVMs)	0.71	0.75 (0.07)	—

Table 5.1: Summary of model evaluation results.

Bypassing the feature selection procedure altogether and fitting the each model using all of the available features achieves an average score across the classifiers just above baseline ($\bar{\mathcal{S}} = 0.54$, $\sigma = 0.015$), demonstrating the necessity and benefit of the selection process.

All models score significantly better than either baseline, both in terms of the score $\mathcal{S}$ and the cross-validation scheme. Despite no interaction terms, the logistic and linear SVM perform comparably well to the kernel SVMs, suggesting that any interaction between these features has a minimal effect on the classification. The ensemble methods also compare favourably, with the collection of SVMs far outperforming any single model, though with a high variance across folds. Twenty-two of the 33 available features are used by at least one model, with no model using `range`, `onset_strength`, mean or maximum interval size, nor several of the dissonance features. `scalar_interval` is also absent from any model, countering results from Falk et al. (2014).

All models use the `stability` feature, and every one except the polynomial SVM utilises the `bayesian_distance` in its predictions. The best performing single model is the SVM with RBF kernel which not only has the highest score and accuracy, but also achieves this using the fewest number of features. Along with the two features mentioned above, this model includes more music theoretic features `i_p5` and `i_M3`, thus aligning well with previous results, and also three of the new dissonance features — `last_max_diss`, `mean_d_order_2` and `sd_d_order_3`.

As for the linear models, the weights in both cases suggest that `percent_pitched` has the most important impact on the transformation, and both models indicate that stimuli with stable pitch contours and melodies close to Western musical phrases improve the likelihood of transformation. They both agree that `last_note_length` should be short, and that the average dissonance in the last half of the stimulus and in the higher order structure be small (negative weights to `mean_diss_3` and `sd_d_order_2`). The linear SVM implies the last note should be the lowest, whereas the logistic model finds the last note should just be lower than the note previous. The weights of the SVM also hints that there should be a small number of interval jumps along with the presence of the perfect fifth is conducive to the illusion. The logistic model differs here and instead uses `onset_variability` as an indicator for regular rhythms, and strongly imposes less variance of the dissonance in the higher order structure (negative coefficients to the `sd_d_order_3` feature).

A full summary of all the feature coefficients for the linear models, along with the selected features for the non-linear kernels are presented in Table C.1.

Chapter 6

Validation Experiment

Repetition legitimises,
repetition legitimises,
repetition legitimises.
— Adam Neely

We conducted an experiment similar to previous studies to collect further data on the illusion and to validate the models from Chapter 5. The experiment is designed to test a diverse range of stimuli from different vocal sources, languages and non-vocal sounds to see how the Speech-to-Song illusion materialises.

6.1 Setup

Experiments from previous Speech-to-Song studies have largely followed a similar format since the original Deutsch et al. (2011) experiments. Typically, the stimulus is repeated between 6 to 16 times and the participant rates how 'song-like' they perceive the speech on some scale after each repetition. Earlier experiments included presenting the phrase in its original context (e.g. Vanden Bosch der Nederlanden et al., 2015) before the repetitions of the isolated segment, whereas some included additional stages to the trial where the participant had to score the stimulus both before and after the repetition stage (Rowland et al., 2019). The scale is often a five point 'Likert'-like scale as used in the original study, but later experiments used a 1 to 7 scale (Margulis and Simchy-Gross, 2016), 1 to 10 (e.g. Tierney et al., 2018a), or a continuous slider (Cornelissen et al., 2016). In some experiments, an additional step was to be completed after the trial, such as in Falk et al. (2014) where the user has to solve a short mathematical equation in order to distract them between successive trials.

Most trials were held at a controlled environment, however a handful were online (including the later experiments of Tierney et al. and the study where the MCG dataset was obtained from). These have the additional challenge of making sure the quality of the data is adequate. Tierney et al. included 'catch trials' where phrases that started out as speech and switched to a sung version of the same phrase so that if participants did not rate these catch trails as sounding like singing then their scores were removed.

Our experiment was held online in a similar fashion, with a basic setup that presented a random selection of the stimuli, one at a time, for the participant to rate on a sliding scale. Recreating the experimental setup of Cornelissen et al. (2016) permits us to reuse the same aggregation scheme and analysis methodology in Chapter 4. The experiment also collected some additional data that is beyond our study but that could be of interest in future research. This includes the results of a short questionnaire before the trials began that collects information on the languages the participant speaks, as well as the sliders position over time during the complete trial.

6.1.1 Material

Ninety-eight hand-selected stimuli were sourced from public domain sources and fall under fair use. The stimuli for this new experiment were selected in part to align with those from past experiments, as well as to the MCG data that the models have been fitted to. This way we a conducting a sort of reproducabilty study to coordinate the results we obtained, and the models are not disadvantaged by being evaluated on data that is vastly different to what they have been . Nonetheless, we include some other non-speech stimuli to test if the models have generalised to other sources of the illusion, specifically environmental sounds as Simchy-Gross and Margulis (2018) found, even though we expect the models to be unreliable on such data.

The speech samples fall under several categories of speech style — natural conversation in an interview setting, spoken word, and poetry, along with two non-English categories Japanese and Mandarin. Although not the main hypothesis of this study, we select different stylistic vocal samples to see if certain manners of speech invoke the illusion stronger than others. In past experiments, vocals were sourced from audiobooks as these recordings provide a clean and professional audio quality. Twenty-one stimuli from two different audiobooks are included in this experiment. However, in such context the speaker does not have a natural speech style — theatrical embellishments that are used by the voice actor to tell the story could account for particular rhythmic or melodic traits within the speech sample that could elicit the illusion. To test this, we also collected twenty-eight stimuli from interviews and podcasts, where the speakers (1 female, 2

category	count	sources	
Vocal	**86**	**9**	
— English	65	7	
—— (interview	28	2	)
—— (audiobook	21	2	)
—— (poetry	10	1	)
—— (MCG	6	2	)
— Non-English	21	2	
—— (Japanese	9	1	)
—— (Mandarin	12	1	)
Non-Vocal	**12**	**12**	
TOTAL	**98**	**21**	

Table 6.1: Breakdown of the stimuli used in our experiment, indicating the number of each audio type and the number of audio sources they were obtained from.

male) have a more casual and natural tone and rhythm in their vocals, whilst still being professionally recorded. We also collected ten stimuli from a recording of poetry (female speaker), as these are expected to have even greater level of intonation and rhythm as the audiobook recordings that could be perceived as musical. It has been shown that language difference between listener and vocal source can boost the illusion (Margulis et al., 2015), so we include twenty-one non-English speech (9 in Japanese, 12 in Mandarin) to test if these effects can be reproduced. The choice of one tonal language and one non-tonal is to align with Leung and Zhou (2018) study that found that tonality did not show an influence in on the illusion, which we will test for here.

We also include 12 environmental sounds in the collection. These include rhythmic sounds (two water dripping sounds, ice cracking, computer keyboard, walking in snow), non-human vocals (sheep, whale and bird song), and other timbral sounds (kitchen door, jungle, car engine and shovelling sounds). Similar sounds have been shown to be highly transforming on repetition in both Simchy-Gross and Margulis (2018) and Rowland et al. (2019), so we tested if these particular samples received higher ratings than the vocal stimuli.

Finally, we include six highly transforming stimuli, 4 from AT and 2 from UVA, for validating the experimental procedure, and to check if the ratings from this experiment align with those of the previous study. We denote this new collection of stimuli AL, summarised in Table 6.1. All audio is encoded as a mono signal WAV file with a sample frequency of 44.1kHz at 16 bit depth, and the average length of a stimulus is 1.26 seconds ($\sigma = 0.63$).

6.1.2 Procedure

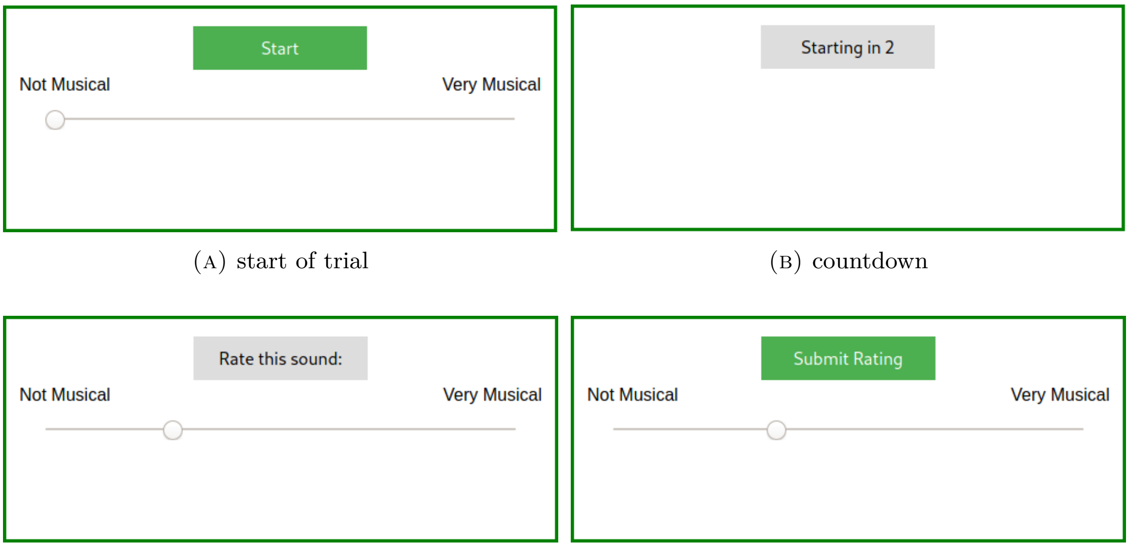
We constructed a publicly accessible, tailor-made webpage specifically for this experiment that is hosted on a private domain secured with TLS certificate, where we have complete control of both the client side frontend and backend server. This way we can control details on how the stimuli are presented to the participant and how we collect and organise the data. The server was created with Node, a Javascript runtime environment suitable for asynchronous web hosting, which serves the experiment and audio files, then collects the user submitting rating data and stores the information in a CSV file for later parsing.

On first entering the webpage, the user was presented a declaration of consent that the participant must agree to before continuing. This declaration contained contact information with the authors and the ethics committee of the Faculty of Humanities of the University of Amsterdam, along with a privacy statement. The checkbox had to be checked to move onto the main experiment. It is at this point a randomly generated, anonymous user ID was generated that was to be sent with the other data so the server can collect the ratings for different participants separately[1].

Next, the user was asked to fill out a short form on their language proficiency. Four checkboxes labelled English, Dutch, Japanese and Mandarin along with a text field labelled ‘Other’ were presented for the participant to enter the languages in which they are fluent in (e.g. able to have a daily conversation with). This information was collected and sent with every rating the user submits.

The next screen briefly introduced the illusion and the procedure of the experiment and the aims of the research. Here, there was a demo of the experiment along with a walkthrough of what the participant should be expected to do. A text encouraged the user to wear headphones and set the volume to a comfortable level, and to make sure they are in an environment where they will not be disturbed. A button labelled ‘Demo’ invited the user to listen to Diana Deutsch’s original transforming vocal sample, with a text hinting that the speaker may begin to sound as if she is singing the phrase. We presented a slider that mimics the slider they will use in the full experiment and asked the participant to move the slider to a position to rate the sample somewhere between *Not Musical* and *Very Musical*, along with a submit button to proceed. This screen served three purposes — first, to get the participant accustomed to the procedure, and by presenting a strong example of the illusion become aware of what to expect from a transforming stimulus. This enabled the user to have some reference on how to rate

[1] We do not store this number in the user’s browser session, so if they refreshed the page or navigate away and return, this number is reset and they will appear as a different participant to the server.



(A) start of trial

(B) countdown

(C) during trial while stimulus is playing

(D) final rating

FIGURE 6.1: Screenshots of the experiment box at the four main stages of a trial.

transforming stimuli and to calibrate their own internal rating scale. Finally, unknown to the user, the score they give Deutsch's sample was recorded and used in our analysis to assess how the other stimuli rate in comparison.

After the demo screen, the user was taken to the main experiment. This page consisted of a detailed set of instructions, and a box that contains a button and a slider, which is the interface to the experiment. Both ends of the scale are explained — *Not Musical* if the sound is exactly like natural speech or some other non-musical sound, and *Very Musical* if it sounds exactly like singing or that it comes from a musical piece. The instructions were as follows: first, the participant must click 'Start' to begin the trial (Figure 6.1a), after which a three second countdown begins before the audio starts playing (6.1b), and the slider is initialised at the 'not musical' end of the scale. A stimulus is selected at random (without replacement) and repeated 8 times with a 1 second pause in between each repetition. During this time, the user was encouraged to drag the slider to the position they feel accurately reflects their perception of the stimulus, and were told that they can change the position at any time and as many times as they feel during the trial (6.1c). After the eighth repetition, the sound stoped and the user was asked to move the slider to their final rating, before confirming this choice with a 'Submit Rating' button (6.1d). This ended the trial, and the process was repeated from the 'Start' phase.

The participant was asked to complete at least 15 trials which would take no more than fifteen minutes to complete, and invited to continue rating more stimuli if they so wished. After the fifteenth trial, a thank you message was displayed for their participation. When the user was finished with the experiment, they simply had to leave the page or close the window to end the experiment.

Data was collected throughout the trial, with the slider position recorded every 100ms during the repetitions of the stimulus. When the user clicked ‘Submit Rating’, a data package containing the user ID, the languages the user speaks, the score they gave to Deutsch’s stimulus, the current stimulus name, the continuous slider rating, and the final rating was sent to the server to be processed and recorded. On the server side, a CSV file was created with the user ID (if the file does not already exist) and this data package dumped into the file. This was done after every trial, such that even if the user does not complete the full fifteen trials (e.g. due to loss of internet connection), we still received the data for the stimuli that they were able to complete.

6.2 Results

6.2.1 Rating Distributions

A total of 55 participants, mostly associates and colleagues of the authors along with some people of the general public, submitted 1198 ratings (average 21.8 stimuli per person). Each stimuli received between 5 and 19 individual scores, with an average number of 12.2. One participant opted to complete all 98 stimuli. Across all the stimuli, there is a wide spread of final rating scores for each one, as illustrated in Figure 6.2, even broader than the MCG data with an average standard deviation of 0.29 (compared with 0.22) over each stimuli. The final score distribution is far more uniform than on MCG, although still skewed to the lower end of the scale (compare with the boxplot in Figure 4.1). Similarly, participants rated Deutch’s stimulus over a wide range between 0.08 and 1.0, with a mean rating of 0.63 ($\sigma = 0.26$). Unsurprisingly, the highest rated stimuli by median were those also from the MCG, however these also received scores across the full slider range. Next, the environmental sounds have the highest ratings from the new stimuli, just above the non-English samples. The vocal recordings collected from audiobooks received the lowest ratings with a mean of 0.30.

The top four of the highest rated stimuli by mean final score in this experiment are the stimuli from the MCG dataset (three also from AT), and another four of the environmental stimuli occupy the top ten, with only two of the new vocal stimuli making it, suggesting that the new vocal stimuli are not particularly strong at invoking the illusion compared to the older dataset. Only five stimuli were on average rated higher than the rating the participant gave to the Deutsch stimulus — three vocals (two from AT, one from UVA) and two environmental sounds (birdsong and water dripping samples), signifying that the original stimulus securely invokes the illusion.

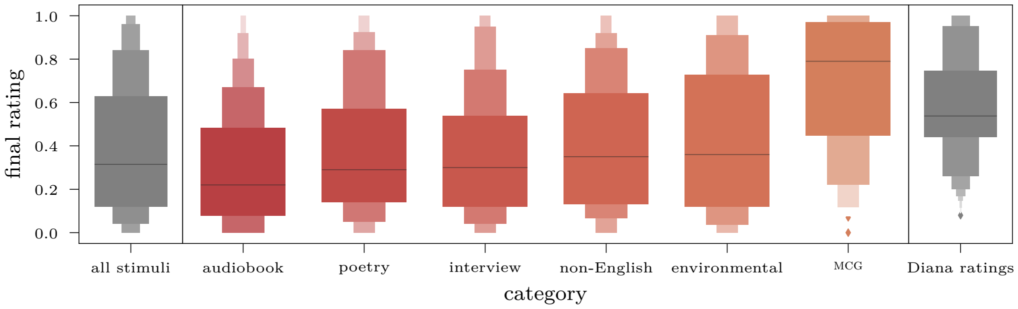


FIGURE 6.2: Enhanced boxplots of all the ratings collected in the experiment, and breakdown of the ratings within each stimulus category. Additionally, boxplot of the ratings given to the Diana Deutsch stimulus on the right.

Using the different aggregation methods and their threshold levels obtained from Section 4.2, the percentage of stimuli that we label as transforming changes considerably — using the mean final rating score with a threshold 0.36, 53% of the AL stimuli transform, whereas taking the median final score (threshold 0.35) identifies only 38% as causing the illusion. For consistency however, we aggregate the ratings in the same manner we prepared the MCG data by taking the mean of the top three scores of each stimuli to use as a final measure of the extent in which it transformed, and use the same threshold value (0.46) to label the transforming stimuli. Under this scheme, 82% of all the experimental stimuli are classed as transforming, much greater than that of MCG (44%), and all environmental stimuli except one (jungle noises) are labelled as such. We utilise the six stimuli that are common to both experiments to verify if the scores align with the previous experiment, and we find that the mean top three score are quite consistent across both studies (Figure 6.3), suggesting that despite the variance in scores, there is little to no difference between the groups of participants in both studies or little effect of systematic differences.

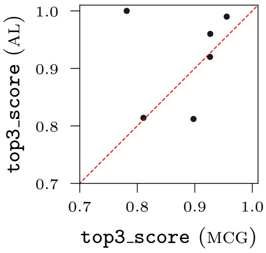


FIGURE 6.3: The mean scores of six stimuli in MCG, plotted against the rating they were given in this experiment. Points on the diagonal line have the same score in both experiments.

We check if the language of the stimulus has any effect on participants ratings. The mean final rating of the English stimuli is 0.37, and 0.41 for the non-English, however a Mann-Whitney test to check if these distributions have the same mean yields a p-value of 0.060

and therefore cannot reject the hypothesis that they are equal (for $\alpha = 0.05$). This holds true for other rating schemes, although for the mean top three score the p-value of 0.054 suggests there is a weak confidence the distributions are different. A similar conclusion is found for the stimuli across the different speech styles of the English phrases — there appears to be no significant differences between the rating distributions for the poetry, audiobook or interview samples when compared with the rest of the English stimuli, suggesting that the illusion is independent of the speech context. However, these subsets are quite small and not highly diverse in themselves (e.g. the poetry stimuli are all from the same source), so a larger study on this line of linguistic enquiry could yield a different result. Between the two non-English language stimuli however, we find that Mandarin has a significantly higher mean to the scores of the Japanese stimuli ($p = 0.004$).

6.2.2 Feature Analysis

Before validating the models on this data, we measure the distribution of the features of these new stimuli and compare them to those from Section 4.1. First, we look at the correlations between the features and the mean top three ratings. Unlike the MCG data, the only significant correlations are in many of the dissonance features (Table B.2, last column), which all have negative values, suggesting more consonance in the transforming stimuli. None of the significant features from MCG carry over to this new set of stimuli, and characteristics such as `stability`, `bayesian_distance` and `key_fit` have no significant correspondence with the score on the complete set of stimuli. Since many of the algorithms in the feature extraction are optimised for the human voice, we conduct the same analysis on just the vocal stimuli, however the conclusions are the same. Changing which aggregate score we make the labelling with also does not alter this result.

Using the Kolmogorov-Sminov test, we find the features that have significantly different distributions between transforming and non-transforming *vocal* stimuli. This includes the mean and maximum dissonance measures for both the whole self similarity matrix and the last note, the `percent_pitched` feature and `mean_diss_3`. This last feature is the only one to also appear significant in the MCG dataset (see Section 4.3). Running this same analysis for all the stimuli (including the environmental sounds), the three dissonance order features also become significantly different.

Finally, we compare directly the two datasets to find any major differences between the two that could explain the discrepancy. Conducting a KS test on the distributions of features in MCG and the new vocal stimuli (AL) reveals that only 3 of the 33 features have significantly different distributions (for $\alpha = 0.05$), namely `stability`, `last_note_length` and `sd_d_order_1`. Further investigation finds that generally the new stimuli have more stable F0 contours than those in the MCG dataset, which coincides with the higher percentage of transforming stimuli in the new set, lending support that this feature has an impact on the scores. On the other hand, `bayesian_distance`, `i_p5` and `percent_pitched` are some of the features used by the models that have a similar enough mean to pass the test. This suggests that the training data is mostly representative of the stimuli obtained for this study, however since more of AL transform there is a larger proportion of stimuli that have features pertaining to the illusion.

6.2.3 Model Predictions

Finally, we evaluate the models that were selected in Section 5.3 by measuring their effectiveness at predicting the new stimuli. This will test not only if the models have generalised well, but highlight if the features we have extracted are indeed sufficient to predict the illusion. Table 6.2 summarises each of the models performance on both the vocal subset and environmental sounds, plus the complete set of stimuli, along with two baselines and the two ensemble models. Evaluation metrics are the (balanced) accuracy and F1 score.

	vocal		environmental		all	
model	B. acc.	F1	B. acc.	F1	B. Acc.	F1
random baseline	0.500	0.317	0.500	0.159	0.500	0.300
logistic	0.548	0.631	0.636	0.429	0.543	0.608
SVM - linear	0.651	0.629	0.636	0.429	0.642	0.605
SVM - RBF	0.497	0.538	0.500	0.000	0.481	0.487
SVM - poly-3	0.513	0.396	0.591	0.308	0.514	0.385
SVM - sigmoid	0.533	0.606	0.682	0.533	0.537	0.597
ensemble (all)	0.526	0.593	0.591	0.308	0.518	0.562
ensemble (SVMs)	0.555	0.598	0.591	0.308	0.546	0.567

TABLE 6.2: Model performances on the vocal stimuli and on all stimuli. Stimuli labelled according to mean top three scheme. Random baseline assigns classification with probability equal to the proportion of positive examples in the data.

As expected, most models score higher on the vocal stimuli than when they are evaluated with the environmental sounds included, with the exception of the polynomial and sigmoid SVMs (although only by a small margin). Every model except the RBF-kernel SVM score higher than random baseline for both metrics. Similarly to the model evaluation in Section 5.3, the ensemble of SVMs outperforms the collective of all the models. The linear models surpass the scores obtained by the non-linear kernels — the logistic model achieves the highest F1 score, whereas the linear SVM has by far the greatest balanced accuracy score, and is the only model to obtain a score comparable to how it performed in the validation procedure (Table 5.1), signifying that it has generalised rather well.

On the environmental stimuli, the models have mixed success. Both linear models achieves an accuracy of 0.636 (F1 = 0.429) each, which is far greater than the random baseline (0.500, 0.153) on this subset, whereas the RBF model simply rejects all these stimuli and gains an F1 score of 0. However, the sigmoid-kernel SVM performs the best on this subset. The reason both ensemble models have the same score is due to the logistic and linear SVM both make the same decisions, so the logistic model has no influence on the outcome of the ensemble.

We can see a modest generalisation of the models according to this validation study, particularly with the linear SVM standing out clearly, showing consistent signs of some predictive inference from the features we have extracted. This model also had the lowest variance in the accuracy score across folds in the evaluation procedure (Table 5.1), and this is reflected here. Other models which showed promise in the selection process failed completely on the new data, specifically the RBF and the ensemble methods. As there is room for some improvement in all of the models scores it is clear there are other factors to the illusion beyond what we have measured from the audio.

Chapter 7

Discussions and Conclusion

Where words leave off, music begins.

— Heinrich Heine

We studied the Speech-to-Song illusion using automatic and computational methods to obtain the music hidden within the voice, and analysed the properties and qualities of the audio in order to assess if one could predict how the illusion manifests from these features alone. This involved developing an algorithm to derive a sequence of notes from the raw audio data, and from this making a series of measurements on this melody. We found that some features established in previous research hold, whilst also demonstrating new measurements that captures some characteristics of the melodic phrase also contained effective information. Data models used a handful of these features to successfully predict if an audio stimulus will transform into song or not, significantly above baseline. We then ran a validation experiment on a fresh set of stimuli to collect a large assortment of new data, and found that one model in particular maintained its predictive power on these new sounds.

7.1 Feature Methods

To start, Chapter 2 detailed an algorithm extended from work by Cornelissen (2015) that performs well at extracting the melody that a listener could hear from natural human speech. Such an algorithm provides useful for the study of the Speech-to-Song illusion. The architecture of the method has two main steps — segmenting and identifying the note boundaries, then computing the potential pitch values of the notes. We parameterised this algorithm and evaluated it to find that it performed better than the original formulation and make agreeable predictions of the melody in the illusion. It is important to make an accurate and objective transcription of the melody as this forms the basis of almost all the features we extract later.

Nonetheless, there are a few possible improvements that can be made. First, it is clear that this procedure is designed and optimised on the human voice and that it is unlikely to perform well on non-vocal sound samples, (as evident in Table C.1). The F0 pitch tracking algorithm in Praat assumes a priori that the pitch will be within the typical vocal range of a speaker, and so will produce unstable results on sources that are well outside of this range. Other more general pitch tracker (for example the neural network algorithm of Kim et al., 2018) are available. The pitch contour itself is susceptible to octave errors that should be dealt with — more advanced melody extraction algorithms also make use of multiple candidate frequencies to optimise the probability of selecting the correct note (e.g. Ryynänen and Klapuri, 2006). Segmentation of notes is built around phonetic features that pertain to physical facets of speech, whereas there are other characteristics of sound that could indicate note boundaries such as rhythmic and temporal ques or sharp changes in pitch or timbre. The algorithm could be made more universal by using information contained in the pitch contour $\boldsymbol{p}$ to further segment notes, for example if there is a large step from one flat region to another then this should also count as a note boundary. Rhythmic qualities are somewhat encoded in the intensity values $\boldsymbol{I}$, where peaks invoke some metric structure, however implied rhythm is not accounted for that could be indicate note onsets. Making the predictions about which notes are perceived in a given pitch contour could be improved further — it was observed that the transcribed note pitch seemed to correlate more with the pitch values at the end of note, and less at the beginning, as if $\boldsymbol{p}$ takes time to 'arrive' at the note value. This could be incorporated by weighting pitches at each time step by their position within the note which could correct unstable and fluctuating pitch curves when taking the mean.

Glissando and pitch slides are not considered by our algorithm — these are stylistic features used extensively by musicians and composers throughout most music and can be imitated by the human voice, however representing this information is not possible in the current formalisation. Although remaining limited to the stable notes is a major simplification of melody, it can be argued that gliding between notes is merely an expressive feature of melody, and not necessary — Meyer (1989, page 14) suggests dynamic changes are a secondary parameter in music, different from the primary features of melody, harmony and rhythm. This makes intuitive sense, as it is possible to play on a piano (where note slides are uncharacteristic) a part written for trumpet and it still be recognisable as the same melody, for example. Hence, there is perhaps little motivation to include this, and to keep the melody representation as simple as possible.

The algorithm also assumes the music that is contained in the audio is strictly monophonic, where there is only one note at any given time with no background harmony or chord constructions. With voice, a human can only produce a monophonic signal so this simplification is valid, but of course this does not extend into the larger domain of sounds and music. Automatic polyphonic music transcription gathers a lot of attention in the MIR field (e.g. Cemgil et al., 2003, Ryynänen and Klapuri, 2005), which is not surprising considering the vast collection of music is polyphonic in nature. While there exists several successful algorithms for this task, they suffer from the same issues as previously mentioned, namely their assumption that the audio source is already music, with some even requiring information on the number and type of instruments that make the sound (as in Grindlay and Ellis, 2011).

Next, we outlined the features that were measured to make the classification of the Speech-to-Song illusion stimuli. These include well established features, notably the measurement of stability and the key fit of the notes, as well as new ideas such as the rhythmic measures and the introduction of dissonance as a metric. Most of these measurements test basic and higher level features of melody (namely rhythm and note values), whereas more detailed qualities such as timbre are not included. For example, Zhang and Ras (2007) outline a large collection of timbral measurements for use in classification of musical instruments, and could provide inspiration for this task, at least for a more general model predicting Sound-to-Music. As Simchy-Gross and Margulis (2018) observed, breaking up and shuffling the environmental stimuli did not break the illusion (when it did to vocals), perhaps suggesting that timbre of the sound could be the cause for the illusion to materialise in this case, and not some melodic or rhythmic characteristic.

A significant measurement we made is the idea of the extracted melody's distance from one that is more likely to be composed. The idea is to measure how much do the notes would have to be shifted to arrive at a more typical melody that could be contained in the audio. We used a Bayesian model of melody to evaluate similar note sequences that are around the one we extract and constrain this evaluation to sequences that are possible given the pitch contour, then used a distance metric to compute the final `bayesian_distance` feature. The Bayesian approach of finding 'typical' music has some inherent bias and issues, as stated at the start of Section 2.3. The most immediate problem is assuming a Western framework of music theory in the construction of keys, and in quantising notes to fit a traditional piano keyboard. Listeners of different cultural backgrounds can perceive different tones that fit better to their traditional musical systems (Curtis and Bharucha, 2009), and so enculturation effects the melody that is perceived. However, we chose to accept this drawback for lack of a more universal model, and stick to Western music theory for consistency. Quantising notes to a 12-tone equal temperament system is also not ideal for similar reasons, however it is the standard approach of discretising note values for representation in a computer model and is used extensively in the digitisation of music. Assumptions about the reference pitch $A_4 = 440\text{Hz}$ could also be relaxed relatively easily by shifting all the notes up in pitch by a small fraction until the distance between extracted notes and quantised notes is minimised[1].

Of course, the Bayesian model of melody can be improved further in several ways, for example extending the sequence of previous notes in the prior (by generalising the probabilities in Equation (2.10) to something like $P(t_i|t_{i-1}, t_{i-2}, \ldots, c, k)$), or extending the key profiles in some way to include other (non-Western) musical scales. The model also only works with quantised pitches, due to Table 2.3 being a discrete probability function. One approach that would allow unquantised, 'out-of-tune' notes is to define a continuous probability distribution that interpolates the points of Table 2.3 and renormalised to produce a proper probability mass function. Rhythmic or metric qualities are also not considered by the model. Presumably, note onsets that fall on some simple grid division pattern could be considered more 'musical' in the sense that that would be more like composed melodies, than those with more 'random' timings. Future work could take this idea and incorporate a way of measuring some statistics on the temporal divisions and note lengths in composed music, and devise a distance measure to make comparisons of the derived melody and the maximum-a-priori melody $\boldsymbol{t}_{\text{MAP}}$. An alternate model of melody, specifically one that captures a broader scope of musical traditions more than just Western folk song, would be highly desirable, as the current model is limited in this way. For example, simple sequence models such as n-grams or Hidden Markov models have been used in previous research for predictive tasks (e.g. Cherla et al., 2015, Conklin and Witten, 1995, Groves, 2013, Pearce and Wiggins, 2004, Whorley and Conklin, 2016, to name a few). IDyOM (Pearce, 2005) is a recent and powerful probabilistic model of the structure in music, that predicts musical events after exposure to some corpus of music. Computationally, checking all tone sequences is the most expensive operation so smarter heuristics in the optimiser could improve performance or allow for broader search space, which in turn would yield more probable tone sequences. The search procedure however is also trivially parrallisable by giving each process a separate chuck of the search space each, and could offer some significant speed-up, permitting a much broader search space.

We also outlined three rhythmic measurements that capture some detail of the intricacies of potential meter in the melody. However, reducing the complex hierarchical nature of metre and the diverse range of rhythmic divisions to a small set of simple measures that quantifies all the information is no easy task. Currently, our measures only really

[1] We would only have to shift the pitches by at most one semitone.

captures the steadiness and consistency of onsets, and not the more intricate complexities that rhythm can contain. Neither simple nor highly structured rhythms can be said to be more or less musical in itself, so it can not be expected that these features alone can signal the Speech-to-Song transformation. The repetition that is necessary to cause the illusion artificially invokes a meter, and since "repetition legitimises", any complex meter can be legitimised in a musical sense, so it remains to be seen if a feature of rhythmic complexity can actually be predictive. Nonetheless, they are included since previous results (e.g. Falk et al., 2014) had some success making the connection of simple rhythms and the illusion.

The motivation of the dissonance measures is to capture something more than just the melodic structure, but to attain some aspect of the 'musical idea' that a melodic phrase tries to convey. It is hard to define exactly what such a meaning could express, but as the build up of *tension and release* are fundamental devices used by composers to attain an emotional response from the listener, dissonance as a measure of tension offers a possible method of quantifying these dynamics. Alternately, models of tension such as Farbood (2012) have shown promise of predicting musical tension as provided by listeners in perception experiments. Nonetheless, some research on tension show correlation with roughness of the sound (Bigand et al., 1996, Pressnitzer et al., 2000), which the dissonance measures we used are based upon. Unfortunately, some of the dissonance features are less intuitive to observe when listening out for them in the stimuli, especially the measurements on the higher order structure, and that it is hard to consciously hear what these are measuring exactly. They are nonetheless designed to be sensitive to note ordering and to quantify in some way not only the internal melodic structure in a musical phrase, such as the harmonic hierarchy, but if there contains a complete musical phrase that could be conveyed as the rise and resolution of musical tension. There are however drawbacks to this measurement. When comparing two notes, these are taken in isolation from the other notes in the sequence, and the measurement is independent from the context of key — in general, two notes can sound more or less harmonious depending on the condition of the surrounding contextual notes and their placement within a key. The quantifying of dissonance in our method is also sensitive to the accuracy of the note extraction algorithm, and assumes the pitches of the notes are precise.

7.2 Data Analysis

In order to classify the stimuli we have as eliciting the illusion or not, we collected the ratings obtained from a previous experiment by Cornelissen et al. (2016, which we named

MCG for convienece) and devised a scheme to assign the correct labels. This required attempting to organise highly noisy data by essentially discarding the lowest scores given by participants and only taking the mean of the top three scores. By only focusing on the top scores we have a very optimistic expectation that the speech transforms, since it takes every participant to give a low rating for our scheme to also rate it low, and we are essentially giving the data the benefit of the doubt. This could produce an over-representative distribution of transforming labels — nearly 50% of stimuli are labelled as such even though in practice a randomly selected segment of speech is unlikely to transform. However, it is unclear the selection process of the MCG study, so there could have been a bias to select samples that are likely to be heard as song. Alternately, instead of using the top three, the top 50% scores could have be used that would have taken more data points into account. We used Figure 4.3 to justify that the strategy preserves the data of the AT dataset, but we could go further with this idea — by using the labels of the AT data we could optimise the thresholds, cutoffs or parameters in an aggregation scheme such that it segments the stimuli as close to the original labels as possible. Removing 30% of the least decisive stimuli to produce a bimodal distribution is also unsatisfying — there could be information contained in these stimuli that would produce better segmentation in feature space. It could be beneficial to instead remove the stimuli that have the highest variance amongst their scores, as in these cases there is a lot of disagreement between the participants whether it transforms or not, so to label these stimuli with any confidence is ambitious. An alternate approach of parsing the data would be to not aggregate the scores for each stimuli, but rather have every rating as a unique data point. This way the model fitting procedure can decide which data points are relevant and carry the necessary information, and which are outliers, in an unsupervised way, rather than creating our own strategy.

After the data was sorted, we checked if any of the significant findings from past studies can be found here too. We conducted the same statistical test as Tierney et al. (2012) and found the same results hold in both their dataset alone and on the full MCG dataset, namely that the distribution of `stability` scores have significantly different means for transforming and non-transforming stimuli. Unfortunately, the distributions fail to pass the Shapiro-Wilk normality tests, so the significance of these results are dubious. By observation, we can see that even if the means differ significantly, this difference is small. As Figure 4.5 illustrates, the full range of `stability` values are represented in both the illusionary stimuli and non-transforming for the MCG stimuli, and the distributions themselves contain a lot of overlap with each other except for the peaks which are offset slightly. This means that given a speech sample with a highly unstable F0 contour, we cannot say with much confidence whether it invokes the illusion or not — only that it is slightly more unlikely. Nonetheless, the more powerful Kolmogorov-Smirnov test that is appropriate for these distributions did find a significant difference between them, so their findings on stability hold up. This test was conducted on all the other features, and found a handful which the test also suggested as having different distribution when measured on transforming or non-illusionary stimuli, including some new features.

Similar conclusions are drawn from the investigation into the correlations between feature values and the final ratings. We find r in all features to be rather weak, even in the best case (the maximum correlation value has a magnitude of 0.27), so we are hesitant to claim any strong evidence that any feature alone is necessary or particularly convincing to the illusion. This is not surprising, as observed above the distributions of the highly rated and lowest rated are largely similar in their support and location of the peaks. It can be seen in Figure 4.6 that the scatter plots are mostly a cloud of points, and that it is a few outliers are actually responsible for the correlation value to be non-zero.

From this we can conclude the not one characteristic is sufficient for the illusion to occur, not even `stability` — at least not from those that we have measured. This suggests the illusion occurs through some interacting combination of characteristics, and that a stimuli requires multiple conditions to be 'just right' for the perceptual shift to occur. We also expect that ultimately, transforming stimuli probably occupy multiple clusters in feature space, as evident from the way non-speech stimuli that have very different features can also invoke the illusion. For example, the raindrops sound in Simchy-Gross and Margulis (2018) is reported as becoming particularly musical on repetition because of its rhythmical content, however we see that none of the rhythmic features are sufficiently telling, demonstrating that our results are limited to vocal stimuli.

7.3 The Models

We fitted a diverse collection of models with both linear and non-linear kernels to describe the data, and used a procedure to obtain a set of features for which each model gave optimal performance. We choose model types that are flexible and that work well with modestly sized, noisy datasets. In particular, a family of SVMs were selected for their performance and generalisability on high-dimensional data, and do not require any assumptions on the distribution of each feature.

There are many other models we could have used — for example Naive Bayes is a simple but effective model that requires very little data to train. Despite in its derivation the assumption that the features are independent, a premise rarely true in real tasks, it can still perform remarkably well even when this does not hold (Zhang, 2005). Decision trees, as used by Graber (2015) work well as a classifier and again require no assumptions about the data, however initial attempts at fitting this type of model succumbed to either extreme over-fitting or woeful under-fitting, and so were not deemed adequate for us. Random decision forests (Breiman, 2001), which is essentially an ensemble of decision trees, elevate this problem by combining multiple trees that 'specialise' on certain inputs, and are used widely in practise as they often out perform other methods, including SVMs (Caruana and Niculescu-Mizil, 2006). While there is also a method to rank feature importance in these models (outlined in the original paper), just as with the other ensemble methods it becomes much harder to interpret how the feature contributes to the classification.

The feature selection method we designed is provisional for the data that we have, and the evaluation metric was composed to obtain desirable properties of the fitted model. We departed from more standard procedures of feature selection as we have the unique opportunity to utilise other data and we take advantage of these additional stimuli to produce a model that not only fits the data well but acts closely with listeners behaviours. The most interesting of these extra material are the manipulated stimuli of Groenveld et al. (2019), where 15 original speech clips are digitally altered such that listeners are ultimately more likely to rate them as illusionary. All the models except the polynomial kernel SVM showed a significant increase in their probabilistic output between the unaltered and highest manipulated stimuli, exhibiting the behaviour we desired. The models are sensitive to some of the features that were altered in the manipulated stimuli, so while it is expected that the classifier's outputs reflects these differences, the fact that the models successfully label the altered stimuli as more likely to transform shows that these features are utilised correctly. We tested if this model evaluation metric is worthwhile by evaluating how the model which maximises the K-fold accuracy score alone behaves. For example, we found a linear SVM that managed a mean accuracy score of 0.70, however not only did this model score the manipulated stimuli only marginally higher than the unaltered versions, but it also highly rated white noise as transforming and did not judge the Diana Deutsch sample as illusionary. Clearly, this is a rather unsatisfying model since it fails at some basic quality checks, and demonstrates that our selection method returns a more favourable model, albeit one with a slightly weaker accuracy score in general. As we wish to model human behaviour, rather than engineer a powerful classifier, this compromise is justified.

That is not to say the evaluation score is ideal, there are certainly improvements and optimisations that could be made. For example, the weights between the different metrics that make the final score (5.2) could be tuned further, as the current formulation gives equal weight to the additional scores that are not calculated from the K-fold evaluation. Quite a large proportion of the final score is based on the Diana Deutsch stimulus —

even though ultimately the model is evaluated on 220 stimuli, this one stimulus contributes one-sixteenth of the final score alone, and the same goes for the white noise sound. This could be too harsh as even the best models succumbs to false negatives, and it could be that there is a slightly better model that happens to falsely reject the Deutsch stimulus for which it is penalised too much and so ranks lower. In hindsight, this could be reduced by creating a small collection of known musical samples (e.g. of actual singing, an instrument playing etc) and taking the average score of these as a metric to be included in the final score. The search procedure itself is quite robust and offers a great speed-up compared to a brute force approach. There are only two hyper-parameters that should be decided beforehand, namely the feature limit (essentially the *depth* of the search) and the top number of candidate feature sets to maintain (the search *breadth*). Of course running the algorithm with higher parameters and exploring a larger space could yield better results, however it is unlikely to find a markedly higher performing model — there appears to be diminishing returns when increasing the number of features as evident from the fact that four of the five models required less than the number we searched up to (see Table 5.1). Besides, increasing the number of features for a minor gain of the final score goes against Occam's razor as a heuristic in scientific modelling, where a simpler model is more desirable.

For both the ensemble models we decided to collect the best of each model type, rather than searching for the best combination of 'expert' models that collaborate together. This would require searching over all possible combinations of feature sets over every model to evaluate how they perform, the total of which is the current number of evaluations raised to the power of the number of models in ensemble — a search space far too large to be feasible. Alternately, we could treat the ensemble as a single model, where each of the component models all work with the same feature vector, and so the feature selection procedure would be the same as the individuals. This is a more typical approach in machine learning, however in our case searching for the best ensemble method is not the priority, as this would not reveal much about how the features combine to make the classification. Instead, we use the ensemble as a way to assess how much the models agree with each other — if the collection of models together performed only at baseline then this would imply that classifiers were at odds with each other and there is little in common among their predictions. On the other hand, if the ensemble achieves a score that is the average score of the individual models then this would imply that all the classifiers are well aligned in their decision boundaries and make similar predictions. In this case the ensemble of SVMs attained the highest score, more than any one model. This indicates that not only are the models coordinated, but that when there are differences in their predictions they somehow combine together in a beneficial way. While this is normally welcome in machine learning practice, unfortunately it makes interpreting its operation difficult. For example, the polynomial SVM is the only model to use the feature `npvi`, however it is hard to know if (or even when) this particular feature is ever used in the ensemble to make a prediction, or if this base model relies on this feature to make a deciding vote.

Of the two linear models where the coefficients can be interpreted we find that both the logistic model and linear SVM both assign parameters with the same sign for the features they have in common (Table C.1), where the sign suggests how the value of this feature impacts the prediction. This indicates some consistency between them and that there is useful information contained in these measurements. For example, a negative coefficient for `stability` means that indeed stable pitch targets push the models decision towards classifying the stimulus as transforming, and the magnitude informs how large of a contribution it makes. We found that the percentage of the stimulus that is pitched is the strongest indicator for both linear models, where the more a pitched note is present in the audio the higher the likelihood of transformation. At first this seems quite intuitive, after all, for a melody to exist there needs to be pitch, and if the entire sample is pitched then it is easy to conclude that it could be musical. However, this leaves little room for rests in the melody, a very important characteristic of musical timing[2]. Rests are typically the indicator of the end to a musical phrase and before the start of the next one, and if we assume an illusionary stimuli is one that is a complete musical phrase then it is unlikely there would be a rest present. Counter to this point however, very percussive and rhythmic sounds would not have a pitch track at all, but could still easily be perceived as musical (e.g. the water dripping sounds of the environmental stimuli used in the experiment), but as the training data is all vocals then the models are not exposed to these types of stimuli.

As expected, the linear models also accepts illusionary stimuli if their melodies are close to those found in Western compositions, as described by the Bayesian model outlined in Section 2.3. The negative weights (and their large magnitudes) of `bayesian_distance` implies this is quite a telling feature. This conforms with the results of the first experiment in Tierney et al. (2018a) who used the same Bayesian model of melody to asses the likelihood of the sequence of notes and found a correlation between the mean rating change of the stimulus and this likelihood. The authors also tested for correlations between some of the components of the melody model, namely the interval size and the conformity to Western keys, and also saw a significant trends in both these features and the participants ratings. However, we do not see this so much in our models — the features we extracted which are analogous to these components `max_jump` or `mean_jump` are not used by any classifier, and `key_fit` is only used by one of the SVMs. It is possible that these features reduce information about the interval structure down too much for

[2] "*The music is not in the notes, but in the silence between.*" — Wolfgang Amadeus Mozart

them to capture any meaningful detail, or that there is little difference between speech and music in these features[3]. The authors then conducted another experiment where they controlled for melodic structure, but could not draw any significant conclusions from these results. It seems then there is a connection, but the nature of how these components interact and influence the musicality rating is complex and perhaps non-linear. The range of note pitches is also not used by any of the models, suggesting that vocal range does not distinguish between speech and singing, an observation also made by List (1963).

There is further strong evidence that a musical phrase is contained in the melody — a new observation in the study of Speech-to-Song illusion. While there is no hard, explicit formulation of what constitutes a musical phrase, we do see some characteristics that are generally common to musical passages, such as the linear models suggestion that the final note is lower than previous note. Huron (1996) analysed the melodies of Western folk songs and found that typically phrases are arched shaped, with the final note more likely to be the lowest of the phrase, and our models appear to have made a similar connection in transforming stimuli. We also find that the models identify illusionary samples from the consonance of the final half of the melody by the negative weights assigned to `mean_diss_3`, implying the presence of some musical resolution to the end of the melody. As several of the models use a selection of the available dissonance measures we can conclude that there is some meaning contained in the structure of the dissonance. Of the weights that can be interpreted, we see that actually a negative weight is assigned to all the dissonance features implying that there should be consonance throughout the whole melody, and not just at the end of the note sequence.

7.4 Experiment Results

We conducted an experiment that aimed to validate our model results further, and to collect a new set of stimuli for additional analysis. We found that the new stimuli are far more likely to transform into music according to the participants of the study, higher than any previous experimental studies. While this could suggest that transforming stimuli are more common than previously realised, there could be other factors at play that account for this. First, the material selected for this experiment was not collected at random — the authors choose the stimuli by hand and did not actively try to obtain an equal balance between non-illusionary and illusionary sounds. This could lead to a bias, where more ‘interesting’ sounding samples are chosen that happen to be transforming. Next, there could be systemic errors in the setup and in the instructional text. While we

[3] Vos and Troost (1989) found that jumps of two or three semitones are by far the most common musical interval, similar to what is found in voice.

gave a positive example of the Speech-to-Song illusion for the participants to familiarise themselves with (namely, the Diana Deutsch phrase), we did not provide a negative example, (e.g. lowest rated stimulus of MCG). The idea was to give the listener some higher expectation of how the illusion sounds like, but as we did not explicitly suggest that some stimuli could not transform at all there could be a tendency to rate sounds higher. We observed that the final rating count across all participants and stimuli was more uniform than in the MCG experiment (compare the boxplots in Figure 4.1 and Figure 6.2) which supports this. Finally, some of the source material for the stimuli are different to those in MCG could account for this difference. For example, generally the non-English samples were rated higher than the other modes of speech. One particular speaker, an interview of a hip-hop music producer, has a notably stable and rhythmic voice that easily transformed. The difference between our set of stimuli and the previous highlights the diversity of vocal styles that have yet to be explored in this illusion.

We tested not only speech but also environmental sounds so the two ends of the slider scale were labelled with the more general text of *Not Musical* and *Very Musical*. While we did offer a description of how these two should be interpreted, the labels themselves are not particularly intuitive or precise — the meaning of 'very musical' could impart an expectation that it should sound exactly like a recording of a fully orchestrated passage for example. In one conversation with a participant after the experiment, she claimed that while she could hear a melody she would not necessarily call it musical, and so scored most of the stimuli rather low. On the other hand, another contributor who is a classically trained, professional singer with experience in non-Western, micro-tonal musical systems found the illusion particularly strong. Interestingly, her strategy involved listening to the melody and assessing how well formed it was to give her final rating, such that melodies that she felt ended abruptly were rated lower than fully completed musical phrases. The diverse range of strategies used by the participants is an ongoing problem of these sorts of perception experiments, where differing ideas and expectations of music result in different ratings and perceptions. We attempted to combat this somewhat by recording the score everyone rated the example stimuli and use this a baseline for 'normalising' their other scores. However, we did not use this as it makes the results incompatible with the results of previous experiments, and did not help in creating a useful threshold in labelling the stimuli as transforming or not. Nonetheless, we believe this is an appropriate step in future experiments since the reason for demonstrating a positive example to the participants is to allow them to calibrate their own internal rating scale, the same should be done to the data they provide.

Since the features of the new stimuli were quite different to the training data, the models did not perform particularly well on the new data, even when evaluated on only the speech stimuli. The top rated models that showed promise during the model selection phase did not manage to maintain their profile here and scored markedly worse, with most only attaining an accuracy slightly above baseline. However, the linear SVM model performs remarkably well despite this and it shows signs of some generalisation by scoring consistently. This suggests that the features it utilises are effective, but that they do not tell the full story as there is still room for improvements. As the non-linear models failed on this data it implies that these over-fitted on the training data, despite a selection procedure that tried to mitigate this. The higher variance across the K-fold cross validation hinted that this could be a problem (especially in the ensemble models, see Table 5.1). It is likely that the kernel SVMs have a greater capacity to over-fit, given their extra degrees of freedom in their construction, and in the way they can manipulate the feature space to be linearly separable.

7.5 Final Thoughts

The result that one of our models is consistent and generalising well is a promising development in the research of this illusion. Not only does it align well to established results, but as we can see some of the new features are utilised that offers a hint on where future research could explore. The features designed to test some aspect of the musical qualities of the song emerged as having some play in how the stimulus transforms that go beyond technical aspects of the sound (e.g. stability, fit to musical key etc), suggesting that when the listener finds music in the stimulus it is not just about these properties, but more about the full musical *idea* that is conveyed by the music that emerges. I believe this is the primary success of our approach — we have demonstrated results that hint at an exciting new perspective onto this illusion. Nonetheless, while these features are inspired by music itself, their use has yet to be tested on actual recordings of music to determine how prevalent, universal or relevant they are in musical compositions, or indeed how they should be interpreted in the musical context.

Ultimately however, these results do not directly address the most fundamental facet to the illusion — the role of repetition in the effect. Although we have identified some general and common characteristics of the speech that elicits the illusion, it is still not clear why repetition is required for it to materialise. The most likely explanation is that multiple repetitions are required for the brain to ‘measure’ some of these features, however there is little study on the nature of how they develop over each iteration. For example, the representation of the melody in the self-similarity matrices (Figure 3.4) for which several of the features are based on could require multiple listens to build this representation in the mind. These matrices are quite dense with information, so it is easy to speculate that this picture ‘grows’ outward from the main diagonal on each loop, as there is more opportunity to make the comparisons between longer distance notes to fill in and complete the matrix, before making the decision of whether the melody is musical or not. Of course, this idea is merely a suggestion to what could be happening in the brain and not based on any previous findings, but research on models of mental representation and memory of melody could be of value in the discussion of this illusion, as presumably the representation of the sound evolves over each repetition. As the Speech-to-Song illusion is stable to the point that once the melody is heard, it cannot easily be 'unheard', as soon as the brain is satisfied with hearing the music the representation ceases to evolve, as if some minima is found and the 'search' for musical information is halted. This, to me, is the heart of the mystery of the illusion.

Unfortunately, while the premise of studying this illusion was to understand further what constitutes music, we are left wondering more about what exactly is the *music* we hear in the Speech-to-Song transformation. The question then is to determine precisely what the brain is listening out for, and I believe that pursuing the answer to this would reveal *why*, exactly, some sounds sometimes behave so strangely.

Appendix A

Feature Summary

Category	Feature	Description
Audio	`stability`	Smoothness of pitch track (only where notes are found), higher values correspond to higher instability
	`length`	Length in seconds of the stimulus, ignoring leading and trailing silence
	`percent_pitched`	Percentage of stimulus where pitch is extracted
Melodic	`max_jump`, `num_jumps`, `mean_jump`, `last_jump`	Basic counts/statistics of note intervals (in semitones) of extracted notes
	`range`	Difference in semitones between highest and lowest note
	`last_note_length`	Length of last note as percentage of stimulus length
	`last_note_lowest`	1 if the last note is the lowest of all notes, 0 otherwise
	`key_fit`	Score of (in range $[0, 1]$) of how well extracted notes fit Krumhansl-Schmuckler key profiles
	`scalar_interval`	Single measure in range $[0, 1]$ on how close all note intervals are to perfect integers, with 1 being all exactly integer, and 0 when all intervals ±0.5 from some integer

	`i_p5`
	`i_3`
	`i_m3`
		Scores (in range $[0, 1]$) on how strong a perfect fifth, major and minor third is present in the notes, where a score 1 means the interval exists exactly

Category	Feature	Description
	`bayesian_distance`	Distance the extracted melody is to the most likely musical melody (according to the Bayesian model)
Rhythm	`onset_variability`	Measure of variability of inter-onset intervals, where 0 means perfectly isochronous onsets, higher values mean greater variety of intervals
	`onset_strength`	Standard deviation of onset strength peaks above 0.5 (after normalising onset envelope)
	`npvi`	Normalised Pairwise Variability Index (nPVI), measures rhythmic variability where 0 means perfectly isochronous onsets, higher values mean greater variety of intervals
Dissonance	`max_dissonance`, `mean_dissonance`, `sd_dissonance`	Statistics on entire dissonance matrix **I**
	`last_mean_diss` `last_max_diss`	Statistics on last column of dissonance matrix **I**

	`mean_diss_1`
	`mean_diss_2`
	`mean_diss_3`
		Average dissonance of notes in first half, between halves, and final half of melody, computed from matrix **II**

	`mean_d_order_1`
	`mean_d_order_2`
	`mean_d_order_3`
	`sd_d_order_1`
	`sd_d_order_2`
	`sd_d_order_3`
		Mean and standard deviation of dissonance between notes $k \in \{1, 2, 3\}$ steps apart (i.e. between each succesive note, every second note, and every third note)

Appendix B

Feature Correlations

Feature	AT	UVA	combined
stability	-0.25 (0.094)	-0.19 (0.015) *	-0.21 (0.002) *
range	-0.39 (0.009) *	-0.10 (0.188)	-0.16 (0.019) *
max_jump	-0.43 (0.003) *	-0.09 (0.228)	-0.16 (0.017) *
num_jumps	-0.17 (0.263)	-0.21 (0.005) *	-0.22 (0.001) *
mean_jump	-0.43 (0.004) *	-0.07 (0.337)	-0.12 (0.068)
max_dissonance	-0.30 (0.046) *	-0.04 (0.572)	-0.08 (0.227)
mean_dissonance	-0.19 (0.217)	-0.07 (0.335)	-0.09 (0.199)
sd_dissonance	-0.18 (0.237)	0.10 (0.181)	0.06 (0.401)
last_mean_diss	-0.04 (0.808)	0.06 (0.435)	0.05 (0.486)
last_max_diss	-0.08 (0.611)	0.06 (0.451)	0.03 (0.637)
percent_pitched	0.04 (0.778)	0.14 (0.072)	0.12 (0.083)
scalar_interval	-0.00 (0.992)	-0.07 (0.359)	-0.06 (0.373)
i_p5	-0.21 (0.161)	-0.14 (0.059)	-0.15 (0.030) *
i_M3	0.02 (0.881)	-0.30 (0.000) *	-0.23 (0.001) *
i_m3	-0.06 (0.719)	-0.14 (0.058)	-0.13 (0.064)
length	-0.21 (0.161)	-0.26 (0.000) *	-0.27 (0.000) *
key_fit	0.11 (0.470)	0.14 (0.074)	0.15 (0.032) *
bayesian_distance	-0.35 (0.017) *	-0.10 (0.181)	-0.15 (0.024) *
onset_variability	-0.20 (0.198)	-0.23 (0.003) *	-0.24 (0.000) *
onset_strength	-0.26 (0.081)	0.05 (0.542)	-0.05 (0.459)
npvi	-0.24 (0.117)	0.05 (0.528)	-0.03 (0.678)
last_note_length	0.06 (0.682)	0.03 (0.692)	0.05 (0.469)
last_note_lowest	-0.00 (0.996)	0.08 (0.301)	0.06 (0.413)
last_jump	-0.08 (0.613)	-0.07 (0.365)	-0.07 (0.287)
mean_diss_1	-0.23 (0.126)	-0.09 (0.229)	-0.10 (0.127)
mean_diss_2	-0.20 (0.187)	-0.02 (0.838)	-0.05 (0.499)
mean_diss_3	-0.02 (0.922)	-0.14 (0.064)	-0.11 (0.112)
mean_d_order_1	-0.22 (0.142)	-0.07 (0.352)	-0.07 (0.286)
mean_d_order_2	-0.03 (0.848)	-0.03 (0.693)	-0.02 (0.722)
mean_d_order_3	-0.20 (0.198)	0.06 (0.470)	-0.00 (0.990)
sd_d_order_1	0.05 (0.758)	0.08 (0.306)	0.06 (0.411)
sd_d_order_2	-0.09 (0.565)	-0.01 (0.872)	-0.04 (0.593)
sd_d_order_3	-0.25 (0.094)	-0.11 (0.154)	-0.14 (0.039) *

Table B.1: Features and their correlations to top3_score of the MCG dataset.

Feature	English	non-English	voice	all
stability	0.11 (0.403)	-0.37 (0.101)	0.03 (0.779)	-0.02 (0.841)
range	-0.04 (0.735)	-0.52 (0.015) *	-0.09 (0.432)	-0.04 (0.663)
max_jump	-0.04 (0.724)	-0.08 (0.734)	-0.03 (0.759)	0.01 (0.955)
num_jumps	0.24 (0.058)	-0.35 (0.120)	0.08 (0.483)	0.00 (0.970)
mean_jump	-0.04 (0.730)	-0.10 (0.654)	-0.01 (0.894)	0.02 (0.879)
max_dissonance	-0.23 (0.069)	-0.54 (0.011) *	-0.33 (0.002) *	-0.34 (0.001) *
mean_dissonance	-0.17 (0.183)	-0.58 (0.006) *	-0.29 (0.008) *	-0.34 (0.001) *
sd_dissonance	-0.15 (0.237)	-0.38 (0.091)	-0.24 (0.026) *	-0.30 (0.003) *
last_mean_diss	-0.11 (0.377)	-0.62 (0.003) *	-0.26 (0.016) *	-0.32 (0.001) *
last_max_diss	-0.20 (0.112)	-0.54 (0.011) *	-0.31 (0.004) *	-0.34 (0.001) *
percent_pitched	0.31 (0.013) *	-0.11 (0.626)	0.28 (0.008) *	-0.00 (0.985)
scalar_interval	-0.07 (0.593)	-0.06 (0.793)	-0.08 (0.447)	-0.08 (0.440)
i_p5	-0.12 (0.338)	-0.38 (0.088)	-0.12 (0.269)	-0.15 (0.147)
i_M3	0.04 (0.780)	-0.02 (0.947)	0.03 (0.795)	0.01 (0.943)
i_m3	-0.18 (0.162)	-0.01 (0.959)	-0.13 (0.219)	-0.19 (0.055)
length	-0.02 (0.879)	-0.04 (0.876)	-0.05 (0.617)	0.08 (0.446)
key_fit	-0.04 (0.780)	0.26 (0.260)	0.04 (0.690)	-0.04 (0.705)
bayesian_distance	-0.20 (0.108)	0.30 (0.181)	-0.05 (0.627)	-0.09 (0.393)
onset_variability	0.09 (0.477)	-0.12 (0.610)	0.02 (0.837)	0.02 (0.834)
onset_strength	-0.29 (0.020) *	0.37 (0.094)	-0.16 (0.142)	-0.19 (0.065)
npvi	-0.14 (0.273)	0.21 (0.358)	-0.05 (0.674)	-0.10 (0.333)
last_note_length	0.02 (0.892)	0.43 (0.052)	0.11 (0.309)	0.08 (0.415)
last_note_lowest	0.06 (0.663)	-0.59 (0.005) *	-0.09 (0.430)	-0.13 (0.216)
last_jump	-0.03 (0.816)	0.28 (0.222)	0.02 (0.891)	0.05 (0.655)
mean_diss_1	-0.08 (0.546)	-0.44 (0.045) *	-0.19 (0.072)	-0.26 (0.010) *
mean_diss_2	-0.15 (0.240)	-0.35 (0.122)	-0.22 (0.039) *	-0.29 (0.003) *
mean_diss_3	-0.18 (0.143)	-0.71 (0.000) *	-0.33 (0.002) *	-0.38 (0.000) *
mean_d_order_1	-0.27 (0.029) *	-0.63 (0.002) *	-0.36 (0.001) *	-0.39 (0.000) *
mean_d_order_2	-0.16 (0.213)	-0.30 (0.184)	-0.21 (0.047) *	-0.29 (0.004) *
mean_d_order_3	-0.17 (0.169)	-0.15 (0.506)	-0.21 (0.056)	-0.28 (0.006) *
sd_d_order_1	0.13 (0.320)	-0.11 (0.625)	0.04 (0.699)	-0.06 (0.549)
sd_d_order_2	0.06 (0.640)	-0.14 (0.543)	-0.01 (0.892)	-0.09 (0.368)
sd_d_order_3	0.14 (0.267)	-0.46 (0.038) *	-0.00 (0.968)	-0.06 (0.586)

Table B.2: Feature correlations of AL for various subsets of stimuli. Environmental stimuli excluded, as there are no significant correlations.

Pearson's correlation coefficient of the relationship between top three score mean and feature values. Brackets contain p-values, measuring the probability of uncorrelated data produces the coefficient, asterisk denotes $p < 0.05$.

Appendix C

Model Features

		SVM kernel			
Feature	logistic	linear	RBF	poly-3	sigmoid
stability	-0.278	-0.411	●	●	●
range	—	—	—	—	—
max_jump	—	—	—	—	—
num_jumps	—	-0.488	—	—	●
mean_jump	—	—	—	—	—
max_dissonance	—	—	—	●	—
mean_dissonance	—	—	—	●	—
sd_dissonance	—	—	—	—	—
last_mean_diss	—	—	—	—	●
last_max_diss	—	—	●	—	—
percent_pitched	0.551	0.703	—	—	●
scalar_interval	—	—	—	—	—
i_p5	—	0.047	●	—	—
i_M3	—	—	●	—	—
i_m3	—	—	—	—	—
length	—	—	—	●	—
key_fit	—	—	—	—	●
bayesian_distance	-0.463	-0.453	●	—	●
onset_variability	-0.278	—	—	—	—
onset_strength	—	—	—	—	—
npvi	—	—	—	●	—
last_note_length	-0.406	-0.545	—	●	●
last_note_lowest	—	0.217	—	●	●
last_jump	-0.154	—	—	—	—
mean_diss_1	—	—	—	—	—
mean_diss_2	—	—	—	—	—
mean_diss_3	-0.242	-0.323	—	—	●
mean_d_order_1	—	—	—	—	—
mean_d_order_2	—	—	●	—	—
mean_d_order_3	—	—	—	—	—
sd_d_order_1	—	—	—	●	—
sd_d_order_2	-0.004	-0.194	—	●	—
sd_d_order_3	-0.439	—	●	—	●

Table C.1: The features used by each model, where the coefficients assigned by the linear models are shown.

References

M. F. Bassett and Charles. J. Warne. On the lapse of verbal meaning with repetition. *The American Journal of Psychology*, 30(4):415–418, 1919. ISSN 00029556. doi: 10.2307/1413679.

Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 41(3):407–434, 2013. doi: 10.1007/s10844-013-0258-3.

Emmanuel Bigand, Richard Parncutt, and Fred Lerdahl. Perception of musical tension in short chord sequences: The influence of harmonic function, sensory dissonance, horizontal motion, and musical training. *Perception & Psychophysics*, 58(1):125–141, 1996.

Sebastian Böck, Florian Krebs, and Markus Schedl. Evaluating the online capabilities of onset detection methods. In *ISMIR*, pages 49–54, 2012.

Paul Boersma and David Weenink. Praat: Doing phonetics by computer (version 5.3. 82)[computer software]. *Amsterdam: Institute of Phonetic Sciences*, 2012.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

Ali T. Cemgil, Bert Kappen, and David Barber. Generative model based polyphonic music transcription. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*, pages 181–184, 2003. doi: 10.1109/ASPAA.2003.1285861.

Srikanth Cherla, Son N. Tran, Tillman Weyde, and Artur S. d'Avila Garcez. Hybrid long- and short-term models of folk melodies. In *Proceedings of the 16th ISMIR Conference, Málaga, Spain*, pages 584–590, 10 2015.

Dante R. Chialvo. How we hear what is not there: A neural mechanism for the missing fundamental illusion. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 13 (4):1226–1230, 2003. doi: 10.1063/1.1617771.

Nathaniel Condit-Schultz. Deconstructing the nPVI: A Methodological Critique of the Normalized Pairwise Variability Index as Applied to Music. *Music Perception*, 36(3): 300–313, 02 2019. ISSN 0730-7829. doi: 10.1525/mp.2019.36.3.300.

Darrell Conklin and Ian H. Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73, 1995. doi: 10.1080/09298219508570672.

Grosvenor W. Cooper and Leonard B. Meyer. *The rhythmic structure of music.* University of Chicago Press, 1960.

Matthew L. Cooper and Jonathan Foote. Automatic music summarization via similarity analysis. In *ISMIR*, 2002.

Marieve Corbeil, Sandra Trehub, and Isabelle Peretz. Speech vs. singing: infants choose happier sounds. *Frontiers in Psychology*, 4:372, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00372.

Bas Cornelissen. Singing speech, or looking for the music in speech. course project, 2015.

Bas Cornelissen, Makiko Sadakata, and Henkjan Honing. Categorization in the speech to song transformation (STS). In *International conference on music perception and cognition*, page 386, 2016.

Meagan E. Curtis and Jamshed J. Bharucha. Memory and Musical Expectation for Tones in Cultural Context. *Music Perception*, 26(4):365–375, 04 2009. ISSN 0730-7829. doi: 10.1525/mp.2009.26.4.365.

Anne Cutler. *Linguistic rhythm and speech segmentation.*, pages 157–166. Macmillan Education UK, London, 1991. ISBN 978-1-349-12670-5. doi: 10.1007/978-1-349-12670-5_14.

Nivja H. de Jong and Ton Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, 2009. doi: 10.3758/BRM.41.2.385.

Diana Deutsch. An auditory illusion. *The Journal of the Acoustical Society of America*, 55(S1):S18–S19, 1974a. doi: 10.1121/1.1919587.

Diana Deutsch. An illusion with musical scales. *The Journal of the Acoustical Society of America*, 56(S1):S25–S25, 1974b. doi: 10.1121/1.1914084.

Diana Deutsch. A Musical Paradox. *Music Perception*, 3(3):275–280, 04 1986. ISSN 0730-7829. doi: 10.2307/40285337.

Diana Deutsch. *Phantom words and other curiosities*. Philomel Records, 2003.

Diana Deutsch, Trevor Henthorn, and Rachael Lapidis. Illusory transformation from speech to song. *The Journal of the Acoustical Society of America*, 129(4):2245–2252, 2011.

John S. Downie, Kris West, Andreas Ehmann, and Emmanuel Vincent. The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005): preliminary overview. In *6th Int. Conf. on Music Information Retrieval (ISMIR)*, pages 320–323, London, United Kingdom, September 2005. URL https://hal.inria.fr/inria-00544675.

Daniel P. W. Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007. doi: 10.1080/09298210701653344.

Leonhard Euler. *Tentamen novae theoriae musicae ex certissimis harmoniae principiis dilucide expositae*. ex typographia Academiae scientiarum, 1739.

Simone Falk and Tamara Rathcke. On the speech-to-song illusion: Evidence from German. In *Speech Prosody 2010-Fifth International Conference*, May 2010.

Simone Falk, Tamara Rathcke, and Simone Dalla Bella. When speech sounds like music. *Journal of Experimental Psychology: Human Perception and Performance*, 40:1491–1506, 2014. doi: 10.1037/a0036858.

Morwaread M. Farbood. A Parametric, Temporal Model of Musical Tension. *Music Perception*, 29(4):387–428, 04 2012. ISSN 0730-7829. doi: 10.1525/mp.2012.29.4.387.

Jonathan Foote. Visualizing music and audio using self-similarity. In *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, MULTIMEDIA '99, pages 77—80, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581131518. doi: 10.1145/319463.319472.

Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*, volume 1, pages 452–455, 2000.

Jonathan Foote and Shingo Uchihashi. The beat spectrum: A new approach to rhythm analysis. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, pages 881–884. IEEE, 2001.

Zhouyu Fu, Guojun Lu, Kai M. Ting, and Dengsheng Zhang. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13(2):303–319, 2011.

Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C. Smith. Query by humming: musical information retrieval in an audio database. In *Proceedings of the third ACM international conference on Multimedia*, pages 231–236, 1995.

Dafydd Gibbon and Ulrike Gut. Measuring speech rhythm. In *Seventh European Conference on Speech Communication and Technology*, 2001.

Masataka Goto. A chorus-section detecting method for musical audio signals. In *2006 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 14, pages 1783–1794, 9 2006.

Emily Graber. Speech to song classification. course project, 2015. URL http://cs229.stanford.edu/proj2015/143_report.pdf.

Emily Graber, Rhimmon Simchy-Gross, and Elizabeth H. Margulis. Musical and linguistic listening modes in the speech-to-song illusion bias timing perception and absolute pitch memory. *The Journal of the Acoustical Society of America*, 142(6):3593–3602, 2017. doi: 10.1121/1.5016806.

Graham Grindlay and Daniel P. W. Ellis. Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1159–1169, 2011.

Gerben Groenveld, John Burgoyne, and Makiko Sadakata. I still hear a melody: investigating temporal dynamics of the speech-to-song illusion. *Psychological Research*, 01 2019. doi: 10.1007/s00426-018-1135-z.

Ryan Groves. Automatic harmonization using a hidden semi-Markov model. In *Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2013.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladmir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002. doi: 10.1023/A:1012487302797.

Goffredo Haus and Emanuele Pollastri. An audio front end for query-by-humming systems. In *Proceedings of International Symposium on Music Information Retrieval*, pages 36–43, 2001.

John L. Hodges. The significance probability of the Smirnov two-sample test. *Arkiv för Matematik*, 3(5):469–486, 1958.

Heike Hofmann, Karen Kafadar, and Hadley Wickham. Letter-value plots: Boxplots for large data. Technical report, had.co.nz, 2011.

David Huron. The melodic arch in Western folksongs. *Computing in Musicology*, 10, 03 1996.

William Hutchinson and Leon Knopoff. The acoustic component of western consonance. *Interface*, 7(1):1–29, 1978. doi: 10.1080/09298217808570246.

Nori Jacoby, Eduardo A. Undurraga, Malinda J. McPherson, Joaquín Valdés, Tomás Ossandón, and Josh H. McDermott. Universal and non-universal features of musical pitch perception revealed by singing. *Current Biology*, 29(19):3229–3243.e12, 2019. ISSN 0960-9822. doi: 10.1016/j.cub.2019.08.020.

Kankamol Jaisin, Rapeepong Suphanchaimat, Mauricio A. Figueroa Candia, and Jason D. Warren. The speech-to-song illusion is reduced in speakers of tonal (vs. non-tonal) languages. *Frontiers in Psychology*, 7:662, 2016. ISSN 1664-1078. doi: 10.3389/fpsyg.2016.00662.

Leon A. Jakobvits. *Effects of repeated stimulation on cognitive aspects of behavior: some experiments on the phenomenon of semantic satiation.* PhD thesis, McGill University, Montreal, April 1962.

Akio Kameoka and Mamoru Kuriyagawa. Consonance theory part II: Consonance of complex tones and its calculation method. *The Journal of the Acoustical Society of America*, 45(6):1460–1469, 1969. doi: 10.1121/1.1911624.

Jong. W. Kim, Justin Salamon, Peter Li, and Juan P. Bello. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165, 2018. doi: 10.1109/ICASSP.2018.8461329.

Anssi Klapuri, Jouni Paulus, and Meinard Müller. Audio-based music structure analysis. In *ISMIR, in Proc. of the Int. Society for Music Information Retrieval Conference*, 2010.

Ron Kohavi. Feature subset selection as search with probabilistic estimates. In *AAAI fall symposium on relevance*, volume 224, 1994.

Carol L. Krumhansl and Edward J. Kessler. Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological review*, 89(4):334, 1982.

Edward W. Large and John F. Kolen. Resonance and the perception of musical meter. *Connection Science*, 6(2-3):177–208, 1994. doi: 10.1080/09540099408915723.

Fred Lerdahl and Ray S. Jackendoff. *A generative theory of tonal music.* MIT press, 1983.

Carole Leung and De-Hui R. Zhou. Pace, emotion, and language tonality on speech-to-song illusion. *Preprints*, 2018. doi: 10.20944/preprints201808.0522.v1.

Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content-based music genre classification. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 282—-289, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581136463. doi: 10.1145/860435.860487.

Joseph C. R. Licklider. A duplex theory of pitch perception. *The Journal of the Acoustical Society of America*, 23(1):147–147, 1951. doi: 10.1121/1.1917296.

George List. The boundaries of speech and song. *Ethnomusicology*, 7(1):1–16, 1963. ISSN 00141836.

Elizabeth Low and Esther Grabe. Prosodic patterns in Singapore English. In *Proceedings of the International Congress of Phonetic Sciences, Stockholm*, volume 3, pages 636–639, 1995.

Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. ISSN 00034851.

Elizabeth H. Margulis. When program notes don't help: Music descriptions and enjoyment. *Psychology of Music*, 38(3):285–302, 2010. doi: 10.1177/0305735609351921.

Elizabeth H. Margulis. Repetition and emotive communication in music versus speech. *Frontiers in Psychology*, 4:167, 2013a. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00167.

Elizabeth H. Margulis. Aesthetic responses to repetition in unfamiliar music. *Empirical Studies of the Arts*, 31(1):45–57, 2013b. doi: 10.2190/EM.31.1.c.

Elizabeth H. Margulis. *On repeat: How music plays the mind.* Oxford University Press, 2014.

Elizabeth H. Margulis and Rhimmon Simchy-Gross. Repetition Enhances the Musicality of Randomly Generated Tone Sequences. *Music Perception*, 33(4):509–514, 04 2016. ISSN 0730-7829. doi: 10.1525/mp.2016.33.4.509.

Elizabeth H. Margulis, Rhimmon Simchy-Gross, and Justin L. Black. Pronunciation difficulty, temporal regularity, and the speech-to-song illusion. *Frontiers in Psychology*, 6:48, 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.00048.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in science conference*, volume 8, 2015.

Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588): 746–748, 1976. doi: 10.1038/264746a0.

Rodger J. McNab, Lloyd A. Smith, and Ian H. Witten. Signal processing for melody transcription. *Computer Science Working Papers*, 1995.

Scott Menard. *Applied logistic regression analysis*, volume 106. Sage, 2002.

Leonard B. Meyer. *Emotion and Meaning in Music.* University of Chicago Press, 1956.

Leonard B. Meyer. *Style and Music: Theory, History, and Ideology.* University of Chicago Press, 1989.

Mehryar Mohri, Pedro J. Moreno, and Eugene Weinstein. Efficient and robust music identification with weighted finite-state transducers. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1):197–207, 2010.

Christoph Molnar. *Interpretable Machine Learning.* Christoph Molnar, 2019. https://christophm.github.io/interpretable-ml-book/.

Nasim Nematzadeh and David M.W. Powers. A quantitative analysis of tilt in the café wall illusion: a bioplausible model for foveal and peripheral vision. In *2016 International Conference on Digital Image Computing: Techniques and Applications (Dicta)*, pages 1–8. IEEE, 2016. doi: 10.1109/DICTA.2016.7796995.

Yizhao. Ni, Matt McVicar, Raul Santos-Rodriguez, and Tijl De Bie. An end-to-end machine learning system for harmonic analysis of music. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1771–1783, 2012.

Aniruddh D. Patel. Language, music, syntax and the brain. *Nature neuroscience*, 6(7): 674–681, 6 2003. doi: 10.1038/nn1082.

Marcus Pearce. *The construction and evaluation of statistical models of melodic structure in music perception and composition.* PhD thesis, City University London, 2005.

Marcus Pearce and Geraint Wiggins. Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, 33(4):367–385, 2004. doi: 10.1080/0929821052000343840.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10, 06 2000.

Reinier Plomp and Willem J. M. Levelt. Tonal consonance and critical bandwidth. *The Journal of the Acoustical Society of America*, 38(4):548–560, 1965. doi: 10.1121/1.1909741.

Daniel Pressnitzer, Stephen McAdams, Suzanne Winsberg, and Joshua Fineberg. Perception of musical tension for nontonal orchestral timbres and its relation to psychoacoustic roughness. *Perception & Psychophysics*, 62(1):66–80, 2000.

Martin A. Rohrmeier and Stefan Koelsch. Predictive information processing in music cognition. a critical review. *International Journal of Psychophysiology*, 83(2):164–175, 2012. ISSN 0167-8760. doi: 10.1016/j.ijpsycho.2011.12.010. Predictive information processing in the brain: Principles, neural mechanisms and models.

Renata L. Rosa, Demsteneso Z. Rodriguez, and Graca Bressan. Music recommendation system based on user's sentiments extracted from social networks. *IEEE Transactions on Consumer Electronics*, 61(3):359–367, 2015.

Jess Rowland, Anna Kasdan, and David Poeppel. There is music in repetition: Looped segments of speech and nonspeech induce the perception of music in a time-dependent manner. *Psychonomic Bulletin & Review*, 26:583–590, 2019. doi: 10.3758/s13423-018-1527-5.

Matti P. Ryynänen and Anssi Klapuri. Polyphonic music transcription using note event modeling. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, pages 319–322, 2005. doi: 10.1109/ASPAA.2005.1540233.

Matti P. Ryynänen and Anssi Klapuri. Transcription of the singing melody in polyphonic music. In *ISMIR*, pages 222–227. Citeseer, 2006.

Pierre Schaeffer. *A la Recherche d'une Musique Concrète [In search of musique concrète]*. Éditions du Seuil, 1952. ISBN 978-2-02-002572-0.

Helmut Schaffrath. The Essen folksong collection in the humdrum kern format (D. Huron, ed.). *Menlo Park, CA: Center for Computer Assisted Research in the Humanities*, 1995.

Donia R. Scott, Steve D. Isard, and Benedicte de Boysson-Bardies. On the measurement of rhythmic irregularity: a reply to Benguerel. *Journal of Phonetics*, 14(2):327–330, 1986. ISSN 0095-4470. doi: 10.1016/S0095-4470(19)30659-X.

William A. Sethares. *Tuning, timbre, spectrum, scale.* Springer Science & Business Media, 2005.

Samuel S. Shapiro and Martin B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965. ISSN 00063444. doi: 10.2307/2333709.

Roger N. Shepard. Geometrical approximations to the structure of musical pitch. *Psychological review*, 89(4):305, 1982.

Rhimmon Simchy-Gross and Elizabeth H. Margulis. The sound-to-music illusion: Repetition can musicalize nonspeech sounds. *Music & Science*, 1, 2018. doi: 10.1177/2059204317731992.

David Temperley. A probabilistic model of melody perception. *Cognitive Science*, 32 (2):418–444, 2008. doi: 10.1080/03640210701864089.

Brian Thompson. Discrimination between singing and speech in real-world audio. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 407–412. IEEE, 2014.

Adam Tierney, Fred Dick, Diana Deutsch, and Marty Sereno. Speech versus Song: Multiple Pitch-Sensitive Areas Revealed by a Naturally Occurring Musical Illusion. *Cerebral Cortex*, 23(2):249–254, 02 2012. ISSN 1047-3211. doi: 10.1093/cercor/bhs003.

Adam Tierney, Aniruddh D. Patel, and Mara Breen. Acoustic foundations of the speech-to-song illusion. *Journal of Experimental Psychology: General*, 147:888–904, 06 2018a. doi: 10.1037/xge0000455.

Adam Tierney, Aniruddh D. Patel, and Mara Breen. Repetition Enhances the Musicality of Speech and Tone Stimuli to Similar Degrees. *Music Perception*, 35(5):573–578, 06 2018b. ISSN 0730-7829. doi: 10.1525/mp.2018.35.5.573.

Christina M. Vanden Bosch der Nederlanden, Erin E. Hannon, and Joel S. Snyder. Everyday musical experience is sufficient to perceive the speech-to-song illusion. *Journal of experimental psychology: General*, 144(2):e43, 2015.

Vladimir N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, 1998. ISBN 978-0-471-03003-4.

Pantelis Vassilakis. Auditory roughness estimation of complex spectra —– roughness degrees and dissonance ratings of harmonic intervals revisited. *The Journal of the Acoustical Society of America*, 110(5):2755–2755, 2001. doi: 10.1121/1.4777600.

Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf. A primer on kernel methods. *Kernel methods in computational biology*, 47:35–70, 2004.

Hermann Von Helmholtz. *On the Sensations of Tone as a Physiological Basis for the Theory of Music.* London: Longmans, Green and Company, 1875.

Piet G. Vos and Jim M. Troost. Ascending and descending melodic intervals: Statistical findings and their perceptual relevance. *Music Perception*, 6(4):383–396, 1989.

Richard M. Warren and Richard L. Gregory. An auditory analogue of the visual reversible figure. *The American journal of psychology*, 1958.

Raymond P. Whorley and Darrell Conklin. Music generation from statistical models of harmony. *Journal of New Music Research*, 45(2):160–183, 2016. doi: 10.1080/09298215.2016.1173708.

Robert B. Zajonc. Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2, Pt.2):1–27, 1968. doi: 10.1037/h0025848.

Harry Zhang. Exploring conditions for the optimality of naive Bayes. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(02):183–198, 2005. doi: 10.1142/S0218001405003983.

Shuo Zhang. Speech-to-song illusion: evidence from MC. *Sino-European Winter School of Logic, Language, and Computation ([SELLC2010SS])*, 2010.

Xin Zhang and Zbigniew W. Ras. Analysis of sound features for music timbre recognition. In *2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*, pages 3–8, 2007. doi: 10.1109/MUE.2007.85.