COMPUTATIONALLY MEASURABLE TEMPORAL DIFFERENCES BETWEEN SPEECH AND SONG

by

David Bruce Gerhard
B.Sc.Comp.E., University of Manitoba, 1996

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in the School
of
Computing Science

SIMON FRASER UNIVERSITY
April 2003

APPROVAL

Name: David Bruce Gerhard

Degree: Doctor of Philosophy

Title of thesis: Computationally Measurable Temporal Differences Between Speech and Song

Examining Committee: Dr. Anoop Sarkar
Chair

Senior Supervisor: Dr. Fred Popowich, Professor
Simon Fraser University, Computing Science

Supervisor: Dr. Pierre Zakarauskas
Chief Technology Officer, Wavemakers Inc.

Supervisor: Dr. Tom Perry, Professor
Simon Fraser University, Linguistics

SFU Examiner: Dr. Binay Bhattacharya
Computing Science

External Examiner: Dr. Philippe Depalle
Faculty of Music, McGill University

Date Approved: ______________________________

Abstract

Automatic audio signal classification is one of the general research areas in which algorithms are developed to allow computer systems to understand and interact with the audio environment. Human utterance classification is a specific subset of audio signal classification in which the domain of audio signals is restricted to those likely to be encountered when interacting with humans. Speech recognition software performs classification in a domain restricted to human speech, but human utterances can also include singing, shouting, poetry and prosodic speech, for which current recognition engines are not designed.

Another recent and relevant audio signal classification task is the discrimination between speech and music. Many radio stations have periods of speech (news, information reports, commercials) interspersed with periods of music, and systems have been designed to search for one type of sound in preference over another. Many of the current systems used to distinguish between speech and music use characteristics of the human voice, so such systems are not able to distinguish between speech and music when the music is an individual unaccompanied singer.

This thesis presents research into the problem of human utterance classification, specifically differentiation between talking and singing. The question is addressed: "Are there measurable differences between the auditory waveforms produced by talking and singing?" Preliminary background is presented to acquaint the reader with some of the science used in the algorithm development. A corpus of sounds was collected to study the physical and perceptual differences between singing and talking, and the procedures and results of this collection are presented. A set of 17 features is developed to differentiate between talking and singing, and to investigate the intermediate vocalizations between talking and singing. The results of these features are examined and evaluated.

To Fred and Hazel, Bruce and Mary.

"Music is the universal language of mankind."

— Longfellow–*Outre-Mer*

"Talk is cheap."

— English Proverb

Acknowledgments

I would like to thank my supervisors, especially Fred, my senior supervisor, for his unending and inimitable help and advice. I would also like to thank my wife, Tricia, for her love and support, and my parents for their continuing encouragement (read “elbow to the ribs”).

Thanks also go to the many people who have helped me with suggestions and advice throughout the course of this work. Without your presence I would still be wandering.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Technical Background

Sound classification, on first analysis, seems a straightforward task. We humans don't often think about how we do it, but we can tell that the source of a pattern of air vibrations behind us is a door closing and not a glass breaking. Sound analysis itself dates back to antiquity, often motivated by the desire to understand musical sound. The relationship between pitch and frequency was determined experimentally by Galileo and Mersenne in 1636 [57].

Teaching computers to recognize sounds has been a popular research task since computers became useful tools for analyzing data. For example, automatic pitch detection methods date back to 1962. Audio analysis research is inter- and multi-disciplinary, and different parts of the sound classification task are interesting for different research domains, including physics, psychology, audiology, music, cognitive science, and philosophy. "If a tree falls in the forest, and there is no-one around to hear it, does it make a sound?" If sound does not exist apart from our perception of it, then the understanding of the physical properties of sound is intimately connected to the understanding of our perception of sound, and any study of sound must also include a study of the perceptual science of sound as well.

We would like to have devices that can hear for us for a number of reasons: to make our lives easier by listening to things that are either too dangerous or too repetitive for us humans; to listen more closely or more accurately than us humans; or to be more able to interact with us humans. A computer could listen for the sound of a crying baby and alert the parent. A computer could listen to and transcribe the speech of an auctioneer for accurate record keeping and accounting. A computer could listen to the radio overnight, and compile a custom program for the listener when the alarm goes off in the morning, containing relevant news, weather, and traffic conditions, devoid of advertisements, or complete with ads, or with a set of ads tailored to the personality of the listener. A computer could index a database of sound effects for retrieval by a sound engineer working on a movie. Sound can also be used as a sensor in factories and industrial areas, providing information about a binding saw blade or a squeaking brake, which otherwise might be missed until failure.

Human utterance classification is a subset of the sound classification problem where the task is to decide which of a set of categories best describes a human vocal sound. This could include such utterances as laughter, coughs and burps, but for the purpose of this work we will restrict human utterances to those containing informational or emotional context, specifically speech, song, and utterances between these two.

1.0.1 Unifying Principles

In selecting features and analysis tools for this classification task, two principles have helped to sharpen the focus and direct the choices made. The first principle is a preference and the second is an observation. Neither of these principles are particularly rigorous in their motivation.

Principle 1. (Time-domain preference) *When two computational methods are available which perform the same task, prefer the one most closely related to time-domain processing.*

Principle 2. (Repetition observation) *Production and observation of human vocal sounds tends to proceed in cycles, the presence and frequency of which define the phenomena being observed or produced.*

Principle 1 is a variant on Ockham's Razor which states that when two equally complete explanations of a phenomenon are available, prefer the simpler. To say that time-domain techniques are simpler is probably not a completely true statement, but the motivation behind this principle is that removing a single stage in a processing pipeline (that of computing a spectral analysis of some form) is beneficial, and if a system could be designed which needed no spectral analysis whatsoever, an efficiency can be realized because the processing could be easily ported to hardware, and a more rapid analysis would result.

An argument against Principle 1 would be that the human auditory system processes sounds in a spectral context, and specifically that the cochlea does frequency band analysis. A system designed specifically to imitate the human auditory system would have to do at least some spectral analysis. The problems with this are twofold. First, as described later in Section 1.3, current spectral techniques are not particularly good at resolving the frequency information in sampled signals. Second, if one of the goals of this type of research is to duplicate human behavior, some would say that the process is unimportant as long as the desired responses are generated.

This thesis concentrates on temporal features as well as temporal methods for feature extraction. This is done primarily for simplicity—if the design target of a system is a solid-state device, temporal methods have the potential to be faster because they do not require the FFT processing step, and because spectral processing often requires higher-dimensional considerations.

Principle 2 is an observation that came from three seemingly independent phenomenon: pitch, vibrato and rhythm. Pitch is the human experience of a sound signal which oscillates at a detectable frequency. When the pitch itself oscillates, this is referred to as vibrato, a technique used by singers and instrumentalists to improve perceptual quality and be heard above an orchestra. Rhythm as a concept is more difficult to quantify computationally, but in terms of human perception of music, rhythm can be defined as the predictable repetition of a series of acoustical events. It should also be noted here, however, that it is often the unpredictable acoustical event within the context of a rhythm that is musically interesting. Rhyme is another human perceptual phenomenon which relies on repetition, but a specific frequency of repetition is not necessary for the perception of rhyme, so this phenomenon does not fall strictly into the domain of Principle 2.

These two principles will be referred to throughout this thesis as a way to narrow the focus within the wide and diverse research area of audio signal classification, human utterance classification, and speech/song continuum classification.

The remainder of this chapter presents motivation for the research presented in this thesis, as well as technical background on sound and hearing, and techniques researchers have used to study auditory signals.

1.1 Motivation

Humans classify audio signals all the time without conscious effort. Recognizing a voice on the telephone, telling the difference between a telephone ring and a doorbell ring—these are tasks that we don't consider very difficult. Problems do arise in the presence of noise, or when the sound is weak or similar to another sound. The classic example problem in audio signal classification is called the "cocktail party problem". This describes the phenomenon where humans are able to carry on a conversation in a room with many other conversations and background noise, the sum of the noise often being louder than the local conversation. As the noise gets louder, the intelligibility of the conversation degrades until the situation becomes what could be called the "techno dance party problem" where the surrounding noise is so great that conversation becomes impossible.

The motivation behind the work presented in this thesis is threefold. One motivation is perceptual and physiological—can we describe what humans do and how they do it? A second motivation is imitative—can we develop a computer program that can emulate human listening. A third motivation is augmentative—can we develop a computer program that can do *more* than humans can do. These tasks are incremental: we must first understand what humans do before we can build a machine capable of imitating or improving on that performance.

The perceptual and physiological task is to understand and quantify human utterances, human perception and categorical classification. How do we produce these sounds, how do they travel from a mouth to an ear, how do we separate these sounds into different categories, are these categories discrete or continuous, how do we interpret and extract information and knowledge from these categories, and how do we use this information and knowledge to plan, make decisions and interact with the world around us.

The difference between speech and song can be likened to the difference between walking and dancing. Both perform the intended action (conveying information; traveling), but a certain "style" or "presence" has been added to one which makes it qualitatively different from the other. Part of the motivation of this research is to try to understand and quantify some of these differences. A difficulty is that these differences seem to be ill-defined and subjective. As will be seen in Chapter 2, people disagree on what is singing when compared to talking, and on what features and characteristics can be used to define the differences. Part of the problem is likely to be linguistic: when describing subjective phenomena, words may mean different things to different people. When describing an utterance, people appear to agree on general concepts, but when they are asked to define the concepts and why they made these decisions, people's understandings start to differ.

The imitative and augmentative tasks are to build devices that can do what humans can do, or can help humans do more or better than they could without the device. Audio signal classification encompasses many different application goals, some of which are imitative and some of which are augmentative. Speech recognition is a particularly difficult imitative goal, since it encompasses many levels of human cognitive processing. Intelligent hearing aids are an augmentative goal since they allow an individual human to hear better than he or she could without the device. This identifies a pair of subsets for the augmentative goal: devices for augmenting the hearing of one specific individual, and devices for augmenting the general ability of human hearing.

Other applications include multimedia database applications such as annotation (for example automatic sound clustering) and retrieval (for example query-by-humming), psychoacoustic therapy, automatic music transcription and consumer electronics applications, as well as speech detection, speaker verification, emotional content analysis, and even auditory user interfaces with conversational awareness, such as the fictional computers in “Star Trek” and “2001, A Space Odyssey”.

1.2 The Components of Sound

In its simplest form, sound is a pattern of air pressure variations transferred to our eardrum, through a series of connected bones, to a frequency resolving organ and then to our brain. While sound itself is a four-dimensional phenomenon (air pressure over time in three-space) human perception of sound is two-dimensional, consisting of amplitude and time. It can become three-dimensional when we measure it in two physical locations, in the case of stereo listening. Stereo listening is principally important for sound localization, and while that can be useful for classification in an auditory scene, for the purposes of this work we will restrict the investigation of sound to a single listening place, and a single ear (or an average between two ears).

In the real world, sound is a continuous function - it is possible to measure the air pressure level and how it changes at any instant in time. This much detail is not necessary when using computers to analyze sounds - humans cannot hear instantaneously fast. For machine listening intended to augment human listening, it may be useful to analyze at higher and lower frequencies than humans can hear, but for imitative applications, it is sufficient to analyze through human perceptible frequencies. The question then is what *can* humans hear, and how can we design computer programs which can "hear" the same things.

Psychological testing shows that the extreme range of human hearing is between 20 Hz[1] and 20,000 Hz. Average human hearing ranges are often limited to a subset of this frequency range. To make a computer program "hear" anything at all, we must first convert the sound waveform into a sequence of numbers. There are two problems with this - if we do not take enough measurements per second, the higher frequencies will not be captured, and if we do not make the numbers accurate enough, the computer will not be able to distinguish between enough air pressure levels. The first phenomenon is called the sampling frequency, and the second is called quantization.

Quantization can be solved by allocating more bits to each sample. 16 bits per sample provides sufficient resolution for $2^{16} = 65,536$ possible air pressure levels at each sample. The sampling rate question is slightly more difficult. The Shanon theory states that to fully represent a waveform up to a frequency of f, a sampling rate of $2 \times f$ must be used. To fully represent a waveform up to 20,000 Hz, or 20 kHz the computer must take a measurement 40,000 times per second, or once every 50 microseconds. Modern CD players use sound sampled at 44 kHz, digital audio tape (DAT) uses a sampling rate of 48 kHz, and digital studios sometimes use 96 kHz. These higher sampling rates are used to allow "filter room" - it is very difficult to design a filter which cuts the sound off exactly at 20 kHz, and the more room between the signal cutoff and the sampling cutoff, the easier it is to design filters and other signal processing systems.

Some listeners have observed that CD sound is not "perfect" even though it is theoretically capable of reproducing all audible frequencies. It is possible that humans can detect frequencies above the theoretical audible limit, when in the context of lower-frequency sounds, or perhaps through bone conduction or other non-auditory perception. The original psychological tests were done with pure sinusoids, waveforms that have only one frequency component in them. Future research will tell if we are, in fact, better at hearing than we thought we were.

[1] One *Hertz* (Hz) is one cycle per second, and corresponds to a waveform which completely repeats itself once in one second.

It is important at this point to differentiate between three closely related terms: *waveform*, *signal* and *noise*. A waveform is the raw data to be analyzed and can contain signal, noise, both or neither. A signal is a waveform or a portion of a waveform considered to contain some desired information and/or properties, while noise is the undesired component of the waveform which often hides or blurs the signal being investigated. The same waveform can have many different signal/noise interpretations, depending on the properties or phenomena being investigated.

1.3 Spectral Analysis Techniques

This thesis concentrates somewhat on time-domain (temporal) signal processing techniques (Principle 1). However, this assumes that a temporal technique is *available* to replace a frequency-domain (spectral) technique. If the spectral technique turns out to be more efficient, or if no temporal technique exists, then spectral techniques will be used, and so it is important to be familiar with spectral techniques as well. Spectral and temporal techniques are compared for a specific task in Chapter 3.

It has been apparent for many years that a useful procedure in the study of sound is the study of the spectral *components* that make up the sound. Frequency analysis of waveforms has been used in many fields other than sound, such as electrical engineering, geology and physics. The amount of information that is available in a one-dimensional waveform at an instantaneous point in time is minimal compared to the amount of information contained in the history of the waveform over time. It is this recent history of the waveform that spectral techniques investigate. The following sections describe some common spectral techniques, their implementation and some evaluation.

1.3.1 Fourier Analysis

A *sinusoid* is a mathematical function that describes the simplest repetitive motion in nature. A ball on a rubber band will descend and slow as the band stretches, stop when the gravitational acceleration equals the restoring force of the rubber band, begin to ascend and stop again when the restoring force is zero and the gravitational acceleration equals the momentum. This system is called a simple harmonic oscillator. The sinusoidal motion that it creates is found in many different forms in nature, and in particular in the varying air pressure of sound waves.

Humans and other vertebrates have an organ called the *cochlea* inside the ear, which analyzes sound by spreading it out into its component sinusoids. One end of this organ is sensitive to low frequency sinusoids, and one end is sensitive to higher frequencies. When a sound arrives, different parts of the organ react to the different frequencies present in the sound, generating nerve impulses which are interpreted by the brain.

Spectral analysis techniques, of which Fourier analysis is the earliest and most commonly used today, represent a waveform as a sum of sinusoids. The sum-of-sinusoids representation displays information about the harmonic makeup of the waveform, something that the human auditory system is particularly good at extracting as well. It is probably because of the human ability to recognize a sound signal using its harmonic makeup that so much work has been put into Fourier-type algorithms.

The Fourier *transform* is an algorithm that generates the Fourier representation of a waveform. The Fourier representation contains a list of sinusoid functions, identified by frequency, and each sinusoid has an associated amplitude and *phase*. The phase of a waveform is the start location of the sinusoid relative to some specific zero. Phase is measured as an angle indicating some part of a complete oscillation. A sinusoid with a phase of 0 radians is identical to a sinusoid with a phase of 2π radians. These waveforms are said to be "in phase". A sinusoid with a phase of π radians is opposite to a sinusoid with a phase of 0 radians. These waveforms are said to be "out of phase" and if added together, would cancel each other out.

It has been shown that the ear is "phase deaf" [16], meaning that two sounds with the same frequency composition but with different phase compositions sound the same to the human ear. For this reason, the phase component of the Fourier representation is sometimes discarded when analyzing sound. However, changes in the phase spectrum of a waveform over time *are* perceptible. This change in phase is perceived as a shift in timbre, not in pitch, so it is not unreasonable to throw away the phase component of a Fourier representation if that representation were to be used only for pitch detection. It should be noted here that rapidly varying phase can result in perceived changes in the pitch, however this rarely happens in human vocal production. Figure 1.1 shows an example of a pair of waveforms with equivalent amplitude spectra but different phase spectra. The fundamental frequency is the same, and so the pitch measure will be the same.

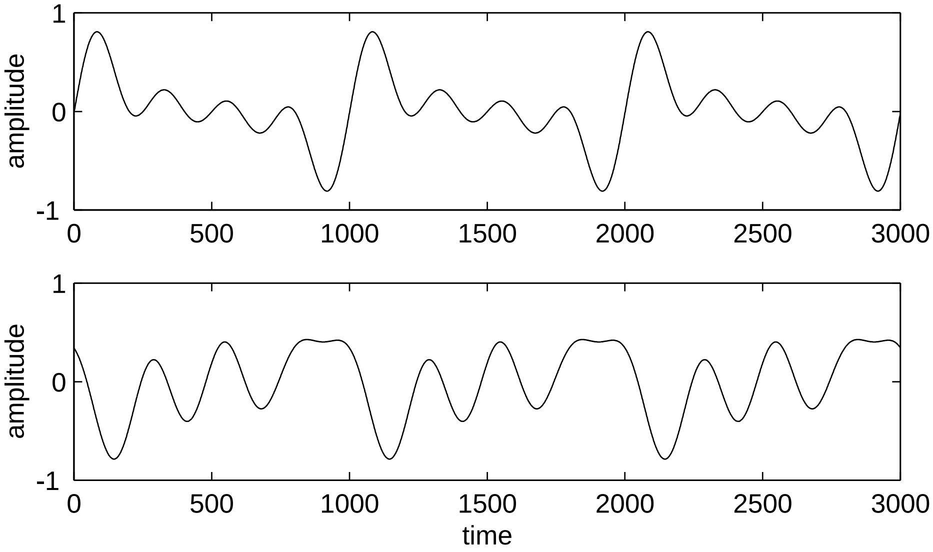


Figure 1.1: Two waveforms with sinusoidal components of the same frequency and amplitude, but with different phase.

The Short-Time Fourier Transform

The theory behind the Fourier transform requires that it be performed on a waveform of infinite length. This is usually accomplished in practice by repeating a finite-length waveform out to plus and minus infinity. If the waveform to be analyzed is long, containing many auditory events, the harmonic information from each of these events will be blurred together. This representation does not give any information about *when* these harmonic events happen.

The short-time Fourier transform (STFT) is an attempt to fix the lack of time resolution in the classic Fourier transform. The input data is broken into many small sequential segments, called frames, and the Fourier transform is applied to each frame in succession. What is produced is a three-dimensional representation, showing the progress of the harmonic spectrum over time. This representation is often referred to as a *spectrogram*.

The finite signal within the frame is repeated to produce the infinite signal required by the Fourier transform. As a consequence, there is often a discontinuity in the waveform at the frame boundaries. This introduces spectral components into the transform that are not present in the original waveform. The solution to this problem is to apply a windowing function to the frame, which gently scales the amplitude of the waveform to zero at each end, reducing the discontinuity at frame boundaries. The windowing functions do not completely remove the frame boundary effects, but they do reduce the effects substantially.

Figure 1.2 shows a simple sine wave with three windowing functions and the corresponding Fourier representations. A single sine wave should have a Fourier representation of a singular component, and as can be seen in Figure 1.2, no STFT window completely removes the boundary effects, but some do better than others. Using no windowing function is equivalent to using a windowing function shaped like a rectangle, and this is referred to as a boxcar window.

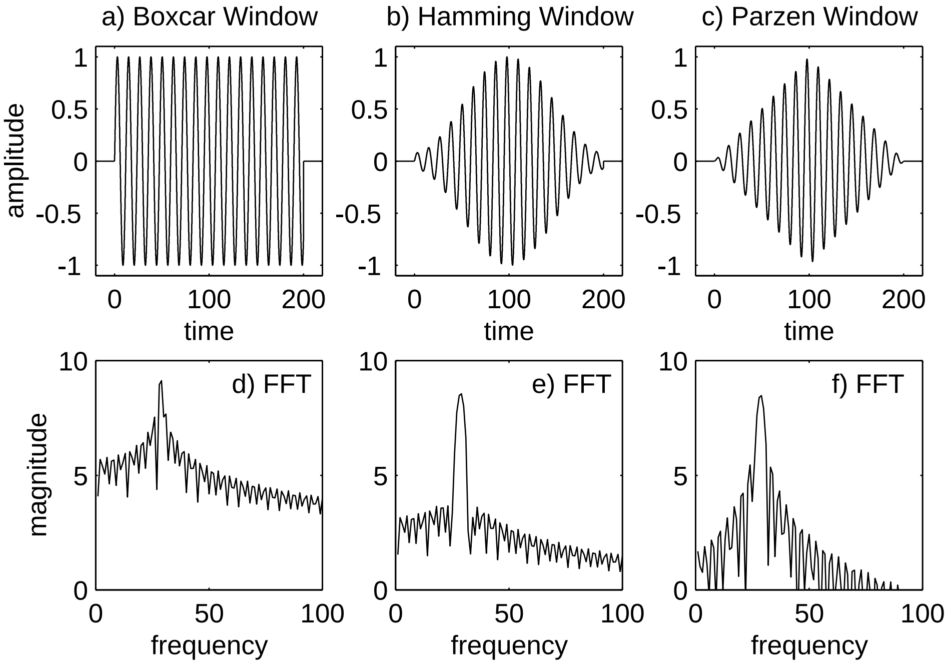


Figure 1.2: The effect of windowing functions on a sinusoid.

Much work has been done to try to design better windowing functions, but as is made clear in [54], the improvements made by these more complicated windows are not worth the extra computation required to produce them. The Hamming window is very simple to implement, takes very little computation time, and yields good results. Gradually reducing the amplitude of the input waveform toward the edges of the frame will substantially reduce the frame boundary artifacts regardless of the window shape used.

When these windowing functions are applied to a waveform, it is clear that some information near the frame boundaries is lost. For this reason, a further improvement to the

STFT is to overlap the frames. When each part of the waveform is analyzed in more than one frame, information that is lost at a frame boundary is retained in the next overlapping frame.

1.3.2 Other Spectral Techniques

The Fourier transform is not the only spectral transform; it is merely the most common. It was one of the original techniques, it is relatively easy to implement computationally, and it has some relevance to the real-world components of audio signals. It is useful for many applications, but there are things that the Fourier representation is not good at, such as time localization and accurate modeling of human frequency perception.

Constant-Q Transform

In the discrete Fourier transform, each frequency band represents an equal fraction of the spectrum. This is based on Fourier theory, and the transform is easy to implement and comprehend. Spectrally rich waveforms that have harmonically related partials appear on the transform as a series of equally spaced peaks.

The human auditory system has long been understood to perform a kind of frequency analysis of the incoming sound. The analysis that is performed by the cochlea is logarithmic in its frequency resolution. Since all studies of sound are, to an extent, studies of the way humans and other vertebrates perceive sound, it makes sense to design a frequency analysis method that models the way the cochlea analyzes frequency.

Thus was born the constant-Q transform [57]. In signal processing theory, Q is the ratio of the centre frequency of a filter band to the bandwidth. The width of each frequency band in the constant-Q transform is related to its centre frequency in the same way, and thus is a constant pitch interval wide, typically $\frac{1}{3}$ or $\frac{1}{4}$ of an octave. This allows for more resolution at the lower-frequency end of the representation and less resolution at the higher-frequency end of the representation, more accurately modeling the cochlear resolution pattern.

The difficulties with this representation are that it is more computationally intensive and it is not necessarily invertible, that is, the result of analysis followed by synthesis might not be exactly the original waveform. For non-real-time analysis-without-synthesis, these problems are tolerable.

Multi-Resolution Transforms

A major drawback of the Fourier transform is that it is a representation that is based completely in the frequency domain. Using the Fourier transform, one can have information about only the frequency behavior of the waveform, without knowing when that behavior occurred, unless a technique like STFT is used.

Multi-resolution techniques look at the spectral makeup of the waveform at many different resolutions, capturing the low-frequency information about the waveform over a large window and the high-frequency information over a smaller window. In the *wavelet* transform, this is accomplished by using a basis function that is expanded and contracted in time [9, 27, 66]. The basis function, called a wavelet, can be thought of as a windowed sinusoid, although this description does not emphasize the mathematical nature of these functions. They are designed to be orthogonal, so that a transform using these wavelets would be reversible.

In the discrete wavelet transform, the wavelet is stretched to fill the entire time frame of the waveform, analyzing how much low-frequency information is present in the frame. The wavelet is then scaled to fit half of the frame, and used twice to analyze the first half and the second half of the frame for slightly higher frequency information, localized to each half. Proceeding by halves, the entire frequency spectrum is covered. High-frequency information is highly localized in time, and low-frequency information is less localized.

Multi-resolution transforms attempt to cross the boundary between a purely time-domain representation and a purely frequency-domain representation. They do not correspond exactly to “time” information *or* “frequency” information; rather the information that they extract from the signal is a kind of time-frequency hybrid. Methods can be employed to extract time or frequency information from a multi-resolution representation such as the wavelet transform.

1.4 Audio Feature Extraction

Feature extraction is typically the first stage in any classification system in general, and in audio signal classification systems in particular. Some researchers have elected to apply a pre-processing module to their system which filters out unnecessary information for the particular application. For example, Kumpf and King [37] use a Hamming window and pre-emphasis in their accent classification system, because the data domain contains only speech.

Researchers attempting more general classifiers typically have not used a pre-processing module, since it could remove useful classification information.

The features used in audio signal classification systems are typically divided into several categories. Perceptual features are based on the way humans hear sound, physical features are based on properties of the physical sound, and signal features are based on statistical and mathematical properties of the waveform itself.

An interesting division of features is for the multimedia database system presented in [79]. The user describes a desired sound, and the authors divide the features used in these descriptions into three categories: acoustical/perceptual features, subjective features, and simile and onomatopœia. Acoustical/perceptual features take into account all of the features we have described so far. A user can request a sound with ZCR in a certain range, or can request a sound with a given pitch track, typically input by singing or humming. Subjective features encompass what the authors call personal descriptive language, which can be more difficult for the system designers to deal with but can often be much more informative. An example of a subjective feature that the authors give is “shimmering”. The simile features allow the user to request a sound by saying it is like another sound. This is often used to select a sub-category, like speech or noise. Onomatopœia is a way to request a sound by imitating the sound, for example making a buzzing noise to look for a sound of bees or electrical hum.

1.5 Physical Features

Physical features are directly related to physical properties of the signal itself. Perceptual features are related to the way humans perceive the sound signals, and as such rely on perceptual modeling. It is because of this that many researchers have elected to base their sound classification systems primarily on physical features. They are easier to define and measure, although they are not as directly relevant to human experience.

1.5.1 Power

Perhaps one of the most straightforward of the physical features, power is a measure of the amplitude of a waveform at any one time. Power is defined as work per unit time, and in the context of audio signals, the power of a signal is related to the amplitude of the waveform. The louder the signal, the more power is in it.

Power measures are used to discover silence in a waveform, as well as dynamic range. The power of a waveform is typically calculated on a short-time basis, by windowing the waveform, as in the STFT, squaring the samples and taking the mean [83]. The square root of this result is the engineering quantity known as the root-mean square value, which has been used by other researchers [63, 79]. Average normalized power (P), for a digital waveform, is equivalent to the average normalized energy (E) per sample. The definitions of these quantities, for a waveform $w(t)$, are presented in Equations 1.1 through 1.4 [31]. The value $p(t)$ will be used throughout this work to represent the various quantities related to power, including amplitude and energy.

$$W_{rms} = \sqrt{|(w^2(t)|} \tag{1.1}$$

$$p(t) = w^2(t) \tag{1.2}$$

$$P = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} w^2(t)dt \tag{1.3}$$

$$E = \lim_{T\to\infty} \int_{-T/2}^{T/2} w^2(t)dt \tag{1.4}$$

Features related to the time-varying power of the waveform have also been used. Power in specific frequency bands, and in particular, the variance of the low sub-band power, is used in [48] to detect silence. Their argument is that the application of a strict power threshold would not detect the difference between frames which contained no signal and frames which contained signal with low power, such as the beginning or end of a fade.

The distribution of power across time has been used to distinguish between speech and music. Speech tends to consist of periods of high power (voiced phonemes) followed by periods of low power (unvoiced phonemes, inter-word pauses), while music tends to have a more consistent power distribution. A measure of the power distribution is used in [62], while a measure of the power modulation rate is used in [63], where the authors claim that speech tends to have a power modulation rate of around 4 Hz.

1.5.2 Fundamental Frequency (f_0)

Only periodic or pseudo-periodic waveforms can have a valid f_0. Perceptually, periodic and pseudo-periodic signals have a pitch. Periodic signals exactly repeat to infinity ($w(t+\tau) = w(t)$) with a period of τ and f_0=τ^{-1} for the largest value of τ. Pseudo-periodic signals *almost* repeat ($w(t+\tau) = w(t) + \epsilon$). There is a slight variation in the waveform from period to period, but it can still be said to have $f_0 = \tau^{-1}$, corresponding to the longest period τ at which the waveform repeats within some tolerance ϵ.

It is clear that extracting f_0 from a signal will only make sense if the waveform is periodic. f_0 detectors often serve a dual purpose in this case—if the detected f_0 makes sense for the rest of the signal, then the signal is considered to be periodic. If the f_0 appears to vary randomly or if the detector provides an impossible or invalid result, the signal is considered to be aperiodic. Often, programmers will build into their algorithms some measure of periodicity detection, and the system will produce an impossible value, such as "0", when the algorithm determines that the waveform is aperiodic.

In a sound or multimedia database such as the one discussed in [79], f_0 is an important feature for distinguishing between pieces of music, or for retrieving pieces of music based on the melody. The authors use the STFT with a peak extractor to identify the f_0 of the waveform. For more on multimedia databases, see [69] and [80].

Speech word boundaries are detected using f_0 in [56]. The idea here is that large variations in f_0 are unlikely to happen in the middle of a word. It is more likely they will happen at the end of the word. The authors discuss the utility of this method for examination of various Indian languages (Hindi, Bengali, Marathi and Telugu) as well as German. However, they do not discuss the f_0 extraction method used.

In many of these systems, there is no differentiation made between pitch and f_0, and although the difference is well understood and easily modeled, it is important to remember that many of these systems do not include perceptual models of pitch detection.

1.5.3 Spectral Features

Many physical features of the spectrum of a waveform can be used for classification, depending on the classification goal. One of the most basic spectral measures is *bandwidth*, which is a measure of the range of frequencies present in the waveform. This feature is used in [62] to discriminate between speech and music. In this case, music typically has a larger bandwidth than does speech, which has neither the low-frequency of the bass drum nor the high frequency of the cymbal. Bandwidth is also used in the system in [79], and in this case the bandwidth is calculated by taking the mean of the difference between the frequency of each spectral component, and the spectral centroid of the waveform. The authors of this paper also use the mean, variance and autocorrelation of the bandwidth as features.

A general feature called *Harmonicity* is used as a feature in several classification systems [62, 79]. Harmonicity refers to relationships between peaks in the spectrum. An object that vibrates, such as the human voice or a musical instrument, creates a sound that has strong frequency peaks at evenly spaced intervals across the spectrum. The harmonicity of a sound can be used to differentiate between voiced and unvoiced speech, or to identify music.

The speech/music classification system presented in [63] uses several features based on statistical measures of the spectrum and spectrogram. These include spectral roll-off point, spectral centroid and spectral flux. The spectral roll-off point is the frequency below which most of the spectral power exists, and is used to distinguish between voiced and unvoiced speech. The spectral centroid is a measure of the mean frequency of the waveform. Music tends to have a higher spectral centroid than speech because of the percussive sounds. The spectral flux is a measure of the rate of change of spectral information, and music tends to have a higher rate of spectral flux than speech.

1.5.4 Formant Location

In general, a *formant* is a peak in the spectral envelope of a sound signal. Usually due to resonance of a filter applied to the driving function, these formants are often related to timbral difference separate from the driving function. Techniques such as multi-resolution spectral analysis and linear predictive coding allow the separation of the driving function from the formants, and thus allow the analysis of the filter used to shape the sound. Formant analysis, and source-filter analysis in general, can be used for many audio analysis problems.

Voiced human speech is produced by a source (vocal cords) generating a periodic function (a glottal pulse) which is shaped by a filter (the vocal tract). The speech has formants at specific frequencies, depending on the phoneme being articulated. In traditional speech recognition, the relative frequencies of the first two formants are typically used to identify the vowel being formed [2, 55]. While formants exist primarily in voiced speech, they also exist in some unvoiced sounds. Whispered speech is completely unvoiced, yet we can understand it as we understand normal speech. This is because whispered speech also contains formants, as shown in Figure 1.3.

Formant location has been used for many years in traditional speech recognition, but it has also been used recently for more specific sound classification. A male/female classification algorithm has been proposed in [76] which uses the location of the first three formants

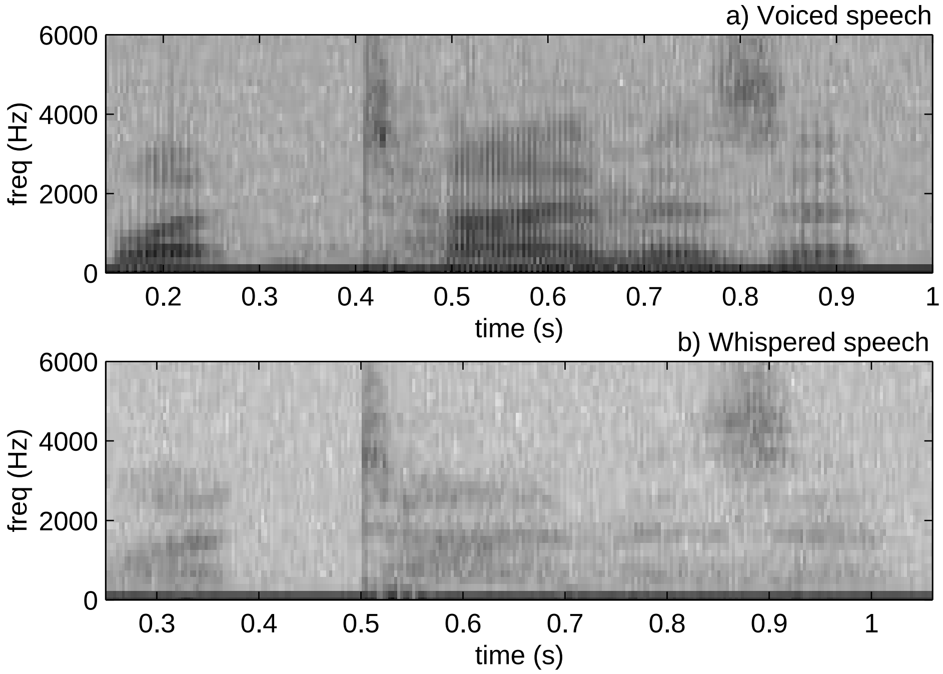


Figure 1.3: A comparison of voiced speech and whispered speech using the phrase "What time is it?"

of the sound signal to classify the gender of the speaker. The authors gathered data about average formant frequencies for males and females, and found that there was sufficient difference to use this as a classification feature. Formants are also used to analyze emotion, prosody, content, language or accent. Accent classification is discussed in [37], where a foreign accent in a known language is identified by the use of foreign phonemes.

1.5.5 Duration and Modulation

The simplest of the time-based features, duration of a sound is simply how much time is taken by the sound. In the multimedia database application described in [79], the duration of the sound is used as a feature. The sound of a finger snap is likely to be shorter than the sound of an explosion, which is again likely to be shorter than the sound of applause. Melody recognition, on the other hand, probably can not make use of duration as a recognition feature, since durations vary between versions of a song, and specifically between the version or versions stored in the database and the version being sung or hummed as input. Duration matching methods can be employed when the length of a sound is likely to be similar to the length of the stored template.

In speech, the duration and spacing of syllables tends to be fairly regular, while in other sounds, and specifically music, tone lengths tend to vary much more widely. This feature, called modulation power in [63] is measured by filtering the signal at 4 Hz (the theoretical modulation rate of speech syllables) and using the power in the 4 Hz band as the feature indicator. In [62], this feature is referred to as tonal duration, and is measured by first finding syllable onsets and offsets, using ZCR to identify fricative consonants, and then finding the time between syllables.

1.6 Perceptual Features

When extracting perceptual features from a sound, the goal is often to identify the features that we as humans seem to use to classify sound. Most perceptual features are somewhat related to a physical feature, and it is usually instructive to investigate the physical counterparts to these perceptual features. When physical features cannot be found that correspond to perceptual features, it is sometimes necessary to extract information by example, and classify based on templates of sounds which have been identified to contain a certain perceptual feature.

1.6.1 Pitch

Pitch seems to be one of the more important perceptual features, as it conveys much information about the sound. It is closely related to the physical feature of f_0. While frequency is an absolute, numerical quantity, pitch is a relative, fluid quantity. A good example of this discrepancy is found in [47], where a system was developed to transcribe sound into a musical score. The system worked correctly, but provided an unexpected result—it placed the musical piece in the key of $C\sharp$ instead of the key of C, because the guitar was tuned slightly sharp.

Humans perceive pitch in situations where current f_0 detectors can fail. One of the most interesting examples of this is the phenomenon of the missing fundamental [58]. When presented with two simultaneous pure tones at a given interval, the human auditory system "hears" the fundamental frequency that would be common to both tones, if it were present. Thus, if two pure sinusoidal tones a fifth apart were played, a pure tone an octave below the lower of these tones would be perceived, as a fundamental to the perceived harmonic series. This implies that for pitch perception, the frequency spectrum of the signal is at least as important as f_0.

1.6.2 Pitch Standards

One of the major problems for automatic pitch detectors is that they often assume that music is based on a specific standard of pitch, for example middle "A" (A_4) being 440 Hz. Such a standard does not really exist, for several reasons.

Musical pitch is understood in a relative scale, with the relationships between notes much more important than their absolute location on a frequency scale. Before 1750, there was no consistent starting point for the musical pitch scale [17]. Absolute pitch had no meaning. In these early times, the number of notes being used was small (sometimes no more than a couple of octaves). Over the centuries, more and more notes have been added to the musical "gamut".

The standard musical pitch system which many musicians adhere to (A_4=440 Hz) is in fact not as standard as it seems. Indeed, between 1636 and 1834, organ manufacturers made instruments with A_4 tuned to anywhere from 392 Hz to 563 Hz [16]. Even today, some orchestras tune slightly higher than 440 Hz, a technique believed to make the music sound "brighter".

The problem of standardized pitch scales manifests itself in many ways for the pitch detection researcher. Music recordings might be based on a different standard pitch from the one assumed by the detection algorithm, which would provide inaccurate results. Even if a recording is made using $A_4 = 440$ Hz, the playback device might not be accurate enough to re-produce this pitch—many old phonograph and tape playback devices have unreliable playback speeds. Beyond the recording device and the pitch standard, the instrument or voice producing the sound might be out of tune.

Most humans have no problem with this situation because our primary pitch detection process is *relative* pitch perception. Some people with *absolute* or perfect pitch perceive specific frequencies instead of frequency ratios, and can have significant difficulty following a musical score when listening to the piece played in a different key [40]. It would be beneficial if researchers could design their pitch detectors with some form of relative pitch. A difficult and somewhat theoretical question is whether there is any physical difference between a piece in the key of $C\sharp$ and a piece in the key of C played on an instrument tuned a half step too high. How could a computer be programmed to detect this difference?

An even more difficult problem is that of the natural drift of amateur unaccompanied singers. A singer may start on key, and as the piece progresses, slowly drift up or down in pitch until at the end of the piece, he or she might be several tones out from the initial key. How does a computer deal with this problem? When does it shift from the key of C to the key of $C\sharp$ to the key of D? What makes these problems even more difficult is that they are usually specific to the domain of the sound input. A piano will not drift during the course of a piece, so a pitch detection algorithm designed for transcribing piano notes might not perform well on unaccompanied singing. Indeed, the pitch of unaccompanied singing often varies so much during the course of a single note that pitch detectors have difficulty determining the note being sung. More discussion on this topic is presented in Section 4.2.

1.6.3 Prosody

Prosody is a characteristic of speech corresponding to changes in pitch and phoneme duration, as well as significant pauses across a spoken phrase, indicating deeper meaning. For example, if the pitch of a phrase rises at the end, it is likely to be a question. Prosody is also used to emphasize certain words in a phrase. The sentence "I took her to the store," can mean different things depending on which part of the sentence has emphasis. "*I* took her to the store," conveys a different meaning from "I took *her* to the store," or "I took her to the *store*." This emphasis can be generated using higher pitch, loudness, increased phoneme duration or a significant pause before the emphasized word.

The analysis of prosody for classification usually assumes that speech recognition has already been performed, and the prosody of the speech conveys further meaning. As an example, in [44] the authors use prosody, among other tools, to identify dialogue acts, or fundamental pieces of speech. Prosody could also be used when speech recognition has not already been performed. An utterance with prosodic raising at the end could be classified as a question, and other characteristics of speech could be identified using prosodic features.

1.6.4 Timbre

When humans discuss sound, they talk of pitch, intensity, and some other well-defined perceptual quantities, but some perceptible characteristics of a sound are more difficult to quantify. We clump these characteristics together, and call them collectively *timbre*, which has been defined as that quality of sound which distinguishes different instruments or voices sounding the same pitch. Most of what we call timbre is due to the spectral makeup of the signal, specifically at the attack of the note. Many spectral characteristics, as discussed above, can be used as classification features, and many of these correspond to the timbre of the sound.

Zhang and Kuo provide a discussion of timbre in [83], and consider it the most important feature in differentiating between classes of environmental sounds, as well as speech and music. Acknowledging that spectral information contained in the attack of the note is important in timbre, they state that the temporal evolution of the spectrum of audio signals accounts largely for the timbral perception. Unfortunately, they do not discuss their method for extracting timbral information from a sound. The extraction of physical features that correspond to timbral features is a difficult problem that has been investigated in psychoacoustics and music analysis without definitive answers.

The authors of the multimedia database system discussed in [79] describe subjective features as well as acoustic and perceptual features of sound. Words used to describe timbre include "shimmering", "bright" and "scratchy", and these ideas can be used in a template matching system, which would classify on the basis of timbre without identifying the corresponding physical features. The system collects examples of sounds that have been classified as "scratchy", clusters them according to the features they have in common, and uses these features to decide if a new sound belongs to this category or not.

1.6.5 Rhythm

Rhythm is a perceptual quantity, and is often defined differently depending on the listener. Rhythm relates to the rate, regularity and pattern of time-level events like drum beats, note and vocal syllable onsets and duration, and linguistic events like prosodic emphasis and rhyme. When a piece of sound is considered rhythmic, it often means that there are individually perceivable events in the sound that repeat in a predictable manner. The *tempo* of a musical piece indicates the speed at which the most fundamental of these events occur. Researchers who are attempting to extract rhythmic information from a piece of sound often look at repetitive events in power level, pitch or spectrum distribution, but musical rhythm is often not as simple as a peak in power every n milliseconds. More likely, there is a complicated series of events, fitting into a repetitive rhythmic framework. The problem is that the tempo of this rhythmic framework often changes throughout a piece of music, for example slowing down just before a chorus or at the end of the piece.

A rhythm detector was employed in [63], in the form of a "pulse metric", using autocorrelation to extract rhythmicity. The waveform is filtered to isolate various frequency bands, and the autocorrelation of each band is taken. The authors indicate that this method is useful to detect a strong driving beat in music, but fails when the rhythm deviates very much from a central time. A common musical technique that results in this deviation is *rubato*[2].

Rubato music has an underlying beat that is intentionally inconsistent in its duration, allowing for emotional expression in the music. This style makes rhythm detection very difficult, for example, when a jazz singer draws out some notes and shortens others. The effect of this on the listener is that if they know the piece, they can appreciate the interpretation of the singer by comparing it with an internal mental model derived from previous experience. Providing a computer with a similar internal model could require coding of experience and context, which is a difficult problem. Conversely, modern dance music uses machine-generated drumbeats, which are usually very rigid in onset time and duration, and because of this the pulse metric performs well in detecting the presence of rhythm in this type of music.

A more general classification system presented in [83] uses rhythm to detect sound effects

[2] The musical term "*rubato*", also called robbed-time, comes from the idea that time is stolen from one note and given to another, so the overall duration is consistent but the rhythm of the individual notes is altered.

such as footsteps, clock ticks and telephone rings. While the authors discuss the effects of rhythm, and why it is a useful feature, they do not discuss the extraction methods used in their rhythm detector.

1.7 Signal Features

Signal features relate to the characteristics of the waveform, which is a representation of the sound within a computer system. Signal features include statistical and mathematical properties of the waveform, and often relate to the manner in which the sound signal has been translated into computer-interpretable information.

1.7.1 Zero-Crossing Rate and Related Features

Since it was made popular in [34], the utility of the zero-crossing rate has often been in doubt, but lately it has been revived. Put simply, the ZCR is a measure of how often the waveform crosses zero per unit time. The idea is that the ZCR gives information about the spectral content of the waveform.

One of the first things that researchers used the ZCR for was f_0. The thought was that the ZCR should be directly related to the number of times the waveform repeated per unit time. It was soon made clear that there are problems with this measure of f_0 [57]. If the spectral power of the waveform is concentrated around f_0, then it will cross the zero line twice per cycle, as in Figure 1.4a. However, if the waveform contains higher-frequency spectral components, as in Figure 1.4b, then it might cross the zero line more than twice per cycle. A ZCR f_0 detector could be developed with initial filtering to remove the higher partials that contaminate the measurement, but the cutoff frequency needs to be chosen carefully so as not to remove the f_0 partial while removing as much high-frequency information as possible. Another possibility for the ZCR f_0 detector would be to detect *patterns* in the zero-crossings, and hypothesize a value for f_0 based on these patterns.

It has since been shown that ZCR is an informative feature in and of itself, unrelated to how well it tracks f_0. Many researchers have examined statistical features of the ZCR. For example, [63] uses the ZCR as a correlate of the spectral centroid, or balance point, of the waveform, which, unless the spectrum is bimodal, is often the location of most of the power in the waveform. If the spectral centroid is of fairly high frequency, it could mean that the signal is a fricative, or an unvoiced human speech phoneme.

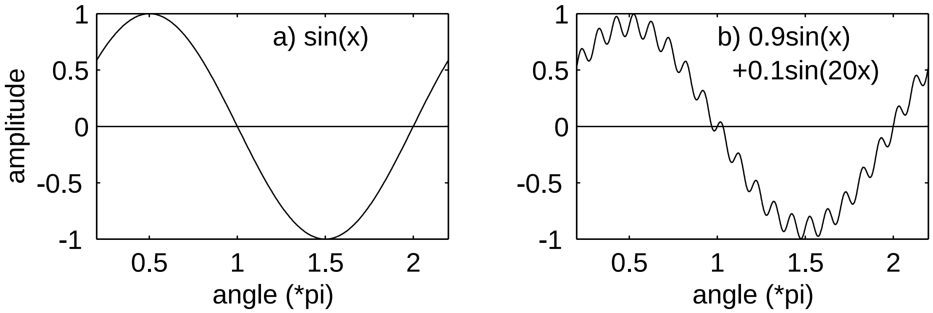


Figure 1.4: Influence of higher harmonics on zero crossing rate. (after [57])

A purely statistical use of the ZCR is found in [62]. John Saunders gathered data about how the ZCR changes over time, and called this a ZCR contour. He found that the ZCR contour of speech was significantly different from that of music, and used this feature to help discriminate between the two. A similar use of the ZCR is the short-time average ZCR feature, used in [83]. Again, the authors used the ZCR as a measure of the spectral characteristics of the waveform, to differentiate between speech and music. These unintuitive uses of the ZCR show an advantage of physical features over perceptual features - that some useful features of sound signals are not immediately evident from human perception.

One of the most attractive properties of the ZCR and its related features is that these features can be calculated very quickly. The ZCR is a time-domain feature, which can be calculated in real time "on the fly," keeping a running total of the zero-crossings as the waveform is received. A system which uses features entirely based on the ZCR would not even need analog-to-digital conversion. It would only need to sense whether the input waveform voltage is positive or negative, and send a pulse whenever the sign of the waveform changes.

1.7.2 Voiced/Unvoiced Frames

One of the fundamental first steps in any speech recognition system is the classification of frames as voiced or unvoiced. On first inspection, this seems like a fairly physical feature: voiced frames tend to be harmonic, and have a lower spectral centroid than unvoiced frames, which tend to be non-harmonic.

The voiced-ness of a sound segment can be used as a classification feature, often detected using a f_0 estimator [32, 75]. The assumption is that the input is a speech signal, and when there is significant power in a frame but no discernible pitch in normal speech range, the frame can reliably be classified as unvoiced. The system is really classifying on the basis of pitches in a certain range, but in the domain of speech recognition, this corresponds to a voiced/unvoiced classification.

1.8 Speech, Music and Song

This thesis deals with human utterance classification, a sub-domain of audio signal classification. Specifically, the topic of interest is the physical and perceptual features that differentiate human talking from human singing. Human singing has many characteristics of what we call music, but the main difference is that in human singing, no instrumental or polyphonic music is present. Singing is often combined with instrumental music to create the full-spectrum songs we hear on the radio, and individual singing voices are often grouped together to create vocal harmonies. The differences between speech and song are often very subtle and much more difficult to quantify than the differences between speech and music.

1.8.1 Speech and Music

When differentiating between speech and music, the research goal is often application-based: Create a system that can separate broadcast radio into segments of speaking and segments of music. The goal of such a system would be, for example, automatic radio tuning, where the user would request either constant music or constant talking, and the radio would find a station with the desired content, and stay there till the content changed, at which point the radio would find another station.

These systems differentiate between speech and song using features that are relevant to the task, including bandwidth, f_0 continuity, and power modulation. While these features are useful for a speech/music system, they fail when applied to the speech/song problem. The bandwidth of a speech signal is much less than that of a full spectrum music signal because the music signal often has bass drum and cymbals, both of which produce sound outside the normal range of human singing or speaking. The bandwidth of an individual singer is the same as that of an individual speaker[3]. A difference between speaking and singing, however, is that singers often use pitches outside of the normal range of speaking.

[3] This is not strictly true for some trained professional singers, who are able to modify the spectral characteristics of their voice to include higher frequencies

The problem with this feature is that pitches inside the normal speaking range can be evidence for speaking *or* singing. Further information on these features is presented in Chapter 4.

Musical instruments often produce constant pitches held for a duration. Instruments such as the piano cannot change the pitch of a single note. Human singing is, on the other hand, not dependent on steady pitches and in fact it is very difficult and often undesirable for human song to have a perfectly constant pitch. Rather, a more perceptually pleasant human song style has a pitch track that uses vibrato, a pseudo-sinusoidal pitch track oscillation.

The speech/music work that has been done recently can be used to inform a speech/song classification project, but the one is not directly extensible from the other. This thesis is a description of the features and techniques that are useful in discriminating between speaking and singing.

1.8.2 Human Utterance Continua

Many human utterances are not strictly classifiable as talking or singing. Utterances like poetry, chant and rap music fall somewhere between speaking and singing, with characteristics of each [41, 43]. Another intermediate utterance is *sprechstimme* or *sprechgesang* (speech-song), developed by the composer Arnold Schoenberg and used later by his student Alban Berg. It is a vocal musical style characterized by widely varying pitches, with the singer approximating the pitch instead of singing the exact note.

When considering a classification domain with intermediate utterances between two classes, there are several ways to proceed, three of which are hard classification, continuum classification, and sub-category hard classification.

Hard classification. This is the traditional classification paradigm, where each new data point must be assigned to exactly one class. In the case of speech versus song, a single two-class discrimination will not accurately describe utterances that fall between speech and song. A two-class paradigm might be a beneficial starting point for classifying intermediate utterances because the two-class features can be extended by assigning a confidence metric to each feature measurement.

Continuum classification. Also called fuzzy classification, soft classification or confidence classification, this method assumes that each incoming data point can have membership in both available classes to some degree. The terminology of these various techniques differs but the result is primarily the same. In confidence classification, each data point is assigned to one class, with an associated confidence metric indicating the "good-ness" of the classification. If all relevant features agree with a classification, the confidence would be high, while disagreement in feature results would result in a lower confidence.

Fuzzy classification considers partial membership in both available classes. Traditional logic states that an object is either a member of a set or is not. Fuzzy logic states that an object can be a member of a set to a degree [35, 45]. As an example, consider terminology such as "tall" or "old" to describe an individual. Because these are relational descriptions (taller, older), the cut-off point between "tall"and "short" is difficult to identify. Two-class Fuzzy logic suggests that each individual be assigned a pair of numbers that indicate membership in each class. A short person might be 25% "tall" and 82% "short", while a very tall person might be 90% "tall" and 7% "short". Note that fuzzy membership in all classes is not required to sum to unity.

Sub-category hard classification. Because the continuum between speech and song can be identified by listing human utterances between speech and song, a third possibility presents itself: instead of allowing a continuum of classification results, create a set of classes between speech and song and require that each incoming data point be assigned to one of these classes. This removes the extra computation that is required with continuum classification while acknowledging the range of possible utterances between speech and song.

In this thesis, hard classification is used to determine the relevant features in the domain and to test the capabilities of each feature. Continuum testing is then performed on the hard classification features, using intermediate utterances, to test the feasibility of using hard classification features for continuum classification.

1.9 Thesis Organization

This thesis proceeds in five chapters and three appendices. Chapter 1 has presented an introduction to the problem as well as some relevant background necessary for understanding the remainder of the thesis. Chapter 2 describes the collection and annotation of the research corpus used in the remainder of the thesis. Pitch has become a fundamental base feature for this research, so Chapter 3 contains a description of several techniques for extracting f_0 from a waveform. Chapter 4 contains a discussion on the features used in the speech/song discrimination task, their motivation, implementation and results. Conclusions and future work are presented in Chapter 5.

The three appendices contain data and documents relevant to the research corpus discussed in Chapter 2. Appendix A is the research protocol for the human subject study performed to collect and annotate the corpus. Appendix B contains copies and examples of the research tools and instruments used in the corpus collection and annotation. Appendix C contains the complete corpus annotation results, including numerical ratings and verbatim written comments.

Chapter 2

Corpus Collection

For audio research in a specific domain, the most desirable option is often to find a corpus of data that has been collected previously and is in use by other researchers, as this provides a place to start and a set of colleagues with whom to compare results. If the domain is new, obscure or specific, such a corpus may not exist. Intermediate utterances between speech and song is such a data domain. There exist many corpora of speaking only [77], and some corpora including sung clips, but searching the current literature did not uncover any corpora containing intermediate clips between speaking and singing, or clips of the same phrase spoken and sung by one individual. Both of these would be valuable to speech/song research.

The first step in collecting a corpus of clips is to consider the options for sources of sounds, of which there are essentially two: find (*extract*) clips from pre-existing sources or record (*solicit*) new clips. Solicited clips have the advantage of control—specific characteristics of content, speaker, subject, style, *etc.* can be sought. The primary disadvantage of solicited clips is the time and effort required to find appropriate subjects and to obtain the necessary permission to solicit such clips. Extracted clips have the advantage of being readily available. However, copyright restrictions may apply, and not all characteristics are controllable. Further, some characteristics concerning source may be unknown and/or unobtainable.

For the corpus used in this thesis, clips were collected in these two ways—some clips were solicited from subjects in a laboratory environment, recorded digitally using a dynamic headset microphone, and some clips were extracted from existing media such as movie soundtracks, broadcast media, and music albums.

2.1 Corpus Domain, Restrictions and Distribution

No corpus can be exhaustive in scope or content. When creating a corpus containing human utterances, a primary concern is the restrictions in the domain, particularly in language, content, and scope. If soliciting clips in more than one language, the subjects must be multilingual or there must be a multi-lingual solicitor. It is easier to find extracted clips in several languages than it is to solicit such clips, although again, some source information may not be available.

If the corpus is restricted to a single language, any conclusions gained from working with the corpus will be applicable only to the content language. On the other hand, language differences in a multilingual corpus may make it difficult to isolate phenomena occurring in only one language. Monolingual segments of a multilingual corpus can be used for monolingual research. The utility of a multilingual corpus must be balanced with the increased time and effort required to collect such a corpus, especially if the research task is, by its nature, monolingual.

This corpus is primarily monolingual, with a small collection of other languages. A larger multilingual corpus may be collected in the future, to expand the current research and results to other languages. The corpus contains 90.3% (756 files) English language utterances with the remainder of the clips (82 files) in languages including French, Italian, Swedish, Gaelic, Japanese, Mandarin Chinese, Rumanian, Hungarian, German, Latin, Iroquois[1], Mon-kmer[2] and Zunian[3]. Some clips have no language, such as whistling or humming.

The corpus can be distributed along three primary axes:

- Extracted clips — Free Solicited Clips — Constrained Solicited Clips
- Spoken utterances — Sung utterances — Intermediate utterances
- Speaker Characteristics

These axes relate to the following characteristics: the collection method of the sound clip (Existing media, human subjects); the content of the sound clip (speech, singing, something in between) and the content source (who is speaking or singing). The spoken/sung

[1] The Iroquois are a collection of aboriginal american tribes from the north-east United States, Quebec and Ontario.

[2] Mon-kmer is the language of the Kmhmu people, an aboriginal tribe from Laos

[3] The Zuni are an american aboriginal tribe who originated in New Mexico and Arizona

categorization is the primary focus of the research for which this corpus was developed, and consists of pure speech clips, pure song clips, and clips that are somewhere between speech and song. These intermediate clips are categorized by listener opinion testing, as discussed in Section 2.2. The speaker characteristics considered include age, gender, and speaking and singing experience, although for the extracted clips, some of this data is unavailable.

2.1.1 Solicited Clips

Clips were solicited from 50 subjects, using a set of 11 prompts. Subjects were instructed to attempt to limit their utterances to 5 seconds. The prompts used were:

1. Please speak or sing anything you like for about 5 seconds.
2. Please sing the first line of your favourite song.
3. In a single short sentence, please tell me what you had for lunch yesterday.
4. Please speak the phrase "When you're worried, will you run away?"
5. Please sing the phrase "Row, row, row your boat, gently down the stream."
6. Please speak the phrase "Row, row, row your boat, gently down the stream."
7. Please sing the phrase "O Canada, our home and native land."
8. Please speak the phrase "O Canada, our home and native land."
9. In a single short sentence, please tell me what you did last weekend, using a voice which is somewhere between singing and speaking.
10. Please utter the phrase "Why is the sky blue?" using a voice which is somewhere between speaking and singing.
11. Please speak or sing anything you like for about 5 seconds.

These prompts are designed to solicit a specific set of utterances, and to control for certain characteristics. There are two types of constraints in the solicited clips—lyric and style. Clips may be constrained in lyric (prompt 4), in style (prompt 2), or both (prompt 7). The first and last prompts solicit unconstrained clips, while prompts 5 through 8 constrain both lyric and style, to provide a baseline characterization of singing and speaking for each subject. Prompts 9 and 10 constrain style, in that subjects are to provide some form of intermediate utterance between speaking and singing. These clips are interesting from the point of view of this research, because they are expected to fill out the continuum between speech and song.

The songs for the clips constrained in lyric and style were chosen to be familiar in content and tune, while containing a variety of phonemes and rhythmic structures. Prompts 5 and 6 are from a common nursery rhyme, containing phonetic glide repetition and few fricatives, and the tune has short notes and small intervals. Prompts 7 and 8 are from the Canadian national anthem, a song familiar to all subjects. This song has longer notes and the intervals are larger, and the lyric for this song contains more fricatives and a repetitive vowel structure.

The prompts were presented to the subjects in one of four orders, to control for familiarity and experiment history. Some subjects were prompted for singing first, then speaking, and some subjects were prompted for the intermediate clips first while some were prompted for intermediate clips last. In each of the four prompt orderings, the first and last clips are unconstrained, using identical prompts.

A total of 550 solicited clips are in the corpus, collected from 50 subjects, consisting of 23 females and 27 males. The mean age is 34.8 years, and the standard deviation of the ages is 13.7 years. The oldest subject is a 70 year old male, and the youngest subjects are 20 years old. Their (claimed) experience ranges from no formal speaking or singing to 60 years combined speaking and/or singing training. The subject set includes professional radio broadcasters and professional singers as well as informed amateurs and novices.

There are an additional 65 unconstrained solicited clips from individuals during conferences, which complement the unconstrained nature of the extracted clips. The prompt used was “Please speak or sing anything you like for 5 seconds,” and each individual gave a single clip or two. Combining these with the 100 unconstrained solicited clips, the first and last from each of the 50 laboratory human subjects, there are a total of 165 unconstrained solicited clips, and 450 constrained solicited clips.

2.1.2 Extracted Clips

The clips extracted from existing media were found by looking through popular movies and music albums for short segments of speech and song, but primarily for intermediate vocalizations[4]. The requirement for these clips is that they are completely monophonic—they have no background noise, music or sound effects. Many of these clips were taken "mid-stream", from the middle of an utterance, to avoid utterance onset and offset artifacts.

A total of 232 extracted clips are included in the corpus, which combined with the 615 solicited clips makes 847 clips total in the corpus. The extracted clips are unconstrained in style and content, although specific style of content was sought for the clips. The entire corpus is available on-line at `http://www.cs.sfu.ca/speech`, including the numerical annotation information presented in the following sections.

2.2 Data Annotation Method

Collecting the data for a corpus is only the first step in producing a useful research tool. The data must be annotated in order to provide statistical information and research targets. A common annotation task in human utterance corpora is to transcribe the words in each clip, since a common research task is automatic speech recognition [4]. For the research presented in this thesis, what people are saying is less important than how they are saying it, so linguistic transcription and part-of-speech tagging is not part of the annotation task for this corpus. The annotation task for this corpus is to develop a classification target for each utterance.

Classification targets for linguistic data are usually fairly universal—humans generally agree on the linguistic content of an utterance even if computers may have problems resolving anaphora (binding pronouns to their related noun phrases) or ambiguities (choosing from one of a set of equally plausible results). Stylistic data such as speech-ness and song-ness of intermediate clips are by their nature more subjective, and so the opinion of a single researcher is not sufficient to determine the classification of such a clip. For this reason, opinions were solicited about the intermediate clips in the corpus both by individuals who had participated in the creation of the corpus by providing clips, and by individuals who had not been involved in the creation of the corpus. This opinion gathering had the added advantage of soliciting opinions on the general differences between speech and song as well as the specific classification of each clip.

[4] Most of the extracted clips come from sources protected under copyright. The Canadian and American copyright acts both allow fair use or fair dealing of copyright material under several situations, two of which apply—first, the clips are being used for research, and second, the clips are too short to be in competition with intended market of the original work [1, 7].

Opinions were solicited from human listeners on an internet-based form containing a representative sub-set of the complete corpus. The web sub-set of the corpus contains all of the constrained solicited clips where the prompt was for an intermediate vocalization (100 clips) as well as 34 unconstrained solicited clips and 76 extracted clips. These 210 clips were reduced to 191 for the web site, by removing 19 clips which were repetitive or unsuitable in content or quality. Some repetitive clips were retained in the web corpus as a control. The clips that are not annotated using the internet-based form are rated as pure speech, pure singing, or rated the same as a similar or identical clip which was rated using the internet form. An additional 12 constrained solicited clips were chosen from other prompts and these were rated by all listeners, bringing the total number of clips on the web-site to 203.

Users began by logging in to the web-site, and providing their name and contact information. Each was assigned a unique user number. This number was used for identification throughout the project. The name and contact information were retained only for later contact if required. The listeners then provided information about their experience with professional or amateur music, singing and speaking, as well as their age and gender.

There are three parts to the web annotation project. Part 1 consists of a simple 1-to-5 rating for all but 30 of the web corpus clips, 1 for pure speech and 5 for pure singing. The extracted clips and the free solicited clips were rated by all listeners, and 20 of the 100 intermediate utterance solicited clips were rated by each listener, 10 from each prompt, to reduce the number of files each listener was required to rate. There were 5 sets of 10 clips from each prompt, randomly chosen for each listener, so all 100 of these clips were rated.

Part 2 contained the remaining 30 clips and a more detailed rating was solicited from the listeners:

1. Rate the file: Talking ○ ○ ○ ○ ○ Singing

2. What is it about the sound that leads you to this judgement?

3. What could the speaker have done to make this clip more speech-like?

4. What could the speaker have done to make this clip more song-like?

If a listener rated a clip as pure speech, s/he was not required to indicate how to make the clip more speech-like, and vice-versa. Part 3 is a free-response form where listeners shared their insights on the project and on the differences between talking and singing.

The data for all three parts were collected using a perl script system hosted from the Simon Fraser University web-site over the course of the month of May, 2001. There were 48 listeners who completed the web annotation project, 19 of whom had been subjects in the data collection phase of the project. Participation was solicited through email, personal contact, invitations to the subjects of the data collection project, and a radio spot on the British Columbia noon radio show "BC Almanac" on the Canadian Broadcasting Corporation.

2.3 Annotation Results: Numerical

This section provides numerical analysis of the subject set and of the ratings provided by the listeners. The data considered here are the numerical statistical data about the subjects themselves, and the 1-to-5 ratings provided by the listeners in Part 1 and Part 2.

2.3.1 Overall Results

The individual speech/song ratings from each listener for each file were averaged to obtain a mean rating (μ) for each clip measured. The estimated probability density function for these mean ratings is presented in Figure 2.1a. For a description of the method for generating this probability function, see Section 4.1.2. Figure 2.1b shows the standard deviation (σ) compared to μ for each file. There are several files for which all of the listeners gave the same rating, resulting in a standard deviation of 0. The standard deviation of the ratings is larger near μ=3.5, and is smaller toward a μ=1 or μ=5. This indicates that in the middle of the speech/song axis, listeners disagree more on their ratings, and the confidence of these ratings is less.

2.3.2 Distribution Analysis

The *Kolmogorov-Smirnov* (K-S) test [54] is a goodness-of-fit test which provides a measure of the distribution deviation between a hypothesized ideal distribution and an observed distribution, or between two observed distributions. This statistical measure of distribution similarity or difference will be used throughout this work. In the first case, the test is used to test the hypothesis that the mean ratings come from a uniform distribution. Later, this test will be used to determine if the population demographics had any influence on the

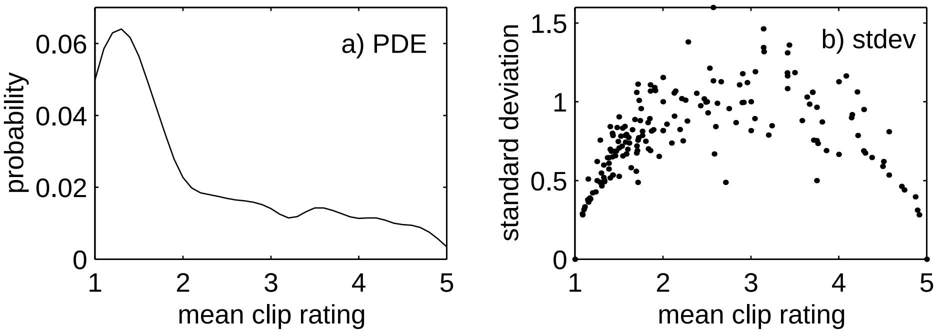


Figure 2.1: Corpus mean rating statistics.

ratings.

The K-S test is based on minimum distribution distance. Two distributions are compared and two statistics are calculated: The K-S *distance* (D) is the largest vertical distance between the two probability distributions, and the K-S *significance level* (α) is a measure, related to the sample size N of the distributions being tested, indicating whether to accept or reject the null hypothesis. In the K-S test, the null hypothesis is that the observed distribution is taken from the ideal distribution, or that the two observed distributions are taken from the same (unknown) ideal distribution. The required distance for a desired significance level is calculated for large N (> 35) using the following formulæ [12]:

$$D_{\alpha=.05} = \frac{1.36}{\sqrt{N}} \qquad D_{\alpha=.01} = \frac{1.63}{\sqrt{N}.} \tag{2.1}$$

If a single distribution is to be compared against an ideal, N is the number of data points in the distribution. If two distributions are to be compared, The effective number of data points N_e is used in Equation 2.1 in place of N:

$$N_e = \frac{N_1 N_2}{N_1 + N_2.} \tag{2.2}$$

For the data used in this measurement, $N = 203$ sound files which were rated, so $D_{\alpha=.05} = 0.0955$ and $D_{\alpha=.01} = 0.1144$. The significance levels are interpreted as the likelihood that this measure could be attributed to chance. Many researchers use a significance level of 0.05 as an assurance that the measurement is not due to chance. The algorithm used to calculate the K-S distance also calculates the significance level for this distance, providing a specific evaluation of the K-S distance.

In this case, if $D = 0.0955$, the likelihood of this distance indicating two distinct distributions is $(1 - \alpha) = 95\%$, and If $D = 0.1144$, the likelihood of this distance indicating two distinct distributions is $(1 - \alpha) = 99.9\%$. If the distance is less, no judgement can be made on whether the distributions are the same or different. The K-S statistic can only indicate the likelihood that the distributions are different. If the statistic does not indicate a difference, that does not mean that the distributions are the same.

If the ideal case is a uniform distribution of clips from speech to song, the K-S distance between the corpus distribution and the ideal distribution is 0.3302, and the significance level of that distance is 2.7226×10^{-10}. This indicates that the mean ratings in the corpus do not come from a uniform distribution. The ratings seem to be somewhat biased toward speech, which could indicate that the original corpus contained more spoken clips, or that people tend to rate intermediate clips more speech-like than song-like, or that there is some other ground truth related to the perception of speech and song. Further research will be necessary to discover whether the bias is an artifact of this particular corpus or if it reflects some more fundamental phenomenon. Written results presented in Section 2.4.2 indicate that the latter is true, although this does not prove that the original sub-corpus content was uniformly distributed.

2.3.3 Results Considering Listener Characteristics

Information was gathered regarding the age, gender, and experience of the listeners. The mean ratings of the extremes of these demographics are compared using the Kolmogorov-Smirnov statistics. The results from the oldest 10% and youngest 10% of the listeners are compared, the results from the men are compared with the results from the women, the results from those listeners with the highest 10% of claimed speaking experience are compared with the lowest 10%, and the same is done for claimed musical experience (a combination of instrumental and singing experience). The results of these K-S statistical measurements are presented in Table 2.1. In this case, N=113 because some of the files in Part 1 of the annotation process were not rated by all listeners. In some cases, the number of listeners in a particular demographic who rated a particular file was small or zero, so these results could not be compared to listener results in the opposite demographic. Because two distributions are being compared, N_e is used.

These statistics show that while there are differences between the results from these demographic groups, the hypothesis that the data come from different distributions (for

Table 2.1: Kolmogorov-Smirnov statistics for demographic groups.

Demographic Group	Kolmogorov-Smirnov distance	Significance level
Young/Old	0.1526	0.1308
Men/Women	0.0464	0.9996
Speaking Experience	0.1429	0.1832
Music Experience	0.0859	0.7805

example, that older people have a different opinion of speaking and singing than younger people) is rejected, indicating that the demographic differences are not significant. In fact, the distance is so small between the ratings of the men and women that the probability is 0.0004 that the two distributions are different. Similar demographic results have been shown in research such as that by Scheirer *et al* [64], where listeners were asked to rate a musical segment on a number of perceptual scales. Scheirer's listener set consisted of 50 participants of varying ages, genders and musical abilities, and the statistical differences between these demographics were studied in more detail than in the current work. In that study, the effect of the demographic differences are characterized as small.

2.4 Annotation Results: Written Comments

Parts 2 and 3 of the study used prompts which solicited written comments from the listeners. Part 2 solicited comments about individual sound clips, and Part 3 solicited general comments about the differences between speech and song. Some representative and pertinent comments are presented here.

The two main features that listeners mentioned are pitch and rhythm. More specific features based on these two quantities are discussed in the following sections. Throughout this thesis, individual subject responses will be indicated using the following format:

clip:subject (rating) "reasoning"

u333:111 (3) "An example reasoning for the rating"

The labels on the responses correspond to the clip label, subject number and the rating given by the subject on the file.

2.4.1 Comments on Individual Sound Clips

Subjects were asked to indicate what about the sound made them rate the sound in the way that they did, and what the speaker could have done to make the sound more speech-like or more song-like. Responses to these questions depended on how the listeners rated the sounds. In this thesis, all quotes from subjects are presented verbatim without corrections in spelling or grammar. Any words or phrases which could be used to identify the subjects have been removed.

Mean rating < 2

These files are considered by most listeners to be very close to speech, and therefore many listeners do not indicate how the clip could be more speech-like. Those who offered opinions mention characteristics like rhythm, regularity and speed. Faster speech, more regular speech, and more rhythmic speech are all considered slightly more song-like. Some listeners also mention pitch. Monotonous utterances are considered speech, although some clips with widely varying pitch are still classified as speech because the pitch varies in what subjects call a "speech-like" way.

One clip with a mean rating of 1.54 was considered to be poetry by several listeners, but the listeners disagreed on whether poetry was more speaking or singing.

n131:211 (4) "poetry"

n131:238 (1) "sounds like the recital of a poem"

n131:317 (1) "poetry is speech not song"

n131:328 (3) "The fine line between reading poetry and singing poetry"

On the same clip, one subject mentioned a subtlety of rhythm that affected the rating.

n131:358 (2) "it could be 100% talking but there's an exaggerated metric pattern that pushed my judgement a tick towards song"

On another clip, some listeners made comments which do not refer to features of the clip, but justify the classification by comparison or analogy.

u162:213 (1) "sounds like Vincent Price story telling"

u162:220 (2) "speaking, but with emphasis which seems song-like"

u162:311 (3) "words are short and clear, but there is a poetry type rhythm"

For a clip taken from a radio presentation of a poem read at a comedy show, some listeners were able to identify the context of the recording, as poetry, comedy or performance. Features such as pitch, rhythm and rhyme are identified, but these features were not sufficient to pull the rating much beyond speech.

u169:221 (2) "sounds like a poem, or stand-up comedy, but hits the same note a few times"

u169:327 (2) "sounds rehearsed, like a performance, not natural speech"

u169:354 (2) "more rhyme than straight speech"

In clips that are rated primarily as speech, the reasons mentioned for not giving a rating of (1), such as rhythm, melody and context, are indications of features likely to be useful to identify song as opposed to speech. More song-like features in a primarily speech utterance drive the ratings toward song.

Mean rating $>$ 4

These files are considered by most listeners to be primarily song. Characteristics that listeners indicate to justify this rating are rhythm, word duration, tone, and in particular tone fitting to a musical scale. Rhyme is also a characteristic that many listeners indicate as evidence for song. Listeners rarely indicate characteristics of a song that move the judgement toward speech. Most justifications for this category of clip are indications of the features of song.

A clip using exaggerated pitch changes in the *sprechstimme* style was characterized primarily as song, but listeners disagreed on the characteristics responsible for the perception of song.

n132:317 (4) "close to singing but lacks tasteful note arrangement."

n132:220 (4) "There are tone changes but I wouldn't say it is exactly singing except perhaps at the end of the sample where the subject held the syllable and changed notes"

Judgements on the *quality* of singing were also made for this same clip.

n132:343 (5) “bad singing, but song-like

n132:251 (3) “that was just painful,not singing, but not speech”

These comments are somewhat surprising, because the clip was recorded by a professional musician and singer.

Several clips were considered song because of the way the clip ended. It seems in these cases, that the way a clip ends can influence the judgement for the entire clip.

u163:213 (4) “started off talking, ended off singing”

u176:231 (5) “sustained last two notes; pitch of last notes not like talking”

Clip u168 is a slightly shorter version of a clip u176, with the “musical” ending removed. Some listeners identified this and commented on it.

u176:329 (4) “almost 2 samples here, starts more speech like, ends with more singing tone”

u176:333 (5) “same as u168, but now it’s long enough to tell that it’s clearly singing”

Compare these ratings and reasonings to those made by the same listeners on the shorter version. The shortened version was presented earlier in the experiment.

u168:231 (3) “rhythm ;different syllable lengths, some tone differences not like speech”

u168:329 (3) “inflected speech, intermediate tension in production”

u168:333 (4) “pitch and loudness pattern of individual syllables seems subordinate to that of the entire phrase”

2 < Mean rating < 4

As is evident from Figure 2.1b, listeners disagree more in this range of mean rating. When one feature of song is present but others are not, as in the case of poetry, with rhythm but no musical melody, listeners disagree on the importance of the feature which is present. When a clip has more features of either speech or song, listeners tend to agree more, and the mean rating tends to approach one end or the other of the scale.

Some extracted sound clips are source-recognizable—listeners can identify the movie, music or television commercial where the sounds come from, and are able to obtain more context for the clip, which in some cases may contribute to their rating choice. Similarly, some solicited sound clips are content-recognizable—When a subject is prompted to sing or to produce an intermediate vocalization, they sometimes choose popular music or a movie quote. If the listener can recognize the context of the quote or the original source of the music, they obtain increased context for the clip and written comments indicate that this extra context influences their rating.

Clip n133 is a solicited clip where the subject is reciting a small segment of a popular song. Listeners who recognize the song rate the clip as singing:

n133:236 (5) “recognizable song”

n133:309 (4) “I know it’s a song. Also, the sound is clearly rythmic”

while those who don’t recognize the song rate the clip more toward talking:

n133:317 (2) “sounds more like poetry recital than someone singing”

n133:251 (2) “words too detached for song, too mashed together for speech”

A clip considered “rap” by several listeners has a mean rating of 3.74. Five subjects used the word “rap” in their justifications, four of whom rated the clip as 4, one rated it as 3:

u166:213 (4) “rap with very specific notes”

u166:311 (4) “has a ‘rap’ beat, words are spoken quickly within the beat, slightly difficult to understand”

u166:330 (3) “sounds like a rap”

u166:(352,360) (4) “rap”

As with many clips with mean ratings between 2 and 4, there is much disagreement in the interpretation of this clip.

u166:221 (5) “sounds like a tune, and regular beat”

u166:308 (3) “intentional rhythm, sentence had a ‘ground’ tone with shifts up and down”

u166:329 (1) “speech inflection, not singing tone”

A clip of chanting in Mon-kmer (from a Kmhmu Highlander) has a mean rating of 3.0. Three listeners used the word “chant” to describe this clip. The standard deviation for the ratings of this clip is 1.23, one of the higher values for this measure, indicating much disagreement in the interpretation of this clip. It is possible that since context and expectation are removed by the use of a presumably unknown language, listeners attend more to features of speech and song understood in their native language.

u172:221 (5) “the ending makes it clear, otherwise could have been a chant”

u172:311 (2) “has foreign language rhythm, with sounds short and concise except the last word at the end”

u172:359 (3) “sounds like someone praying, tyical in-between thing to me”

2.4.2 Comments on speech versus singing in general

In Part 3 of the study, listeners are invited to comment on their general observations of speech and song. This prompt is used:

> “In the following field, please write some general observations that you might have made over the course of this study about the characteristics of speech and singing, as well as the similarities and differences between them.”

Most of these comments describe singing in relation to speaking. This might indicate that listeners consider speaking to be the default human utterance, while singing might be considered to be a modification of speaking.

Several listeners provide their opinion of a definition of singing:

317 “Singing has melody and rhythm.”

328 “Rhyming when combined with musical scales is singing.”

346 “[. . .] song, which I understand to be rhythmic speech with distinct, sustained pitches that follow a musical scale of some sort.”

349 “Singing has rhythm, a certain sound and use of the vocal cords, a tune and emphasis on syllables not given such stress in speech.”

333 "[. . .] singing is imposing a nonlinguistic (presumably musical) pattern on speech."

One listener gives a definition of speech:

354 "Speech is the conveyance of meaning without rhythm, rhyme or deliberate pitch for enjoyment's sake."

It is unclear whether it is "the conveyance of meaning" or "deliberate pitch" which is (or is not) for enjoyment's sake according to this listener.

Features identified by listeners

Many listeners describe features that they consider relevant for speech or singing, usually in the context of singing as it compares to speech. The most common features are pitch (also described as tone and melody) and rhythm (also described as beat, speed, and patterns of rhythm). Some listeners also consider rhyme and repetition as features of song. Other listeners chose to describe speech as it compares to singing, but the same basic features were used, and the listeners described speech as the absence of the features necessary for singing, like pitch, rhythm, and vibrato. These features are described in more detail in Section 2.5.

Some listeners describe the differences between speech and song using more elusive terms which do not relate to a definable feature. These include "emotion", "flow" and "feeling". Some example comments are:

232 "There's something in the amount of 'feeling' behind what is being said that pushes it toward speech. Good luck quantifying that!"

236 "The process of identifying speech and singing is subjective to the listener as well as to the person who's voice is being heard."

347 "Near the end I became aware of a dimension of clarity and sharpness of pitch that characterizes singing."

The statistics in Section 2.3 show that there is greater agreement when a clip is rated closer to pure speech or pure song. There is greater subjectivity, as listener 236 puts it, when a clip is closer to the middle of the spectrum.

Other observations

An interesting observation that some listeners present is that the perception of a clip as speech or singing can vary temporally through the course of the clip, and is sometimes related to the expectation of the listener:

231 "Expectations seem to play a role—if we hear a pitch change we don't expect in regular speech, we might call it song."

308 "My opinion was often based on where I thought the speaker/singer was going next. I needed to extrapolate from the short clip."

359 "Some sounds show continuously intermediate stuff and some keep switching between one and the other."

Some listeners made comments about the distribution of the clips, the setup of the experiment, or the correctness of the proposed classification division:

330 "Most of them sounds more like talking. I think to be qualified as singing, it has to sound much nicer, more rhythmic and with tone changes (but not like speaking in a drama)."

343 "Being a singer, I thought the majority of the sounds were speaking. I believe that a lot of them would be labeled song by a non-musician."

346 "Most of the examples were of speech, but many were not examples of ordinary 'talking'."

357 "I don't believe that it is necessarily appropriate to set up a two-part division of these stimuli. I may have grouped items differently, if it were not set that they must fall along a single axis from speech to song."

Some observations that listeners make do not agree with experimental results, making the task of extracting information from the opinion of listeners even harder. One listener described speech as monotone and non-rhythmic in comparison to song, but while English speech does tend to be more constrained that song in pitch range, there are numerous observed phenomenon where pitch in speech varies significantly for a communicative purpose, such as end-of-sentence raising for questions, and prosodic emphasis. Observations of these phenomenon are limited to the English language utterances in the corpus.

2.5 Features of Speech and Song

From the comments provided by the listeners, and by the observations of the author of the similarities between similarly rated files, the following features are identified as potentially useful in the classification of human utterances in the speech/song domain.

Pitch is a primary feature in the speech/song axis. Speech utterances sometimes have a monotone pitch, but often have a pitch track that moves in what listeners refer to as a "speech-like" way, varying across syllables and indicating speaker intent or other prosodic characteristics. Sung utterances have a noticeable melody, and target pitches adhere to some form of musical scale. Vibrato is a key feature that can provide evidence for song, although it is important to note that, like all of these features, it is not universal—there are circumstances where an utterance with a vibrato-like pitch track may be classified as something other than pure song.

Rhythm is a feature that many listeners indicate as evidence for a speech/song classification. Speech with a discrete rhythm may be classified closer to song. This feature is more difficult to measure—amplitude cues that humans use to lock on to a rhythmic pattern are difficult to identify amid other power-based features that humans might ignore for a particular rhythmic pattern. Power fluctuations in specific frequency bands may be useful for this task.

Rhyme, usually indicative of poetry or singing, is another feature difficult to quantify. Rhyme is a higher-level repetition than rhythm, taking phonetic information into account. It is unclear at this point wether it will be possible to detect rhyme in an utterance without a phoneme recognition engine such as those used in automatic speech recognition systems.

Many listeners justify their classifications with comparative statements, such as "it uses speech-like rhythms", as well as indicating context, as in "I've heard this before" and expectation, as in "it depends on where he's going next." These are more difficult to quantify and especially difficult to emulate with computational algorithms. One specific piece of future work will be to compare sound files with similar descriptions, and try to identify features that might be responsible for these classifications.

2.6 Discussion

The collection and annotation of this speech/song corpus serves two important purposes. It provides a research tool for examining the differences among human utterances between speech and song, and serves as a starting ground for the development of tools which will be able to perform this fuzzy classification. It also provides insight into the human perception of the differences between speech and song, and human utterances in general.

Features that are important in the classification of speech and song include pitch, rhythm, rhyme, repetition, vibrato, expectation and context. Some of these features are measurable from statistical observations of the sound waveform. Expectation and context are difficult to quantify and measure, since they relate to world-view and heuristic knowledge.

Chapter 3

f_0 Estimation

Fundamental frequency (f_0) estimation, also referred to as pitch detection, has been a popular research topic for many years, and is still being investigated today. At the 2002 IEEE International Conference on Acoustics, Speech and Signal Processing, there was a full session on f_0 estimation. The basic problem is to extract the fundamental frequency (f_0) from a sound signal, which is usually the lowest frequency component, or *partial*, which relates well to most of the other partials. In a periodic waveform, most partials are harmonically related, meaning that the frequency of most of the partials are related to the frequency of the lowest partial by a small whole-number ratio. The frequency of this lowest partial is f_0 of the waveform.

Most research into this area goes under the name of pitch detection, although what is being done is actually f_0 estimation. Because the psychological relationship between f_0 and pitch is well known, it is not an important distinction to make, although a true pitch detector should take the perceptual models into account and produce a result on a pitch scale rather than a frequency scale.

Current speech recognition engines often discard the pitch information as irrelevant to the recognition task. While it is true that individual phonemes are recognizable regardless of the driving pitch, or even in the absence of pitch (recall Figure 1.3), this does not imply that pitch information is not useful. Much semantic information is passed on through pitch that is above the phonetic and lexical levels. In tonal languages, the relative pitch motion of an utterance contributes to the lexical information in a word. In this case, speech recognition algorithms must attend to the pitch or the context of the utterance to avoid ambiguity.

3.1 Theory of Pitch

The musical pitch of an audio signal is a perceptual feature, relevant only in the context of a human listening to that signal. The musical pitch scales that are used today were developed before people knew about frequency and spectral content, and was based on the similarity or dissimilarity of the note. Pitch is loosely related to the log of the frequency, perceived pitch increasing about an octave with every doubling in frequency. However, frequency doubling below 1000 Hz corresponds to a pitch interval slightly less than an octave, while pitch doubling above 5000 Hz corresponds to an interval slightly more than an octave [16, 30]. This relationship also changes with intensity. The perceived pitch of a sinusoid increases with intensity when the sinusoid is above 3000 Hz, and a sinusoid with frequency below 2000 Hz is perceived to drop in pitch as the intensity increases [8].

It is important to note that these measurements of the differences between frequency and the perception were made on isolated sinusoids. Real-world sounds have many harmonics above the fundamental frequency. The perception of pitch changes with this harmonic content as well. A richer spectrum seems to reinforce the sensation of the pitch, making the octave seem more “in-tune”. The more sine-like a waveform is, the more distinct the notion of frequency, but the less distinct the perception of pitch [73]. This sensation also varies with the relationship between the partials. The more harmonically related the partials of a tone are, the more distinct the perception of pitch. Pitch perception also changes with intensity, duration and other physical features of the waveform.

There is some controversy as to how the human auditory system perceives pitch [5, 46, 81]. One group of people have traditionally used pure tone pitches to measure phenomena like critical bands, masking, and pitch perception. The other group of people use more complex tones to see how humans perceive groups of sounds and dissect the “scene” of sound around them. There are also important observations arising from the psychology, psychoacoustics and psychophysics being researched around the perception of tones and pitch, which provide insight into the problem of automatic f_0 estimation. For our purposes, it is less important to decide which general theory of audition is right, and more important to glean information about how humans perceive pitch from each group of researchers.

3.2 Automatic f_0 Estimation

Fundamental frequency estimation has consistently been a difficult topic in audio signal processing. Many context-specific attempts have been made, and many of them work well in their specific context, but it has been difficult to develop a "context-free" f_0 estimator. f_0 estimators developed for a particular application, such as musical note detection or speech analysis, are well understood, but depend on the domain of the data: a detector designed for one domain is less accurate when applied to a different domain. The result is that there are many f_0 estimators currently on the market, but few that are appropriate to more than one domain.

Therefore, choosing a f_0 estimator for a speech/song discrimination is a difficult task because detectors that work well for music, and hence for song, work less well for speech, and vice versa. Three possible solutions to this problem are: find a detector that is reasonably good for both speech and song; build a detector that works very well for both speech and song; or use two f_0 estimators, one suited to speech and one suited to song, and compare the results. The latter generates two positive outcomes: the f_0 estimation is more reliable, and the differences between the f_0 estimations can be used as a classification feature between speech and song. For this thesis, f_0 estimators developed for speech and for instrumental music were found, but not specifically for vocal music. For this reason, it was decided to evaluate a set of f_0 estimators and choose one which was mostly accurate for both speech and song.

3.2.1 Evaluating f_0 estimators

It is difficult to empirically measure the performance of a f_0 estimator for several reasons. First, performance depends on domain, as discussed above. A f_0 estimator will almost certainly behave better in the context for which it was developed. Second, it is difficult to automatically rate the result of a f_0 estimator against expected outcomes, precisely because it is difficult to measure f_0 in the first place. We humans are good at it, and so we can listen to a file and judge the accuracy of a f_0 estimation engine, but to lend credibility to this measure, we must have many people, both expert and lay, judge the f_0 estimation result on a large number of sound files. Once a measure like this is taken, however, it can be used to evaluate the results of other f_0 estimation methods. Another way to evaluate f_0 estimators is to compare the results of multiple detectors on a common corpus. If, for a set of n detectors, $k \approx n$ of them agree, it is likely that the remaining $n - k$ are incorrect.

This third method of comparison is what will be used in this work. Section 3.7 presents an evaluation of three f_0 estimators by comparing their results. Errors in one f_0 estimator provide evidence that the other two are likely to be more accurate, and visual inspection of the f_0 tracks which are significantly different provide further insight into which f_0 track techniques may be better than others. The f_0 tracks will be evaluated based on the corpus discussed in Chapter 2. This is reasonable since it is this data on which the f_0 estimators will ultimately be used.

3.2.2 Measuring Frequency

There are a number of standard methods that researchers use to extract f_0, based on various mathematical principles. Since pitch is a perceptual quantity related to f_0 of a periodic or pseudo-periodic waveform, it should suffice to determine the period of such oscillation, the inverse of which is the frequency of oscillation. The problem comes when the waveform consists of more than a simple sinusoid. As harmonic components are added to a sinusoidal waveform, the appearance of pitch of the waveform becomes less clear and the concept of "fundamental frequency" or f_0 must be considered. The goal of a f_0 estimator is to find f_0 in the midst of the other harmonically related components of the sound.

The difficulty of finding the f_0 of a waveform depends on the waveform itself. If the waveform has few higher harmonics or the power of the higher harmonics is small, the f_0 is easier to detect, as in Figures 3.1 and 3.2. If the harmonics have more power than the f_0, then the period is harder to detect, as in Figures 3.3 and 3.4. Figure 3.4 is an example of the phenomenon of the missing fundamental.

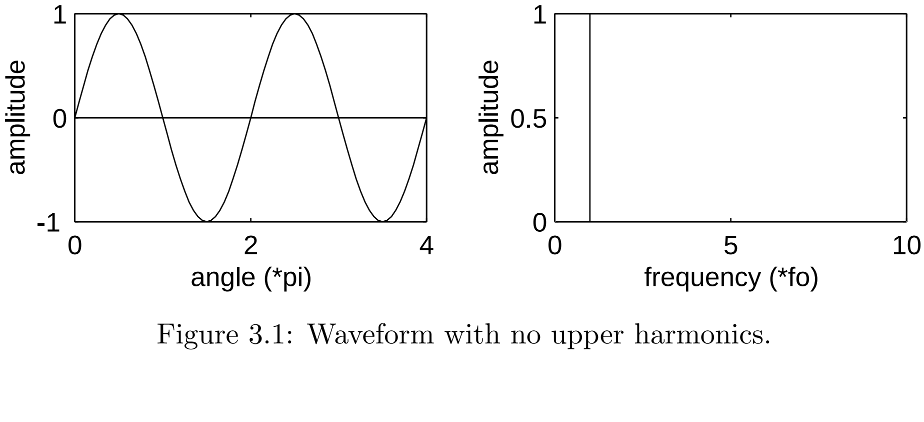


Figure 3.1: Waveform with no upper harmonics.

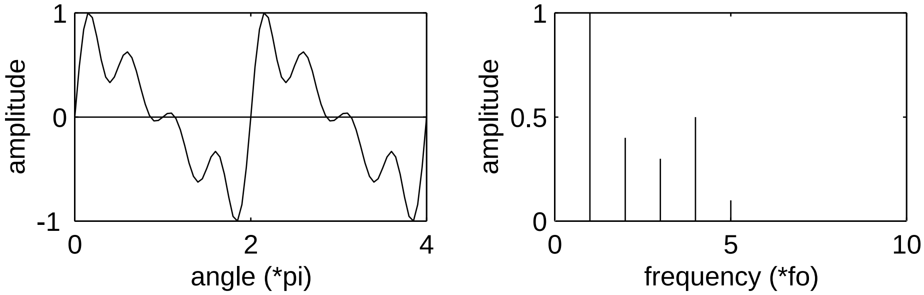


Figure 3.2: Waveform with lower power upper harmonics.

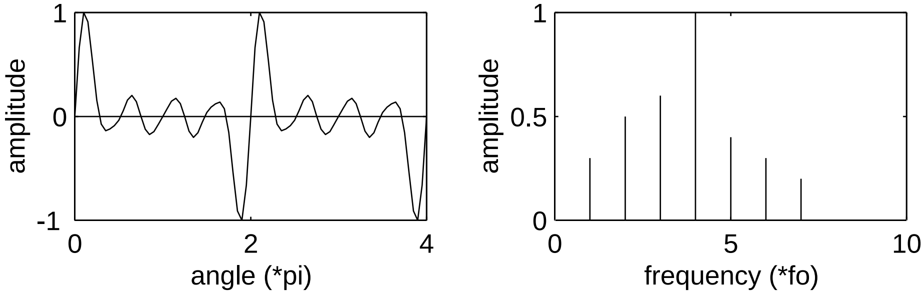


Figure 3.3: Waveform with higher power upper harmonics.

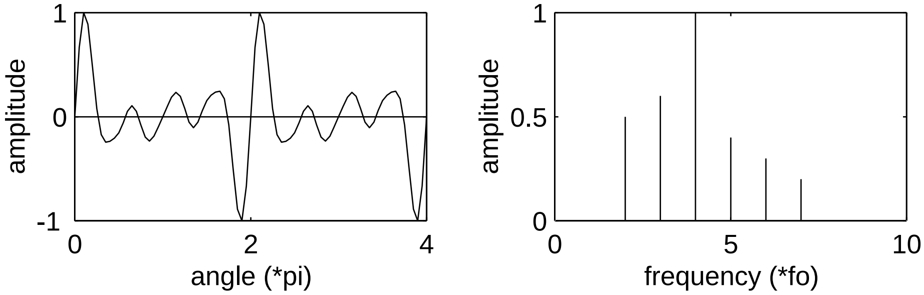


Figure 3.4: Waveform with high power upper harmonics and no fundamental.

The next sections in this chapter discuss three general domains of f_0 estimation algorithms, organized by the type of input and the processing paradigm. Time domain methods are presented first, as they are usually computationally simple. Frequency domain methods, presented next, are usually more complex. Statistical methods use probability theory to aid in a decision. After this, Section 3.6 discusses improvements that can be applied to any f_0 estimation algorithm, and Section 3.7 presents a comparison and evaluation of some freely available algorithms. The chapter concludes with a discussion.

3.3 Time-Domain Methods

The most basic approach to the problem of f_0 estimation is to look at the waveform that represents the change in air pressure over time, and attempt to detect the f_0 from that waveform.

3.3.1 Time-Event Rate Detection

There is a family of related time-domain f_0 estimation methods which seek to discover how often the waveform fully repeats itself. The theory behind these methods is that if a waveform is periodic, then there are extractable time-repeating events that can be counted, and the number of these events that happen in a second is inversely related to the frequency. Each of these methods is useful for particular kinds of waveforms. If there is a specific time-event that is known to exist once per period in the waveform, such as a discontinuity in slope or amplitude, it may be identified and counted in the same way as the other methods.

Zero-crossing rate (ZCR). As discussed in Section 1.7.1, the ZCR of a waveform is the number of times that the waveform changes sign. If the power of the waveform is concentrated in the fundamental frequency, then it should cross zero twice per cycle, once from positive to negative and once from negative to positive. Variation on this method include counting only positive-slope zero crossings, and measuring the distance between the zero-crossings.

ZCR detection has been used in the context of f_0 estimation as recently as [60], where the mean and the variance of the zero crossing rate were calculated to increase the robustness of a feature extractor. The feature is not used to measure the f_0 directly, but is used instead to track the constancy of the f_0 across time frames. If the waveform is steady-state or slowly varying, as is the case in most pseudo-periodic musical signals, the mean and variance of the ZCR will be consistent over the course of a note, and thus these features can be used to detect note boundaries, glissade and frequency modulation effects.

Peak rate. This method counts the number of positive peaks per second in the waveform. In theory, the waveform will have a maximum value and a minimum value each cycle, and one needs only to count these maximum values (or minimum values) to determine the frequency of the waveform. In practice, a local peak detector must be used to find where the waveform is locally largest, and the number of these local maxima in one second is the frequency of the waveform, unless each period of the waveform contains more than one local maximum. Similar alternatives are available for this method as are available for the zero-crossing rate detector—the distance between the local maxima gives the wavelength which is inversely proportional to the frequency.

Slope event rate. If a waveform is periodic, the slope of the waveform will also be periodic, and peaks or zeros in the slope can be extracted in the same way as the ZCR. In some cases, zeros or peaks in the slope might be more informative than zeros or peaks in the original waveform, or the detection of these events might be more robust, depending on the domain of the signal.

Discussion

The major difficulty with time-event rate detection methods is that spectrally complex waveforms rarely have just one event per cycle. Waveforms with rich harmonic spectra may cross zero many times or have many peaks in a cycle (recall Figure 1.4).

There are some positive aspects of time-event rate detection algorithms. These methods are exceedingly simple to understand and implement, and they take very little computing power to execute. If the nature of the signal is known, a method can be implemented which is tailored to the waveform, reducing the error. Peak counters have been the implementation of choice for hardware frequency-detectors for may years, because the circuit is very simple, and coupled with a simple low-pass filter, provides a fairly robust module.

3.3.2 Autocorrelation

The correlation between two waveforms is a measure of their similarity. The waveforms are compared at different time intervals, and their "sameness" is calculated at each interval. The result of a correlation is a measure of similarity as a function of time lag between the beginnings of the two waveforms. The *auto*correlation function is the correlation of a waveform with itself. One would expect exact similarity at a time lag of zero, with increasing dissimilarity as the time lag increases. The mathematical definition of the autocorrelation function is shown in Equation 3.1, for an infinite discrete function $x[n]$, and Equation 3.2 shows the mathematical definition of the autocorrelation of a finite discrete function $x'[n]$ of size N.

$$\begin{aligned} R_x(\nu) &= \sum_{n=-\infty}^{\infty} x[n]x[n+\nu] && (3.1) \\ R_{x'}(\nu) &= \sum_{n=0}^{N-1-\nu} x'[n]x'[n+\nu] && (3.2) \end{aligned}$$

The cross-correlation between two functions $x[n]$ and $y[n]$ is calculated using Equation 3.3:

$$R_{xy}(\nu) = \sum_{n=-\infty}^{\infty} x[n]y[n+\nu]. \tag{3.3}$$

The value of $R_{xy}(0)$ gives a measure of the similarity of two separate functions. This measure will be used to evaluate feature extractors in Chapter 4.

Periodic waveforms exhibit an interesting autocorrelation characteristic: the autocorrelation function itself is periodic. As the time lag increases to half of the period of the waveform, the correlation decreases to a minimum. This is because the waveform is out of phase with its time-delayed copy. As the time lag increases again to the length of one period, the autocorrelation again increases back to a maximum, because the waveform and its time-delayed copy are in phase. The first peak in the autocorrelation indicates the period of the waveform.

Problems with this method arise when the autocorrelation of a harmonically complex, *pseudo*periodic waveform is taken. One can imagine the output of an autocorrelation applied to the waveform in Figure 1.4b. The first peak would not be at the period of the full waveform, but at the period of the 20th harmonic overtone. The first "large" peak would indeed occur at the fundamental period of the waveform, but it reduces the robustness and increases the computational complexity to have the algorithm try to distinguish between "large" and "small" peaks.

The YIN f_0 estimator

The YIN f_0 estimator [10], developed by Alain de Cheveigné and Hideki Kawahara, is named after the oriental yin-yang philosophical principal of balance, representing this author's attempts to balance between autocorrelation and cancellation in the algorithm. The difficulty with autocorrelation techniques has been that peaks occur at sub-harmonics as well, and it is sometimes difficult to determine which peak is the fundamental frequency and which represent harmonics or partials. YIN attempts to solve these problems by in several ways.

YIN is based on the difference function, which, while similar to autocorrelation, attempts to *minimize the difference* between the waveform and its delayed duplicate instead of *maximizing the product* (for autocorrelation). The difference function is presented in Equation 3.4.

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2 \tag{3.4}$$

In order to reduce the occurrence of subharmonic errors, YIN employes a cumulative mean function which de-emphasizes higher-period dips in the difference function:

$$d'_t(\tau) = \begin{cases} 1, & \tau = 0 \\ \frac{d_t(\tau)}{\frac{1}{\tau}\sum_{j=1}^{\tau} d_t(j)} & \text{otherwise} \end{cases} \tag{3.5}$$

Other improvements in the YIN f_0 estimation system include a parabolic interpolation of the local minima, which has the effect of reducing the errors when the period estimation is not a factor of the window length used (in this case, 15 ms). For a more complete discussion of this method, including computational implementation and results, see the cited paper.

3.3.3 Phase Space

The phase space signal representation is a way of observing the short-time history of a waveform in a way that makes repetitive cycles clear. The basic phase space representation is to plot the value of the waveform at time t versus the slope of the waveform at the same point [28]. A periodic signal should produce a repeating cycle in phase space, returning to a point with the same value and slope. Higher dimension phase space representations plot the value and $n-1$ derivatives of the signal in n dimensions.

Pseudo-phase space, also called embedded representation, is a simpler form of phase space. The value of the incoming waveform is plotted against a time-delayed version of itself. The representation plots the points $(x, y) = (f(t), f(t-\tau))$, and in the n-dimensional case, $(x_0, x_1, \ldots, x_{n-1}) = (f(t), f(t-\tau_1), \ldots, f(t-\tau_{n-1}))$. Often, for simplicity, $\tau_k = k\tau_1$.

In the remainder of this discussion, "phase space" refers to the general class of representations that include multi-dimensional phase space and pseudo-phase space representations, unless otherwise stated. For a more detailed discussion of a theoretical phase space f_0 estimator, see [22, 71, 72].

Phase Space for f_0 estimation

Any periodic signal forms a closed cycle in phase space, and the shape of the cyclic path depends on the harmonic composition of the signal. The f_0 of a signal is related to the speed with which the path completes the cycle in phase space. The task then becomes detecting the difference between new values in phase space crossing the old path, and new values intersecting and re-tracing the old path. The simplest solution would be to compare distances between points in phase space, and detect when the distance becomes minimal. An initial point would be selected, and the distance from that point would be traced as a function of time. When this distance became zero (or a minimum value) the waveform may have repeated.

This solution is akin to the problem of zero-crossing rate detection, with the associated problems. The phase space cycle might be retracing itself, or it might be crossing itself. It is clear that a simple distance measure will not be sufficient to measure the repetition rate. The distance in higher dimensions might yield a better result—it is conceivable that paths which overlap in two-dimensional space will not overlap in higher dimensions. The question to ask is how many dimensions are required to ensure that the only time the path of the signal intercepts itself is when it begins to repeat itself. The answer to this question will depend on the type of data being investigated, but for band-limited periodic signals, this dimension will be finite. A proof of this statement follows.

Theorem 3.1. *Given a band-limited periodic signal, a phase space representation can be* *constructed requiring a finite number of dimensions.*

Proof. A band-limited signal can be represented as a discrete time series sampled at twice the maximum frequency of the signal (shanon). Since the given signal is periodic, the corresponding time series can be represented by a finite number of samples, repeated infinitely many times. Consider a time series f of n samples. For this series, n difference measures d can be made for each sample, corresponding to the first n derivatives of the continuous signal. For $f(0)$, these are:

$$\begin{aligned} d_1(0) &= f(1) - f(0), \\ d_2(0) &= f(2) - f(0), \\ &\dots \\ d_n(0) &= f(n) - f(0). \end{aligned}$$

No further difference measures can be made since for the periodic signal, $f(n+1) = f(0)$, and $d_{n+1}(0) = f(n+1) - f(0) = f(0) - f(0) = 0$. In general, $d_{n+1}(k) = f(k+n+1) - f(k) = f(k) - f(k) = 0$, and differences above d_{n+1} cycle back to the values of the original differences.

Since the number of difference measures is finite, the number of dimensions required to define them is also finite, and the set of n differences represents a unique point in the n-dimensional space, which will be passed through only once per cycle. □

It is important to note that this proof amounts to a sufficient condition: It is possible to fully represent the phase space of all derivatives in a finite-dimensional hyperspace. It may not be necessary to use all of these dimensions to fully represent the waveform in a non-intersecting hyperspace path. If the signal is band-limited, fewer dimensions are necessary, and in the degenerate case of a sinusoid, only two dimensions are necessary to fully represent the cycle in a non-intersecting hyperspace path. If the amplitude and first derivative of a sinusoid are plotted against each other, the result is a circle.

While the number of dimensions may be finite, the window size must be kept small. Otherwise, the dimensionality of the fully represented phase space will be unwieldy. If the window size is smaller than a complete cycle of the periodic waveform, there would be insufficient information to determine the frequency.

Phase Space of Pseudo-Periodic Signals

A bigger problem with phase space f_0 estimation is how to deal with *pseudo*-periodic signals. In a phase-space representation, the path of a pseudo-periodic signal will never re-trace itself, although it will follow a closely parallel path.

A *Poincaré section* of a phase space plot is a lower-dimensional orthogonal slice through the plot which produces a cross-section of a path being considered. A Poincaré section of a periodic signal will be one or more discrete points, indicating the locations that the path intersects the section.

A pseudo-periodic signal will generate a cloud of points in a Poincaré section, localized in one or more clusters. If these clusters are separate, the mean location of each cluster can be treated as the intersection point for that cluster, and the period can be calculated by the time lag between successive points in the same cluster.

A problem arises when two clusters of points are close together, such that for some points it is not clear which cluster they should belong to. In this case, higher-dimension phase-space representations should be employed until the clusters are shown to be disjoint. There are many potential problems with this suggested method, but it may provide another alternative to the many f_0 estimation algorithms that are currently available.

3.4 Frequency-Domain Methods

There is much information in the frequency domain that can be related to the f_0 of the signal. Pitched signals tend to be composed of a series of harmonically related partials, which can be identified and used to extract the f_0. Many attempts have been made to extract and follow the f_0 of a signal in this manner.

3.4.1 Component Frequency Ratios

As early as 1979, Martin Piszczalski was working on a complete automatic music transcription system, the first step of which would be pitch detection [52, 53]. His system would extract the pitch of the signal (assuming that a single note was present at each point in time) and then find note boundaries, infer pitch key, and present a score.

Piszczalski's original procedure began with a spectral transform and identification of the partials in the signal, using peak detection. For each pair of these partials, the algorithm finds the "smallest harmonic numbers" that would correspond to a harmonic series with these two partials in it. As an example, if the two partials occurred at 435 Hz and 488 Hz, the smallest harmonic numbers (within a certain threshold) would be 6 and 7, respectively. Each of these harmonic number pairs are then used as a hypothesis for the fundamental frequency of the signal. In the previous example, the pair of partials would correspond to a hypothesis that the fundamental frequency of the signal is about 70 Hz. After all pairs of partials are considered in this way, the hypothesis most strongly suggested by the pairs of partials is chosen as the fundamental frequency. Some pairs of partials are weighted higher, meaning that their "vote" for the fundamental frequency of the signal counts for more than other pairs of partials. The weighing factor depends on the amplitude of the signals—higher amplitude pairs are counted more than lower amplitude pairs.

This method does not require that the fundamental frequency of the signal be present, and it works well with inharmonic partials and missing partials.

Dorken and Nawab presented an improvement to Piszczalski's method in [13]. They suggest "conditioning" the spectrum using a method they had previously used for principal decomposition analysis. This conditioning had the effect of identifying the frequency partials more accurately, and hence making the entire transform more accurate. Another improvement that they propose is to perform the entire transform in a constant-Q domain, making lower-frequency partials better defined, in an attempt to make the transform closer to human perception.

3.4.2 Filter-Based Methods

Filters are used for f_0 estimation by trying different filters with different centre frequencies, and comparing their output. When a spectral peak lines up with the passband of a filter, the result is a higher value in the output of the filter than when the passband does not line up.

Optimum Comb Filter

The optimum comb f_0 estimator [47] is a robust but computationally intensive algorithm. A comb filter has many equally spaced pass-bands. In the case of the optimum comb filter algorithm, the location of the passbands are based on the location of the first passband. For example, if the centre frequency of the first passband is 10 Hz, then there will be narrow pass-bands every 10 Hz after that, up to the shanon frequency.

In the algorithm, the input waveform is comb filtered based on many different frequencies. If a set of regularly spaced harmonics are present in the signal, then the output of the comb filter will be greatest when the passbands of the comb line up with the harmonics. If the signal has only one partial, the fundamental, then the method will fail because there will be many comb filters that will have the same output amplitude, wherever a passband of the comb filter lines up with that fundamental.

Tunable IIR Filter

A more recent filter-based f_0 estimator suggested in [38], this method consists of a narrow user-tunable band-pass filter, which is swept across the frequency spectrum. When the filter is in line with a strong frequency partial, a maximum output will be present in the output of the filter, and the f_0 can then be read off the centre frequency of the filter. The author suggests that an experienced user of this tunable filter will be able to recognize the difference between an evenly spaced spectrum, characteristic of a richly harmonic single note, and a spectrum containing more than one distinct pitch. The paper also presents suggestions for automating this search procedure, as a computer would be faster at scanning the frequency spectrum and more accurate at identifying the difference between a richly harmonic single note and multiple concurrent notes.

This f_0 estimation method is in some ways similar to the operation of the stroboscope, a tool used by piano tuners. The tool consists of a spinning disk with black and white marks, illuminated by a strobe light. The strobe is connected to a microphone, and emits a pulse of light as the input signal peaks, once per period. The spinning disk can be sped up or slowed down until the disk is illuminated once every rotation. This can be seen when the black and white marks on the disk appear stationary.

3.4.3 Cepstrum Analysis

Cepstrum analysis is a form of spectral analysis where the output is the Fourier transform of the log of the magnitude spectrum of the input waveform [18]. This procedure was developed in an attempt to make a non-linear system more linear. Naturally occurring partials in a frequency spectrum are often slightly inharmonic, and the cepstrum attempts to mediate this effect by using the log spectrum.

The name cepstrum comes from reversing the first four letters in the word "spectrum", indicating a modified spectrum. The independent variable related to the cepstrum transform has been called "quefrency", and since this variable is very closely related to time [57] it is acceptable to refer to this variable as time.

The theory behind this method relies on the fact that the Fourier transform of a pitched signal usually has a number of regularly spaced peaks, representing the harmonic spectrum of the signal. When the log magnitude of a spectrum is taken, these peaks are reduced, their amplitude brought into a usable scale, and the result is a periodic waveform in the frequency domain, the period of which (the distance between the peaks) is related to the fundamental frequency of the original signal. The Fourier transform of this waveform has a peak at the period of the original waveform.

Figure 3.5 shows the progress of the cepstrum algorithm. Figure 3.5b shows the standard spectral representation of a periodic harmonic signal (whistling at A_4). Figure 3.5c shows the log magnitude spectrum of the same signal. Note the periodicity of both spectra, and the re-scaled nature of the log magnitude spectrum.

The cepstrum method assumes that the signal has regularly-spaced frequency partials. If this is not the case, such as with the inharmonic spectrum of a bell or the single-partial spectrum of a sinusoid, the method will provide erroneous results. As with most other f_0 estimation methods, this method is well suited to specific types of signals. It was originally developed for use with speech signals, which are spectrally rich and have evenly spaced partials.

3.4.4 Multi-Resolution Methods

An improvement that can be applied to any spectral f_0 estimation method is to use multiple resolutions [19]. The idea is relatively simple: If the accuracy of a certain algorithm at a certain resolution is somewhat suspect, confirm or deny any f_0 estimator hypothesis by using the same algorithm at a higher or lower resolution. Thus, use a bigger or smaller time window to calculate the spectrum. If a frequency peak shows up in all or most of the windows, this can be considered a confirmation of the f_0 estimator hypothesis. However, each new analysis resolution means more computational expense, which is why multi-resolution Fourier analysis is slower than a dedicated multi-resolution transform such as the discrete wavelet transform.

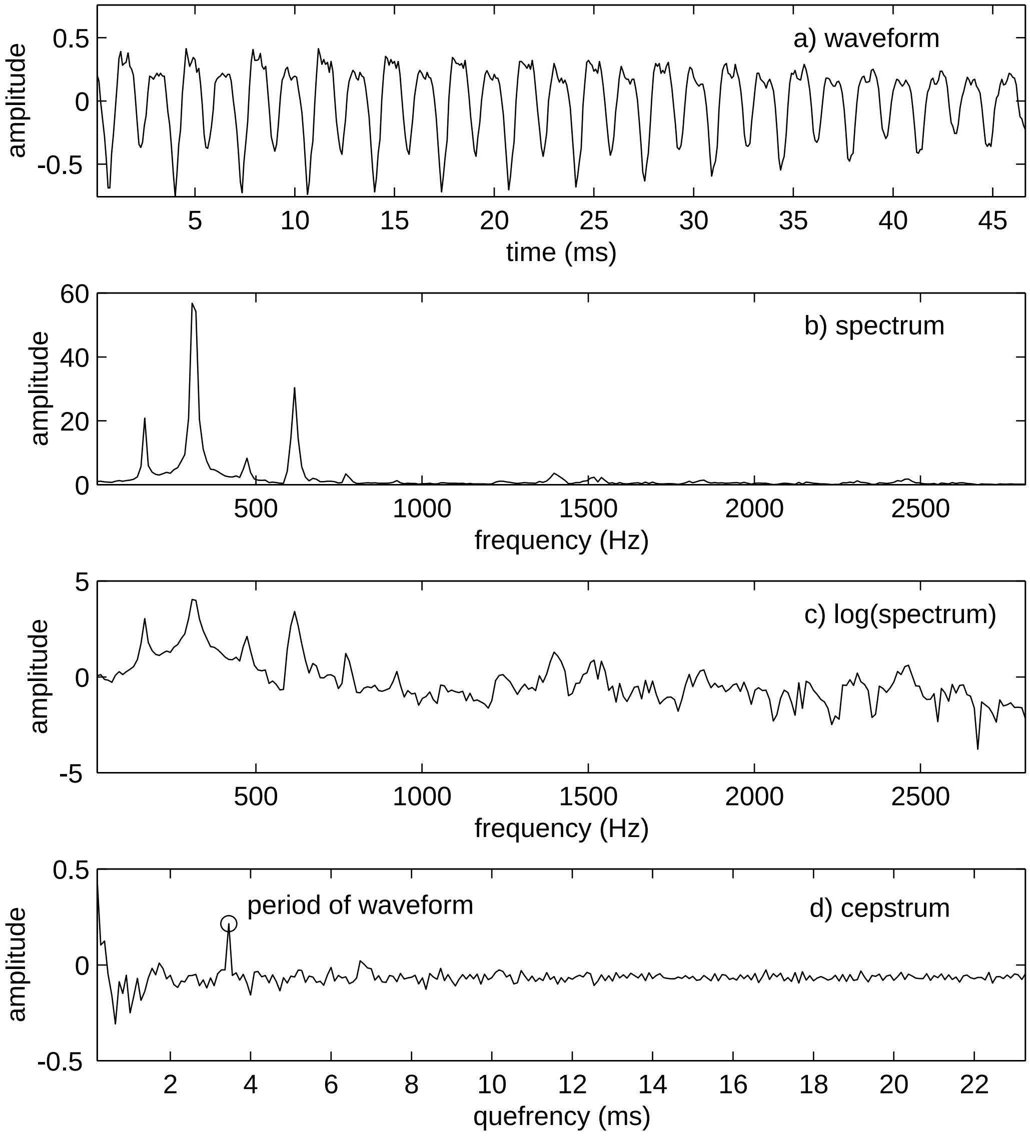


Figure 3.5: Stages in the cepstrum analysis algorithm.

3.5 Statistical Frequency Domain Methods

The problem of automatic f_0 estimation can be considered, in some ways, a statistical one. Each input frame is classified into one of a number of groups, representing the f_0 estimator of the signal. Many researchers have thought that modern statistical methods might be applied to the problem of f_0 estimation. Two such methods are presented here.

3.5.1 Neural Networks

Connectionist models, of which neural nets are an example, are self-organizing pattern matchers, providing a classification output for messy or fuzzy input. Logically, they consist of a collection of nodes, connected by links with associated weights. At each node, signals from all incoming links are summed according to the weights of these links, and if the sum satisfies a certain transfer function, an impulse is sent to other nodes through output links. In the training stage, input is presented to the network along with a suggested output, and the weights of the links are altered to produce the desired output. In the operation stage, the network is presented with input and provides output based on the weights of the connections.

The choice of the dimensionality and domain of the input set is crucial to the success of any connectionist model. A common example of a poor choice of input set and test data is the Pentagon's foray into the field of object recognition. This story is probably apocryphal and many different versions exist on-line, but the story describes a true difficulty with neural nets. As the story goes, a network was set up with the input being the pixels in a picture, and the output was a single bit, yes or no, for the existence of an enemy tank hidden somewhere in the picture. When the training was complete, the network performed beautifully, but when applied to new data, it failed miserably. The problem was that in the test data, all of the pictures that had tanks in them were taken on cloudy days, and all of the pictures without tanks were taken on sunny days. The neural net was identifying the existence or non-existence of sunshine, not tanks.

A connectionist model for the recognition of pitch might take as input a set of spectral partials, or the time-domain waveform, or the phase space representation of the signal. It would likely output a frequency hypothesis, which could then be translated to pitch.

Another approach to using connectionist models for f_0 estimation is the modeling of the human auditory system, as in [61], where the authors present a neural network model based on the cochlear mechanisms of the human ear. Other neural network models could be based on the functioning of the neural pathways (although a good model of this activity has not yet been developed) or could be based on the psychological reaction to pitch. Whatever the case, for a connectionist model, input domain and training data must be chosen carefully.

Another problem with connectionist models is that even if a good model is found, it does not provide any understanding of how the problem is solved. All of the algorithmic information in the model is stored in the weights of the connections, and in large models with thousands or millions of connections, it is prohibitively complicated to translate these weights into a description or algorithm. One must be happy with the "black box" doing what it does without knowing why or how.

3.5.2 Maximum Likelihood Estimators

Boris Doval and Xavier Rodet have presented a series of papers on f_0 estimation using maximum likelihood estimators [14, 15]. This statistical technique compares different variable value hypotheses based on the likelihood of their being correct in context with the past values of these variables. The intent is to recognize and deal with the slight inharmonicity of naturally occurring frequency partials in a pitched signal.

The model they present is set up as follows: an observation O consists of a set of partials in a short-time Fourier transform representation of a sound. Each observation is assumed to have been produced by a sound with a particular fundamental frequency f_0, and each spectrum contains other information including inharmonic and non-sinusoidal partials (noise). This model is a simplification of the general sound model, assuming that a sound consists primarily of harmonic partials at integer multiples of f_0, with a minority of inharmonic partials and noise.

For a set of candidate fundamental frequencies, the algorithm computes the probability (likelihood) that a given observation was generated from each f_0 in the set, and finds the maximum. The choice of the set of fundamental frequencies is important, because theoretically, the observation could originate from *any* f_0.

3.6 General Improvements

Most of the models described can be improved by pre-processing the input, reducing the input domain, or by increasing the frequency or time resolution of the input depending on whether the input data is time or frequency information. There are two more major improvements that can be employed by most of these methods, and these are described below.

3.6.1 Human Auditory Modeling

Because pitch detection (and hence f_0 estimation) is, by its nature, a perceptual problem, any algorithm designed specifically for pitch should be able to be improved by adding some characteristics of the human auditory system. A simple improvement that can be added to any frequency-domain method is to use a constant-Q spectral transform instead of a basic Fourier spectrum. As described in Section 1.3.2, a constant-Q transform is more computationally demanding, but is more faithful to the human auditory perceptual system.

Two factors must be considered when deciding whether or not to use human auditory modeling. First, the application for which the detector be used. If the goal is simply to detect the fundamental frequency of the signal without consideration of the pitch, human perceptual factors are probably not very important. However, if the goal is to detect the pitch for a transcription application, human factors are more relevant. The second factor is computational complexity. Human auditory modeling often results in a significant increase in the computation time required for the application. If computation time is a domain constraint, it may be necessary to forego auditory modeling in favor of a method which is faster but less physiologically accurate.

If properties of the human auditory system are to be used in any application, including f_0 estimation, we must first understand the human perceptual system much better than we currently do. Presently, the most we can do is make the computer system provide the same type of results that the human system does, and hope that these improvements will make the system more accurate and robust.

3.6.2 f_0 estimator Tracking

An improvement that several researchers have implemented, applicable to any f_0 estimation algorithm, is tracking [15]. A f_0 estimation based on a single spectral window, no matter how high the resolution of the spectral representation or how robust the algorithm, is the f_0 estimation of a single frame of time. The human system tracks the pitch of an incoming waveform, allowing us to identify such phenomena as *glissandi* (a smooth transition from one pitch to another) and pitch intervals. A time window containing a definite pitch of a small number of cycles is often very difficult for a human to identify [57], but when many time windows are played one after another, a sensation of pitch becomes apparent.

A simple modification to a f_0 estimation algorithm which can improve performance without increasing the computational burden is to give preference to f_0 hypotheses that are close to the f_0 hypothesis of the last time frame. Storing the f_0 hypothesis of the n previous time frames requires only n more memory words, and the comparison to the present hypothesis is a simple operation for each past time frame considered.

A more involved comparison method is the use of hidden Markov models (HMMs), statistical models which track variables through time [11]. These models have been used to solve many problems in linguistics and circuit theory as well as f_0 estimation. HMMs are state machines, with a hypothesis available for the output variable at each state. At each time frame, the HMM moves from the current state to the most likely next state, based on the input to the model and the state history which is represented in the current state. The state transition properties of HMMs are calculated using input-output pairs, consisting of (in the case of f_0 estimation) a set of spectral windows (or a set of spectral partials) and the corresponding best f_0 hypothesis.

3.7 Evaluation of Implementations

Because there has been much f_0 estimation research lately, many researchers have designed and implemented their own f_0 estimators, and some have made these available to the wider research community. Using an off-the-shelf f_0 estimator is a good place to start because the algorithm is already implemented, and the researcher can begin immediately by analyzing results of the algorithm and designing add-on or sub-feature analysis components. One drawback is that the algorithm has been designed for a particular research problem and might not be appropriate for the problem at hand, although the algorithm could me modified to apply more closely if needed.

3.7.1 Common Problems with f_0 estimators

When a signal is pseudo-periodic with a low-power fundamental, it is possible to mistake an upper harmonic for the fundamental. Humans do this as well, and it is a result more of the signal itself than of the recognition algorithm. A period-k signal can become a period-$2k$

signal through a process called period doubling [65, 29]. At the transition point, it is unclear whether it is appropriate to count the period as k or $2k$. This transition point is unstable, so it is uncommon to hear signals of ambiguous pitch in nature. However, it does indicate that period doubling errors may be a difficult problem to overcome.

Subharmonic errors can lead to misleading results because they often occur within the context of a single pitch event, causing the f_0 estimation to jump back and forth between two (or more) subharmonics of the "true" fundamental frequency. The challenge then is to improve the f_0 estimation algorithm to deal with these problems. The YIN improvements attempt to rectify subharmonic errors, and have some success over less computationally complex algorithms.

3.7.2 Off-the-Shelf f_0 estimators

For this thesis, three off-the-shelf f_0 estimators are evaluated and compared. The first two f_0 estimators are part of a speech analysis software package called Colea, developed by Philip Loizou [42] for the MATLAB programming environment. This package contains tools for analyzing speech using f_0 estimation, formants, and spectral content. There are two f_0 estimators built into this package, one based on autocorrelation and one based on the cepstrum.

The third off-the-shelf f_0 estimator is the YIN algorithm described in Section 3.3.2.

3.7.3 Evaluation

The three f_0 estimators were tested on the speech/song corpus and the f_0 estimations were compared. Since the f_0 estimations were based on different frame rates, the first task was to match the f_0 estimations on a normalized time scale by interpolating between the frame measurements of the f_0 estimations to match the highest frame rate. Figure 3.6 shows an example of the three f_0 estimation techniques compared on a common scale. Figure 3.7 shows an example of a situation where the three f_0 estimation techniques did not agree.

These three f_0 estimators were compared using two criteria:

- Consistency between detectors
- Visual inspection of results

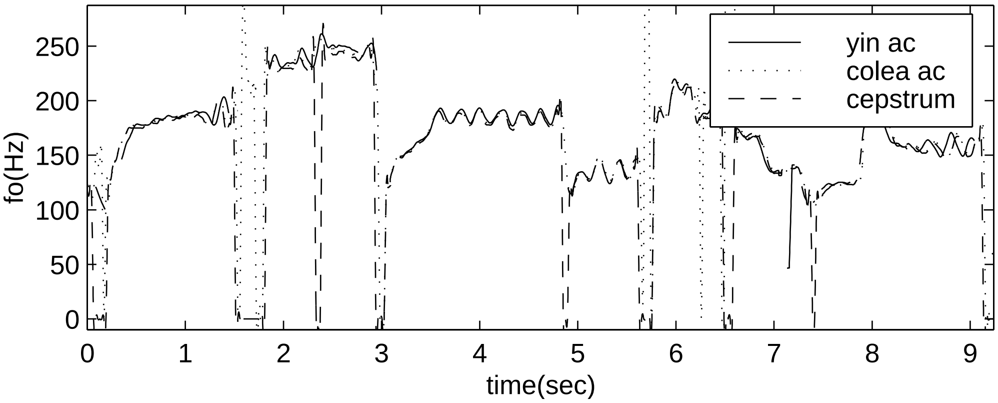


Figure 3.6: Comparison of three f_0 estimation methods - all methods near agreement (file b226).

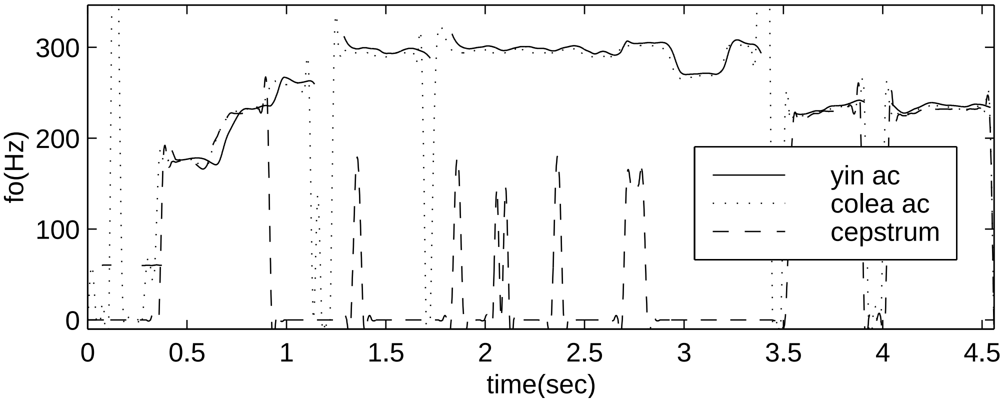


Figure 3.7: Comparison of three f_0 estimation methods showing differences among methods(file b212).

It should be noted that consistency between detectors, as an evaluation technique in isolation, is not particularly rigorous. It is not unreasonable to expect two detectors to agree on an erroneous f_0 estimation. This evaluation method becomes acceptable when combined with visual inspection. Files with one method in disagreement with the other two were inspected visually, and the f_0 estimations were compared to the perceived pitch track. In instances where two methods agreed, in the majority of cases, visual inspection showed that the agreeing methods were correct and the method not in agreement was in error.

A further comparison method could be to generate a manual (and presumably accurate) f_0 track for each file and compare these tracks to the results generated by each method. This evaluation technique was considered too labour-intensive for this work, and the results gained from the three presented criteria are sufficient for a comparison among the methods. If f_0 track accuracy were of paramount importance in comparing the methods (as in a transcription project), annotated corpora currently exist with Electroglottogram (EGG) f_0 track targets which could be used to evaluate the f_0 accuracy of the method. Synthetic signals have often been used to test f_0 detectors, although care must be taken to use synthetic signals that closely resemble the real-world signals that the estimator is likely to encounter.

Relative accuracy is a sufficient measure for this work because this provides an evaluation of the kinds of errors we are interested in, being subharmonic errors and existence errors. Subharmonic errors are described in Section 3.7.1. Existence errors are generated when a pitched frame in a sound is considered by the detector to not have a pitch, or when a f_0 hypothesis is presented for an un-pitched frame. The three criteria used here are sufficient for evaluations based on these measures.

The first criterion is measured by finding the difference between each pair of f_0 tracks. For each file, the difference between the three f_0 estimations are calculated according to Equation 3.6:

$$D = \frac{1}{N} \sum_{P_1^v, P_2^v} |P_1 - P_2|, \tag{3.6}$$

where N is the length of the f_0 track and P^v is a notation for the valid portions of a f_0 track P. The mean difference over the set of files is calculated, and the results for the entire corpus as well as the talking files only and the singing files only are presented in Table 3.1.

It can be seen from these results that the two f_0 estimation techniques based on autocorrelation had more similar results than the f_0 estimator based on cepstrum. This is

Table 3.1: Mean f_0 estimation difference between three f_0 estimators.

	YIN / Colea AC	YIN / Colea Cepstrum	Colea AC / Cepstrum
All files	13.33 Hz	19.68 Hz	41.22 Hz
Singing files	11.11 Hz	17.80 Hz	31.83 Hz
Talking files	14.00 Hz	20.27 Hz	43.75 Hz

perhaps to be expected, since the base algorithm is the same. This measure is enough to support the hypothesis that the Colea cepstrum f_0 estimator is not as accurate as the two autocorrelation f_0 detectors.

It is also important here to look at how the f_0 estimators compared in specific tasks. Since f_0 estimators are usually designed for a specific task, one would expect the f_0 estimator to perform better for that task (e.g. the estimation of the pitch of speech) than another task (e.g. the estimation of the pitch of song). Table 3.1 shows that the difference between the autocorrelation methods is lower for singing files than for talking files, but the difference between the cepstrum method and the two autocorrelation methods is higher for singing files than for talking files.

The second evaluative criterion is visual inspection of the three f_0 tracks. Files were selected with low and high relative error rates, and these were visually inspected for consistency errors. An example of a file with high difference is in Figure 3.7, where the cepstrum f_0 estimator failed to detect the f_0 of the signal through the time range of about 1 second to 3.5 seconds. The two autocorrelation methods agree well on this sample. It is important to notice the slight delay between the two autocorrelation f_0 tracks. This is because of the extra processing steps in the YIN detector, which seem to introduce a slight delay in the measured f_0 track. This can be corrected by re-aligning the f_0 track with the time scale of the original utterance, but this again is another computational step.

The visual inspection provides no rigorous results, although a count could be made of the files in which subharmonic and existence errors occurred, and the detector responsible for the error. The results that were provided by the visual inspection are that in most cases, differences between the f_0 tracks are due to errors in the cepstrum f_0 track, especially in singing utterances. Subharmonic errors show up between the autocorrelation f_0 estimators, and these errors are more or less equally distributed among the YIN, Colea and/or both.

It should be noted that the Colea f_0 estimators provided no measure of the confidence of the f_0 estimation. When the utterance is non-periodic, the f_0 estimation becomes erratic, jumping to zero or to a higher value out of range. These jumps, combined with the application of a power threshold, can be used to detect the presence or absence of f_0 which is an important feature in the speech/song comparison. The confidence metric of the YIN estimation means that this extra post-processing is not required.

Based on the results of this evaluation, it was decided to use YIN exclusively for the remainder of the f_0 measures in this thesis. YIN was more accurate and provided a confidence measure.

3.8 Discussion

f_0 estimation algorithms tend to be based on a number of fairly strict assumptions:

1. The input waveform consists of a single pitched signal, segmented into frames, and the waveform is homogeneous throughout the time frame being considered.
2. The input is limited to a specific audio domain, for which the algorithm is designed.
3. f_0 estimation is the same thing as pitch detection.

These assumptions are acceptable for initial development, and many successful algorithms have been developed using these assumptions. Indeed, without severely limiting the domain at the beginning of research, it would be impossible to achieve anything at all. Many researchers who accept that assumption 3 is theoretically incorrect continue to cite their work as pitch detectors rather than f_0 estimators. Given the slightly non-logarithmic transfer function from frequency to pitch, and also given some considerations about the base frequency used to create the music (e.g. $A_4 = 440$ Hz), a simple transformation can be developed to accurately map the frequency of a signal to its musical pitch.

Assumption 2 is another necessity for the introductory design of an algorithm. As the algorithms become more robust and more accurate, the domain for which the algorithm is useful will expand until assumption 2 can perhaps be relaxed. It is equally possible, however, that the nature of audio signals is such that certain algorithms are good for certain input and not others, and there is no “silver bullet” algorithm that will handle every periodic input without error. It is even conceivable that the human perceptual system uses more than one analysis method for deducing pitch from the vibrations of the eardrum.

This leaves assumption 1. f_0 estimation of *multiple* auditory streams is not difficult for the human auditory system, although it is difficult to concentrate on more than one stream at a time. Work on auditory stream separation is proceeding, but it would perhaps be more fruitful if the f_0 estimation community would work with the stream separation community, and vice versa. Clearly, each has much to learn from the other.

Chapter 4

Feature Discovery, Extraction and Evaluation

This chapter describes the features that are expected to perform well in distinguishing between speech and song, and why. Feature extractors are developed based on the opinions collected in Chapter 2. The results of the feature extractors are compared with the desired ratings for the sounds in the corpus, and the relative feature performance is evaluated.

4.1 Introduction

For the differences between speaking and singing to be measurable, features must be identified which produce different results when presented with different utterances. The discovery of these features is discussed in Section 4.1.1. Once the features are identified and extractors are developed, the results of these feature extractors must be evaluated. The techniques used for evaluation are presented in Section 4.1.2. The development of each individual feature extractor is discussed in Sections 4.2 to 4.4. Section 4.5 presents techniques for and results of the evaluation of the feature models.

4.1.1 Feature Sources

In determining relevant features to be studied for speech/song discrimination, there are a number of resources. First is the personal experience of the researcher, which as always is subjective and must be considered carefully and with a certain amount of suspicion.

Humans are the experts at discriminating between speech and song, and a human who has been thinking about this problem for a long time is likely to have some accurate insight into these differences.

However, since the experiences of a single person are by nature personal and subjective, a larger sample of people must be consulted for their experiences and opinions. The responses gathered in the corpus collection described in Chapter 2 provide suggestions for potential perceptual features. The complete list of responses is in Appendix C.

Another source of ideas for potential features is the research literature. People who are working on the differences between speech and song [83] as well as the differences between speech and music [63] have gathered potential features which may be useful. These features must be examined first to determine their suitability for speech/song discrimination, since they are often specific only to the task for which they were developed.

Once the set of features is chosen, and feature extractors are developed for each, the results of these feature extractors must be analyzed and evaluated. The following section presents methods used to analyze the feature results.

4.1.2 Analysis

There are three main motivations for analyzing the results of the feature extractors:

- Is the feature extractor accurately measuring the phenomenon of interest?
- Is the feature a useful measure of the differences between speech and song?
- Given a set of feature extractors, are they measuring different phenomena or merely different characteristics of one underlying phenomenon?

This third point is of particular interest because if several feature extractors are measuring the same phenomenon, only one of them is necessary in the complete system.

Feature extractor accuracy is impossible to judge without *a-priori* knowledge of the expected values of the phenomenon being measured. Generating these target values requires annotating the data by hand, a labour-intensive task which is beyond the scope of this thesis. The second two evaluations are sufficient to determine the suitability of the features being tested. Pragmatically, if the feature is shown to be a useful measure of the differences between speech and song, it is unimportant whether it is measuring the intended phenomenon or not. The next two sections describe the methods used to evaluate individual features and to compare feature results.

Probability Density Estimations

Once a feature algorithm has been developed (as presented in later sections of this chapter), a feature model is built which is used to classify new clips. The feature model is developed by collecting talking and singing feature values from the development corpus. To compare these two feature value sets, the system must determine which class (talking or singing) is more likely for each feature value. This is done by estimating the *probability density function* (PDF) for the talking files, P_t and the singing files, P_s. For a random variable, the probability density function is a measure of the likelihood that a measurement of that variable will fall within a specified range. Equation 4.1 shows the calculation of a probability from a PDF $f(x)$:

$$P(a < X < b) = \int_a^b f(x)dx. \tag{4.1}$$

The estimation of a PDF is the opposite problem: given a set of measurements $x_1, x_2, \ldots$ of a random variable X, estimate the probability density at every point in the range of possible values of X. The resulting function is called a *probability density estimation* (PDE) and is notated by "$\hat{}$". There are several methods to calculate the PDE, including the histogram (probably the most common) and the kernel method, which is used in this thesis [67]. Both methods are described here.

The *histogram* method divides the range of the variable into bins, and estimates the PDF by counting the number of measurements that fall within each bin. This method is often preferred because it is easy to implement and well understood, but it has two drawbacks. First, the resolution of the histogram PDE is limited to the bin width, and second, the resulting PDE is discrete, which is often not desirable for presentation or calculation. Specifically, if two PDEs are to be compared (as in a talk/sing feature model), the crossover point between them would occur somewhere between the limits of a bin, and the exact crossover point would be difficult to discern. As the bin width is reduced to improve the resolution, the number of measurements per bin is reduced and variance increases within each bin.

Several improvements to the histogram have been suggested including multi-resolution histograms, smoothed histograms and variable bin width, but these increase the computational complexity of the algorithm, decrease the simplicity (which was originally one of the attractive features of the histogram) and move away from theoretical accuracy.

A relatively simple alternative to the histogram is called the *kernel* method. The set of measurements (represented by a series of delta functions) is convoluted with a gaussian kernel, $w(x)$, of appropriate size, as in Equation 4.2. The effect is that a gaussian range of probability is added to the PDE for each measurement. These gaussians "pile up" where they are close together, indicating high probability density, and where the measurements are far apart, the gaussians are separate, indicating low probability. The gaussians can be approximated by any easily-generated bell-shaped curve, and for the talking and singing PDEs in this thesis, a hann window[1] [50] was used.

$$\hat{P_X} = \frac{(\delta \times \{x1, x2, \ldots\}) \otimes w[n]}{\sum_{-\infty}^{\infty}((\delta \times \{x1, x2, \ldots\}) \otimes w[n])} \tag{4.2}$$

The denominator term is added to satisfy the requirement that $\int(\hat{P}_X)dx = 1$. A number of improvements to the kernel method are presented in [67], but it was judged that the improvements in theoretical accuracy are small compared to the required increase in computational and intellectual complexity.

Evaluation: Kolmogorov-Smirnov Testing and Feature Correctness

Once the PDEs have been calculated, the next step is to compare them. First, Kolmogorov-Smirnov statistics are calculated to determine if there is a significant difference between the probability distributions. If there is no significant difference, the feature is considered not useful for a classification scheme. For those features with a significant K-S distance, the PDEs are compared by taking the log difference between them, as shown in Equation 4.3:

$$\hat{P}_{s-t} = \log(\hat{P}_s) - \log(\hat{P}_t) = \log\frac{\hat{P}_s}{\hat{P}_t}. \tag{4.3}$$

In some situations, one or the other of the PDEs may be equal to zero, in which case $\hat{P}_{s-t}$ would approach $\pm\infty$. To avoid this, the comparative PDE is hard-limited to ± 1, with the following justification: $\hat{P}_X \leq 1$ since $\int(\hat{P}_X)dx = 1$. $\log(a \leq 1) \leq 0$. In comparing the

[1] The hann window is similar to the hamming window described in Section 1.3.1. These windows are calculated as $w[n] = a - b\cos(2\pi n/M),\ 0 \leq n \leq M$ with $a = 0.5, b = 0.5$ for the hann window and $a = 0.54, b = 0.46$ for the hamming window.

two PDEs we are interested in their relative values only where one PDE does not clearly dominate over the other. If $\hat{P}_{s-t} > \pm 1$, one PDE clearly dominates and the limit of ±1 is justified. The improved formula $\hat{P}'_{s-t}$ is presented in Equation 4.4:

$$\hat{P}'_{s-t} = \begin{cases} 1, & \log(\hat{P}_s) - \log(\hat{P}_t) \geq 1, \\ -1, & \log(\hat{P}_s) - \log(\hat{P}_t) \leq -1, \\ \log(\hat{P}_s) - \log(\hat{P}_t), & \text{otherwise.} \end{cases} \tag{4.4}$$

This comparative PDE ($\hat{P}'_{s-t}$) is used to evaluate each feature model against individual sound files with *a-priori* known ratings. For each evaluation file being tested, each feature extractor is applied to obtain a set of feature values. Each feature x generates a speech/song rating $\hat{P}'_{s-t}(x)$. This rating is evaluated in two ways: absolute and relative correctness.

The *absolute correctness* of each speech/song rating is simply a measure of whether or not the individual feature extractor, in isolation, produced a rating identical to the target rating. The absolute correctness of $\hat{P}'_{s-t}(x)$ is calculated thus: If the *a-priori* class of the file is talking, and $\hat{P}'_{s-t}(x) < 0$, the feature model is considered to have behaved correctly for that file, and is given a value of 1. If not, the feature model is given a value of 0. The mean absolute correctness rating over all files is calculated for each feature, and these results are presented as each feature is described in detail.

The *relative correctness* of $\hat{P}'_{s-t}(x)$ is calculated by comparing the value of $\hat{P}'_{s-t}(x)$ to the *a-priori* rating for the file. If the *a-priori* rating of the file is 5 (pure singing) and $\hat{P}'_{s-t}(x) = 4.5$ for a given feature, that feature is given a correctness of 0.9, because the difference between the target rating and the PDE value is 10%.

4.2 Vibrato

Sung utterances often have an associated f_0 track which oscillates in a characteristic way. Sometimes when people sing they add this oscillation to the pitch of the note they are singing, for stylistic or other reasons. This phenomenon, known as *vibrato* is characterized by a stationary pitch modulated by a 4–8 Hz pseudo-sinusoidal waveform.

4.2.1 Perceptual Motivation

Figure 4.1 shows an example f_0 track of a sung utterance with vibrato and discrete f_0 levels, and Figure 4.2 shows an example f_0 track of a spoken utterance. These f_0 tracks are of the same individual speaking and singing the phrase "Row, row, row your boat, Gently down the stream."

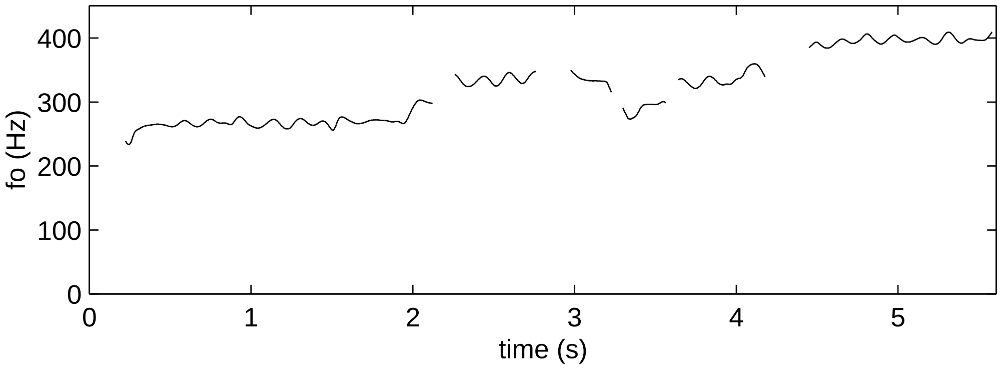


Figure 4.1: Example of a sung utterance f_0 track (file e255).

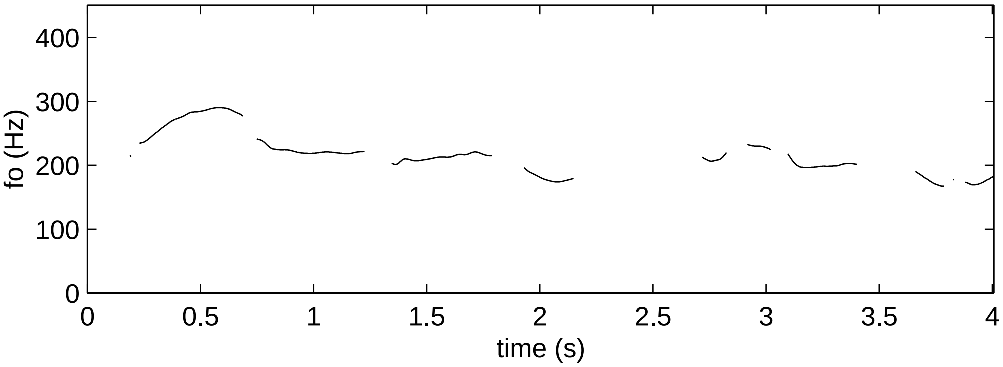


Figure 4.2: Example of a spoken utterance f_0 track (file f255).

Many sung utterances in the corpus displayed the presence of vibrato, and vibrato was cited as a decision factor for many subjects. Following are examples of subject responses indicating vibrato.

u174:314 (5) "He held on to certain syllables that were (I think) at the ends of words. He also seemed to be using a little bit of vibrato."

u175:314 (5) "The pitch is high and she uses vibrato"

u175:325 (5) "the wobbly voice"

u175:352 (5) “melody and vibrato”

The third response in this list shows that people can perceive vibrato and associate it with singing without being able to name it.

4.2.2 Physical Realization

As can be seen in Figure 4.1, the perceptual feature of vibrato is physically manifest as a frequency modulation of the vocal signal. Several researchers in various fields have studied this phenomenon [59], and from their work, it is determined that on average, the modulating frequency varies between 4 and 8 Hz. The shape of the vocal tract does not change during vibrato pitch oscillation, so the formant locations stay constant, as shown in Figure 4.3, using the clip s101, which contains the lyric “O Freunde” sung in opera style.

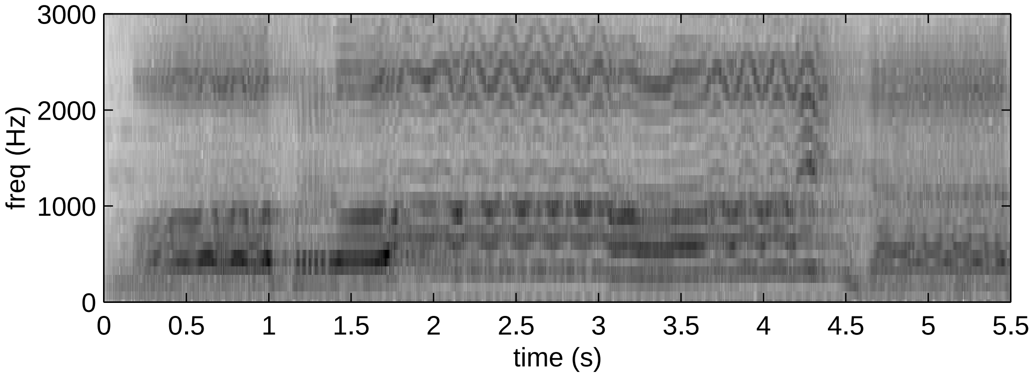


Figure 4.3: Spectrogram of a vocal clip with vibrato showing partials moving in and out of formats.

Vibrato blurs the pitch realization, making it more difficult to determine the intended pitch target. This means that traditional methods of music detection which make use of pitch constancy fail when presented with a pitch track sung with vibrato. Music detection methods, and specifically traditional methods of automatic music transcription, rely on a relatively consistent f_0 across the note being transcribed. If the f_0 of the note changes by a small amount, the transcription engine is able to make assumptions about the note and assign a discrete value to the pitch, but if the f_0 changes by more than a quarter tone across the note, or if it does not adhere to the expected scale defined in the transcription engine, the pitch will not be correctly recognized and the transcription will be incorrect.

4.2.3 Feature Extraction

Once the frequency of the vocal signal has been extracted, detecting a frequency modulation is equivalent to detecting local periodicity in the f_0 track. As in other domains where the task is to detect periodicity, there are essentially two approaches: time-domain (temporal) methods and frequency domain (spectral) methods. A common temporal method of periodicity detection is based on autocorrelation, and a common spectral method is based on the Fourier transform. The following section describes implementations for both of these methods, and following that is a comparison of the results of these methods.

The goal of a vibrato detector is to produce a rating based on both the strength and frequency location of any frequency modulation. Frequency modulation lower than 1–3 Hz is likely to be due to f_0 changes between phonemes rather than vibrato within an individual phoneme. Frequency modulation higher than 9–15 Hz borders on the audible range, as well as being very difficult or impossible for the human vocal track to produce. Modulation frequencies within this range (4–8 Hz) should therefore have higher scores than modulation frequencies outside of this range. The second characteristic is the magnitude of the modulation, a stronger modulation resulting in a higher vibrato score.

4.2.4 Common Pre-Processing

Both algorithms begin in the same way, by accepting a f_0 track as input. The first derivative of the f_0 track is calculated, so that the vibrato detection operates on the slope of the track instead of the track itself. Since vibrato-like modulations tend to be sinusoidal, the first derivative does not alter the vibrato information. The first derivative is used to identify the areas of high f_0' which indicate utterance segment breaks. Figure 4.4 shows the first derivative of an example frequency track. Compare with Figure 4.1, which shows the original f_0 track from the same file.

The next pre-processing step is to separate the utterance into segments. These are identified by periods of unvoiced utterance and changes in f_0 and power. This is done to isolate a series of vibrato measures in an utterance, so that a single segment with vibrato in an otherwise flat utterance will give a significant vibrato measure in the final evaluation. Figure 4.5 shows a single speech segment of the utterance used in Figures 4.1 and 4.4. Each speech segment is then normalized to control for differences in the range of the frequency modulation at different base frequencies.

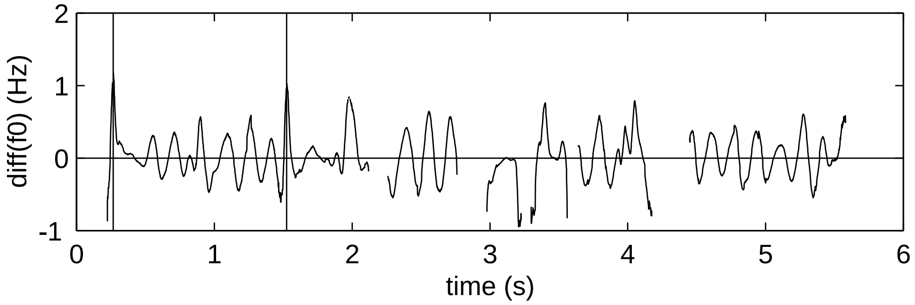


Figure 4.4: Vibrato pre-processing: $f(0)'$ for the sound in Figure 4.1.

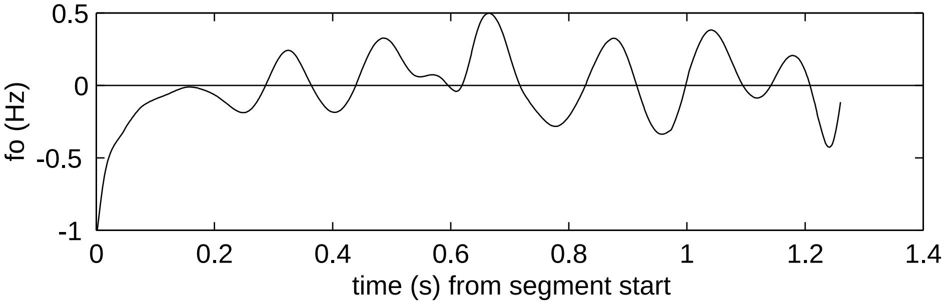


Figure 4.5: Vibrato pre-processing: utterance segmentation.

In the first vibrato detection method, the fast Fourier transform (FFT) is used to detect the presence of low-frequency modulation in the f_0 track. The FFT is applied to the f_0 track, and the largest peak in the FFT is taken to represent the location of the modulating frequency, if any. The FFT vibrato feature value is equal to the frequency band in which the maximal spectral peak is found. Figure 4.6 shows the FFT vibrato measure of the utterance segment presented in Figure 4.5.

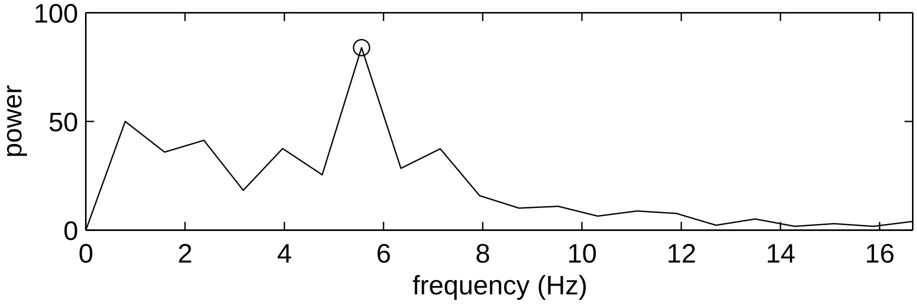


Figure 4.6: Vibrato measure using FFT of the f_0 track.

The second method for detecting the presence of vibrato uses the autocorrelation of the f_0 track to identify the presence and strength of a vibrato-like frequency modulation. The vibrato measure is found by taking the time location of the first non-zero peak and multiplying it by the amplitude of the peak. In this way, strong peaks at shorter lags are rewarded, and weak peaks at longer lags are suppressed. Figure 4.7 shows the autocorrelation vibrato measure of the utterance segment presented in Figure 4.5.

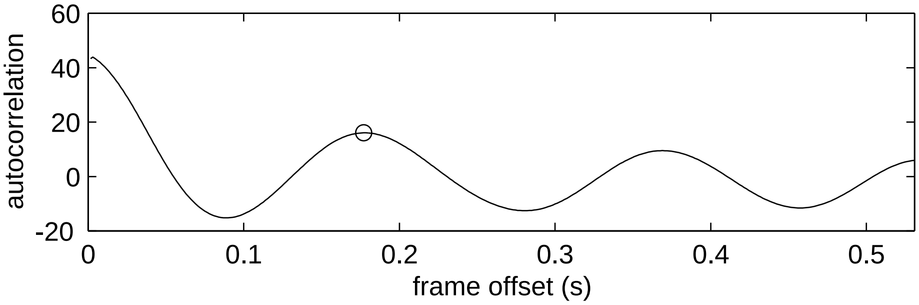


Figure 4.7: Vibrato measure using autocorrelation of the f_0 track.

A more accurate vibrato model could be implemented using emphasis for peaks in the recognized vibrato frequency range, but the methods already described were found to provide adequate results at this point. Other vibrato detection methods are presented in [59].

4.2.5 Comparing Spectral and Temporal Vibrato Detection Algorithms

In the temporal algorithm, the vibrato measure is calculated by finding the location of the first peak in the autocorrelation corresponding to a periodicity frequency above 3 Hz. The amplitude of this first peak is then multiplied by the location of the autocorrelation peak. This scale factor tends to favor lower-frequency vibrato.

In the spectral algorithm, the FFT is calculated and the phase information is discarded by taking the absolute value. The vibrato measure is then calculated as the amplitude of the first high-magnitude peak in the FFT above the threshold of 3 Hz. If there is no vibrato present in the signal, the power in the FFT will be concentrated in the lower frequency bands and the amplitude of the first peak above 3 Hz will be smaller than that of a signal with vibrato. For both algorithms, the vibrato measure for a complete clip is taken as the maximum vibrato measure across all of the segments in the clip.

The feature models for the two vibrato features are presented in Figures 4.8 and 4.9. The correctness for these feature models are presented in Table 4.1. Throughout this chapter, shorthand notational labels will be used to present tabular information about the features. These labels will be defined as they are used, and the full set of labels is presented in Table 4.7 near the end of this chapter. In this case, V_{AC} is the feature label for vibrato measured using the autocorrelation method, and V_{FT} corresponds to vibrato measured using the fast Fourier transform.

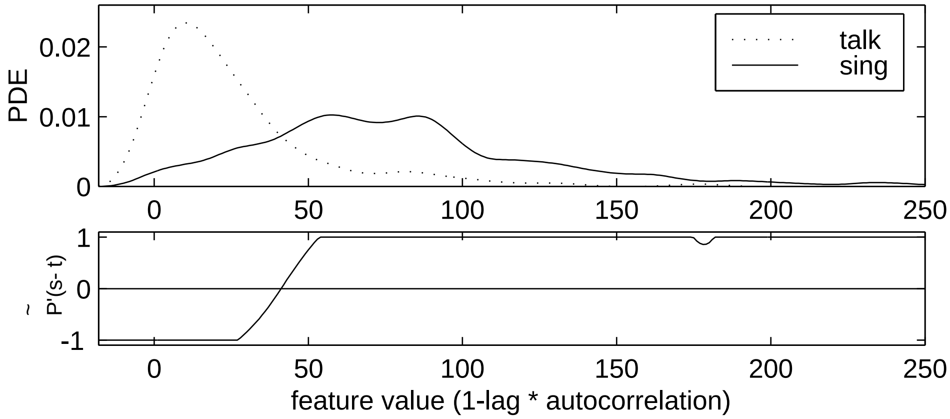


Figure 4.8: Feature model of autocorrelation-based vibrato measure, V_{AC}.

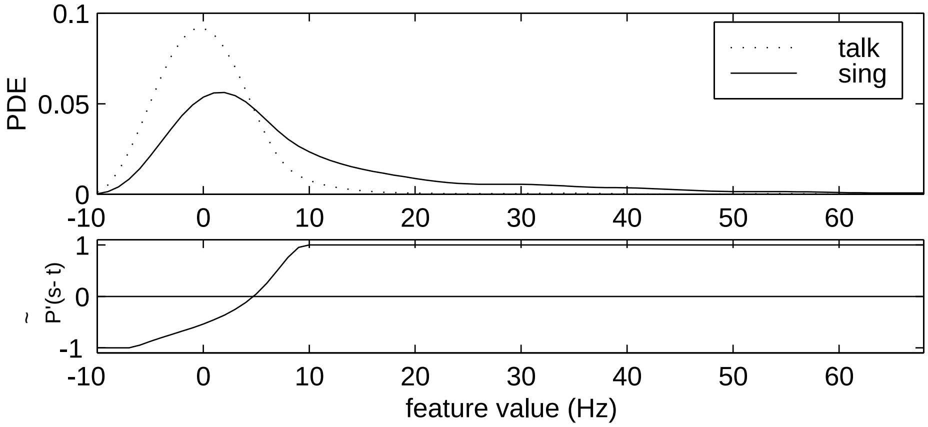


Figure 4.9: Feature model of FFT-based vibrato measure, V_{FT}.

Table 4.1: Correctness of vibrato features.

Feature	V_{AC}	V_{FT}
Relative	0.8021	0.6487
Absolute	0.8147	0.7262

4.3 Simple f_0 Statistics

As discussed in earlier sections, the f_0 track of an utterance is a key starting point for a speech/song discrimination. Depending on the analysis used, f_0 track based features can be quite computationally intensive. Statistics of f_0 are an exception to this. f_0 statistics features are extracted by taking a suitable window length and calculating the mean, the standard deviation, maximum and minimum of the f_0. A suitable window length for f_0 statistics is between 2 and 5 seconds. A shorter segment would not contain sufficient data for a reasonable statistical judgement, and a segment longer than 5 seconds could contain utterances of more than one class, leading to blurred statistics. For this work the complete utterance file was used for each clip.

4.3.1 Perceptual Motivation

Many subjects observed statistical f_0 features when quantifying the speech/song difference. An abnormally high pitch can indicate song, as can a pitch that varies greatly. Following are examples of subject responses indicating statistical f_0 features which are evidence for song.

n132:212 (5) “Great variance in pitch. . . ”

n134:236 (5) “high pitch and expressive talking”

Similarly, certain statistics can be evidence for speech, for example when there is “not enough” pitch variation in the utterance. Following are examples of subject responses indicating pitch statistical features which are evidence for speech.

n133:212 (3) “There’s rhythm but not enough difference in pitch”

u161:310 (2) “very little change in tone”

The difficulty with f_0 statistics as evidence for speech is that there are cases where a value that might be considered song-range is in fact from a speech utterance. For example, highly prosodic speech has high f_0 variance but most subjects rated such utterances more toward speech than song:

n134:251 (2) “Talking, but with a lot of change in pitch”

4.3.2 Physical Realization and Extraction

The usefulness of f_0 statistics is at first not immediately obvious, because the human trial annotation revealed some disagreement, but the fact that people mention features like pitch range and pitch variance is sufficient to investigate various simplistic measures of the statistical distribution of f_0 across a clip.

Once the f_0 has been extracted, the calculation of the statistics is straightforward, using standard algorithms. The maximum, minimum, mean and standard deviation were calculated for the f_0 track of each clip in the development corpus, and the results were compiled into feature models for these four statistics. The feature models are presented in Figures 4.10 through 4.13.

4.3.3 Discussion

Statistical features of f_0 are features which are inspired by human observation but whose final usefulness will be determined by how well the results of each feature fit with the desired classification scheme. As such, each feature must be investigated independently of the initial motivation, and if the feature is found not to separate well, it should not be used.

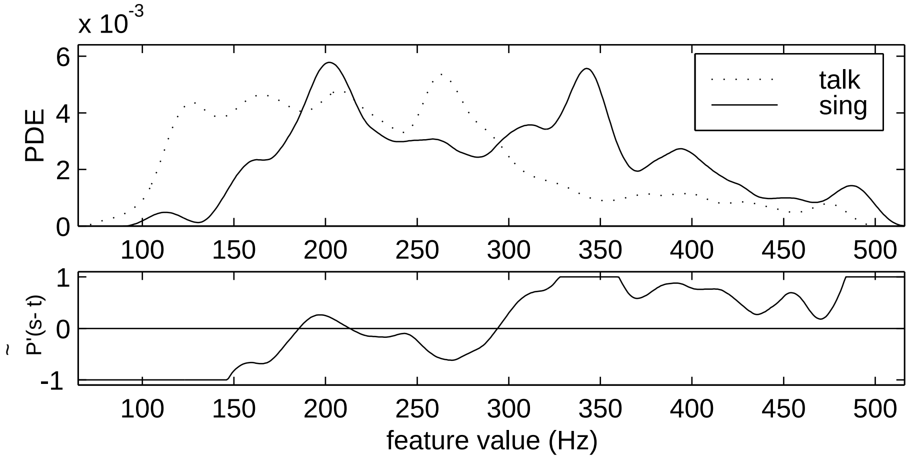


Figure 4.10: Feature model of maximum f_0, $M(f_0)$.

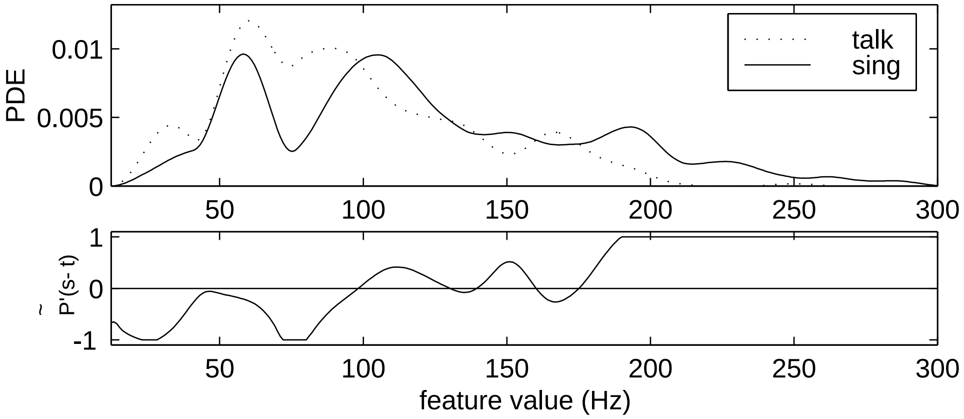


Figure 4.11: Feature model of minimum f_0, $m(f_0)$.

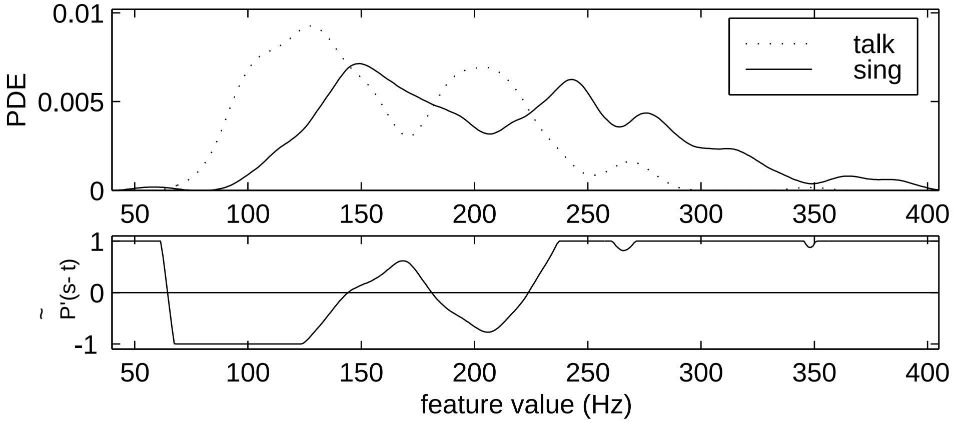


Figure 4.12: Feature model of mean f_0, $\mu(f_0)$.

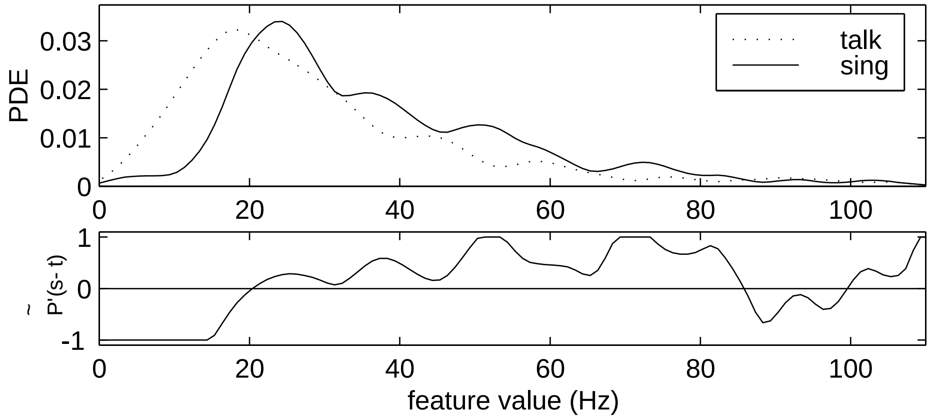


Figure 4.13: Feature model of f_0 standard deviation, $\sigma(f_0)$.

The four statistical features described above are maximum, minimum, mean and standard deviation. The maximum f_0 feature separates well at the tails, with singing having an approximately bimodal distribution, perhaps attributable to the differences in speaking and singing frequency range between males and females, although this was not directly investigated. However, between feature values of 175 and 300, the feature does not separate the classes reliably, since both classes have appreciable probabilities in this range. The minimum f_0 feature seems to follow a somewhat common contour for both features, with the PDE of the singing class shifted up the feature value scale. The feature separates well at the 75 Hz dip in the singing PDE and above 175 Hz. The mean f_0 has an approximately bimodal distribution for talking and singing, with the valley in the talking PDE, again near 175 Hz, lining up with a peak in the singing PDE, and a peak in the singing PDE near 200 Hz lining up with a valley in the talking PDE. This feature is expected to separate well for most feature values. The standard deviation of f_0 shows probabilities which are similar for most feature values. This feature is not expected to separate well. The measured correctness of these features is presented in Table 4.2. The feature labels used here are $M(f_0)$ for maximum f_0, $m(f_0)$ for minimum f_0, $\mu(f_0)$ for mean f_0 and $\sigma(f_0)$ for the standard deviation of f_0.

Table 4.2: Feature correctness for f_0 statistics

Feature	$M(f_0)$	$m(f_0)$	$\mu(f_0)$	$\sigma(f_0)$
Relative	0.6216	0.5945	0.6779	0.5743
Absolute	0.6561	0.6277	0.7078	0.5977

4.3.4 Statistics based on f_0 of speech segments

An expected improvement to the f_0 statistics is to investigate only on speech segments, where segment boundaries are determined by beginning and ending of valid f_0 areas, as well as areas of high f_0 slope—song clips often have high slope pitch transitions between notes. Figures 4.14 and 4.15 show the mean and standard deviation feature models based on speech segments.

The relative and absolute correctness for these feature models are presented in Table 4.3. The labels used here are $\mu_s(f_0)$ for the segment-based mean f_0 and $\sigma_s(f_0)$ for the segment-based standard deviation of f_0.

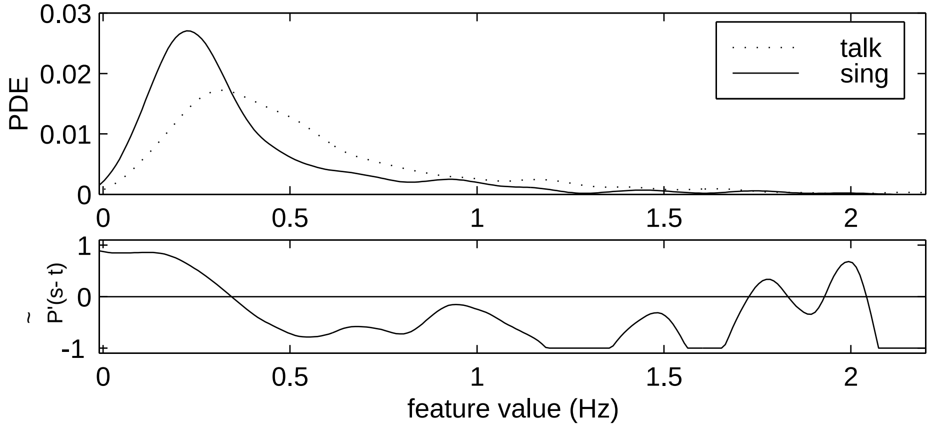


Figure 4.14: Feature model of segment-based mean f_0, $\mu_s(f_0)$.

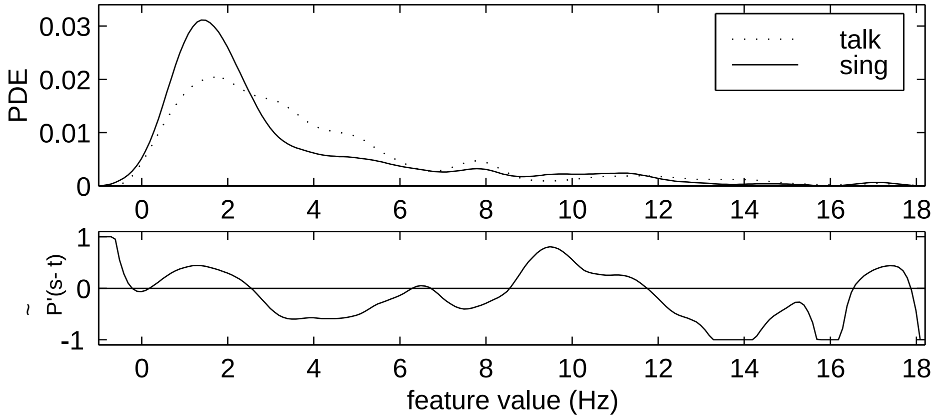


Figure 4.15: Feature model of segment-based f_0 standard deviation, $\sigma_s(f_0)$.

Table 4.3: Feature correctness for speech segment f_0 statistics.

Feature	$\mu_s(f_0)$	$\sigma_s(f_0)$
Relative	0.5958	0.5457
Absolute	0.6511	0.5910

4.3.5 f_0' Statistics

As with f_0 statistics, the slope of the f_0 track can be instructive in the decision between talking and singing. f_0' statistics are calculated in the same way as those for f_0. The slope of the f_0 is calculated by taking the first difference of the f_0, which is the discrete version of the first derivative:

$$f_0'[n] = f_0[n] - f_0[n-1]. \tag{4.5}$$

The following statistics are used for analyzing the slope of the maximum, μ and σ. Minimum f_0' is not used because in many cases, the YIN f_0 extractor produced consecutive frames with identical estimations, resulting in a minimum f_0' of 0 Hz for many clips. As with f_0, higher order statistics and different averages of f_0' could also be used. Figures 4.16 to 4.18 show the feature models for the f_0' statistics.

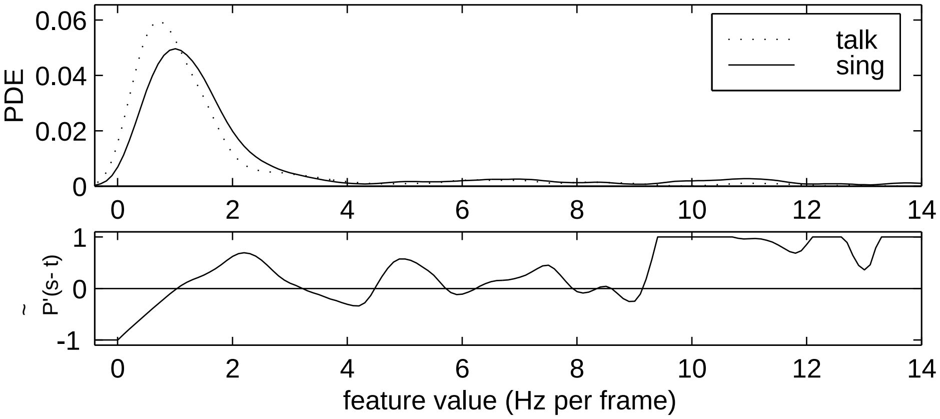


Figure 4.16: Feature model of maximum f_0', $M(f_0')$.

The correctness of these features is presented in Table 4.4. The feature labels are $M(f_0')$ for maximum f_0', $\mu(f_0')$ for mean f_0', and $\sigma(f_0')$ for the standard deviation of f_0'.

Table 4.4: Feature correctness for f_0' statistics.

Feature	$M(f_0')$	$\mu(f_0')$	$\sigma(f_0')$
Relative	0.5418	0.5576	0.5185
Absolute	0.5910	0.5860	0.5493

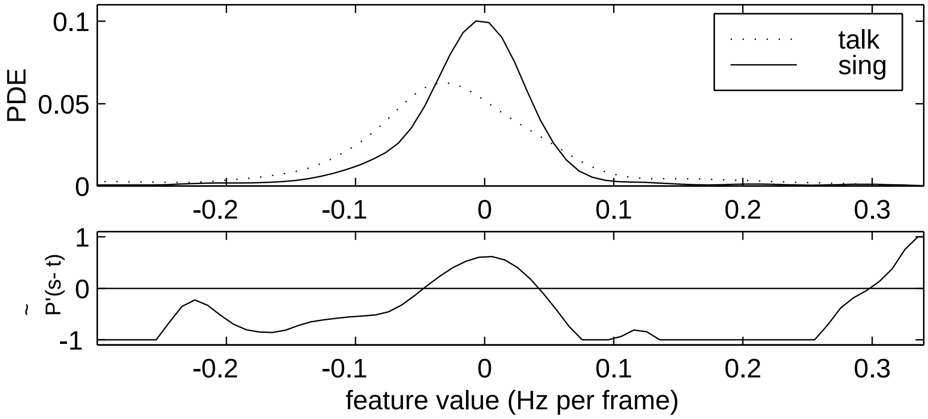


Figure 4.17: Feature model of mean f_0', $\mu(f_0')$.

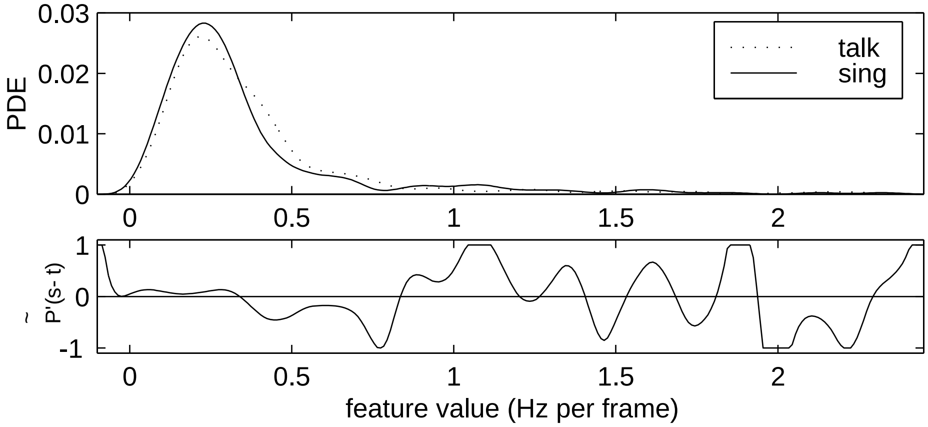


Figure 4.18: Feature model of standard deviation of f_0', $\sigma(f_0')$.

4.4 Rhythm

When asked what makes an utterance speech-like or song-like, many listeners identified rhythm as an important feature, although their descriptions of the specifics of rhythm in the characterization often degrade to circular reasoning: speech is characterized by speech-like rhythm and song is characterized by song-like rhythm. To discover what it means for a rhythm to be song-like or speech-like, we must first do a data reduction to extract relevant features for a rhythm measure, and then study the results of this measure on speech clips and song clips.

Preliminary research shows that rhythm in song involves phrase repetition and repetition of power patterns, as well as phoneme repetition. Feature extractors designed to extract repetition, such as autocorrelation and other f_0 extraction techniques, would be useful for this task.

The difficulty with doing work on rhythm is that perceptually, it seems as though rhythm is a single concept similar to pitch. Many subjects would simply cite "rhythm" or "lack of rhythm" as a reason for rating a file as speech or song. Understanding the physical characteristics of rhythm is beyond the scope of this thesis, but two features were investigated that relate to rhythm. These features were identified by the observation of the author and are not based on the perceptual annotation from the corpus.

4.4.1 Utterance Segment f_0 Track Match

The first feature examined in trying to model rhythm is the similarity between the f_0 track of individual speech or song utterance segments. If the f_0 is similar in two different segments, this is possibly an indication of word or phrase repetition, which could indicate rhythm. Therefore, it is suspected that f_0 track matching can be used as a discriminatory feature between speech and song. The observation is that in speech, f_0 tracks are not correlated from segment to segment, whereas in song they often are correlated.

This feature was extracted by first segmenting the sound file into individual sound segments as described in Section 4.2.4. Short segments are discarded, and the remaining segments are correlated for all segment pairs, with the largest correlation taken to be the feature value for that file. Figure 4.19 shows the feature model for the segment f_0 track match. The relative accuracy for this feature is 0.7908, and the absolute accuracy is 0.8063. The label for this feature is R_s.

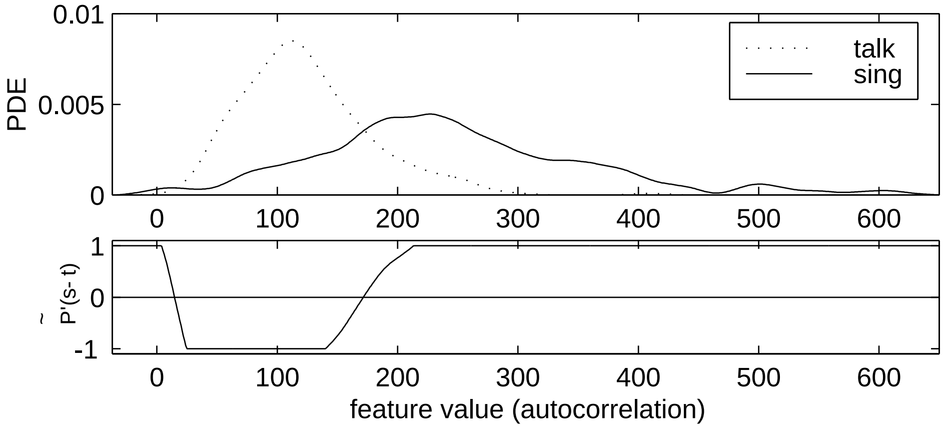


Figure 4.19: Feature model of speech segment repetition, R_s.

This feature separates well at all feature values. Spoken utterances have segment f_0 correlations concentrated around 100, while sung utterances tend to have higher segment f_0 correlations, centered around 210 with a wider deviational spread. It should be noted that there are other possible ways to measure the correlation between segments. The spectrum, power and ZCR are also expected to be correlated in a pair of repeated segments.

4.4.2 Voiced and Unvoiced Parts of Speech

Human speech can be divided into three categories relating to the driving function of the vocal tract. If the vocal chords are held taut, the air forced through them by the lungs sets up a periodic vibration which results in a pitched sound. This happens with all vowels, liquid consonants like “r” in “road” and “l” “lunch”, and nasals like “m” in “match” and “n” in “north”. These sounds together are called “voiced” and are distinguished from “unvoiced” by the fact that they all have a pitch. Unvoiced sounds include fricatives like “sh” in “shower” and “th” in “thesis”, stops like “p” in “pound” and “d” in “duck”. These sounds can add to the perception of rhythm in an utterance, but not to the perception of melody, because they have no pitch.

Perceptual Motivation

When a person sings, a common phenomenon is *voiced lengthening*, where the voiced segments of the utterance are longer than they would be in normal speech, and the unvoiced segments are shorter. The voiced segments produce the notes, and the unvoiced segments are used between the notes to define the semantics of the utterance. Human perception of the meaning behind an utterance is concentrated at the “edges” of the utterance, where the phonemes change.

Another phenomenon in sung utterances is *silence filling*. In normal speech, there are segments of silence before or after stops like “t” in “stream” and “feet”. Silence is also used for punctuation, where a comma or a period might be used in written text. These silences are compressed or removed completely when singing, because of voiced lengthening, and because punctuation in song is often achieved by the music. It is important to note, however, that singing can often include large segments of silence between verses or phrases.

Physical Realization

There are several ways to distinguish between voiced, unvoiced and silence frames in an utterance, and these rely on the physical manifestation of voiced and unvoiced utterances. Unvoiced utterances are very often fricative, which means that the driving function is chaotic. Instead of a steady stream of air through tight vocal chords producing a pitched sound, the vocal chords are held loose and air is allowed to flow freely through them, creating a chaotic broadband driving function.

A second physical characteristic of unvoiced speech is that it is spectrally broadband, and the power is often concentrated in higher frequency bands, as compared to voiced speech which has power concentrated in a lower-frequency spectral region. Also, since voiced speech is pitched, the power spectrum of voiced speech shows peaks at harmonically related frequencies.

A third method is to consider the result of a f_0 extractor. If a f_0 extractor is sufficiently good at detecting the *presence* of pitch as well as the value of f_0, this information can be used to determine the distribution of voiced frames by considering all frames with a valid f_0 measure to be voiced, and all remaining frames to be unvoiced or silent depending on a supplementary measure such as power. These physical differences between pitched and non-pitched speech can be used to develop feature extractors for the speech/song discrimination task.

It is interesting to note that in a whispered utterance, all phonemes are unvoiced, and hence aperiodic, and this is why it is impossible to add a melody to a whisper[2]. Humans can perceive pitch in an aperiodic waveform if it contains spectral peaks and/or troughs. If there is more power concentrated around a particular frequency, a vague notion of pitch will be perceived which will increase as the bandwidth around that frequency decreases. In human whispering, the broadband aperiodic source signal is filtered into formants by the vocal tract, and outside of the context of speech, these formants can lead to the perception of pitch. If one starts to whisper and then makes random motions with one's mouth, one will discover that the more open the mouth and the oral cavity, the higher the perceived pitch. In the context of speech, however, this information is used to decipher the phonemes being generated. The characterization of these whispered utterances is beyond the scope of this thesis, so for the purposes of this work, the periodicity of an utterance frame will be used as a measure of whether the frame is voiced or unvoiced.

Feature Extraction

The feature extraction techniques for the distribution of voiced frames depend on the physical phenomenon being extracted. The first technique relies on detecting the chaotic nature of the driving function. This is done with the zero-crossing rate (ZCR). The ZCR has been discussed in Chapter 3 as a method to detect the frequency of a waveform, and reasons were given there as to why ZCR by itself is not a successful f_0 extractor. The task here is to detect the *existence* of a pitch instead of detecting the value. If there is no pitch, the ZCR will indicate this with higher values and more erratic distributions. A periodic waveform will have a comparatively low ZCR and a more even distribution.

The ZCR voiced frame distribution detector looks at the waveform frame by frame and counts the number of times the waveform crosses zero in that frame, each frame being 15 ms long. The mean ZCR for the entire file is then calculated, and frames with ZCR above the mean are taken to be unvoiced, with frames below the mean being voiced.

This initial algorithm was then extended to include silence detection, where the power of the waveform dropped below a threshold, and a measure of the ZCR statistical distribution was added to capture instances where the ZCR value alone was not enough to determine the difference between voiced and unvoiced utterances.

[2] Recall Figure 1.3 for a comparison of normal and whispered speech.

Once the file was segmented into voiced, unvoiced, and silent frames, totals were calculated for each and translated into a proportion for the file. Three features were extracted from this: proportion of voiced frames (PV_Z), proportion of unvoiced frames (PU_Z), and proportion of silence frames (PS_Z), the Z in the feature labels indicating the use of ZCR in the feature value calculations. These three feature models are presented in Figures 4.20 to 4.22, and the corresponding correctness results are presented in Table 4.5.

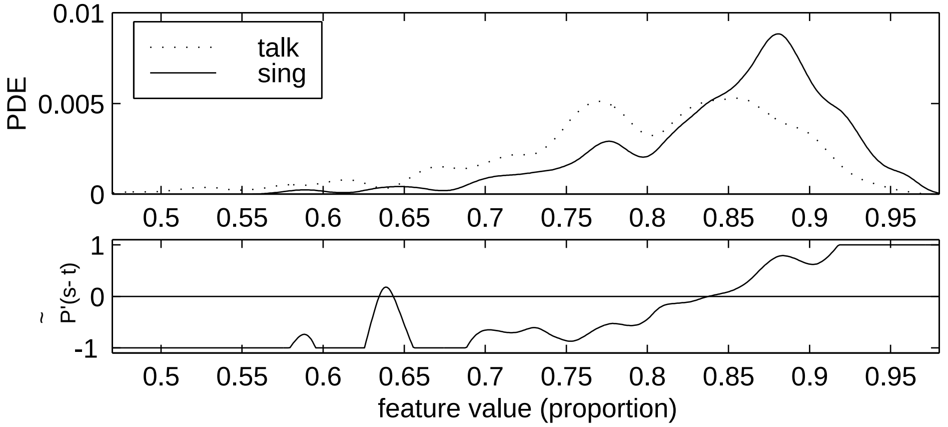


Figure 4.20: Feature model of ZCR-based proportion of voiced frames, PV_Z.

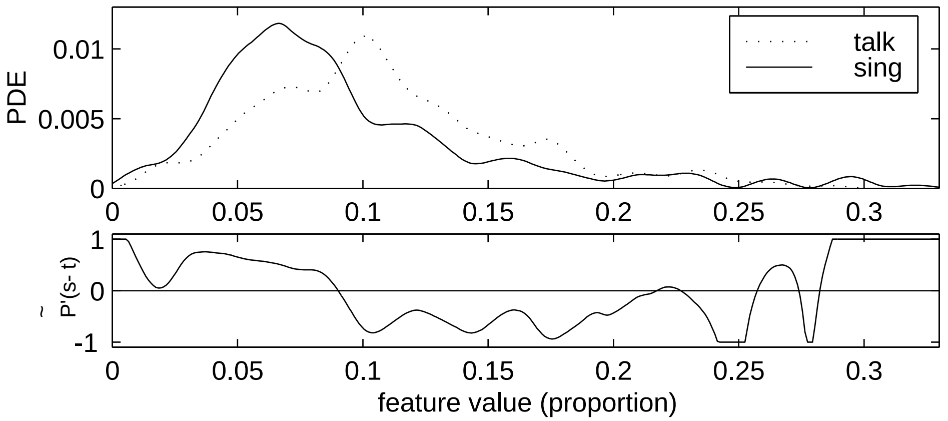


Figure 4.21: Feature model of ZCR-based proportion of unvoiced frames, PU_Z.

The second method of voicedness detection involves investigation of the power distribution of each frame. If the power in a frame is concentrated in the higher frequency bands, the frame is considered to contain an unvoiced utterance, and if the power is concentrated in the

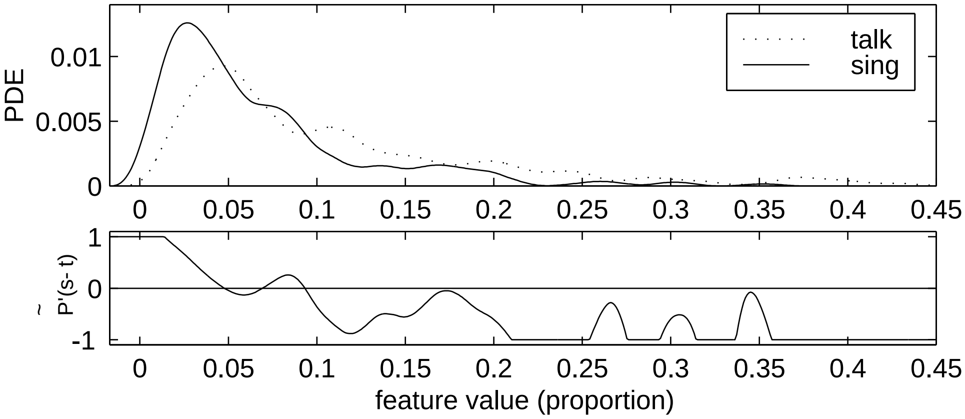


Figure 4.22: Feature model of ZCR-based proportion of silence frames, PS_Z.

Table 4.5: Feature correctness for ZCR voiced frame ratio features.

Feature	PV_Z	PU_Z	PS_Z
Relative	0.6003	0.5531	0.5647
Absolute	0.6361	0.6394	0. 5877

lower frequency bands, the frame is considered to contain a voiced utterance. Other power characteristics are relevant as well, such as the power spectral stratification that comes with periodic waveforms, but the high-low differentiation was found to be sufficient in preliminary testing. This method wasn't implemented because it was considered too computationally intensive.

The third method assumes that the f_0 extractor being used is capable of not only detecting the value of the f_0, but also the presence or absence of pitch. The f_0 extractor (YIN) being used in the current set of feature extractors gives a confidence measure along with the f_0 value, and this confidence measure indicates whether or not YIN "believes" that the waveform is periodic. With some additional computation, this confidence metric can be used as a voiced/unvoiced measure.

The confidence measure is augmented with a heuristic f_0 track analysis algorithm that detects anomalies in the f_0 track. This anomaly detector is sensitive to values of f_0, f_0' and higher order derivatives of f_0. When any of these measures pass beyond perceptually defined thresholds, the frame is considered to be non-pitched and hence unvoiced or silent.

Once again, the voicedness detector is combined with a power measure to determine if the non-pitched segments are unvoiced or silent. The feature models of the f_0 track voicedness detector are presented in Figures 4.23 and 4.24. The correctness results for these features are presented in Table 4.6. The labels for these features are PU_{f_0} for the proportion of unvoiced frames using the f_0 extractor, and $PV_{f_0}$for the proportion of voiced frames.

From the ZCR-based voiced frame distribution measure, three features were derived: proportion of voiced, unvoiced and silent frames. Since all frames can only be one of these three, It was decided that for the f_0 extractor based voiced frame distribution, it is sufficient to measure the two quantities: unvoiced and voiced frames, since the proportion of silent frames can be derived from the proportion of unvoiced and voiced frames.

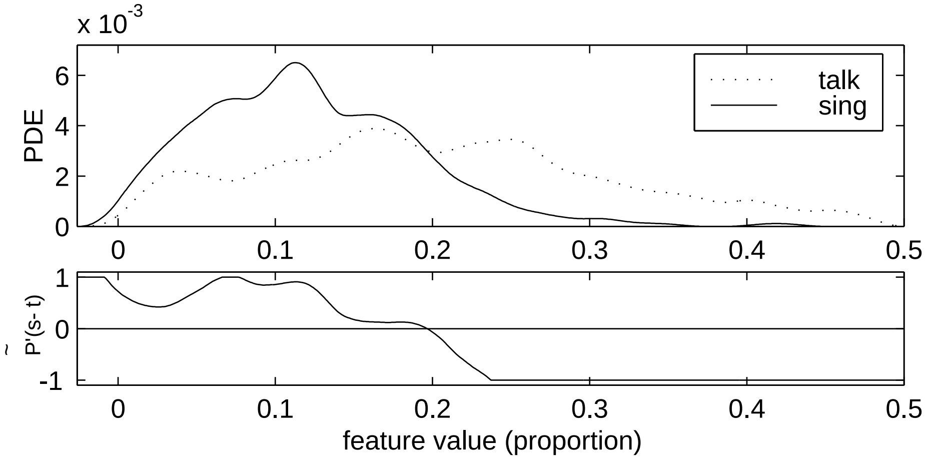


Figure 4.23: Feature model of f_0-based proportion of unvoiced frames, PU_{f_0}.

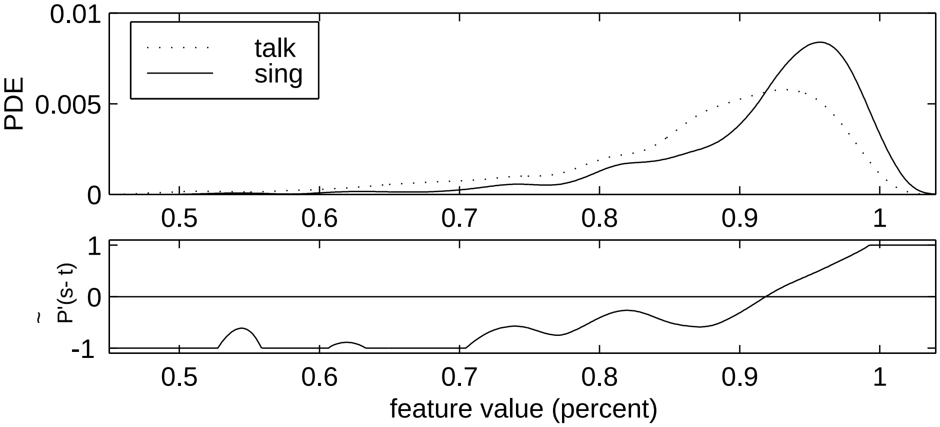


Figure 4.24: Feature model of f_0-based proportion of voiced frames, PV_{f_0}.

Table 4.6: Correctness for f_0-based voiced frame ratio features.

Feature	PU_{f_0}	PV_{f_0}
Relative	0.6472	0.5656
Absolute	0.6678	0.6260

4.5 Feature model evaluation and comparison

The features presented in the above sections separate speech from singing with reasonable accuracy, considering that they are isolated features measuring individual phenomena. It is informative to evaluate these features to see if they are in fact measuring the same or related phenomena, and to see the statistical significance of the differences between the PDEs for each feature. To investigate these two concepts, two measures are employed on the overall results of these feature models: Kolmogorov-Smirnov distances and cross-correlation between features. The correctness results from all feature models are also collected and analyzed. The feature labels for the names of the features presented in previous chapters are collected for reference in Table 4.7.

Table 4.7: Feature labels.

Label	Feature
V_{AC}	Vibrato, using autocorrelation
V_{FT}	Vibrato, using fast Fourier transform
$M(f_0)$	Maximum f_0
$m(f_0)$	Minimum f_0
$\mu(f_0)$	Mean f_0
$\sigma(f_0)$	Standard deviation of f_0
$M(f_0')$	Maximum f_0'
$\mu(f_0')$	Mean f_0'
$\sigma(f_0')$	Standard deviation of f_0'
R_s	Segment f_0 track repetition
$\mu_s(f_0)$	Segment-based mean f_0
$\sigma_s(f_0)$	Segment-based standard deviation of f_0
PV_Z	ZCR-based proportion of voiced frames
PU_Z	ZCR-based proportion of unvoiced frames
PS_Z	ZCR-based proportion of silent frames
PV_{f_0}	f_0-based proportion of voiced frames
PU_{f_0}	f_0-based proportion of unvoiced frames

4.5.1 Kolmogorov-Smirnov Evaluation

To evaluate the feature models individually, it is informative to apply the Kolmogorov-Smirnov test described in Section 2.3. In this case, the probability distributions from the *a-priori* speech and song files are compared, and the question is asked whether the two distributions in fact came from the same distribution (indicating that the feature has poor separation) or from different distributions (indicating that the feature may have good separation). In this case, $N_t = 326$ pure talking files and $N_s = 273$ pure singing files, so for the K-S calculations,

$$N = \frac{N_t N_s}{N_t + N_s} = 148.6. \tag{4.6}$$

Using Equation 2.1, we have

$$D_{\alpha=.05} = 0.1116, \qquad D_{\alpha=.01} = 0.1337. \tag{4.7}$$

Recall that $D_{\alpha=.05}$ and $D_{\alpha=.01}$ are the distances required for the null hypothesis (the distributions are the same) to be accepted to the corresponding significance level ($\alpha = .05$ and $\alpha = .01$ respectively). The K-S test was applied to all 17 features in the feature set, and the results are presented in Table 4.8, sorted by decreasing K-S distance. These results show that most of the feature extractors provide sufficient distances between the distributions of the talking files and the singing files for them to be from statistically different distributions.

For three features, $\mu_s(f_0)$, $\sigma(f_0')$, and $\sigma_s(f_0)$, the null hypothesis is accepted at a significance level of $O(0.01)$, and the feature values for the talking and singing files cannot be considered to have come from the different distributions. Visual inspection of the feature models shows that this is not an unreasonable assessment of the usability of these features. The feature $M(f_0')$ shows a K-S distance which is small but worth noticing, and visual inspection of the feature model does indeed show that this feature may not separate as well as some with higher K-S distances.

4.5.2 Cross-Correlation Evaluation

In addition to confirming that the classes within a feature model come from statistically different distributions, another question to consider is whether several features are measuring the same phenomenon, or are they all measuring different phenomena. This can be determined by calculating the zero-lag cross-correlation ($R_{F_1,F_2}(0)$) between each pair of features.

Table 4.8: Kolmogorov-Smirnov results for each feature model.

Feature	K-S Distance	Significance (α)
PU_Z	0.7549	1.5253×10^{-75}
PU_{f_0}	0.6325	3.7513×10^{-53}
V_{AC}	0.6279	2.1580×10^{-52}
R_s	0.5937	7.0015×10^{-47}
PS_Z	0.5714	1.8420×10^{-43}
V_{FT}	0.4661	4.6000×10^{-29}
$\mu(f_0)$	0.3395	1.2799×10^{-15}
$\mu(f_0')$	0.3270	1.6075×10^{-14}
$M(f_0)$	0.2900	1.6370×10^{-11}
PV_{f_0}	0.2657	9.9910×10^{-10}
$\sigma(f_0)$	0.2357	9.5202×10^{-08}
$m(f_0)$	0.2328	1.4317×10^{-07}
PV_Z	0.2191	9.3605×10^{-07}
$M(f_0')$	0.1827	7.9608×10^{-05}
$\mu_s(f_0)$	0.0955	1.2526×10^{-01}
$\sigma(f_0')$	0.0941	1.3624×10^{-01}
$\sigma_s(f_0)$	0.0774	3.2402×10^{-01}

$R_{F_1,F_2}(0)$ is calculated using Equation 3.3. Each pair of features is cross-correlated and divided by the second moment of the first feature, $\overline{{F_1}^2(n)}$, so that the zero-lag autocorrelation is equal to unity, $R_{F_1,F_1}(0) = 1$.

The cross-correlation between features can be interpreted in the following way:

- $R_{F_1,F_2}(0) = 1$: the features produce identical results for each file, and are likely to be measuring the same phenomenon.

- $R_{F_1,F_2}(0) = -1$: the features produce identically opposite results, which also indicates that the features are likely to be measuring the same phenomenon.

- $R_{F_1,F_2}(0) = 0$: the features are considered orthogonal, and therefore are likely to be measuring different phenomenon.

Table 4.9 presents the cross-correlation results for each pair of features, and Figure 4.25 presents a graphic visualization of these results, with darker gray levels indicating smaller cross-correlation results. The cross-correlation results have been ordered so that feature pairs with higher $R_{F_1,F_2}(0)$ values are positioned close together. This also means that higher

$R_{F_1,F_2}(0)$ values occur close to the diagonal. Features with lower overall cross-correlation values are on the left and top of the table and figure.

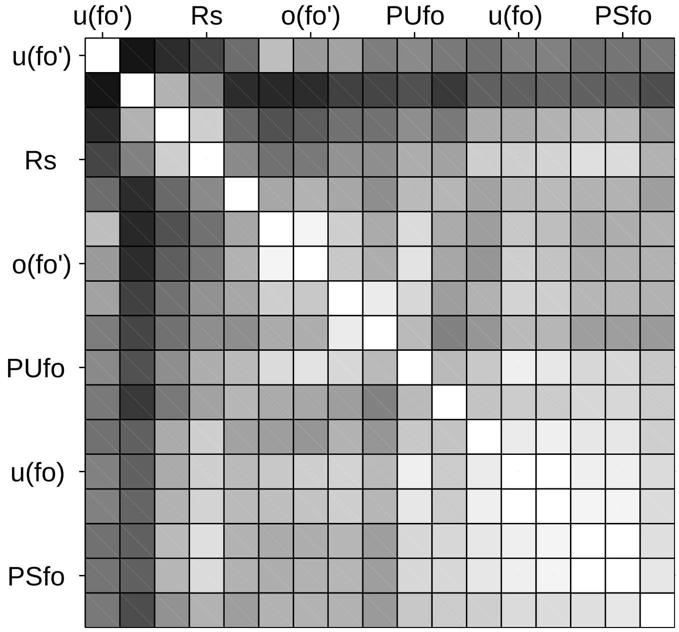


Figure 4.25: Graphical representation of feature pair cross-correlation results from Table 4.9.

These cross-correlation results show that some features are in fact measuring similar or related information. There appear to be groups of related features with high cross-correlation between them: $M(f_0)$, $\mu(f_0)$, PV_Z and PV_{f_0} seem to be related in their correlation distribution, as well as $\sigma_s(f_0)$, $\sigma(f_0')$, $M(f_0')$ and $\sigma(f_0)$. These two groups seem to be uncorrelated with each other. The features that measure voiced frame distribution seem to be clustered together, however PS_Z is not strongly correlated with this group. The two vibrato measures are not strongly correlated with each other or any other features. It is interesting to see that $\sigma(f_0')$ and $M(f_0')$ are highly correlated but $\mu(f_0')$ is not highly correlated with either of these features. Recall, from the K-S distance measures, that $\mu(f_0')$ has a much more significant separation than $\sigma(f_0')$ or $M(f_0')$.

A further instructive observation is the mean of the cross-correlations of a feature with the other features. This feature is calculated according to Equation 4.8:

Table 4.9: Cross-correlation results for all feature pairs.

	$\mu(f_0')$	V_{FT}	V_{AC}	R_s	PS_Z	$\mu_s(f_0)$	$\sigma_s(f_0)$	$\sigma(f_0')$	$M(f_0')$	$\sigma(f_0)$	PU_{f_0}	$m(f_0)$	$M(f_0)$	$\mu(f_0)$	PV_Z	PV_{f_0}	PU_Z
$\mu(f_0')$	1.000	0.083	0.187	0.272	0.435	0.738	0.596	0.640	0.491	0.545	0.470	0.440	0.515	0.505	0.451	0.456	0.480
V_{FT}	0.083	1.000	0.697	0.500	0.182	0.165	0.182	0.256	0.276	0.321	0.222	0.377	0.384	0.396	0.376	0.376	0.303
V_{AC}	0.187	0.697	1.000	0.809	0.409	0.324	0.368	0.441	0.453	0.560	0.481	0.664	0.668	0.688	0.719	0.713	0.564
R_s	0.272	0.500	0.809	1.000	0.544	0.440	0.481	0.568	0.558	0.686	0.639	0.801	0.804	0.819	0.865	0.859	0.699
PS_Z	0.435	0.182	0.409	0.544	1.000	0.650	0.690	0.644	0.559	0.734	0.714	0.631	0.726	0.724	0.696	0.696	0.616
$\mu_s(f_0)$	0.738	0.165	0.324	0.440	0.650	1.000	0.950	0.808	0.660	0.851	0.658	0.615	0.776	0.742	0.669	0.676	0.690
$\sigma_s(f_0)$	0.596	0.182	0.368	0.481	0.690	0.950	1.000	0.772	0.677	0.890	0.647	0.589	0.799	0.759	0.684	0.690	0.692
$\sigma(f_0')$	0.640	0.256	0.441	0.568	0.644	0.808	0.772	1.000	0.907	0.830	0.621	0.689	0.821	0.802	0.716	0.718	0.698
$M(f_0')$	0.491	0.276	0.453	0.558	0.559	0.660	0.677	0.907	1.000	0.734	0.507	0.585	0.732	0.712	0.623	0.623	0.597
$\sigma(f_0)$	0.545	0.321	0.560	0.686	0.734	0.851	0.890	0.830	0.734	1.000	0.725	0.776	0.936	0.903	0.834	0.836	0.777
PU_{f_0}	0.470	0.222	0.481	0.639	0.714	0.658	0.647	0.621	0.507	0.725	1.000	0.756	0.791	0.789	0.841	0.842	0.795
$m(f_0)$	0.440	0.377	0.664	0.801	0.631	0.615	0.589	0.689	0.585	0.776	0.756	1.000	0.915	0.935	0.904	0.903	0.810
$M(f_0)$	0.515	0.384	0.668	0.804	0.726	0.776	0.799	0.821	0.732	0.936	0.791	0.915	1.000	0.986	0.933	0.934	0.850
$\mu(f_0)$	0.505	0.396	0.688	0.819	0.724	0.742	0.759	0.802	0.712	0.903	0.789	0.935	0.986	1.000	0.939	0.939	0.855
PV_Z	0.451	0.376	0.719	0.865	0.696	0.669	0.684	0.716	0.623	0.834	0.841	0.904	0.933	0.939	1.000	0.995	0.868
PV_{f_0}	0.456	0.376	0.713	0.859	0.696	0.676	0.690	0.718	0.623	0.836	0.842	0.903	0.934	0.939	0.995	1.000	0.893
PU_Z	0.480	0.303	0.564	0.699	0.616	0.690	0.692	0.698	0.597	0.777	0.795	0.810	0.850	0.855	0.868	0.893	1.000

$$\mu(R_{F_n}) = \frac{1}{N} \sum_{k=1}^{N} R_{F_n,F_k}(0), \tag{4.8}$$

where $N = 17$ is the number of features in the set. The results of this measure are presented in Table 4.10, sorted from lowest to highest value. A lower value of $\mu(R_{F_n})$ indicates a feature that is less correlated with the rest of the feature set. This information can be combined with the correctness and separation measures presented earlier, to select a set of features with potentially high overall utility. The correctness ratings are collected in Table 4.11 for convenience.

Table 4.10: Mean feature cross-correlation results.

Feature	Mean Cross-Correlation
V_{FT}	0.359
$\mu(f_0')$	0.488
V_{AC}	0.573
PS_Z	0.626
$M(f_0')$	0.629
R_s	0.667
$\mu_s(f_0)$	0.671
$\sigma_s(f_0)$	0.674
PU_{f_0}	0.676
$\sigma(f_0')$	0.702
PU_Z	0.717
$m(f_0)$	0.729
$\sigma(f_0)$	0.761
PV_Z	0.771
PV_{f_0}	0.773
$\mu(f_0)$	0.794
$M(f_0)$	0.798

The correctness ratings presented here are calculated by evaluating the feature models on the same data used to develop them. This is the most straightforward evaluation method, but the risk is always present that the feature models will reflect intrinsic biases in the development data that will not be discovered by evaluating on the same data—recall the tank detector anecdote from Section 3.5.1. If the tank detector system, accidentally trained with tanks on sunny days only, had only been tested with the same data used to develop the system, the error would never have been detected.

One way to more rigorously verify a developed system when a limited amount of data

Table 4.11: Collected correctness results for each feature.

Feature	Relative Correctness	Absolute Correctness
V_{AC}	0.8021	0.8147
R_s	0.7908	0.8063
V_{FT}	0.6487	0.7262
$\mu(f_0)$	0.6779	0.7079
PU_{f_0}	0.6472	0.6678
$M(f_0)$	0.6216	0.6561
$\mu_s(f_0)$	0.5958	0.6511
PU_Z	0.5531	0.6394
PV_Z	0.6003	0.6361
$m(f_0)$	0.5945	0.6277
PV_{f_0}	0.5656	0.6260
$\sigma(f_0)$	0.5743	0.5977
$M(f_0')$	0.5418	0.5910
$\sigma_s(f_0)$	0.5457	0.5910
PS_Z	0.5647	0.5877
$\mu(f_0')$	0.5576	0.5860
$\sigma(f_0')$	0.5185	0.5493

is available is to break the data set into a development subset and a test subset. The development subset is used to develop the system and generate the models, and the test subset is used to verify that the system performs well on data it has not seen yet. To fully verify the system design without biasing on any one portion of the data set, the full data set is divided into several development subsets, and each is used to develop a version of the system, evaluated using the corresponding leftover test subsets.

This testing method was employed to more rigorously verify the feature model development algorithms. The original talking and singing corpus was divided into equal sections, and the talking and singing files from each section were used to develop separate versions of the system. This testing was done in two sessions, first with four sections and then with ten sections. With four sections, four separate systems were developed (using 75% of the corpus) and tested (on the remaining 25%) for absolute and relative correctness, as above. The mean results from these four systems are presented in Table 4.12, sorted by absolute correctness. In the same way, ten separate systems were developed and tested, and the mean results are presented in Table 4.13, sorted by absolute correctness. These results show that, while the correctness results are lower, there are no significant biases in the feature model development, and the features that performed well when developed on the full corpus continued to perform well when the development and test data were separated.

Table 4.12: Feature correctness results for separate development and test data, four sets.

Feature	Relative Correctness	Absolute Correctness
V_{AC}	0.7937	0.8071
R_s	0.7800	0.7886
V_{FT}	0.6415	0.7064
$\mu(f_0)$	0.6549	0.6779
PU_{f_0}	0.6234	0.6560
$\mu_s(f_0)$	0.5798	0.6376
PV_Z	0.5958	0.6342
PV_{f_0}	0.5556	0.6191
PU_Z	0.5377	0.5956
PS_Z	0.5597	0.5923
$M(f_0)$	0.5785	0.5839
$m(f_0)$	0.5665	0.5755
$\mu(f_0')$	0.5423	0.5755
$\sigma(f_0)$	0.5542	0.5722
$M(f_0')$	0.5194	0.5571
$\sigma_s(f_0)$	0.5207	0.5537
$\sigma(f_0')$	0.4860	0.4513

4.5.3 Useful Feature Models

From these results, some conclusions can be drawn. In terms of correctness, the best three individual feature models are V_{AC}, R_s and V_{FT}. The feature models which are most independent of the rest of the features are V_{AC}, $\mu(f_0')$ and V_{FT}. The features which provide the best separation between classes are PU_Z, PU_{f_0} and V_{AC}. The individual feature evaluations among these six feature models indicate that they would be useful in the design of a full classification engine. These feature models are discussed further in Section 5.1.1.

4.5.4 Feature Models Applied to Intermediate Vocalizations

As a final evaluation of the developed feature models, intermediate vocalizations from the speech-song corpus are tested. Recall that these clips were rated between speaking and singing by listeners in the user study in Chapter 2. The mean ratings from the experiment are compared to the feature models generated using the entire speech/song data set.

For each intermediate utterance clip, a feature value is calculated according to the feature

Table 4.13: Feature correctness results for separate development and test data, ten sets.

Feature	Relative Correctness	Absolute Correctness
V_{AC}	0.7921	0.8034
R_s	0.7823	0.7966
V_{FT}	0.6421	0.7153
$\mu(f_0)$	0.6616	0.6881
PU_{f_0}	0.6309	0.6627
$\mu_s(f_0)$	0.5786	0.6339
PV_Z	0.5960	0.6339
PV_{f_0}	0.5564	0.6186
$M(f_0)$	0.5850	0.6170
PU_Z	0.5411	0.6170
$m(f_0)$	0.5754	0.6000
PS_Z	0.5583	0.5848
$\mu(f_0')$	0.5504	0.5814
$\sigma(f_0)$	0.5491	0.5678
$M(f_0')$	0.5178	0.5492
$\sigma_s(f_0)$	0.5182	0.5458
$\sigma(f_0')$	0.4955	0.5017

extractor algorithm. The feature value is then applied to the feature model to generate a computed measure (M_c) between -1 and 1. This rating is compared to the human measure (M_h), being the mean listener rating for that clip, a value between 1 and 5. The mean rating result is scaled to match the feature result, and the euclidean distance is calculated between these ratings, using Equation 4.9:

$$D = \left| M_c - \left(\frac{M_h - 3}{2} \right) \right| . \qquad (4.9)$$

If the computed measure and the human measure match, the distance will be zero. If they are opposite, *e.g.* if $M_c = -1$ indicating talking and $M_h = 5$ indicating singing, the (maximal) distance will be 2.

The human measure is compared to the computed measure for each intermediate file applied to each feature, and the mean distances for all intermediate files are presented in Table 4.14.

These distances show that while no feature model duplicates the human perception of intermediate vocalizations between speech and song, some features do provide encouraging results. It is interesting to note that the features that most closely approach the human

Table 4.14: Feature model results compared to human ratings of intermediate files.

Feature	Mean Distance
V_{FT}	0.4122
$\mu_s(f_0)$	0.5189
V_{AC}	0.5241
PU_Z	0.5333
R_s	0.5498
PV_Z	0.5722
PU_{f_0}	0.5762
PV_{f_0}	0.5804
PS_Z	0.5830
$\sigma(f_0')$	0.5956
$\sigma_s(f_0)$	0.6007
$\mu(f_0')$	0.6212
$m(f_0)$	0.6643
$M(f_0')$	0.7162
$M(f_0)$	0.7958
$\sigma(f_0)$	0.8068
$\mu(f_0)$	0.8148

ratings are some of the same features that have performed well in other evaluations. These intermediate results are preliminary, but they do show that it should be possible to develop features that can model intermediate vocalizations.

The final chapter in this thesis describes some procedures for the development of a multi-feature classification system, as well as some conclusions and future directions for this research.

Chapter 5

Conclusions

This chapter presents a summary of this thesis, and identifies areas where more work could be done. Suggestions are made to augment the research corpus, including more languages and more specific intermediate utterances. Feature improvements are suggested, including the addition of new feature extractors and improvements to existing algorithms. The chapter ends with some closing observations.

5.1 Summary

This thesis sought to address the question: "Are there measurable differences between the auditory waveforms produced by talking and singing?" The work was divided into two main goals with the ultimate task of designing a set of algorithms to extract relevant features from the auditory waveform.

The first goal was to find a set of relevant features for the speech/song discrimination task. Three sources were used to discover phenomena which may be relevant for the task. First, the principal researcher proposed some features from his own research experience. Second, a set of human listeners were asked to provide their opinions on the differences between speech and song, based on their observations of specific sound clips and their general observations. Third, features were taken from current auditory classification research.

The second goal was to develop feature extractors for each of these phenomena. The feature extractors fell into three general classes: vibrato features, statistical f_0 features, and rhythm features. Vibrato was a phenomenon cited by many listeners and researchers, and two methods were described to extract vibrato information from the waveform. Statistical

f_0 features were extracted based on the f_0 track, the slope of the f_0 track, and utterance segments. Rhythmic features included proportion of voiced, unvoiced or silent frames, and utterance segment repetition.

Once the feature extractors were developed, the features were evaluated using three measures. First, the features were tested on *a-priori* labeled data, to determine if the features were classifying correctly. Second, the features were tested using Kolmogorov-Smirnov distances to determine if the features were separating the two target classes with statistical significance. Finally, the cross-correlation of the feature results was used to determine whether the features were measuring separate phenomena or a small number of underlying phenomena.

The two principles presented in Chapter 1 have helped to guide this research. Principle 1 led to the consideration of techniques such as autocorrelation and zero-crossing rate, both of which resulted in effective feature extraction techniques for several phenomena. Principle 2 provided insight into the cyclical nature of utterance and perception, specifically in the understanding of vibrato and rhythm.

5.1.1 Summary of Feature Model Results

From Section 4.5.3, the feature models that are most likely to be useful in a classification engine are V_{AC} and V_{FT}, the autocorrelation- and FFT-based vibrato measures; PU_Z and PU_{f_0}, the ZCR- and f_0-based proportion of unvoiced frames; R_s, the correlation between utterance segments; and $\mu(f_0')$, the mean f_0 slope. It should be noted that although $\mu(f_0')$ has low cross-correlation with the rest of the feature models, the correctness results for this feature are not as good as some of the other feature models. It should also be noted that some of the remaining feature models not discussed here do show some promising results, and further study should be done on these features as well.

The vibrato measures V_{AC} and V_{FT} both performed well, and the cross-correlation between them was lower than what might be expected for two features designed to measure the same phenomenon. Despite the preference for time-domain techniques (Principle 1), it seems as though both time- and frequency-domain algorithms are useful here.

The same is true for PU_Z and PU_{f_0}, two feature models designed to measure the proportion of unvoiced frames in an utterance. Both models had high K-S distances, with reasonable correctness and cross-correlation results.

The R_s feature model, which measured the similarity between the f_0 track of utterance segments, performed well in correctness and K-S distance, and had reasonable cross-correlation measures with the rest of the feature set.

The following sections describe techniques which can be used to develop a full classification engine to separate human vocal utterances into speech and song, or along a continuum between speech and song.

5.2 Corpus Improvements

Because the corpus was collected and annotated fairly early in the project, many somewhat arbitrary assumptions had to be made about the type of data to collect and the type of annotation that would be relevant. Having completed the study and feature analysis, an obvious next step would be to collect and annotate a new corpus taking the findings of this study into account, and then apply the algorithms to the new corpus and see what happens.

A set of sound files has been found that could be used as an additional test corpus - a CD called "Music Play" developed for an early childhood music curriculum[74] contains many samples of chant-like utterances with rhythmic and melodic components, and it would be interesting to see how humans evaluate these clips, as well as to examine the results of the feature extractors on these clips.

The corpus contained sounds from professional and amateur singers, as well as professional and amateur speakers, although no attempt was made to isolate the differences between professional and amateur utterances. An interesting research project would be to gather more data and analyze differences in speaking and singing style which may come with training or experience.

The f_0 track evaluations based on the corpus were strictly comparative. A corpus improvement that would add credibility to the f_0 tracks used would be to incorporate some form of f_0 validation, such as electroglottogram or annotated f_0 values.

5.2.1 Language

The files in the initial corpus are primarily English, with some representative selections from other languages. This is because the study intentionally concentrated on the English language to keep it manageable. An interesting improvement to the corpus and to the entire research project would be to expand the investigation to include many other languages and cultures. In tonal languages such as Mandarin Chinese, lexical information is encoded in the pitch contour as well as the phoneme, and it would be particularly interesting to study the way in which this affects the experience of song in these cultures.

People who listen to a language they do not speak sometimes identify the language as somewhat song-like. For example, many non-Swedish speaking people identify the Swedish language as "sing-song". Some listeners from the study in this thesis mentioned that when the utterance language was unfamiliar, the speech/song judgement was made based on the the presence of features from a familiar language.

5.2.2 Prosodic Speech

Cross-language perception is one example of spoken language being perceived as song-like, and prosodic speech is another. A study of the properties of prosodic "warning" speech, for example, would be very interesting because many people identify the characteristic rise and fall of a "warning" utterance as song-like. As an example, imagine (when you were young) a parent or caregiver saying "you're gonna be sor—ry" and drawing out the "sorry" into almost a descending major third interval. A teacher acquaintance, when told about this research project, related an anecdote where whenever she starts to get impatient with the students, they will accuse her of singing to them. This may be another example of prosodic speech becoming song-like.

Another class of utterances that lies between speech and song is skipping rhymes or clapping rhymes. These play chants are common in younger people and an interesting project would be to investigate the perception of these rhymes by people who use them frequently as compared to people who hear them as if for the first time. These rhymes often have frequently repeated pitch tracks which would move them more toward the song end of a speech/song continuum.

Other school-yard examples of song-like speech are taunting utterances, such as "neener neener neener" and Nelson Muntz's infamous "haw haw" from the "Simpsons" cartoon television series, as well as play utterances such as the "olley-olley oxen free" at the end of a game of hide-and-seek. An investigation of the similarities and differences between all of these prosodic utterances could lend understanding to the human (and the computer) experience of song and speech.

5.2.3 Specific Intermediate Classes

The corpus extraction in this thesis presented general prompts to collect intermediate classes, and the search for intermediate utterances in existing media was not directed to any specific type of intermediate utterance. Throughout this work, many specific intermediate classes have been identified, including poetry, liturgical chant, rap music, playground rhyme, auctioneering, and highly prosodic speech such as warnings, lectures and sermons. Many research projects could stem from the study of the song-ness of any of these utterances, or the identification of features relevant to differentiating between any or all of them.

5.2.4 Context-Free Utterances

One problem with the current corpus is that, as mentioned in Section 2.4.1, some samples were source- or content-recognizable, implying that extra meaning was obtained based on the previous experience of the listener and not on the characteristics of the sound.

One way to isolate the effect of expectation or context would be to use clips based on the *Harvard sentences*, which are phonetically balanced and contextually neutral. An example is "The boy was there when the sun rose." These sentences contain a well-balanced collection of phonemes and provide consistent lyrics. An experiment could be developed which would solicit samples from subjects using a Harvard sentence or a set of Harvard sentences, spoken in a particular style. Styles could include normal speech, read speech, song based on a random melody, song based on a well-known melody and various intermediate utterances.

5.3 Feature Improvements

The feature extractors developed in this work provide reasonable individual separation, and cross-correlational studies between the features show that it is reasonable to believe that combining features will result in better classification results. Regardless, the features presented are not definitive and there is no reason to believe that they are as good as they can be. Future work related to features can be divided into two categories: Improvement of current feature models and extractors, and development of new features.

A standard set of features used for many sound analysis research areas is called *mel frequency cepstral coefficients* (mfcc). No work was done in this thesis on mfcc for speech/song classification, and it would be interesting to apply mfcc to this problem. An improvement on mfcc, called human factor cepstral coefficients (hfcc) has recently been presented [68]. An interesting research project would be to compare these two feature sets on the intermediate speech/song domain.

A set of features which have been evaluated individually can be combined into a multi-dimensional feature model. Dimensionality reduction techniques can be used to isolate orthogonal axes which simplify the classification problem.

The features investigated in this thesis are primarily temporal in motivation. Spectral differences between speaking and singing are also evident, and future work could also include the investigation of spectral differences between speech and song, for example higher-frequency power augmentation and so-called "singers formant".

5.3.1 Improvement of Feature Evaluation

Some critics of Kolmogorov-Smirnov testing have indicated that the significance indicator may become unreliable at large values of N (greater than 100). The significance indicator gives a measure of the likelihood that even though there is a distance between the distributions of two random variables, they come from the same original distribution. If the significance is 0.05, then it can be interpreted that 99.5% of the time, the two random variables behave independently of each other, and 0.5% of the time, the random variables behave identically.

Critics theorize that as N becomes large, the measurements taken when the random variables behave identically will be overshadowed by the measurements taken when the random variables behave independently of each other, artificially inflating the results and making the two distributions appear more different than they actually are. One way to confirm or deny this theorem for the current work would be to randomly divide the data into smaller sets (as in the separation of the development and test sub-corpora), and perform K-S testing on all set pairs. The comparison could then be made between the K-S distances of set pairs from the same original distribution (*i.e.* two sets from talking or two sets from singing) and K-S distances of set pairs from the theoretically different distributions (*i.e.* one set each from talking and singing). If the distances are significantly smaller in set pairs from the same distribution, this would confirm that the significance of the original K-S distance is valid.

5.3.2 Improvement of Current Features

Some of the feature extractors developed for this work may be improved with further research, although any modification would need to be compared to the current algorithm for that feature. A specific improvement that may be of use is related to the pre-processing of the segment-based features. Currently, segments are identified based on f_0 track and power fluctuations. This does not allow for the separation of segments which have homogeneous power and/or f_0, such as segments separated by a non-fricative consonant. Formant analysis could be employed to detect changes between such segments and effectively separate the signal into individual phonemes.

The statistical measures used to extract the f_0 features in this work are simple first order statistics. Further research could include analysis of higher order statistical measures such as skew and kurtosis, as well as f_0 range and f_0 slope. Mode and median of the f_0 and the f_0 slope may also be useful in this determination.

5.3.3 Development of New Features

Some of the features identified by the listeners in the corpus study have not been studied or developed in this work. These additional features include nebulous concepts such as expectation and context, as well as features which would require higher-level linguistic analysis such as rhyme or lyrical repetition. A discussion of some of the issues involved in the development of these features is presented here.

Rhyme

Phonetic information would be very useful in detecting patterns of rhyming words, and this would require formant extraction and F1:F2 characterization, similar to the preliminary steps of a speech recognition engine. Rhyme is likely to correlate well with rhythmic structures, so these two features could relate to and inform one another. Rhyme information could be extracted through formant, phoneme or orthography information, depending on the level of analysis available in the rest of the system.

Expectation and Context

Some listeners from the corpus annotation project identified the fact that simple perceptual features may not be sufficient to characterize the speech-song continuum axis. Especially in ambiguous utterances, context and expectation play an important role as well. If the lyrics in the utterance are ambiguous, but remind the listener of a song once heard, this additional context may be sufficient to nudge the listener's opinion in the direction of song. If a listener hears the the lyrics of a familiar song in an unfamiliar environment, either spoken or sung, the classification may be different than for unfamiliar words. Similarly, the expected ending of an utterance can lend weight to one end or the other of the scale.

Expectation and context are listener-specific difficult to quantify, but expectational probabilities have been used in speech recognition engines in the past and could be applied to this problem as well.

5.4 Related Research Areas

The classification of human utterances is a specific research sub-domain, combining audio signal processing, psychology, audiology, speech analysis, music analysis and other disciplines. Each of these areas present concepts and research ideas that are worthy of more study.

5.4.1 Audio Signal Classification

The work presented in this thesis pertains to classification of a specific audio domain. Automatic classification in other domains and more general systems could be developed based on this work and current research being performed elsewhere.

Musical instrument classification is currently a research topic of interest. Given a monophonic musical sound, the classification output could be considered on several levels: Is the instrument a wind, string, reed, brass or percussion? Within the string instruments, is the instrument a violin or a cello? Within the cellos, is the instrument a Stradivarius or a mass-produced schoolroom model? Is the player a professional or a beginner? Within the context of polyphonic music, is the musical sound produced by a 60-piece orchestra or a small ensemble? Is the chord being played a major 7th or a minor 9th?

Animal sound classification would be similar in many ways to musical instrument classification. Given an animal sound, is the animal that produced it a mammal, a bird or a fish? Is the mammal a cat, a dog, a horse or a rodent? Is the cat a lion, a panther or a house cat? Current methods of speaker identification could be applied to identify specific animal individuals for migratory tracking and environmental or zoological research.

Sound effect classification is an area of interest which includes classification of many different types of sounds. Many sound effects are short and percussive, so f_0 based features would not work well for these. Other features based on attack and decay, as well as perhaps fractal or wavelet-based techniques, would work well in this context.

5.4.2 Psychological Studies

When trying to imitate, emulate or improve upon the human auditory system, it is important to understand that system in the context of the behaviors being emulated or improved. Further psychological and audiological studies will help to uncover perceptual phenomena helpful to the development of computer listening systems.

Many audiological questions present themselves in the context of this work. How quickly do humans classify sounds? How do humans perceive auditory ambiguity, and how can that be coded into or improved upon using computer perception? How do humans use experience, context and expectation in perception of audio signals? What steps are there between the low-level perception of sound and the attachment of meaning? What happens to the human auditory system at the extremes of pitch and loudness perception? Some or all of these questions may already be answered by researchers in audiology and conative psychology.

5.5 Closing Remarks

Understanding and being able to measure the differences between talking and singing is significant because it increases our understanding of human utterances and improves our ability to design computer programs which may be able to interact more easily with humans and the audio environment.

The measurable differences between speech and song are both perceptual and physical, and in this thesis, computational algorithms to extract these differences have been shown to be possible.

As with any research project, parts of this work seemed to invite almost limitless investigation. Each question that was answered prompted two more; each feature developed suggested another. There is a great deal of fascinating research ahead.

Appendix A

Corpus Research Protocol

This appendix contains the protocol document submitted to the research ethics committee of Simon Fraser University when applying for ethical approval for the collection of the corpus used in this thesis. Although some features of the collection and annotation protocol have changed, specifically the scale used, the collection and annotation process followed the proposed protocol.

A.1 Introduction

In this document I will present an outline, specifications and discussion on a proposed protocol for collecting a corpus of sound files containing human utterances. The corpus collection is primarily for my thesis work on fuzzy classification of human utterances on a speech/song axis, but I am planning to build the corpus in such a way that it can be published and used for other research. For that reason I will be making the corpus domain more general and the annotations more informative than perhaps is necessary for my thesis alone.

The building of this corpus will proceed in two stages. Stage 1 is the collection of appropriate sound files from various pre-recorded sources including internet, radio, published sources such as music and spoken word CDs and movie soundtracks, as well as collection from live sources, in the form of solicited utterance samples from human subjects.

Stage 2 of the corpus building procedure is to annotate the corpus. This stage will consist of going through the corpus and transcribing the words, as well as soliciting human subject opinions of the sounds in the corpus. The human opinions will give the corpus validity in the speech/song classification, especially in the fuzzy intermediate domain which will contain utterances such as poetry and chant.

A.2 Corpus Design

This section presents a discussion on the proposed structure of the corpus as well as the limitations and restrictions that will be applied to the corpus design. A summary of this discussion will be presented at the end of the section.

A.2.1 Corpus Domain

The FSS (fuzzy speech song) corpus is intended for a specific research domain: human utterance classification in the domain of speech and music. The primary limitations on the corpus are that it will contain only monophonic (with no background or noise) human utterances, containing speech, song, or some intermediate vocalization.

Some secondary restrictions are that the FSS corpus will contain primarily English when a language is used, although the corpus is not restricted to English; samples that contain song will be primarily in the 12-tone equal tempered music system (commonly referred to as the “western” music system) but again the corpus is not formally restricted to the western music system. The corpus will contain a few samples of other languages and other music systems in the corpus for comparison, especially tonal languages and aboriginal music systems.

The corpus will include samples that reflect different characteristics of human speech, as well as different intentions for the corpus itself. The corpus will be able to be segmented along three axes:

- Constrained utterances — Free utterances
- Spoken utterances — Sung utterances
- Speaker Characteristics

A.2.2 Constrainedness

In order to make the corpus useful for the specific context of fuzzy speech/song classification, but at the same time still be valid for real-world samples, the corpus will contain solicited human utterances of two types. Constrained human utterances will have one or more restrictions placed on the utterance during recording. The proposed constraints fall into four categories:

- Constraints on content of utterance
- Constraints on style of utterance
- Constraints on both content and style
- No Constraints

Content constraints. The constraints on the utterance content consist of requiring the speaker to utter a specific phrase. The phrases to be uttered are chosen to reflect certain expected features of the speech/song classification. Two features that will be investigated in this manner are voiced/unvoiced distribution and formant constancy.

It is expected that song will show a higher percentage of voiced segments of speech (vowels, etc.) and a lower percentage of unvoiced segments (fricatives, plosives etc.). Indeed, preliminary experiments have shown this to be true. All English lyrical (spoken or sung) utterances contain voiced phonemes, but not all contain fricatives. The voiced/unvoiced distribution feature extractor would behave as if all utterances with no unvoiced segments are song. For this reason, the corpus should contain a spoken utterance with only voiced phonemes to make sure the full system can handle such an utterance. It is proposed that one of the spoken utterances solicited from subjects be:

> “When you’re worried, will you run away?”

Another feature expected in song is that the glide of diphthongs will be suppressed till the beginning or ending of the phoneme. To test this, it is desired to have utterances with many diphthongs. For this reason, it is proposed that one of the utterances solicited from subjects be:

> “Row, row, row your boat, gently down the stream.”

The diphthongs in this utterance are expected to be short and rhythmic. As a contrast, the following utterance will also be solicited:

> “O Canada, our home and native land.”

Both of the above utterances will be solicited spoken as well as sung.

Style constraints. This corpus is being designed with a particular piece of work in mind, that is an attempt to characterize the distinction between speech and song, with investigations also directed toward intermediate vocalizations. Because of this, part of the corpus will contain utterances where the subject is prompted to sing or is prompted to speak. As indicated above, some samples will be requested in both spoken and sung styles, so the differences between speaking and singing in these samples would not be obscured by differences in content or in subject characteristics.

There would be samples taken of unconstrained content with constrained style as well. The purpose of these samples would be to expand the corpus beyond constrained utterances, which test particular characteristics and features, but are not appropriate for design of a system to operate on "real-world" data.

The style-constrained samples would allow the speaker to choose the content (lyrics) of the utterance, but would insist on a particular style of utterance. Example prompts are:

> "Sing the first line of your favourite song."
>
> "What did you have for lunch yesterday?"

A further style constraint which will attempt to illicit samples in the middle ground between speaking and singing would allow the speaker to utter any lyric in any style so long as it is *neither* speaking *nor* singing. A prompt for this style constraint would be:

> " In a single sentence, Tell me what you did last weekend, using a voice which is somewhere between singing and speaking."

A similar prompt using constrained content would be:

> "Utter the phrase 'Why is the sky blue?' using a voice which is somewhere between singing and speaking."

The phrase "Why is the sky blue?" has many characteristics that are desirable for this corpus. It contains a good distribution of fricatives, both voiced and unvoiced, and two diphthongs which rhyme.

No constraints. This section of the corpus will consist of samples that are unconstrained in any way. These samples include all "found" samples (samples not directly solicited from human subjects), for example samples taken from radio or from movie soundtracks. The corpus will also include some unconstrained samples solicited from subjects. The majority of the corpus that I currently have falls into this category. The richness and variability of completely free samples fills out the structured nature of the rest of the corpus.

A.2.3 Utterance Class

The second way of dividing the corpus is in the perceived style of the utterance itself. Since the majority of the research is concentrating on speech and song, many of the samples will fall clearly into one of these two categories, with the remainder falling into the category of "Fuzzy speech/song", indicating that the sample has characteristics of both speech and song, but is not clearly one or the other. Some samples will be specifically designed to fall within this category, such as the manipulated sample corpus described in Section A.3.4, as well as some of the style-constrained utterances. Some found utterances will end up in this category as well. The possible utterance classes are:

- Purely speech utterances
- Purely song utterance
- Fuzzy speech/song

It is important to note that this classification will rely on human opinion testing of the corpus, and not from any characteristics of the corpus files. The entire corpus, once collected, will be labeled on a fuzzy scale between speech and song, using the results of the human opinion testing described in Section A.4. It is expected that there will be many samples characterized as pure speech or pure song, which is why the corpus collection protocol is biased toward samples which are expected to fall into the fuzzy category between speech and song.

A.2.4 Speaker Characteristics

Human speech is varied because human speakers are varied. Since the purpose of the proposed corpus is to aid in the design and testing of a system that will operate on human speech, it is important that the corpus contain a balance of human speaker characteristics. For this reason, it will be important to make sure that the subject base contains a good balance of individuals on the basis of the following characteristics:

- Age
- Gender
- Musical/Speech training

Young people, especially children, speak with higher pitch than do adults so the pitch range feature extractor proposed in the classification system should be able to handle speech from children. Older people have different voice characteristics, as do children at the verge of puberty. The proposed corpus would do well to have samples from representatives of each of these age groups. To avoid collecting samples from individuals under the age of consent, all child samples will be found rather than solicited, taken primarily from movie soundtracks.

Men and women have different pitch aspects of speech. The proposed corpus will have a balance of male and female subjects.

A characteristic of speech that is especially relevant for this corpus is musical training. Song is a faculty that all humans possess, but those that are trained in singing have the ability to make their voice do exactly what they want. These speakers will be able to give samples of very high quality song, and might be more able to give samples in the middle-ground between speech and song, or samples that are neither speech nor song. Individuals who are trained in speech, such as actors or radio personalities, also have the ability to manipulate their voices as desired. The corpus should have a portion of samples solicited from trained users of speech and song.

A.3 Corpus Collection

This section describes the protocol for collecting the samples which will populate the FSS corpus as described above. There will be four categories of collection:

- Free samples
- Constrained Samples
- Found Samples
- Manipulated Samples

Each category is described here, including proposed collection protocol. The solicited samples will be collected from human subjects using a protocol approved by Simon Fraser University according to the university research ethics guidelines.

The subjects will be selected randomly, with intention to fill out the categories described in Section A.2.4.

A.3.1 Free Sample Subcorpus

An important sub-corpus is the corpus of solicited samples with no constraints. As discussed above, these samples are necessary to fill out the otherwise structured nature of the corpus, and also provides some "real-world" samples for a system designed on this corpus to deal with.

The proposed collection protocol for unconstrained samples is this: Two unconstrained sample phrases will be collected from each subject, using the following prompt for both samples:

> "Please speak or sing anything you like for about 5 seconds."

A.3.2 Constrained Sample Subcorpus

This subcorpus will be gathered from human subjects, in the same way that the free sample subcorpus will be collected, using various constraints as described in Section A.2.2. The samples will be constrained in style, in content or in both style and content. The proposed prompts, as stated above, are:

> "Sing the first line of your favourite song."
>
> "What did you have for lunch yesterday?"
>
> "Please speak the phrase 'When you're worried, will you run away?' "
>
> "Please sing the phrase 'Row, row, row your boat, gently down the stream.' "
>
> "Please speak the phrase 'Row, row, row your boat, gently down the stream.' "
>
> "Please sing the phrase 'O Canada, our home and native land.' "
>
> "Please speak the phrase 'O Canada, our home and native land.' "
>
> " In a single sentence, Tell me what you did last weekend, using a voice which is somewhere between singing and speaking."

> "Utter the phrase 'Why is the sky blue?' using a voice which is somewhere between singing and speaking."

As with the unconstrained samples, the subjects will be encouraged to limit their utterances to about 5 seconds. For prompts that require a specific phrase, the user will be encouraged to read, remember, then speak the phrase as if they were talking or singing to another human. For prompts requesting an utterance that is neither speech nor song, the subject will be encouraged to practice a couple times before recording, to get a feel for what a non-speech, non-song sound might be like. Because singing a phrase first may influence how the subject then speaks the same phrase, the order of the prompts will be varied.

A.3.3 Found Sample Subcorpus

This subcorpus will be populated by extracting short segments of sound from publicly available audio, such as radio, published music, movie soundtracks, and .wav and .mp3 files available on the internet. Copyright laws allow reproduction of copyright material for research purposes.

I will be scouring the net, radio and movies for sounds that would be appropriate for this corpus. Examples of sounds that I am expecting to acquire:

- "Daisy, daisy, give me your answer, do" (HAL 9000, "2001")
- "Good morning vietnam!" (Robin Williams, "Good Morning Vietnam")

Various vocalists have been suggested to me as well including Mark Knopfler, Yoko Ono and Bob Dylan. The challenge with collecting found samples will be to find samples of people singing and speaking without any background noise or music. Stationary background noise is acceptable, because the system will be able to filter it out as long as there is a couple seconds of silence (with the background noise) before the human utterance begins.

A.3.4 Manipulated Sample Subcorpus

This subcorpus will consist of samples that have been deliberately manipulated to fool a specific feature extractor. Three of the feature extractors that are expected to be successful in detecting the presence of song are:

- Pitch outside of normal pitch range.

- Presence of vibrato in pitch track.
- Larger proportion of voiced segments.

To design sounds that would fool each individual feature detector, I would begin with a sound that would clearly be classified as speech, and then manipulate characteristics of the sound using granular synthesis. The goal of this manipulation would be to create a sound that a human would consider to be speech, but which has one of the features of song as expected from the feature extractor being designed.

As an example, I would take a sound sample of someone speaking, with pitch inside the normal pitch range for speaking, and granularly increase the pitch so that the pitch range feature extractor would classify it clearly as song. Opinions of this file would be solicited in the usual manner (see Section A.4) to determine what effect the pitch range has on the perceived class of the sound.

This procedure would be repeated with the other features: adding a harmonic ripple to voiced segments of a speech sound; extending the voiced segments and compressing the unvoiced segments; and performing similar manipulations with other features.

The same procedure would be performed in the other direction—making song samples sound like speech for a particular feature extractor. For example, removing spectral ripple from a song sample; bringing the pitch into normal speaking range; compressing the voiced segments and extending the unvoiced segments; and performing similar manipulations with other features.

The goal of this sub-corpus would be to verify the individual success or failure of each feature for a speech/song classification, as well as testing the robustness of the overall system in the presence of one divergent feature result.

A.3.5 Corpus Summary

Figure A.1 shows the proposed corpus by subdivision criteria. Table A.1 summarizes the proposed subcorpora by collection procedure, along with the related collection methods, expected sizes, and purposes.

Each sample in the corpus can be classified on each axis. if a sample were solicited under the constraint of style = song, for example, the source would be “solicited” with the gender, age and training characteristics corresponding to the subject, and the class would be indicated by the opinion gathering after the corpus is fully collected.

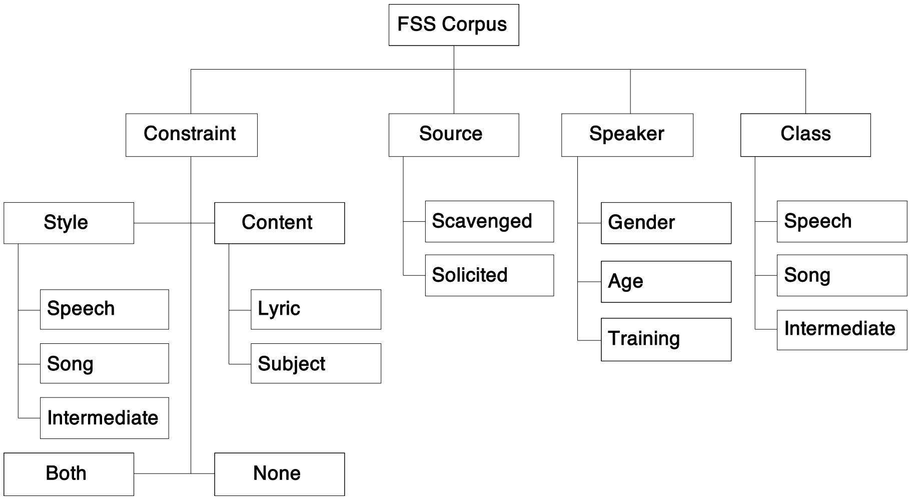


Figure A.1: The FSS Corpus subdivisions and categories.

Table A.1: FSS collected subcorpora characteristics.

Subcorpus	Collection Method	Size	Purpose
Free	Solicitation	100	“real-world” samples
Constrained	Solicitation	500	boundary conditions
Found	Extraction	100	existing samples
Manipulated	Design	50	individual feature testing

A.4 Opinion Solicitation

The second stage of building the corpus is to annotate the corpus. This consists of transcribing all lyrics used in the speech and song samples, as well as soliciting human opinion scores for all samples in order to label the corpus on the “speech/song”axis. The opinions will be solicited in a manner similar to the solicitation of the samples, and opinions will be solicited from the subjects who provided the samples, as well as other subjects who did not.

A.4.1 Opinions on the Full FSS Corpus

Depending on the size of the corpus, subjects will be asked to provide an opinion for some or for all of the corpus. 750 samples of 5 seconds each would take 1 hour, 2 minutes to listen to, without pauses between samples. If we predict 15 seconds for each sample, listening and classifying, the time to complete 750 samples would be 3 hours, 12 minutes. I expect that it will be prudent to break the opinion collection into 30 minute sessions.

The opinions will be recorded with the age, gender and musical or speech training level of the subject, along with whether or not the subject provided samples for the corpus in the corpus collection stage.

The subjects will be asked the following questions about each sample:

> "Please rate this sample on a scale between speaking and singing. A sample of pure speech should be rated '0' and a sample of pure song should be rated '10' "
>
> "Please rate the quality of speech or song, from 0 to 10. A 'bad' quality sample should be rated '0' and a 'good' quality sample should be rated '10'. Please rate the quality of the voice only, not the quality of the recording itself."
>
> "Please choose on e of the words on the response form to describe the sample. Choose one of: speech; whisper; yell; poem; babble; monotone; chant; rap; song; or write another word in the space provided."

A.4.2 Opinions on a Selected Sub-set of the FSS Corpus

Along with these general opinions, a small sub-set of the corpus will be selected for further opinion gathering. This subset will consist of solicited, found and designed samples which fall into the following categories:

- Clearly speech
- Clearly song
- "Rap" style utterance
- Poetry
- Chant
- Babbling, and singing without words
- Whispering
- Yelling
- Monotonous speech (as in a university lecture)

Also in this sub-set will be samples from the corpus that are difficult to categorize, or perhaps samples that fall into the fuzzy middle ground between speech and song.

I will solicit more detailed opinions on this sub-corpus. The subjects will be asked the following questions about each sample:

> "Please rate this sample on a scale between speaking and singing. A sample of pure speech should be rated '0' and a sample of pure song should be rated '10' "
>
> "Please rate the quality of speech or song, from 0 to 10. A 'bad' quality sample should be rated '0' and a 'good' quality sample should be rated '10'. Please rate the quality of the voice only, not the quality of the recording itself."
>
> "Please indicate what the speaker might have done to make this utterance more speech-like"
>
> "Please indicate what the speaker might have done to make this utterance more song-like"

The first two questions are identical to the first two questions for the full corpus, and are included in the sub-corpus opinion testing to test for opinion consistency. The second two questions are free-response, and are included to extract a general intuition about speech, song, and the middle-ground between them.

A.5 Summary

This document describes the proposed protocol for collecting and annotating the FSS (Fuzzy Speech Song) corpus, intended for research on human utterance classification, specifically speech, song and the fuzzy intermediate domain between speech and song.

Stage 1 of the corpus collection protocol consists of acquiring utterance samples from human subjects and from available media, in four categories: Constrained utterances, Unconstrained utterances, Found utterances, and Designed utterances.

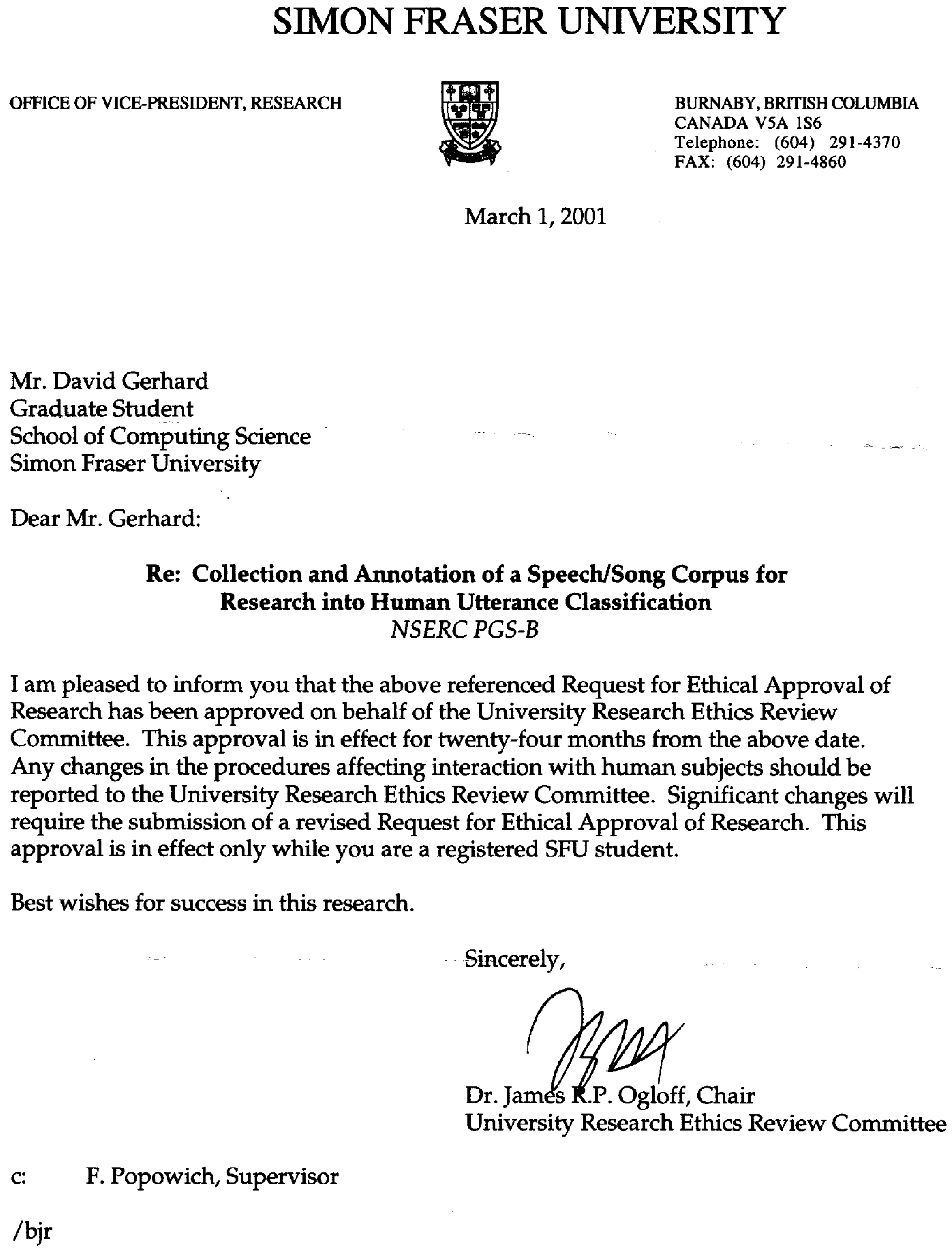
Stage 2 of the corpus collection protocol is the annotation of the corpus by human subject opinion. Human subjects will be asked to listen to the samples in the corpus and provide opinions based on a series of questions. Subjects will also be asked to provide more specific opinions on a subset of the corpus.

Appendix B

Corpus Research Instruments

This appendix contains the ethics approval letter, as well as examples of the collection and annotation tools for the corpus used in this thesis. Included are the consent form for the collection stage (the consent form for the annotation stage is similar), the information sheet provided to subjects, the research instruments, and web form used for the corpus annotation.

B.1 Ethics Approval Letter



SIMON FRASER UNIVERSITY

OFFICE OF VICE-PRESIDENT, RESEARCH

BURNABY, BRITISH COLUMBIA
CANADA V5A 1S6
Telephone: (604) 291-4370
FAX: (604) 291-4860

March 1, 2001

Mr. David Gerhard
Graduate Student
School of Computing Science
Simon Fraser University

Dear Mr. Gerhard:

Re: Collection and Annotation of a Speech/Song Corpus for Research into Human Utterance Classification
NSERC PGS-B

I am pleased to inform you that the above referenced Request for Ethical Approval of Research has been approved on behalf of the University Research Ethics Review Committee. This approval is in effect for twenty-four months from the above date. Any changes in the procedures affecting interaction with human subjects should be reported to the University Research Ethics Review Committee. Significant changes will require the submission of a revised Request for Ethical Approval of Research. This approval is in effect only while you are a registered SFU student.

Best wishes for success in this research.

Sincerely,

Dr. James R.P. Ogloff, Chair
University Research Ethics Review Committee

c: F. Popowich, Supervisor

/bjr

B.2 Informed Consent Form: Collection

SIMON FRASER UNIVERSITY
INFORMED CONSENT BY SUBJECTS TO PARTICIPATE
IN A RESEARCH PROJECT OR EXPERIMENT

The University and those conducting this project subscribe to the ethical conduct of research and to the protection at all times of the interests, comfort, and safety of subjects. This form and the information it contains are given to you for your own protection and full understanding of the procedures. Your signature on this form will signify that you have received a document which describes the procedures, possible risks, and benefits of this research project, that you have received an adequate opportunity to consider the information in the document, and that you voluntarily agree to participate in the project.

Having been asked by David Gerhard of the School of Computing Science of Simon Fraser University to participate in a research project experiment, I have read the procedures specified in the document. I understand the procedures to be used in this experiment. I understand that my voice will be recorded, and that my recorded voice will be included in a data corpus that may be made available to other researchers. I understand that it may be possible for others to identify me by my voice, and I understand that no personal information will be connected to my voice apart from my age, my gender and how well I am trained in speaking and singing.

I understand that I may withdraw my participation in this experiment at any time.

I also understand that I may register any complaint I might have about the experiment with the researcher named above or with the Director of the School of Computing Science, Binay Bhattacharya, at 291-4277.

I may obtain copies of the results of this study, upon its completion, by contacting David Gerhard.

I have been informed that the results of this research may be published, and that my identity will be kept confidential.

I understand that my supervisor or employer may require me to obtain his or her permission prior to my participation in a study such as this.

I agree to participate by having my voice recorded as I respond to a series of prompts.

NAME (please type or print legibly):
ADDRESS:
SIGNATURE:
WITNESS:
DATE:

B.3 Information Sheet

COLLECTION AND ANNOTATION OF A SPEECH/ SONG CORPUS FOR RESEARCH INTO HUMAN UTTERANCE CLASSIFICATION
INFORMATION SHEET

This is a brief introduction to the research project. In describes what the research is for and what will happen in the experiments.

I am collecting a corpus, or a group of files, of people using their voice. I want to find out what the differences are between speaking and singing, and I want to find out what the middle-ground between speaking and singing looks like. There are two steps to this corpus collection.

Stage 1 is recording the sounds. I will be asking people like you to speak, or sing, in response to a series of prompts. The prompts are designed to get you to use your voice in the ”middle-ground” between speaking and singing. If youre not sure what this sounds like, you can practice a bit before I record your voice.

Stage 2 is finding out what you think of the sounds. I will call you back when I have collected all the sounds I need, and if you want, you can participate in this second step as well. I will be playing the sounds in the corpus for you, and you are invited to rate the sounds on a scale from speech-ness to song-ness. That way, I am not just going on my own opinion. There are other questions I will be asking as well, and you can look over the question sheet before you start, if you want.

It is important for you to understand that this corpus will be used for research I will listen to the sounds and I will program a computer to try to identify which sounds are speech and which sounds are singing, and which are somewhere in between.

Other researchers may want to use this corpus so that we are all working from the same set of data and we can all compare our results. Your personal information will in no way be associated with your voice, but if you do not want your voice to be heard by other scientists, please let me know. The samples are short, and for the most part everyone is saying the same thing, but if you are concerned about other scientists hearing your voice, you dont have to participate.

If you have any questions about the research, please feel free to ask. You may withdraw at any time and your data will be discarded.

B.4 Research Instrument: Collection

COLLECTION AND ANNOTATION OF A SPEECH/ SONG CORPUS FOR RESEARCH INTO HUMAN UTTERANCE CLASSIFICATION
STAGE 1: CORPUS COLLECTION

In this stage of the corpus collection and annotation project, you are asked to provide voice samples for the corpus. Samples should be limited to 5 seconds, if possible. Please read each prompt before you begin, and when recording, please use a natural voice, as if you were talking or singing to a friend. It might help to read and remember the prompt, then voice the sample without looking at the prompt. Feel free to practice any sample before your record it.

1. Please speak or sing anything you like for about 5 seconds.
2. Please sing the first line of your favourite song.
3. In a single short sentence, please tell me what you had for lunch yesterday.
4. Please speak the phrase "When you're worried, will you run away?"
5. Please sing the phrase "Row, row, row your boat, gently down the stream."
6. Please speak the phrase "Row, row, row your boat, gently down the stream."
7. Please sing the phrase "O Canada, our home and native land."
8. Please speak the phrase "O Canada, our home and native land."
9. In a single short sentence, please tell me what you did last weekend, using a voice which is somewhere between singing and speaking.
10. Please utter the phrase "Why is the sky blue?" using a voice which is somewhere between speaking and singing.
11. Please speak or sing anything you like for about 5 seconds.

B.5 Example Web Annotation Form

Collection and Annotation of a Speech/ Song Corpus for Research into Human Utterance Classification
Simon Fraser University
Principal Researcher: David Gerhard, dbg@cs.sfu.ca

Your Subject Number is 9999. If this is not correct, please stop the experiment and email the principal researcher

Part 1

This page contains a lot of data, and may take a few minutes to load, depending on your connection speed.

Please listen to and rate each sound on a scale between talking to singing. When you have listened to and rated all the files, click "Submit" to continue.

Section A of E

Sound file	Hear it	Rate it
n106:	🔊 ▶	Talking ○ ○ ○ ○ ○ Singing
n107:	🔊 ▶	Talking ○ ○ ○ ○ ○ Singing
n108:	🔊 ▶	Talking ○ ○ ○ ○ ○ Singing
n109:	🔊 ▶	Talking ○ ○ ○ ○ ○ Singing
n110:	🔊 ▶	Talking ○ ○ ○ ○ ○ Singing
⋮	⋮	⋮
n126:	🔊 ▶	Talking ○ ○ ○ ○ ○ Singing
n127:	🔊 ▶	Talking ○ ○ ○ ○ ○ Singing

Submit Reset

For more information, contact David Gerhard, dbg@cs.sfu.ca

Appendix C

Corpus Annotation Results

This appendix contains the data collected from the web annotation form for the corpus. Section C.1 contains the numerical results from Part 1 of the collection experiment. Section C.2 contains the numerical and written responses to Part 2 of the experiment, where subjects were asked to comment on aspects of speech and song regarding particular files. Section C.3 contains the written responses to Part 3, where subjects were asked to comment on their experiences of speech and song, as well as their experience of this experiment.

All quotes from subjects are presented verbatim without corrections in spelling, grammar or punctuation. Any words or phrases which could be used to identify the subjects have been removed.

The following abbreviations are used in this appendix:

N Number of ratings made for this file.

μ Mean of all ratings for this file.

σ Standard deviation of ratings for this file.

H Highest rating for this file.

L Lowest rating for this file.

C.1 Numerical Results: Part 1

C.1.1 Part 1.A

sbj : files n106–n127																						
211	2	3	3	4	2	4	2	4	2	4	4	3	2	4	4	2	2	2	4	4	5	2
212	2	2	1	1	3	1	1	2	2	1	1	3	1	5	5	3	4	1	4	4	5	3
213	3	2	2	2	1	1	1	3	1	4	1	3	2	5	5	3	2	1	3	3	5	1
220	1	2	1	3	2	2	1	2	2	2	2	1	2	4	4	1	3	1	4	3	5	1
221	1	2	2	4	2	4	1	1	2	4	2	3	1	5	4	1	2	1	5	5	5	3
222	2	1	2	1	1	2	1	2	1	1	1	1	2	3	1	1	1	1	3	3	3	2
223	1	2	2	4	1	2	1	2	1	1	1	1	1	4	2	1	1	1	4	2	4	1
231	1	2	1	3	1	2	1	1	1	3	1	1	1	5	4	1	2	1	4	3	4	1
232	2	3	3	1	2	3	1	1	1	1	2	2	1	4	2	2	2	1	3	3	4	1
236	2	3	2	5	1	2	1	3	1	2	1	1	1	5	5	2	3	1	5	4	5	1
238	2	1	1	4	3	2	2	2	1	3	1	1	1	5	5	3	2	1	4	4	5	2
242	3	3	3	4	1	4	2	4	3	4	3	3	4	5	5	3	4	2	5	5	5	3
243	4	2	4	3	2	1	1	3	1	1	1	1	1	5	5	1	2	1	4	4	5	1
246	3	1	4	2	1	1	1	1	1	4	1	1	1	5	3	1	2	2	5	5	3	1
249	3	2	2	4	1	1	1	3	1	2	1	1	1	5	1	1	1	1	5	3	5	1
251	1	2	2	4	2	1	1	1	1	1	1	1	1	3	3	1	3	1	4	2	5	1
253	2	1	2	1	1	2	1	1	1	2	1	1	2	4	2	1	1	1	2	2	4	1
254	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅	∅
308	1	2	1	3	2	1	1	1	1	1	2	1	1	4	3	1	2	1	3	3	5	2
309	1	2	1	2	2	1	1	1	1	2	1	1	1	5	4	2	1	1	4	3	5	1
310	2	2	2	3	1	2	2	1	1	2	2	1	1	4	3	1	2	1	3	3	4	2
311	2	1	2	3	1	1	1	1	1	1	1	1	1	5	3	1	1	1	4	3	3	1
312	1	1	1	2	1	1	1	1	1	2	1	1	1	5	3	1	1	1	4	4	4	2
314	2	2	1	2	2	1	1	1	2	1	1	2	2	4	3	2	2	1	3	3	4	2
316	2	3	2	4	2	2	2	1	2	2	2	2	2	5	5	2	2	2	4	4	5	2
317	2	1	2	3	2	1	1	1	1	1	1	1	1	3	2	1	2	1	3	4	3	1
320	1	2	1	2	1	2	1	1	1	2	1	1	1	4	2	1	1	1	4	3	5	1
324	4	2	3	3	1	1	1	1	1	1	2	1	1	4	5	1	1	1	5	4	4	2
325	2	3	2	4	2	3	2	1	2	3	3	1	1	4	3	2	2	1	4	4	5	1

328	2	1	3	1	1	1	1	1	2	2	1	1	2	4	3	1	2	1	4	2	3	1
329	1	1	2	3	1	1	2	1	1	3	1	1	1	4	4	2	3	1	4	3	4	1
330	2	1	2	1	1	1	1	1	1	1	1	2	1	5	4	1	2	1	3	2	4	1
333	1	2	1	4	2	3	1	1	1	1	1	1	1	4	3	1	1	1	4	2	5	1
335	2	2	2	4	2	2	1	1	2	4	2	2	1	5	5	1	1	1	3	3	5	1
340	1	1	2	1	1	2	1	2	1	1	1	1	2	4	2	1	1	1	3	3	3	1
343	1	1	2	3	1	1	1	1	1	1	1	1	1	3	3	1	1	1	4	4	4	4
346	2	2	2	4	1	1	1	1	1	1	1	2	1	4	2	1	2	1	3	4	4	1
347	2	1	3	4	1	2	1	1	2	3	1	1	1	3	4	1	4	2	2	4	4	3
348	1	2	4	4	3	1	1	1	2	1	1	1	2	5	4	1	1	2	4	3	5	3
349	2	1	3	3	2	2	1	1	1	3	2	1	1	4	3	1	1	1	4	3	5	1
352	3	4	3	4	1	2	1	2	1	1	1	1	1	5	5	1	2	1	4	2	3	1
353	3	2	3	2	2	1	1	1	1	2	2	1	2	4	3	2	1	1	3	3	3	2
354	1	1	3	3	1	3	1	1	1	2	2	1	1	4	1	1	1	1	4	3	4	1
357	2	2	2	2	2	2	2	1	2	1	2	2	2	4	3	1	3	2	4	4	3	2
358	3	2	4	3	1	2	1	1	2	1	1	2	1	5	4	1	3	1	4	2	4	1
359	1	1	1	2	3	1	1	1	1	1	1	1	1	3	4	1	1	1	4	2	4	1
360	2	1	1	3	1	1	1	2	1	1	1	2	1	4	4	1	1	1	3	3	3	1
N	46	46	46	46	46	46	46	46	46	46	46	46	46	46	46	46	46	46	46	46	46	46
μ	1.9	1.8	2.1	2.9	1.5	1.7	1.2	1.5	1.3	1.9	1.4	1.4	1.3	4.3	3.4	1.4	1.8	1.2	3.8	3.2	4.2	1.5
σ	0.8	0.7	0.9	1.1	0.7	0.9	0.4	0.8	0.5	1.1	0.7	0.7	0.6	0.7	1.2	0.6	0.9	0.4	0.7	0.8	0.8	0.8
H	4	4	4	5	3	4	2	4	3	4	4	3	4	5	5	3	4	2	5	5	5	4
L	1	1	1	1	1	1	1	1	1	1	1	1	1	3	1	1	1	1	2	2	3	1

C.1.2 Part 1.B

sbj : files u115–u137																							
211	1	4	2	2	2	5	3	2	5	2	3	2	4	2	2	3	4	2	5	2	2	3	2
212	1	3	1	1	2	5	2	1	5	1	1	2	1	2	1	1	1	1	5	1	1	1	1
213	2	1	2	2	2	5	4	2	5	1	1	4	3	2	1	4	3	5	5	2	4	2	4
220	2	3	2	2	1	5	3	2	5	2	2	2	2	2	2	3	1	4	5	2	2	2	2
221	2	3	1	2	1	5	4	3	5	1	1	2	3	2	1	4	4	5	5	2	3	1	1
222	1	3	2	2	1	4	1	1	5	1	1	1	2	2	1	2	2	3	5	2	2	1	2
223	1	3	2	1	1	5	1	2	5	2	2	1	1	2	1	3	2	5	5	1	4	2	2
231	1	2	1	1	1	5	2	2	5	2	1	2	3	2	1	3	2	2	5	2	1	2	2
232	3	4	3	2	2	4	2	2	5	2	2	3	3	2	2	4	3	4	5	2	4	3	3
236	3	2	3	2	1	4	3	1	5	1	2	4	2	5	3	5	4	5	5	1	3	2	2
238	2	4	2	1	1	5	1	1	5	1	1	2	2	3	3	5	4	5	5	2	4	2	2
242	4	5	4	4	3	5	5	5	5	∅	4	5	5	5	5	5	5	3	5	5	4	2	1
243	1	3	1	1	1	5	1	2	5	1	3	2	2	2	1	2	2	2	5	1	1	1	2
246	1	2	1	1	2	5	1	2	5	1	1	2	3	2	1	2	2	5	5	2	1	1	1
249	1	1	1	1	1	5	1	1	5	1	1	1	4	1	1	5	3	5	5	1	1	1	1
251	2	2	1	1	2	5	1	1	3	1	1	3	1	2	1	4	3	3	5	2	2	2	2
253	1	2	2	1	1	5	2	1	4	2	1	2	1	2	1	2	2	2	5	2	2	2	1
254	2	3	4	3	2	5	4	3	4	1	1	2	3	4	2	4	5	5	5	2	3	3	2
308	1	3	1	1	2	5	3	2	5	1	1	3	3	2	1	2	1	∅	5	3	1	1	1
309	1	1	1	1	1	4	1	1	4	1	1	2	3	1	1	3	2	4	5	1	2	1	1
310	1	2	1	1	2	5	2	1	5	1	1	1	3	2	1	3	2	4	5	2	1	2	2
311	1	2	1	1	1	5	2	2	5	1	1	3	3	2	1	3	3	5	5	1	3	1	2
312	1	2	1	1	1	5	1	1	4	1	1	1	2	1	1	3	1	2	5	1	1	1	1
314	2	3	3	2	1	5	3	2	4	2	1	1	3	3	2	3	2	5	5	2	2	2	1
316	2	2	2	2	2	4	2	2	4	2	2	3	2	2	2	3	3	3	5	2	3	2	2
317	1	3	1	1	3	5	1	1	4	1	1	3	1	1	1	3	2	4	5	1	2	3	1
320	1	1	1	1	1	5	1	1	5	1	1	1	3	1	1	2	2	3	5	1	1	1	1
324	2	2	3	1	2	5	4	2	5	2	2	1	2	4	1	3	4	5	5	2	2	1	1
325	2	4	3	2	1	4	3	2	5	2	2	4	3	3	2	3	4	5	5	2	3	3	2
328	1	2	1	1	1	5	2	2	3	1	1	3	2	2	1	3	2	4	5	1	2	1	1

329	2	3	2	2	1	4	2	2	4	2	1	1	2	2	2	2	2	2	5	2	1	2	3
330	2	2	1	1	1	5	1	2	4	1	2	1	1	2	1	1	3	1	5	1	1	1	1
333	1	1	1	1	1	5	1	1	5	1	1	2	1	1	1	3	2	4	5	1	1	1	1
335	2	2	2	2	4	5	2	1	5	1	2	4	5	1	1	5	3	5	5	1	3	2	1
340	1	1	1	1	1	5	2	1	4	1	1	2	1	1	1	2	3	2	5	1	4	1	1
343	1	2	2	2	1	5	3	2	5	1	2	2	3	3	1	3	3	3	5	2	2	1	1
346	1	3	1	1	2	4	2	1	4	1	1	2	3	3	1	2	2	1	5	2	3	2	1
347	2	4	3	1	1	5	3	3	4	1	1	3	3	4	3	4	3	2	5	3	3	4	4
348	1	3	2	1	1	5	3	1	5	2	3	2	4	4	2	4	3	5	5	1	2	3	1
349	1	3	1	1	2	4	3	1	4	1	2	2	2	2	1	4	1	2	5	2	3	1	2
352	1	2	1	1	1	5	1	1	5	1	1	3	1	2	1	2	2	4	5	1	2	2	1
353	2	1	1	1	1	4	2	1	3	2	2	2	2	2	2	4	3	3	5	2	2	2	2
354	1	3	1	1	1	4	2	∅	4	1	1	2	1	1	1	2	3	4	5	1	3	1	∅
357	3	4	3	2	2	4	3	3	4	2	1	3	3	3	1	2	3	3	5	3	3	2	2
358	2	3	1	1	2	5	2	1	4	1	3	4	2	2	1	2	3	4	5	2	1	2	1
359	1	2	1	1	1	5	1	1	5	1	1	1	1	1	1	1	2	1	5	2	2	1	1
360	1	1	1	2	1	5	1	1	4	1	1	2	2	2	1	1	2	2	5	1	3	3	2
N	47	47	47	47	47	47	47	46	47	46	47	47	47	47	47	47	47	46	47	47	47	47	46
μ	1.5	2.5	1.7	1.4	1.5	4.7	2.1	1.7	4.5	1.3	1.5	2.3	2.4	2.2	1.4	3	2.6	3.4	5	1.7	2.3	1.8	1.6
σ	0.7	1	0.9	0.7	0.7	0.4	1.1	0.8	0.6	0.5	0.7	1	1.1	1	0.8	1.1	1	1.4	0	0.8	1	0.8	0.8
H	4	5	4	4	4	5	5	5	5	2	4	5	5	5	5	5	5	5	5	5	4	4	4
L	1	1	1	1	1	4	1	1	3	1	1	1	1	1	1	1	1	1	5	1	1	1	1

C.1.3 Part 1.C

sbj : files u138–u159																						
211	4	3	3	2	4	5	2	2	2	3	4	4	4	3	2	3	5	4	5	5	5	4
212	3	2	1	1	1	4	2	1	1	3	1	3	1	2	2	1	1	3	5	5	5	4
213	4	4	3	1	1	1	1	2	5	1	3	2	1	1	1	1	5	5	5	5	5	5
220	2	2	2	2	2	4	2	1	4	3	3	2	1	2	1	2	3	3	4	5	5	5
221	4	2	1	1	1	4	2	3	4	5	5	2	1	2	2	1	4	5	5	5	5	5
222	3	2	2	2	3	2	3	2	2	3	2	2	1	3	1	1	1	3	5	5	5	3
223	3	1	1	1	1	4	1	1	3	2	3	1	1	1	1	1	1	4	5	5	5	5
231	3	2	3	1	1	4	1	2	3	3	2	1	2	3	1	1	3	4	4	5	5	5
232	3	3	2	2	2	3	2	1	2	3	3	3	2	3	3	2	4	4	3	5	5	2
236	5	3	3	4	3	5	3	1	2	2	3	2	3	4	2	3	4	4	5	5	5	2
238	5	3	3	2	4	5	5	1	1	4	1	1	3	4	3	2	5	5	5	5	5	5
242	∅	5	5	4	3	5	4	1	1	3	1	3	1	1	1	1	4	3	3	5	5	4
243	2	4	2	1	2	5	1	1	4	3	4	3	2	3	1	1	2	4	5	5	5	5
246	2	1	1	2	2	4	2	3	4	4	4	2	2	3	2	1	2	5	5	5	5	∅
249	4	5	1	1	3	3	1	1	5	5	5	1	1	1	1	1	2	5	5	5	5	5
251	4	2	2	2	1	4	2	1	2	2	3	2	2	2	3	1	3	3	5	5	5	4
253	3	3	2	2	2	3	2	1	1	3	2	1	1	2	1	2	1	2	4	5	5	3
254	3	4	3	2	3	4	3	1	1	3	4	1	2	3	1	1	2	4	5	5	5	4
308	3	3	1	1	2	4	2	1	4	4	4	2	1	4	1	1	2	5	5	5	5	2
309	2	3	2	2	1	4	1	3	4	3	4	2	2	2	3	1	3	4	4	5	5	5
310	3	2	2	2	2	3	3	1	3	2	2	1	2	3	1	2	2	3	4	5	5	4
311	3	2	1	1	2	4	1	1	1	2	2	2	1	1	1	1	3	2	5	5	5	5
312	1	1	1	1	1	4	1	1	1	2	1	1	1	3	1	1	4	5	5	5	5	5
314	3	2	2	2	2	4	2	2	3	3	4	2	2	3	3	2	4	5	5	5	5	5
316	3	3	2	2	3	4	2	2	3	3	3	3	2	2	2	2	4	4	4	5	5	4
317	4	3	2	1	1	2	1	1	1	3	3	1	1	2	1	1	3	1	4	5	4	4
320	2	2	2	1	1	5	1	1	1	1	1	1	1	2	1	1	4	5	5	5	5	5
324	4	2	2	2	2	4	4	2	3	3	3	4	2	4	1	1	3	4	5	5	5	5
325	4	3	2	2	2	4	4	2	3	4	3	2	3	4	1	2	3	3	4	5	5	4
328	2	3	3	1	1	2	1	1	3	3	4	1	1	2	2	1	2	3	5	5	4	4

329	4	2	2	1	2	3	3	1	3	3	2	2	2	2	1	2	5	4	5	5	5	3
330	2	2	1	1	1	2	1	1	2	3	2	1	1	3	1	2	1	3	3	5	5	3
333	1	2	2	1	1	2	1	1	2	1	1	1	1	1	1	1	5	4	4	5	5	3
335	2	1	1	1	1	5	1	1	5	3	3	1	2	5	1	1	5	5	5	5	5	5
340	3	4	3	1	1	4	2	2	2	3	1	2	1	1	2	1	4	3	4	4	4	3
343	3	2	2	2	2	3	2	1	2	3	3	2	1	3	1	1	1	4	4	5	5	5
346	2	2	1	1	3	4	3	1	3	3	2	1	2	3	1	1	4	3	5	5	5	4
347	4	2	2	2	2	3	3	1	2	4	2	2	3	3	1	2	4	2	4	4	5	3
348	4	3	4	1	1	5	1	2	4	2	4	2	2	3	2	1	4	5	4	4	5	5
349	3	2	3	3	1	3	3	1	1	2	2	1	1	1	1	2	4	3	4	5	4	4
352	3	2	2	1	1	4	2	1	2	3	2	2	1	2	2	1	5	4	4	5	5	5
353	3	2	2	1	2	4	2	2	2	2	2	2	2	2	1	2	2	3	4	4	5	4
354	3	3	2	1	1	1	1	1	2	3	2	1	1	2	1	1	2	1	5	5	5	5
357	3	2	2	2	2	3	2	2	4	3	4	2	2	3	1	2	4	4	4	5	5	5
358	3	2	2	2	3	4	2	1	3	3	3	1	2	4	1	1	4	4	5	5	5	4
359	3	2	1	1	1	4	1	1	2	2	2	1	1	2	2	1	4	5	5	5	5	4
360	3	3	2	1	2	4	2	1	1	2	1	1	1	1	1	1	1	3	5	5	4	4
N	46	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	46
μ	3	2.5	2	1.6	1.8	3.6	2	1.4	2.5	2.8	2.7	1.8	1.6	2.5	1.4	1.4	3.1	3.7	4.5	4.9	4.9	4.2
σ	0.9	0.9	0.9	0.7	0.9	1	1	0.6	1.2	0.9	1.1	0.8	0.7	1	0.7	0.6	1.3	1.1	0.6	0.3	0.3	0.9
H	5	5	5	4	4	5	5	3	5	5	5	4	4	5	3	3	5	5	5	5	5	5
L	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	4	4	2

C.1.4 Part 1.D

sbj :	u224	u225	u226	f216	f217	f218	g244	g245	g246	h213	h214	h215
211	2	1	1	1	3	2	5	2	5	1	2	2
212	1	2	1	1	2	1	5	3	5	1	2	1
213	1	2	1	1	3	1	5	4	5	1	1	1
220	1	1	1	1	2	1	5	3	5	1	2	2
221	1	1	1	2	3	2	5	3	5	1	1	1
222	1	2	1	1	2	1	5	1	5	1	1	1
223	1	1	1	1	1	1	5	1	5	1	1	1
231	1	2	1	1	2	1	5	2	5	1	2	2
232	2	3	1	1	4	3	5	3	5	1	2	1
236	1	2	1	1	1	2	5	3	5	1	2	2
238	2	3	2	2	2	1	5	3	5	1	1	1
242	1	1	1	2	3	1	5	4	5	4	5	2
243	1	3	1	1	2	1	5	2	5	1	1	1
246	1	2	1	2	3	3	5	3	5	1	2	1
249	1	1	1	1	3	1	5	3	5	1	1	1
251	2	2	1	1	3	2	5	3	4	1	1	1
253	1	1	1	1	1	1	5	3	5	1	2	1
254	1	2	1	3	4	3	5	3	5	1	1	1
308	2	1	1	1	1	1	4	1	5	1	1	1
309	1	2	1	1	2	1	5	2	5	1	1	1
310	1	2	1	1	2	1	5	2	5	1	2	2
311	1	1	1	1	2	1	5	2	5	1	2	1
312	1	1	1	1	1	1	5	1	5	1	1	1
314	1	2	1	1	2	1	5	2	5	1	2	1
316	2	2	2	2	2	2	5	3	5	2	3	2
317	1	1	1	2	2	1	5	2	5	1	1	1
320	1	3	1	1	1	1	5	2	5	1	1	1
324	1	2	1	1	2	1	5	2	5	1	2	1
325	1	2	1	1	2	1	4	3	4	2	1	1
328	1	2	1	1	2	2	5	2	5	1	2	1

329	2	3	2	2	2	1	5	2	5	1	1	1
330	1	1	1	1	2	1	5	2	5	1	1	1
333	1	1	1	1	1	1	4	1	5	1	1	1
335	1	1	1	1	3	1	5	1	5	1	1	1
340	1	1	1	3	4	3	5	4	5	2	2	2
343	1	3	1	1	2	1	5	3	5	1	1	1
346	1	1	1	1	1	1	5	1	5	1	1	1
347	2	3	2	2	3	2	5	3	4	1	2	2
348	1	2	1	1	2	1	5	1	5	1	1	1
349	1	1	1	1	3	1	5	3	5	1	1	1
352	1	1	1	1	1	1	5	1	5	1	1	1
353	1	1	1	1	2	2	3	2	3	1	1	1
354	1	2	1	2	3	2	5	2	5	2	2	2
357	1	2	1	1	2	1	4	3	4	1	2	2
358	1	2	1	1	3	1	5	1	5	1	1	1
359	1	1	1	2	2	2	5	2	5	1	1	1
360	1	1	1	1	2	1	5	2	5	1	1	1
N	47	47	47	47	47	47	47	47	47	47	47	47
μ	1.2	1.7	1.1	1.3	2.2	1.4	4.9	2.3	4.9	1.1	1.5	1.2
σ	0.4	0.7	0.3	0.5	0.8	0.6	0.4	0.9	0.4	0.5	0.7	0.4
H	2	3	2	3	4	3	5	4	5	4	5	2
L	1	1	1	1	1	1	3	1	3	1	1	1

C.1.5 Part 1.E

sbj	files i211–i220									
212	1	3	2	4	1	4	2	2	1	1
220	2	3	3	3	2	4	2	2	2	1
222	2	1	2	2	1	4	1	1	2	1
325	3	3	3	3	2	3	2	1	2	1
329	4	3	2	1	1	4	3	2	2	1
333	1	1	1	2	1	3	1	1	2	1
335	2	1	1	2	1	5	1	1	1	1
N	7	7	7	7	7	7	7	7	7	7
μ	2.1	2.1	2	2.4	1.3	3.9	1.7	1.4	1.7	1
σ	1.1	1.1	0.8	1	0.5	0.7	0.8	0.5	0.5	0
H	4	3	3	4	2	5	3	2	2	1
L	1	1	1	1	1	3	1	1	1	1

sbj	files i221–i230									
232	4	5	4	2	1	1	3	4	2	2
236	1	4	4	3	2	2	1	4	4	5
238	1	5	5	3	1	1	2	5	5	5
242	2	5	5	5	1	1	1	5	5	4
308	1	2	1	2	1	1	1	2	2	1
310	1	3	4	2	1	1	1	3	3	3
312	1	5	5	2	1	1	1	4	4	3
314	1	5	5	4	1	1	2	4	3	4
343	1	5	5	2	1	1	2	4	4	4
347	2	3	4	4	3	3	2	4	4	5
349	2	3	4	3	1	1	2	3	3	3
353	1	3	3	3	1	1	1	2	2	2
N	12	12	12	12	12	12	12	12	12	12
μ	1.5	4	4.1	2.9	1.3	1.3	1.6	3.7	3.4	3.4
σ	0.9	1.1	1.2	1	0.6	0.6	0.7	1	1.1	1.3
H	4	5	5	5	3	3	3	5	5	5
L	1	2	1	2	1	1	1	2	2	1

sbj	files i231–i240									
211	4	2	2	1	4	2	3	2	4	2
213	4	1	1	1	4	2	1	3	3	2
221	3	1	1	2	5	1	2	1	4	2
223	2	1	1	1	4	1	1	2	3	1
246	2	2	2	2	5	1	1	1	3	2
324	4	2	3	2	5	1	2	2	4	2
328	2	1	1	1	3	1	1	2	2	1
330	3	2	1	1	5	1	2	1	2	1
357	3	1	1	3	4	1	2	2	3	3
359	3	1	1	1	4	1	1	1	1	1
N	10	10	10	10	10	10	10	10	10	10
μ	3	1.4	1.4	1.5	4.3	1.2	1.6	1.7	2.9	1.7
σ	0.8	0.5	0.7	0.7	0.7	0.4	0.7	0.7	1	0.7
H	4	2	3	3	5	2	3	3	4	3
L	2	1	1	1	3	1	1	1	1	1

sbj	files i241–i250									
231	1	3	1	2	1	3	1	1	1	1
309	1	2	2	4	1	4	3	3	1	1
340	1	4	1	2	1	4	1	3	1	1
346	1	2	1	2	1	4	2	1	1	1
N	4	4	4	4	4	4	4	4	4	4
μ	1	2.8	1.3	2.5	1	3.8	1.8	2	1	1
σ	0	1	0.5	1	0	0.5	1	1.2	0	0
H	1	4	2	4	1	4	3	3	1	1
L	1	2	1	2	1	3	1	1	1	1

sbj	files i251–i260									
243	1	4	1	1	1	3	1	3	4	1
249	1	5	1	1	1	1	1	1	1	1
251	2	4	1	∅	∅	1	2	3	3	2
311	1	4	1	1	1	1	1	3	1	1
317	1	3	1	1	1	1	1	2	1	1
348	1	4	1	1	1	3	1	4	2	2
352	1	5	1	1	1	1	1	2	1	1
354	1	3	1	1	1	1	2	3	2	1
358	1	4	1	2	1	1	1	3	1	1
360	1	4	1	1	1	1	1	2	1	1
N	10	10	10	9	9	10	10	10	10	10
μ	1.1	4	1	1.1	1	1.4	1.2	2.6	1.7	1.2
σ	0.3	0.7	0	0.3	0	0.8	0.4	0.8	1.1	0.4
H	2	5	1	2	1	3	2	4	4	2
L	1	3	1	1	1	1	1	1	1	1

sbj	files j211–j220									
212	1	1	1	3	2	2	1	1	1	1
220	2	2	1	3	2	3	∅	2	∅	2
243	1	1	1	3	3	2	1	1	1	1
310	2	3	2	4	3	4	2	2	1	2
312	1	1	1	3	1	5	1	1	1	1
314	1	2	1	2	2	4	2	2	1	1
333	1	1	1	4	3	4	1	2	∅	1
335	1	1	1	4	2	5	1	1	1	1
358	1	2	1	4	2	4	2	2	1	1
360	1	1	1	2	1	4	1	1	1	1
N	10	10	10	10	10	10	9	10	8	10
μ	1.2	1.5	1.1	3.2	2.1	3.7	1.3	1.5	1	1.2
σ	0.4	0.7	0.3	0.8	0.7	1.1	0.5	0.5	0	0.4
H	2	3	2	4	3	5	2	2	1	2
L	1	1	1	2	1	2	1	1	1	1

sbj	files j221–j230									
222	1	3	3	2	1	2	1	2	4	3
249	1	3	5	1	1	1	1	4	5	5
251	1	2	3	2	2	1	1	2	5	4
324	1	3	5	2	2	1	1	4	5	5
343	1	3	5	3	2	1	1	4	5	5
347	2	3	4	1	4	3	2	3	4	4
349	1	2	5	3	1	1	1	2	4	3
N	7	7	7	7	7	7	7	7	7	7
μ	1.1	2.7	4.3	2	1.9	1.4	1.1	3	4.6	4.1
σ	0.4	0.5	1	0.8	1.1	0.8	0.4	1	0.5	0.9
H	2	3	5	3	4	3	2	4	5	5
L	1	2	3	1	1	1	1	2	4	3

sbj	files j231–j240									
211	3	1	2	1	4	1	4	3	2	2
213	5	1	1	2	4	1	3	3	2	1
232	3	1	1	3	4	2	3	3	3	4
236	4	1	2	2	4	2	4	3	2	2
238	5	1	1	2	4	1	5	3	2	2
309	5	1	1	1	5	1	4	3	1	1
311	4	1	1	1	4	1	5	3	1	1
328	2	1	1	1	4	2	5	1	1	1
330	2	1	1	1	4	1	2	2	1	
353	2	1	1	1	3	1	3	2	1	1
357	3	2	1	3	3	2	4	3	2	3
359	3	1	1	1	2	1	3	2	1	1
N	12	12	12	12	12	12	12	12	12	11
μ	3.4	1.1	1.2	1.6	3.8	1.3	3.8	2.6	1.6	1.7
σ	1.2	0.3	0.4	0.8	0.8	0.5	1	0.7	0.7	1
H	5	2	2	3	5	2	5	3	3	4
L	2	1	1	1	2	1	2	1	1	1

sbj	files j241–j250									
221	2	2	1	4	2	1	4	2	1	1
223	1	1	1	2	1	1	1	2	1	1
242	4	5	4	5	2	1	2	3	3	1
246	1	3	2	4	2	1	4	2	2	1
317	1	2	1	2	1	1	2	2	2	1
340	2	1	2	1	1	2	2	1	1	1
346	1	2	1	4	1	1	3	1	1	1
N	7	7	7	7	7	7	7	7	7	7
μ	1.7	2.3	1.7	3.1	1.4	1.1	2.6	1.9	1.6	1
σ	1.1	1.4	1.1	1.5	0.5	0.4	1.1	0.7	0.8	0
H	4	5	4	5	2	2	4	3	3	1
L	1	1	1	1	1	1	1	1	1	1

sbj	files j251–j260									
231	2	4	3	1	1	1	1	2	4	1
308	1	1	1	1	1	1	1	1	4	1
325	1	2	2	1	1	1	1	1	3	2
329	2	3	2	1	2	2	3	2	4	2
348	1	5	2	1	1	2	1	1	3	2
352	1	4	2	1	1	1	1	1	5	1
354	1	3	1	1	1	1	1	1	3	1
N	7	7	7	7	7	7	7	7	7	7
μ	1.3	3.1	1.9	1	1.1	1.3	1.3	1.3	3.7	1.4
σ	0.5	1.3	0.7	0	0.4	0.5	0.8	0.5	0.8	0.5
H	2	5	3	1	2	2	3	2	5	2
L	1	1	1	1	1	1	1	1	3	1

C.2 Numerical and Written Results: Part 2

The following questions were asked on the web form:

1. Rate the file: Talking ○ ○ ○ ○ ○ Singing
2. What is it about the sound that leads you to this judgement?
3. What could the speaker have done to make this sample more speech-like?
4. What could the speaker have done to make this sample more song-like?

The responses are presented here in the following manner:
subject number (rating) response

C.2.1 Rating and Question 1

file = n128; N = 24; μ= 1.46, σ= 0.66; H = 3; L = 1

211 (2) rythm, no tone change

213 (1) sounds like a new reporter

212 (1) Sounds like a tougue-twister. No rhythm and words unclear.

220 (1) not much intonation and little tone change

236 (3) the rhythmic nature

238 (1) There is no tone or hint of melody.

242 (1) No lilt to the words, monotone

243 (1) The last accent

251 (2) there doesn't seem to be a tune

308 (2) There's a rhythm, but no melody

309 (1) No distinct rythym, monotonic

310 (2) It sounds like speaking, but there is a bit of change in the tone

311 (2) There is a bit of rhythm, but not the tones don't vary enough to be singing.

312 (1) Short concise duration of each word.

317 (2) there's no melody

324 (1) pretty flat, no rhythm/melody

328 (1) it is almost monotone, no variation

335 (1) Flat spectrum and rate of speech

343 (1) a more connected lang., not English which would be more broken

346 (1) Sounds like spoken French

349 (1) The words are said as spoken rather than with irregular empasis on different syllables

353 (2) It has a slight rhythm

357 (3) The connectedness (sustained quality) of the utterance

358 (1) no melody, no rhythm

file = n129; N = 24; μ= 4.50, σ= 0.59; H = 5; L = 3

211 (5) rythm, tone change

213 (4) rhythmic jazz

212 (5) The rhythm.

220 (3) it's not really singing but it's rhythmic

236 (4) use of rhythm

238 (5) It has a rhythm and a beat.

242 (5) Beat to speech

243 (5) The way he pronouned the sound.

251 (5) sounds very rhythmis, kind of like a precussion line

308 (5) rhythm, melody

309 (4) Extremely rythmic, unlikely to be coincidental

310 (4) change in tone and rhythm

311 (4) lots of rhythm and a beat

312 (4) Rythmic.

317 (5) no words

324 (5) presence of rhythm, no talk

328 (5) It has a rythum to it

335 (5) Rhythm

343 (5) connected, rap-like, another culture's singing?

346 (4) Rhythmic in a musical way

349 (4) it has rhythm and uses the vocal cords as not normally used in speaking

353 (4) the beat

357 (4) Primarily the rhythmic quality, along with the lack of sustained vowel sounds, as well as the quality of the final syllable

358 (5) said nothing, hence no speech. Beat, rhythm, more like music, so I call it "singing"

---

file = n130; N = 22; μ= 1.95, σ= 0.65; H = 3; L = 1

211 (3) poetry, rythm, few tone changes

213 (2) spoken in verse

212 (2) Funny tone for each word.

220 (2) it is definitly speaking but there is emphasis and a lilt to it

236 (2) it seemed like expressive talking

238 (3) It definitely had a rhythm.

243 (2) The beginning of the second sentence

251 (2) sounds like a Shakepearian play

308 (1) changes in pitch were slight and came at expected points in a sentence

309 (1) Sounds like someone reading from a book

310 (1) Sounds like someone is reading

311 (2) no beat, sing-song rhythm like in a chant

312 (1) No musical sounds.

317 poetry is speech, not song

324 (3) continuity and intonation

328 (3) the fine line between reading poetry and singing poetry

335 (2) Slightly poetic

343 (1) sounds like reading a story

346 (2) Sounds like a poem being recited

349 (2) it has unnatural emphasis creating rhythm but not a tune

353 (2) he is talking but not monotone

357 (2) The exaggerated width of the intonation range

358 (2) overly “colored” speech, some melodic (exaggerated) prosody at the end

---

file = n131; N = 24; μ= 1.54, σ= 0.83; H = 4; L = 1

211 (4) poetry,

213 (2) spoken in verse

212 (3) Emotion, feelings and varying speed.

220 (2) it is speaking but the rhythym of poetry lends itself to certain sing-like qualities

236 (1) sounded like talking

238 (1) Sounds like a recital of a poem

242 (1) More like a recitation

243 (1) The accent and cotent

251 (1) still sounds like peotry, but words enounciated well

308 (1) even tones throughout, no melody

309 (1) Again, sounds like reading from a book

310 (1) Sounds like reading, not variation in tone

311 (1) the words are short and clear and spoken in the same tone of voice

312 (1) Straight speech—not music.

317 (1) poetry is speech, not song

324 (2) continuity and little intonation

328 (3) the fine line between reading poetry and singing poetry

335 (1) Pauses between words

343 (1) reading

346 (2) Poetic speech

349 (1) it’ show you would speak the words

353 (1) she is talking

357 (2) Rhythmic patterning, as well as the stylized rising antecedent, and falling consequent phrase intonation, and finally the exaggerated slowing down of the tempo toward the end of the phrase.

358 (2) it could be 100% talking but there's an exaggerated metric pattern that pushed my judgement a tick towards song

---

file = n132; N = 24; μ= 4.38, σ= 0.65; H = 5; L = 3

211 (4) tone change, sustaining notes

213 (5) notes associated with words

212 (5) Great variance in pitch. There's rhythm.

220 (4) There are tone changes but I wouldn't say it is exactly singing except perhaps at the end of the sample where the subject held the syllable and changed notes

236 (5) the melody, various pitches used and rhythm

238 (5) it was a song

242 (5) Varied pitch, lilt

243 (4) The tone

251 (3) that was just painful,not singing, but not speech

308 (4) rhythm and a kind of melody

309 (4) Clearly not normal speech, widely varied tones

310 (5) sounds like a song because there is varied tone

311 (5) words are stretched out and the tones go up and down

312 (4) Musical tones.

317 (4) close to siging but lacks tasteful note arrangement

324 (5) continuity, melody, long vowels

328 (5) holds to a musical scale

335 (5) Varied tones and running words together

343 (5) bad singing, but song-like

346 (4) This is sing-song speech

349 (4) it had tune but the voice wasn't used in a singing style ie like opera

353 (4) He can't sing

357 (3) The aberrant quality of the intonation. It seems undirected, and inconsi stent from beginning to end.

358 (4) it's almost pure singing, but the melody sounds too fake and contorted

---

file = n133; N = 24; μ= 3.58, σ= 0.88; H = 5; L = 2

211 (4) rythm

213 (3) sounds like butterfly

212 (3) There's rhythm but not enough difference in pitch.

220 (3) rhythmic and tonal

236 (5) recognizable song

238 (4) there's a rhythm and a beat

242 (5) Beat

243 (3) The tone and cotent

251 (2) words too detached for song, too mashed together for speech

308 (4) rhythm (past experience suggests that it will move into a melody)

309 (4) I know it's a song. Also, the sound is clearly rythmic

310 (3) no variation in tone but there is rhythm

311 (4) there is a beat

312 (3) Very rhythmic but not melodic.

317 (2) sounds more like poetry recital than someone singing

324 (5) rhythm, melody, long vowels

328 (5)

335 (4) Familiar lyric and rhythm in voice

343 (4) some song quality, rap-like, rhythmic

346 (4) Sounds like a "pitch-less" or "spoken" song

349 (3) repeated words are songlike but lyrics are spoken

353 (3) she's not really singing

357 (3) The repetitions and almost sustained quality

358 (3) not sung, because the pitch doesn't quite fall on regular intervals, and the beat is off, but definitelly in the direction away from talking and into singing

---

file = n134; N = 23; μ= 1.70, σ= 0.56; H = 3; L = 1

211 (2) inflection

213 (2) speech with inflection to express heightened emotion

212 (2) Regular speed but abnormally high pitch at the end.

220 (2) it was talking but the emphasis produced tone changes

236 (3) high pitch and expressive talking

238 (2) it's an emphatic statement

243 (1) tone, cotent

251 (2) talking, but with a lot of change in pitch

308 (1) pitch suggests excitement, but not musical intentionality

309 (1) Sounds like someone speaking normally

310 (2) sounds like she's talking

311 (2) words are short and clear

312 (1) Not melodic or rhythmic.

317 (1) perso sounds excited

324 (2) sounds more like an interjection

328 (2)

335 (1) Very brief

343 (1) excited, happy, therefore raised pitch

346 (1) Sounds like excited speech

349 (2) i don't know

353 (2) her ptich

357 (2) Emphasis seems to follow meaning, but is exaggerated in pitch

358 (2) it's not signing, but no plain talking has so wide pitch excursions, well except maybe motherese

---

file = u160; N = 23; μ= 1.57, σ= 0.84; H = 4; L = 1

211 (2) rythm

213 (1) the crying doesn't add anything to the singing

212 (2) Talking with an abnormal tone and strange pauses between words.

220 (2) it was talking, but with enough empahsis to seem a wee bit song-like

236 (4) rhythm and various pitches

238 (2)

242 (3) More a 'noise' than speech

243 (1) tone, content

251 (1) words enounciated in certain spots

308 (1) pitch changes consistent with normal speaking by older person (what's that about?)

309 (3) I can't tell if it is someone struggling to speak or to sing

310 (1) little change in tone

311 (2) words are mostly clear and short, but there is a bit of a rhythm

312 (1) No musical notes.

317 (1) person talking

324 (1) lack of fluidity

328 (1)

335 (1) Pauses

343 (1) telling something, sounds scared

346 (1) Sounds like an old person with a speech impairment

349 not a clue what they are saying!

353 (1) no comment

357 (2) The unstable quality of the sound indicates either weeping, or some physiological problem

358 (1) broken talking, but not towards singing

---

file = u161; N = 24; μ= 1.83, σ= 0.70; H = 3; L = 1

211 (2) rythm, inflection

213 (2) sounds rythmic, like poetry

212 (1) Regular tone in a question statement.

220 (3) rhythmic delivery and intonation

236 (3) different pitches used and the common rhythm of speach

238 (3)

242 (1) No lilt to words

243 (1)

251 (2) more variation in pitch and slurring together of words than normal speech

308 (1) there's a rhythm, but not emphasized for rhthym's sake

309 (2) Some indications of musical tone, but VERY little

310 (2) very little change in tone

311 (2) words are short and clear , bit of a rhythm

312 (1) No tune.

317 (1) sounds like person giving directions

324 (2) no rhythm, no melody

328 (1)

335 (2) Non sensical words and pauses between words

343 (1) whining sound which is usually connected

346 (2) Rhythmic speech

349 (2) rhythm of song but voice of speaker

353 (2) the whining

357 (3) Patterning of rising pitch toward end.

358 (2) almost the same as in 134 above

---

file = u162; N = 23; μ= 1.87, σ= 0.81; H = 3; L = 1

211 (2) rythm

213 (1) sounds like Vincent Price story telling?

212 (3) Has a rhythm not like talking but no big difference in pitch like a song.

220 (2) speaking, but with empahsis which seems song-like

236 (1) no melody

238 (1) sounds like a recital

243 (2)

251 (3) unusual enounciation, but tone stays fairly constant

308 (2) rhythm and held notes, but small changes in tone

309 (1) Again, sounds like someone reading

310 (1) little variation in tone

311 (3) words are short and clear, but there is a poetry type rhythm

312 (1) No tune or holding of notes.

317 (1) not much rhythmic content

324 (3) has some rhythm and fluidity

328 (3)

335 (1) Poetic sound

343 (1) reading some literary work, play, etc.

346 (3) Poetic reading, sing-song-like

349 (2) lilting tune-like sound but still spoken

353 (2) no comment

357 (2)

358 (2) mostly talking, but there is a distinct rhythmic element and the pitch excursions are exaggerated to warrant a bit of “song”-like rating

---

file = u163; N = 24; μ= 4.21, σ= 1.06; H = 5; L = 1

211 (4) rythm, tone change, sustained notes

213 (4) started off talking, ended off singing

212 (5) Has a rhythm and change in pitch.

220 (5) every word had a tone associated

236 (5) melody, high pithes

238 (5) it sounds like a part of a song

242 (5) Beat, varied pitch, lilt

243 (4) the tone

251 (4) words sound unusual, pitch varies

308 (4) melodic, “umm” is clearly a note

309 (5) Speaker is clearly using deliberate major chords

310 (5) Starts on a high note, ends on a low, varied rhythm

311 (5) there is a beat, some words are drawn out and also spoken quickly to fit within the beat

312 (4) Specifically, the word “umbrella” sounds musical.

317 (1) someone talking as if makig a point

324 (5) fluidity, melody, long vowels

328 (2)

335 (5) Blending of one word to next

343 (5) changing pitch and connected sound

346 (4) Sustained pitches

349 (5) unnatural emphasis and singing voice

353 (3) no comment

357 (3)

358 (4) see 133 above, only here there is not much talking element, just not entirely decisive singing either

---

file = u164; N = 24; μ= 1.71, σ= 0.69; H = 3; L = 1

211 (2) sustained

213 (1)

212 (2) Emotion(excitement) in the sound leads to higher pitch not exactly like talking.

220 (2) very empahtic but not distinct notes

236 (2) expression and tone

238 (2) Sounds like a rock star asking the audience a question

242 (1) No variation in pitch

243 (1) tone

251 (3) sounds like it’s a musical, just finished the dialogue part that leads into a song

308 (3) intentional pattern of tone and rhythm, but tone and rhythm are not the point

309 (1) I know what it's from..also just sounds like someone yelling

310 (2) little variation in tone

311 (2) words are short and clear, little variance in tones

312 (1) No musical sound.

317 (1) someone shouting not singing

324 (3) sounds like yelling

328 (1)

335 (1) Lack of rhythm

343 (2) yelling

346 (1) Shouting, like an actor in a movie

349 (2) yelling voice but it carries on at the end rather than cutting off like a normal person would say it

353 (2) the change in tone

357 (2)

358 (1) sounds more like an emotional rather than a sung utterance to me despite the pitch range and protracted "finale"

---

file = u165; N = 23; μ= 2.22, σ= 0.75; H = 4; L = 1

213 (2) what is it about plays written in meter?

221 (2) could be that the word "frequently" does not fall on a beat?

223 (2) lack of melody

231 (2) Speaking, but with exagerrated tone changes

232 (3) Wide range of pitches, but not a lot of change in pace

246 (2) declamatory style

249 (1) sounds like theater

308 (2) tone shifts for emphasis, not for aesthetics

310 (2) little variation in tone, no rhythm

311 (2) little variance in tones, no beat

314 (2) Certain syllables were emphasized more than others.

325 (3) The elongation of certain sounds.

327 (2) theatric intonation and rhythm

329 (3) changes in pitch between syllables

330 (2) sounds like talking in a drama

333 (2) non-speech-like pitch and amplitude peaks imposed on regularly spaced syllables

340 (4) Intonation rather than meaning

347 (4) Frequency range and pattern

348 (2) intonation

352 (2) no melody

354 (2) absence of melody

359 (1) it's just plain talk with a "funny" accent

360 (2) intonation

---

file = u166; N = 22; μ= 3.74, σ= 0.91; H = 5; L = 1

213 (4) rap with very specific notes

221 (5) sounds like a tune, and regular beat

223 (4) catchy flow, rhyme like

231 (3) rhythm, not tone changes of normal speech

232 (4) change in pitch between strophes

246 (5) rhythm and harmony

249 (5) rythmic

308 (3) intentional rhythm, sentence had a "ground" tone with shifts up and down

310 (4) lots of rhythm, quite varied tone

311 (4) has a "rap" beat, words are spoken quickly within the beat, slightly difficult to understand

314 (4) The tempo and pitch didn't sound like a natural speaking voice.

325 (4) the speed

327 (4) high for a male voice, rhythmic

329 (1) speech inflection, not singing tone

330 (3) sounds like a rap

333 (4) external rhythm imposed on the speech, words are stressed differently than in speech

340 (3) Tone important

347 (4) Frequency range, tempo, and pattern.

348 (4) rhythm

352 (4) rap

354 (4) rhyme and attempt to pitch the words

360 (4) rap

---

file = u167; N = 21; μ= 2.95, σ= 1.24; H = 5; L = 1

213 (1) couldn't really tell what this was - sounded like yelling/growling?

221 (2) voice is rough (sounds angry)

223 (3) musical elements but drum/bass like

231 (2) no rhythm/accents, tone/pitch changes

232 (4) doesn't sound like words of any language, but rather a rythmic utterance

246 (5) rhythm and harmony

249 (5) speaking in syllables and rythmic

308 (2) intentional rhythm, lack of resonance (note)

310 (1) nothing rhythmical or musical

311 (2) no rhythm, sounds like someone is speaking loudly or yelling

314 (2) It sounded too forced to be speaking, but it didn't sound like singing either!

325 (4) I am not totally sure...mostly the fact that it is so fast

327 (4) words indiscernable, rhytmic, fast

329 (1) speech inflection

330 (3) weird sound

333 (3) loudness and pitch variation seem highly stylized

340 (4) Rhythm

347 (3) Tempo and frequency range

352 (4) no words

354 (3) its more rhythmic than straight speech

360 (4) rap

---

file = u168; N = 23; μ= 3.52, σ= 1.62; H = 5; L = 1

213 (4) sounds like mystic prayer/chanting

221 (3) like speech, except last note sound oddly high

223 (2) sounds like preaching

231 (3) rhythm ;different syllable lengths, some tone differences not like speech

232 (4) wide pitch range and note length, especially at end.

246 (4) rhythm not completely right

249 (4) tone an end of sample

308 (5) patterned tones and rhythm

310 (5) varied tone

311 (2) sounds like the rhythm of a foreign language being spoken

314 (5) The variation in pitch and tempo

325 (2)

327 (5) high pitch for a male voice, tones held, pitch variation

329 (3) inflected speech, intermediate tension in production

330 (1) sounds like

333 (4) pitch and loudness pattern of individual syllables seems subordinate to that of the entire phrase

340 (2) Not much rhythm

347 (3) Tempo and frequency range.

348 (4) melody

352 (5) different pitches

354 (4) use of pitch

359 (4) sounds like someone who can't sing, trying to.

360 (3) ambiguous

---

file = u169; N = 23; μ= 1.65, σ= 0.57; H = 3; L = 1

213 (2) talking with a rythym

221 (2) sounds like a poem, or stand-up comedy, but hits the same note a few times

223 (1) lack of melody

231 (1) rhythm and tone changes of regular speech - no recognizable musical intervals

232 (2) Speach that has character

246 (2) declamatory style

249 (1) sounds like a speech

308 (2) rhythm, pitch variations close to normal conversation

310 (1) little variation in tone

311 (2) no beat, tone varies, but no consistency to it.

314 (2) The words are said at a quick pace and the pitch doesn't vary much

325 (1) the speed and clarity

327 (2) sounds rehearsed, like a performance, not natural speech

329 (3) large inflection in pitch

330 (1) sounds like

333 (2) Though the intonational pattern doesn't sound like conversation, it doesn't sound like there's any attempt to impose some kind of external pattern of durational or pitch alternations on the individual syllables excep perhaps onthe last word sounds at all

340 (2) Rhyming

347 (2) Tempo and frequency range

348 (2) it is speech like

352 (1) public speech

354 (2) more rhyme than straight speech

359 (1) sounds like plain talk in some canadian accent

360 (1) a speech

---

file = u170; N = 22; μ= 1.86, σ= 1.08; H = 4; L = 1

213 (1) it's talking

221 (1) sounds more like excitement or anger, words add to this

223 (1) sounds like cartoon character in surprise

231 (1) clipped; no identifiable tone; nothing sustained

232 (3) speach with so much character, it's on its way to singing

246 (1) sounds natural

249 (1) sounds like surprise

308 (1) pitch changes reflect excitment

310 (1) sounds like screaming

311 (2) no beat, tone varies, but no consistency to it.

314 (4) The pitch sounds unnaturally high and changes in intervals that don't sound like speaking.

325 (3) the pitch

327 (2) sounds like an exclamation

329 (4) siren like sliding from syllable to syllable

330 (2)

333 (1) Although the intonation is not normal, it is more consistent with some kind of emotional or stylistic imposition on the regular pattern, rather than a musica one

340 (1) Communication intent

347 (4) High pitch and melodic pattern

352 (2) scream

354 (1) shrill irritation

359 (2) someone speaking happy, adds little 'tune' to speech

360 (2) intonation

---

file = u171; N = 22; μ= 2.5, σ= 1.6; H = 5; L = 1

213 (1) sexy whisper = talking

221 (1) sounds like pensive pause in speech

223 (4) drum in background

231 (1) wasn't "musical" - unidentifiable tone

232 (5) can't understand words, wide pitch range, wide note length range

246 (5) the music in the background

249 (2) sexy voice

308 (5) held note, rhythm

310 (4) starts with some sort of note

311 (3) can't tell if there is a beat, but it doesn't sound like talking (That's helpful huh?)

314 (5) The drum beat in the background, and the lingering on the first syllable.

325 (1) the speed

327 (4) recognize the song; hear background music

329 (2) gutteral, hard sound

330 (1)

333 (1) same as previous sample (u170)

340 (1) Communication intent

347 (2) Narrow frequency range

352 (2) whisper of satisfaction!

354 (1) lack of pitch

359 (1) nothing in it suggests singing to me

360 (3) vocal not speech sounds

---

file = u172; N = 22; μ= 3.0, σ= 1.23; H = 5; L = 1

213 (5) dunno

221 (5) the ending makes it clear, otherwise could have been a chant

223 (2) sounds like reciting verses

231 (4) the last two notes - otherwise classified as talking

232 (3) chant like.

246 (2) the last two syllables

249 (4) the end of the sample

308 (1) normal conversational rhythm and tone for language

310 (2) little variation in tone until the end

311 (2) has foreign language rhythm, with sounds short and concise except the last word at the end

314 (4) There were certain syllables where the pitch jumped up above the rest.

325 (1) the fact that it is another language

327 (4) sounds like chanting

329 (4) quite a glottal form of singing

330 (1)

333 (2) same as u169

340 (3) Tone pattern

347 (4) Regular tempo & frequency range plus the inflection at the end

352 (4) rythm and last note

354 (3) intentional use of pitch

359 (3) sounds like someone praying, tyical in-between thing to me

360 (3) rhythm

---

file = u173; N = 22; μ= 1.81, σ= 1.10; H = 4; L = 1

213 (1) kid talking- sounds disorganized

221 (1) slight surges in pitch on each syllable, also laugh at end?

223 (1) sound like baby talk

231 (1) normal tonal changes of speech

232 (2) no clue. just sounds that way. can't recognise a song pattern

246 (1) sounds natural for a kid

249 (3) can't tell with short sample

308 (4) more than one voice in same rhythm, pitch variation

310 (1) just talking with little variation in tone

311 (1) words sound short and concise, with a bit of laughter at the end.

314 (1) The variation in pitch came in places that you would expect from someone who is speaking.

325 (1) it is a child's voice. often has more pitch and intonational range

327 (4) rhythmic

329 (3) reverb in sample connects sounds to make it seem more like singing

330 (1)

333 (3) pitch pattern seems more phrase-length rather than from individual syllables

340 (1) Communication intent

347 (3) Melodic pattern and frequency range

352 (3) laugh and strange language

354 (1)

359 (2) sounds like a recitation, slightly sung speech

360 (1)

---

file = u174; N = 22; μ= 4.68, σ= 0.48; H = 5; L = 4

213 (4) more lounge-chanting. I give it a B, Jack

221 (5) discrete jumps in notes, in beat

223 (5) changes in pitch

231 (5) some of the pitch changes were not usual to speech

232 (5) variable pitch and note length

246 (5) looks like native indian singing

249 (5) goes up and down, all trilly

308 (5) patterned shifts in tone and rhythm, held notes

310 (5) there is rhythm

311 (4) words sound drawn out, variance in tones, consistent with a beat.

314 (5) He held on to certain syllables that were (I think) at the ends of words. He also seemed to be using a little bit of vibrato.

325 (5) seems to be a beat to the voice

327 (4) sounds like chanting

329 (5) pitch changes, very tight vocal production

330 (5)

333 (5) Phrase-level patterns clearly dominate

340 (4) Stress pattern

347 (4) The regular tempo

352 (5) slow and melodic

354 (4) use of pitch and rhythm

359 (4) it's like a better-sung prayer

360 (5)

---

file = u175; N = 22; μ= 5, σ= 0; H = 5; L = 5

213 (5) clear notes

221 (5) very melodious, high, discrete pitches

223 (5) changes in pitch

231 (5) musical interval changes;sustained notes

232 (5) pitch descent, changes in note length

246 (5) rhythm and harmony

249 (5) up and down

308 (5) clear notes and melody

310 (5) music is in measures, varied tone, rhythm

311 (5) a beat, rhythm, variance in tones that sound irregular in spoken language

314 (5) The pitch is high and she uses vibrato

325 (5) the wobbly voice

327 (5) higher pitch, long clear vowel sounds

329 (5) pitch changes, medium/high air

330 (5)

333 (5) same at u174

340 (5) Tone pattern

347 (5) Tempo and clarity of pitch

352 (5) melody and vibrato

354 (5) intentional use of pitch

359 (5) this is typical singing to me.

360 (5)

---

file = u176; N = 22; μ= 4.59, σ= 0.80; H = 5; L = 2

213 (5)

221 (4) holding the note at the end

223 (5) drawn out words towards end

231 (5) sustained last two notes; pitch of last notes not like talking

232 (5) variable pitch and note length

246 (5) last two syllables

249 (5) same as u168, end of sample makes sure this time

308 (5) pattered rhythm and tone shifts

310 (5) rhythm, varied tone

311 (3) some words short and clear, others drawn out

314 (5) This sounded even more like singing than File u168 because of the raised pitch and vibrato at the end.

325 (2) not quite sounding like speech, but more of a combination between the two

327 (5) rhythmic, long vowels, high pitch

329 (4) almost 2 samples here, starts more speech like, ends with more singing tone

330 (5) ending sound?

333 (5) same as u168, but now it's long enough to tell that it's clearly singing

340 (5) Tone pattern

347 (5) Melodic tempo and clarity of pitch

352 (5) long note plus rythm

354 (5)

359 (4) it's a song, so badly sung that it sounds like speech

360 (4) aged?

C.2.2 Question 2: More speech-like

file = n128

211 less regularity

236 used less rhythm

242 Slow down

251 enounciate more clearly (yes I realize that that probably wasn't english)

308 emphasized a meaningful "word"

310 spoken with no change in tone

317 less repetitive sounding

353 slow down

357 Slowed down, added pauses

file = n129

211 less of both

212 Get rid of the rhythm.

220 less rhythm

236 used full words

238 Actually say some words instead of just sounds.

242 Steady pitch

243 Use words and change the tone.

251 had more variation in sound to make it sound like words

308 less resonance, smaller interval between tones?

309 Spoken slower, used actual words

310 left out the varied rhythm and not variation in tone

312 Less rythmic. Less variance in tones.

317 remove the melody

335 Used vocabulary and non repetition

346 Make it less rhythmic

349 said the syllables in a talking voice

353 Not to be so rhytmic

357 Made word-like sounds, or alternately pure sustained vowels. These had all the seeming of being nonsense syllables, but also created a sense of imitation of non-vocal sounds.

358 say some words, use phonemes instead of drum sounds, and get off the steady beat

---

file = n130

211 less rythm

212 Speak the words with a normal tone.

220 less intonation

236 used less expresssion

251 avoid speaking it as a metre (as in poetry)

317 remove the rhythm

324 break down in fluidity

328 use a more monotone speech

335 No rhythm in speaking

346 Make it less rhythmic

349 take out all unnatural emphasis

353 talk in monotone

357 Reduce emphasis to fewer words.

358 keep a flatter/falling prosody

---

file = n131

211 less rythm

212 More uniform speed.

220 flat delivery, de-emphasise rhythm

242 Nothing

317 less rhyme

324 break down in fluidity

328 use a more monotone speech

346 Make it more casual, less rhythmic

353 not to be so harmonic

357 Mark the words with pitch and loudness based on their infomational qualities rather than on their place in the formal structure.

358 not emphasize prosodically the metric pattern

---

file = n132

211 not holding notes

212 Small changes in pitch and no dangling note at the end.

220 less tone changes, less lilt

236 used less differences in pitches

242 Same pitch

251 kept words closer to their normal spoken length

308 “flatten”

309 Less tone variation, no need to draw out last syllable

310 spoken in mor e of a monotone, less drawn out words

311 said the words shortly and clearly without drawing them out

312 Don’t hold sounds as long.

317 remove the melody

324 remove intonation/melody

335 Paused between words

343 break up the words and don’t change pitch so much

346 Make it more natural, follow speech intonation

349 taken out the tune

353 talk

357 Avoid the odd melisma on “-onds,” and mark emphasis based on meaning.

358 use sentential prosody, not jump around with the pitch

---

file = n133

211 less rythm

212 A uniform rhythm.

220 less rhythm

236 used a song that I didn’t recognize

238 not be so obvious on the rhythm

242 Slow down, no beat

251 words detached

308 lowered pitch

309 Not rhyme so obviously, be less rythmic

310 less rhythm

312 Speak less rhythmically.

317 remove the rhythm

324 remove rhythm, flatten style

335 More pauses and less rhythm

343 not so rhythmical

346 Make it less rhythmic

349 emphasized the words as if speaking to someone

353 no comment

357 not slur the words together as much, and take pauses

358 use sentence prosody: cut the “melody” and get off the beat

---

file = n134

211 less inflection

212 Not as high pitch at the end.

220 flatter delivery

251 kept tone more constant

310 less variation in tone

317 lowered tone

324 flatten style

353 no comment

357 give weight to either “so” or “-cit-” rather than such great emphasis on the first, and even more on the second.

358 keep the pitch curve within regular speech bounds

---

file = u160

211 less rythm

212 Enunciate the words in a regular speed.

220 less intonation

236 used a straight rhythm

242 Be coherent

309 Less tone variation, especially at the end

317 lower tone near end

357 have greater stability of pitch on each syllable

file = u161

211 less

220 flatter delivery, less rhythm

236 change pitches

242 Nothing

251 kept each word more separate

309 Don’t scale upwards so much in tone at the end

310 no variation in tone

317 lowered tone

335 Spoken slower and flatter

346 Less rhythmic

349 said it straight

353 no comment

file = u162

211 less

212 Reduce the change in pitch at the first part.

220 less intonation, more monotonic

251 put more time between each word

308 less resonance, less “held notes”

317 lowered fluctuations in tone

324 remove continuity

346 Make it more natural, less exaggerated

353 no comment

358 same as above

file = u163

211 less

213 sung no words

212 More regular speed and tone.

220 less tone, more monotonic

236 lower pitch and straight rhythm

242 Same pitch, equal stress on words

251 kept each word more separate

308 don't change pitch

309 Don't use major chords!

310 no variation in tone

311 not drawn out the words

312 Don't hold sounds so long.

317 not hold word own so long

324 remove melody and fluidity

335 Paused between words

343 break up the sounds

346 Follow natural speech intonation

349 not said Ummmmmmmmbrella

353 talk and slow down

358 see 133 above

---

file = u164

211 less

212 Lower the pitch.

220 monotonic enunciation, less tone change

236 not as much expression

242 Nothing

251 kept each word more distinct

308 less tone drop at end of sentence

310 no variation in tone

317 lowered volume of voice

324 remove intonation

343 less connected

349 not had the suspension at the end

353 slow down

---

file = u165

221 allow voice to vary off the 2 notes

223 monotone

231 less dramatic pitch changes

232 Less pitch modulation

246 speak faster

308 flattened tones

310 more monotone

314 Not lingered on some syllables (like the word “mean”)

325 word durations could be shorten, less intonational chance.

327 flatter intonation, less melodramatic,

329 less pitch inflection, less airstream

330 no dramatic tone changes

333 kept stressed syllable parameters within a more normal range

340 Reduce the rhythm

347 Narrow the frequency range

348 say it less pronounced rhythmically

352 less intonation

354 used less pitch

---

file = u166

213 quit rappin’

221 keep a more constant tone throughout (no jump half way)

223 monotone

231 less rhythm, more minor tone changes

232 not changed pitch; used more natural rythm (not forced to a meter)

246 more casual pauses

249 lengthed longer words?

308 flatten out tone changes

310 less variety in tone, less rhythm

314 Lowered the pitch and not used such a straight tempo.

325 slow

327 slower, flatter tone, lower pitch

330 no such rythm?

333 slow down the tempo and make it more irregular following more standard alternations of longer (stressed) and shorter/reduced (unstressed) syllables

340 reduce the stress pattern

347 Narrow the frequency range, more uneven tempo.

348 less fast

352 less rythm and prosody

354 less pitch

---

file = u167

221 don't stay on beat

223 less rhythm

231 rhythm & pitch changes of normal speech,

232 not changed pitch

246 more casual pauses

249 be monotone and not rythmically

308 broken up rhythm

314 Varied the pitch just a little more, and not sounded so rough and growly

325 use words

327 slower

330 I don't know

333 reduce range of variation

340 Reduce the stress pattern

347 More irregular tempo

352 speak to words with no rythm

354 more meaning, less rhythm

---

file = u168

221 let tone drop off at end

223 less melody

231 lengthen the short sounds; more speech-like pitch changes

232 more monotone end; less range in pitch

246 more speech-like prosody

249 monotone

308 monotone

310 spoken in a monotone

314 Used a straighter tempo and a more natural pitch variation (i.e. the last three syllables use pitch intervals that sound like song, not speaking)

327 rhythm less regular

329 less inflection, less air & more tension in production

333 broken up the phrase-length pattern of pitch change by emphasizing the individual syllables more

340 The final sound made it slightly songlike

347 Less regular tempo

348 pronounce more individual words

352 less difference between the pitches

359 avoid this melody-ish tonal progression

---

file = u169

213 break rythym

221 allow more drop-off in pitch at the end, don't keep using the same note

223 nothing

232 more monotone

246 speak faster

308 broken up rhythm

314 Let his voice drop at the end of each phrase

327 slower, more irregular, natural rhythm and less dramatic emphasis & pitch variations

329 less inflection, more gutteral production

333 The natural increases of duration, pitch and loudness on individual syllables are exaggerated – bringing them into line with normal speech would be sufficient

340 Reduce the stress patterning

347 More irregular tempo

---

file = u170

221 lower tones, more relaxed

223 lower pitch

232 monotone, especially at end

314 Lowered the pitch

325 lower the pitch

327 less melodramatic - reduce pitch variation within a single vowel

329 less pitch change, more gutteral production

347 Lower the pitch

352 be more monotone and low-voiced

359 avoid the brisk height change in "jo-ob"

---

file = u171

223 less drawn out speech

232 keep notes/syllables short

246 no music

249 talked normally

308 flattened tone

310 no music at beginning

314 Not conveyed so much emotion and relaxation in her voice (i.e. could have used a stronger voice)

327 less pitch variation in a single held vowel

329 less connection

347 Narrower frequency range

352 shorter

---

file = u172

221 remove last 1/2 second of piece

223 less flow and rhythm to speech

231 change the last two tones

232 eliminate pitch change and holding of last note

246 keep last two syllables shorter

249 not gone up at the end

310 no variety of tone

314 Not let the pitch jump up in the middle and end

327 less rhythmic, less pitch variation

329 less pitch change, less free air

340 Reduce the tonality

347 Less regular tempo

352 stay to a small range of pitches

359 avoid brisk changes in tones

---

file = u173

223 nothing

232 make monotone

249 if words made sense?

308 spoken alone, more even rhythm

327 less rythmic, less regular pattern of pitch variaiton

329 less connection, more glottal production

333 might be easier to decide with a longer phrase

347 Less regular pattern narrower frequency

352 less rythm

359 speak in a more “declarative” fashion

---

file = u174

221 monotone

223 monotone

231 change pitch of syllable 3,4,5 syllable(s) don’t sustain 5

232 make monotone and consistent length of syllables

246 speak faster

249 ....

308 broken up rhythm, flattened tones

310 had no rhythm

314 Not held onto the ends of words, and used a less syncopated rhythm.

325 flatten the rythm

327 less rythmic, less regular pitch variation

329 more glottal, less air

333 syllable duration and pitch contours should be more regular, less affected by larger-scale patterns

340 Ditch the long stressed syllables

347 Less regular tempo

352 no vibrato

359 avoid the scale-like toneal changes

---

file = u175

221 use only 1 or 2 notes

223 monotone

231 don’t sustain notes; change pitch of a few to normal speech rise and fall

232 monotone

246 speak faster

249 monotoneish

308 monotone

310 no rhythm, varied tone

314 Used less vibrato and let her voice drop to a lower pitch.

325 slow down, not so much pitch change!

327 shorter vowel sounds

329 more glottal, less air

340 Reduce the emphasis on pattern

347 Less regular tempo and fuzzier pitch

352 be faster and monotone

359 not sing

file = u176

221 don't hold the note at the end

223 shorter drawl in words. speak "snappily"

231 change pitch of last notes, and make them shoetr, more clipped

232 make monotone and consistent length of syllables

246 keep last syllable much shorter

249 spoke monotone

308 flattened out melody

310 no rythm, no varied tone

314 Used a straighter tempo, less verbrato, and a more natural pitch variation (i.e. the last three syllables use pitch intervals that sound like song, not speaking)

325 slow down, without so much variation in tone

327 irregular rhythm, shorter vowels

329 tenser, glottal production and less air

333 kept syllable-level patterns more prominent (as in the beginning of the phrase)

340 Ditch the long terminal syllable

347 Irregular tempo and fuzzier pitch

352 laugh!

359 not sing

C.2.3 Question 3: More Song-Like

file = n128

211 tone change

213 more rythmic

212 Add pauses between words and vary the pitch.

220 used more of a lilt when speaking, more intonation

236 used more tones and pitches

238 Made it sound more like a poem?

242 Vary pitch of voice

243 Use more pitches

251 have a tune, or recognizable notes

308 more intentionally held notes

309 Tones help...

310 more variety in tone and rhythm 317 more melodic

324 slow down and add rhythm

335 Varied tone and spoke slower

343 change pitch

346 Make pitches follow a musical scale

349 given it a tune and long notes

353 sing

357 Slowed down, and sustained the vowels and voiced consonants longer

358 vary the pitch, get on a beat

---

file = n129

220 more tone change

236 various pitches

242 Nothing

309 Musical notes...tones again.

310 could have held the notes longer

312 Hold notes longer.

317 hold some of the notes longer

343 disconnect and more monotone 346 Make pitches more distinct

349 ?

353 Sing

357 If it were hummed, it would seem more song-like, by avoiding the sense of nonvocal imitation.

358 add a melody line

---

file = n130

211 more tone chnage

212 Add more rhythm and more variation on the pitches.

220 more intonation

236 used various pitches and tones

251 more variations in pitch, more sustained words

308 intentionally held notes, larger interval in tones

309 Emphasize the tone variations already present

310 varied their tone, rhythm, put the words into measures

312 More mucical.

317 more melodic

324 add melody

328 vary his voice to a scale

335 Used more varied tones and ran words into each other

343 connect the words and change pitch

346 Make pitches follow a musical scale

349 give it a tune

353 sing

357 Sustain the individual sounds longer.

358 sing more :-) OK, get a melody and start off earlier with the beat

---

file = n131

211 more tone change

212 More variation on pitches.

220 more intonation and tone change

236 used a noticeable melody

238 put a tune to it

242 Varied pitch/speed

251 varied the pitch more, included a tune

308 held notes and varied pitch

309 already has ryming, could easily insert rythym or tones.

310 more variation in tone, different notes

311 stretched the words out

312 Add a tune.

317 less mootone

324 add melody

328 vary her voice to a scale

335 More varied tones and run words together

343 same as n130

346 Make pitches distinct, follow musical scale

349 tune, unusaul emphasis

353 sing

357 Change the length of syllables to a greater extent. It sounds as if all syllables have the same length, except the two repetitions of “row”, which are longer than all the others.

358 use a melodic pitch line

---

file = n132

211 consistency

220 more purity? (can’t describe it) in the notes

242 Nothing

251 had a tune

308 “clearer” on notes?

309 Be on key

312 More musical.

317 more cohesive note selection

346 Make pitches distinct, follow musical scale

349 used an operatic voice

353 sing

357 Sustain the vowels in preference to closing most syllables toward the final consonants.

358 put the pitch on a melodic line instead of just jumping about with it

---

file = n133

211 more toal changes

213 add accompaniament

212 Should change pitch.

220 more purity? in the tone and more modulation

238 sing it

242 Nothing

251 mashed words together more

308 more variety of tone

309 Even more emphasis on the rythym, maybe louder voice

310 more variation in notes

312 Longer duration of sounds.

317 add some melody

335 More tonal variety

343 not so monotone

346 Make pitches more distinct, follow musical scale

349 used a singing voice

353 no comment

357 sustain pitches on the vowels.

358 the opposite from above: take the pitch where it's supposed to go (on the note intervals) and keep with the rhythm

---

file = n134

211 more rythm tone change

212 Add rhythm.

220 more tone

251 slowed down first part slightly

308 repeated phrase with same pitch changes, creating a rhythm

309 Would need musical tones

310 more musical with notes

312 Add a melody.

317 slowed down

324 add rhythm/melody

335 Slower more varied tone

343 connect the words, intervals there

346 Make pitches distinct (as above)

353 no comment

357 Sustain the emphasized syllables on stable pitches.

358 put a melody on the pitch line

---

file = u160

211 more tone change, rythm

212 More variance in pitch.

220 more tonal quality in each word and more rhythm

236 used various tones and pitches

242 More beat to 'noise'

251 have some tune

308 held notes

309 Less broken speaking, voice would have to be stronger

310 more rhythm, put the words into measures with notes

312 Add a tune.

317 lengthen duration of words

324 add continuity/fluidity

335 Run words together

343 change pitch, connect the sound

346 (as above)

353 Sing

357 not have so many breaks, and create greater connection between syllables.

358 melody!

---

file = u161

211 moe tone change

212 Add rhythm.

220 more tonal quality in words, more purity in the tone

236 change the rhythm

242 Varied the pitch, speed

251 mashed words together more

308 identical repitition

309 Make the second half as rythmic as the first half

310 make it musical with notes, rhythm

312 Use a tune.

317 more melodic

335 Less pauses and lengthen connecting componenets of signal

346 (as above)

349 opened throught to use singing voice

353 no comment

---

file = u162

211 tone change

212 More variance in rhythm.

220 more purity in the tone

236 used various pitches

251 not finished pronouncing eaach word so clearly

308 clearer patterns in tone

309 Add deliberate musical tones

310 put it into measures with notes, varied tone

312 Use a melody.

317 hold some words longer

324 add melody

335 Strung words together

346 (as above)

353 no comment

---

file = u163

211 more tone change

213 sung more words

220 i.e. could have not held onto the ‘um’ in umbrella

242 Nothing

251 varied the sound (pitch) of the first part of the phrase

308 if the notes of “I must get my” were as distinct as the “umm”

312 Use a tune for the whole sentence.

317 more melody

346 Make it more rhythmic, follow musical scale

353 sing

358 see 133 above

---

file = u164

211 tone, rythm

212 Add rhythm.

220 words could have had notes associated

236 used various pitches

242 More variation in pitch, beat

251 spend more time pronouncing each word

308 more tone drop at end of sentene

309 Don’t be so monotonic

310 made it musical with a variety of rhythms, notes

312 Add a tune.

317 more melodic

324 revise tone

335 shortened last word

343 last sound should not fall, but hold it

346 (as above)

349 not used a yelling voice

353 sing

---

file = u165

221 user more than 2 notes, stay on beat

223 flowing melody

231 different endings on the final words of each phrase - up, not down, and held

232 change length of emphasis on some syllables

246 more harmony

249 sung

308 held emphasized word on clear note

310 more musical with notes and varied tone

314 Varied the pitch more

325 range the pitch

327 held tones longer

329 more airstream

330 maybe improve the rythm and ending sound?

333 phrases seem to end too abruptly for real singing

340 More emphasis on tones produced

347 More melodic pattern

348 use more melody

352 sing the prosody as a melody

354 used more related pitch

359 utter a more scale-like tonal progression

---

file = u166

213 quit rappin’

221 not much, make note change more sudden

223 more melody

231 hold vowels, musical tone changes

232 more variable length of note holds

308 held a note for variation

310 more variety of notes

314 Varied the pitch more.

327 more steadily held tones, longer vowels

329 more airstream

330 more tone changes

333 maybe just a longer sample would have been enough

340 more musical tonality

347 Make it longer

348 more melody

352 all the syllables on a certain pitch

354 more pitch

---

file = u167

221 higher voice, more varying notes

223 more melody

231 rhythm, recognizable pitches, sustained notes

232 don't know

308 more variation in tone

310 variety of tones, rhythms, more clarity

314 Not sounded so rough and growly.

327 vowels cleaner, less gutteral

329 less harsh gutteral production,

330 don't know

333 voice quality is very unlike singing

340 more musical tonality

347 Increase the frrequency range and length

352 less drum-sound, more melodic

354 more pitch

---

file = u168

213 need more lounge singing in temples

221 notes are not well defined... too close in pitch

223 more melody/flow

231 sustain some; make pitch changes musical intervals

232 more variation in note length in first third of strophe

246 right rhythm

249 ...

329 more air, less tension in production

330 add rythm?

333 again, probably just having a second phrase would have been enough

340 More rhythm.

347 More frequency range and longer

348 more melody

359 actually tune in right with the notes he suggests

---

file = u169

221 emphasis on word “Canada we eat” doesn’t sound song-like

223 more melody

231 change pitch of last syllables of each phrase - up or down by some standard musical interval, and sustain

232 more note length changes

246 more harmony

249 not so harsh?

308 held notes

310 had rhythm, varied tone

314 Not cramed so many words into each “beat” (assuming is a straight rhythm to this sound clip)

325 enlongate the vowels and words

327 longer vowel tones, different pattern of pitch

329 more air, more relaxed production

330 don’t know

333 do something besides just over-emphasizing the naturally stressed syllables

340 Increase the stress patterning

347 Increase the frequency range and longer

352 exagerate and differentiate the pitches

359 same; tonal progression

---

file = u170

213 (name that tune? deadbeat club)

221 don't slur the notes together, more abrupt pitch changes

223 more flow in melody

231 lower the pitch; get some greater range of pitch; sustain

232 ??? don't know

246 fill pause

249 longer words?

308 held notes

310 made it musical with many notes and some rhythm

314 Not dropped pitch at the end

325 more range

327 don't know

329 more relaxed production and more air

333 put some music into it?

340 More patterning of rhythm and tones

347 Make it longer

352 more voice

359 stick to some (western?) scale

---

file = u171

213 perhaps if the sample was longer, i would change my mind

221 use more notes, don't wisper

223 more melody

231 different pitch -recognizable musical tone; hold the note

249 ...

310 no speaking at the end

325 more change in intonation and pitch

327 less breathy, whisper-like

329 less glottal stroke, more air, & less gutteral production

340 More patterning of stress and tone

347 Wider frequency range and longer

352 more voice and different pitches

359 sing

---

file = u172

221 not much, maybe make notes easier to distinguish in first part

223 more melody

231 have some different pitch changes in other syllables; sustain some notes;rhythm

232 more pitch and note length changes throughout

246 more rhythm

249 more at start

308 held notes longer

310 varied rhythm and tone

314 Some syllables (i.e. the first second or so) sounded too monotone and were at a pitch that sounded like the person's natural voice (i.e. sounded low and comfortable)

325 more pitch change

327 more pitch variation in withing and between held vowel sounds

329 freer air, less glottal production

340 More stress patterning

347 Broader frequency range

352 more different pitches

359 not utter a “plainly-spoken” couple words here and there

---

file = u173

221 get rid of not fluctuations within syllables

223 more melody and flow

231 different pitch changes;hold some notes

232 don't know

246 slower rhythm

249 ...

308 spoken longer with similar rhythm

310 more rhythm and varied tone

314 Changed the pitch in a less natural way (i.e. raised the pitch near the beginning)

325 not sure

327 extend vowel sounds

329 less glottal production, more air

333 might be easier to decide with a longer phrase

340 More rhythm

347 More melodic pattern and longer

352 not laugh

359 stick to some scale

file = u174

223 more melody

327 less strictly regular rhythm

329 less tension in larynx, more air

340 More tonal patterning

347 Broader frequency range

359 take singing lessons

file = u175

223 not much

329 more relaxed larynx, less glottal, more air

file = u176

221 sounds off tune, notes aren't easy to make out at beginning

223 more melody

325 more movement in pitch and tone

329 more relaxed larynx, more air

359 take singing lessons and medicine for his throat

C.3 General Comments

The following questions were asked on the web form:

1. In the following field, please write some general observations that you might have made over the course of this study about the characteristics of speech and singing, as well as the similarities and differences between them. You may be as brief as you wish, and you may leave this section blank if you wish.
2. In the following field, please enter any comments or observations you may have had on your experience participating in this research study. Again, you may leave this section blank if you wish.

C.3.1 Speech/Song Observations

211 singing vs poetry rythm is important sustained notes help to define singing much of speach starts towards singing in some aspect

220 Great topic to study, Dave. Now I know how difficult it is to distinguish bewteen speech and singing. I see rhythm, pitch and emotion are important factors. They interact with each other depending on the situation and context. Very hard to draw a clear line between speech and singing.

220 I felt that both rhythm and tone had a large contribution to how much like singing I felt a sample was.

221 The “pitch surges within syllables” that made one clip sound speech-like, was present in a native-Indian clip that I judged as pure singing.

223 speech: -short spirts of sound -more monotone than singing -less flow, rhythm, melody (sounds that fit nicely one after another) singing: -comforting flow -smooth transition between sounds -melody -longer drawn out sounds...words that drag a bit when heard

231 Expectations seem to play a role—if we hear a pitch change we don’t expect in regular speech, we might call it song. Normal speech has lots of pitch change. The pitch at ends of phrases is used to give information. (“uptalk”, turning a statement into a question or not). Sustaining notes seems to happen only in singing—except for long drawn-out “Hmmmmmmmmm” when thinking, which doesn’t sound like singing usually if it’s followed by speech, but does sound like singing if you sing afterwards.

232 There’s something in the amount of “feeling” behind what is being said that pushes it toward speach. Good luck quantifying that!!

236 The process of identifying speech and singing is subjective to the listener as well to the person who’s voice is being heard.

238 Talking and singing at the far end of the scale are very distinct. In the middle, anything with a beat, rhythm, and intonation sounds like a song.

242 I had not realised that pitch played such an important role in making a difference between talking and singing.

243 Singing contains more pitches and is more harmonious than speech.

249 That it is not always that clear what the difference betweent he two is. With short samples it can be difficult. Also different languages sometimes sound melodic.

251 words in song tend to be pronounced slightly differently than speech, and do not seem

as distinct as words, ie, there seems to be a more constant stream of air leaving our mouths, as opposed to speech, where the sound seems to be cut off slightly between words

308 My opinion was often based on where I thought the speaker/singer was going next. I needed to extrapolate from the short clip.

309 It's interesting that I had a tendancy to include rhyme so much as a requirement for "singing"...but only when the sound was in english. When the sound was in another language, I had to rely much more on detecting tone and rythym.

311 there are a lot of similarities, but perhaps the biggest differences are the rhythm and presence of a beat in singing, whereas spoken language often is irregular in that way. Tones in spoken language can go up and down without a consistent pattern to it and words are just said as they are, without having to shorten them or draw them out so they fit in the measure

312 I am not trained at all in music, but it seems if I can recognize a clear musical note in a sound, I rate it as singing, otherwise, I don't.

314 I noticed how hard it is to put the characteristics of speech into words. There's something about a person's speaking voice that is natural. When speaking, the pitch just sounds "normal" for them; the sound isn't forced, there is no vibrato, the intervals at which the pitch changes sound "normal"....it's just really hard to describe!

317 Singing has melody and rhythm. In our culture voice is pretty monotone.

328 Is reading poetry singing? Ryming when combined with muscical scales is singing.

329 as an elementary music teacher, I'm faced with kids who have never sung before, and I try to coax singing out of them. It was interesting trying to mimic the samples and figure out the mechanism to recreate them.

330 Most of them sounds more like talking. I think to be qualified as singing, it has to sound much nicer, more rythmic and with tone changes (but not like speaking in a drama)

333 At first I thought there would really be only a binary distinction—either it's singing or it's not. By the end, I started to think that what matters most for "singing" is the imposition of some kind of pattern on the phrase where that pattern is not derived purely from the linguistic and/or pragmatic/emotional demands of communication. In other words, singing is imposing a nonlinguistic (presumably musical) pattern on speech.

340 Song emphasizes *patterns* of rhythm, stress, and tone more strongly.

343 Being a singer, I thought the majority of the sounds were speaking. I believe that a lot of them would be labelled song by a non-musician.

346 Most of the examples were of speech, but many were not examples of ordinary "talking". They included exaggerated intonation, poetic speech with sing-song (musical) qualities, foreign languages with unfamiliar intonation patterns or lexical tones (Chinese), etc. There were few examples of genuine song, which I understand to be rhythmic speech with distinct, sustained pitches that follow a musical scale of some sort.

347 Near the end I became aware of a dimention of clarity and sharpness of pitch that characterizes singing

349 Singing has rhythm, a certain sound and use of the vocal cords, a tune and emphasis on syllables not given such stress in speech.

352 Our prosody can be very musical but it stays easy to imagine the context and know it's just talking. Melody (extent of the variations between pitches), rythm and speed seem the best criterions. Hard part is to know where to put counting-out rhymes and rap...

354 Speech is the conveyance of meaning without rhythm, rhyme, or deliberate pitch for enjoyment's sake.

357 I don't believe that it is necessarily appropriate to set up a two-part division of these stimuli. I may have grouped items differently, if it were not set that they must fall along a single axis from speech to song.

359 the nature of the diffenrences is very diverse; some sounds show countinuously intermediate stuff and some keep switching between one and the other; aiming at spotting that difference could be relevant

C.3.2 Methodology Observations

211 somewhat subjective (is that the point?)

213 Part 2 was too long and labour intensive. I would say that most of your subjects will not fill in every field for every sample. I was getting tired after 5 of them! Otherwise- great work, David! :) [name].

220 I had a difficult time verbalizing what I felt. When I said that a sample could have been made more like talking by a flatter delivery, I found it hard to break that down into more direct observations.

221 It was difficult to explain why I made my choices in such a small space. I think that some of my intuition was too lengthy to explain, so I didn't.

223 Where do you classify rap music or rapping? It has flow and rhythm but is generally known as talking. It may not have melody but the flow of the speech may move it more towards the musics side. Also, high pitched children and cartoon characters that change their pitch often when speaking may be confused with singing.

231 Whew! Hard work, makimg all those decisions!

232 I'm frustrated trying to identify why I catagorized something one way or the other. I was tempted at points to go for the extremes so I wouldn't have to write something in one of the fields. I didn't give in to that temptation though.

238 Some people's voices just naturally sound like they're singing. It was obvious where people tried to monotonize phrases to force them to sound 'speech-like'.

242 I found this study very interesting, as did my witness (my daughter). I had never listened so closely to the various speech patterns before and found the study fascinating.I am glad I participated.

246 Phase 2 was a little too long. The judgments towards the end may not be as accurate as those towards the beginning.

249 It was fun :)

312 Easy to do this research. Great use of the Internet. I heard about this on CBC radio.

314 This was a lot harder than I thought it would be! I enjoyed it though. Good luck with the research.

317 best wishes in your research

324 an observation: in the *first* page of testing the browser appears to be stalled while in fact it is loading data (although there's no sign whatsoever), perhaps putting an initial

message would avoid confusing some more test users.

325 That it is quite hard to determine what is the middle between speech and singing.

329 great study, Dave. Hope you are doing some vocal physiology so voice people can use your results.

333 I found the tiny little comment windows in part 2 to be really annoying. These bigger ones in part 3 are much better. I hope my answers in that section were not too long. What kind of answers were you expecting? Despite this, I really enjoyed the experiment. I found the non-English examples very interesting, because of course I had less of (or no) idea what the actual linguistic pattern might be, nor could I be sure I knew much about the musical pattern possibly being imposed on the speech. I noticed some Putonghua (Mandarin Chinese), possibly one other Chinese dialect, (if it was Cantonese, perhaps it was sung after all, because I couldn't understand it), at least one S. Asian language (Hindi? Tamil?) and one N. American language (Navajo?) I'd be interested to know more about your results!

343 It would be interesting to know the results of this. I am entering Speech-Lang. Pathology in the fall and this is interesting for me. Good luck!

346 No additional comments, except that this is an issue that has interested me and that I had planned to do some research on (perhaps similar to your study) around 1990. However, my grant proposal was turned down and I did not pursue this line of research further. I presume you know List's old paper on the topic. Good luck with this project! The sound samples are interesting and fun to listen to.

347 I would be interested in followup about the experimental dimentions

352 Very well done but the order is not random.

353 It was fun. Good Job man.

357 Many of the stimuli appeared clearly artificial, and seemed out of place. Also, some of the questions seemed to be asking another thing entirely. For instance, "what could the speaker have done to make this more speech-like" is really asking, what quality of the sound would have to change to render the perception of a more speech-like sound. That is, the question is properly one about how the perceptual domain is influenced by the acoustic domain, but the question is posed in terms of the productive domain.

358 Part B became tiring after a while, because I felt I was writing the same thing over and over, with small wording differences. Perhaps I didn't understand what I was supposed to do (in this case sorry for the unusable data)—anyway, I didn't quite fill in the last few, I hope that's not too bad. The first part was almost fun. I'm not sure I've been very

consistent, it might be interesting to see if my ratings became more polarized in the course of the experiment. Did you have any repeated stimuli to check for reliability?
359 Lots of fun!

Bibliography

[1] Canadian Copyright Act. Chapter C-42, Section 29: Exceptions: fair dealing. [Online] Retrieved March 14, 2003, from `http://laws.justice.gc.ca/en/C-42`, 2001.

[2] Claudio Becchetti and Lucio Prina Ricotti. *Speech Recognition, Theory and C++ Implementation*. John Wiley & Sons, Toronto, 1999.

[3] Morten Bek, Troels Grosbøll-Poulsen, and Mads Ulrik Kristoffersen. Evolutionary trained kohonen networks as classifiers for human utterances. Master's thesis, Department of Computer Science, University of Aarhus, 2002.

[4] Steven Bird and Jonathan Harrington. Speech annotation and corpus tools (editorial). *Speech Communication*, 33:1–4, 2001.

[5] Albert Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge, 1990.

[6] Brian Clarkson and Alex Pentland. Unsupervised clustering of ambulatory audio and video. In *International Conference on Acoustics, Speech and Signal Processing*, volume VI, pages 3037–3040. IEEE, 1999.

[7] United States Code. Title 17: Copyright, Section 107: Limitations on exclusive rights: fair use. [Online] Retrieved March 18, 2003, from `http://www4.law.cornell.edu/uscode/17/107.html`, 2000.

[8] Stanley Coren, Lawrence M. Ward, and James T. Enns. *Sensation and Perception*. Harcourt Brace College Publishers, Toronto, 1994.

[9] Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005, 1990.

[10] Alain de Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4), 2002.

[11] Phillipe Depalle, Guillermo García, and Xavier Rodet. Tracking of partials for additive sound synthesis using Hidden Markov Models. In *International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 225–228. IEEE, 1993.

[12] ERI distance learning center. Internet based benefit and compensation administration. [Online] Retrieved March 18, 2003, from `http://www.eridlc.com/onlinetextbook/appendix/table7.htm`

[13] Erkan Dorken and S. Hamid Nawab. Improved musical pitch tracking using principal decomposition analysis. In *International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 217–220. IEEE, 1994.

[14] Boris Doval and Xavier Rodet. Estimation of fundamental frequency of musical sound signals. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3657–3660. IEEE, 1991.

[15] Boris Doval and Xavier Rodet. Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs. In *International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 221–224. IEEE, 1993.

[16] John M. Eargle. *Music, Sound and Technology.* Van Nostrand Reinhold, Toronto, 1995.

[17] Alexander Ellis and Arthur Mendel. *Studies in the History of Musical Pitch.* Frits Knuf, Amsterdam, 1982.

[18] James L. Flanagan. *Speech Analysis, Synthesis and Perception.* Springer-Verlag, New York, 1965.

[19] Edouard Geoffriois. The multi-lag-window method for robust extended-range f_0 determination. In *Fourth International Conference on Spoken Language Processing*, volume 4, pages 2239–2243, 1996.

[20] David Gerhard. Computer music analysis. Technical Report CMPT TR 97-13, Simon Fraser University, 1997.

[21] David Gerhard. Automatic interval naming using relative pitch. In *Bridges: Mathematical Connections in Art, Music and Science*, pages 37–48, August 1998.

[22] David Gerhard. Audio visualization in phase space. In *Bridges: Mathematical Connections in Art, Music and Science*, pages 137–144, August 1999.

[23] David Gerhard. Audio signal classification: an overview. *Canadian Artificial Intelligence*, 45:4–6, Winter 2000.

[24] David Gerhard. A human vocal utterance corpus for perceptual and acoustic analysis of speech, singing and intermediate vocalizations. *Journal of the Acoustical Society of America*, 112(5):2264, November 2002.

[25] David Gerhard. Perceptual features for a fuzzy speech-song classification. In *International Conference on Acoustics, Speech and Signal Processing*, volume IV, page 4160. IEEE, Spring 2002.

[26] David Gerhard. Pitch-based acoustic feature analysis for the discrimination of speech and monophonic singing. In *Journal of the Canadian Acoustical Association*, pages 152–153. CAA, September 2002.

[27] David Gerhard and Wiltold Kinsner. Lossy compression of head-and-shoulder images using zerotrees of wavelet coefficients. In *Canadian Conference on Electrical and Computer Engineering*, volume I, pages 433–437, 1996.

[28] Vincent Gibiat. Phase space representations of acoustical musical signals. *Journal of Sound and Vibration*, 123(3):537–572, 1988.

[29] James Glieck. *Chaos: Making a New Science.* Penguin, New York, 1987.

[30] Stephen Handel. *Listening.* MIT Press, Cambridge, 1989.

[31] Leon W. Couch II. *Digital and Analog Communication Systems.* Maxwell Macmillan Canada, Toronto, 1993.

[32] Léonard Janer, Juan José Bonet, and Eduardo Lleida-Solano. Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms. In *Fourth International Conference on Spoken Language Processing*, volume 2, pages 1209–1212, 1996.

[33] Christian Kaernbach and Laurent Dernay. Psychophysical evidence against the autocorrelation theory of auditory temporal processing. *Journal of the Acoustical Society of America*, 104(4):2298–2305, October 1998.

[34] Benjamin Kedem. Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE*, 74(11):1477–1493, November 1986.

[35] Bart Kosko. *Fuzzy Thinking: The New Science of Fuzzy Logic.* Hyperion, New York, 1993.

[36] Dejan Kulpinski. LLE and Isomap analysis of spectra and colour images. Master's thesis, Simon Fraser University, 2002.

[37] Karsten Kumpf and Robin W. King. Automatic accent classification of foreign accented austrialian english speech. In *Fourth Internalional Conference on Spoken Language Processing*, volume 3, pages 1740–1743, 1996.

[38] John E. Lane. Pitch detection using a tunable IIR filter. *Computer Music Journal*, 14(3):46–57, Fall 1990.

[39] Fred Lerdahl and Ray Jackendoff. *A Generative Theory of Tonal Music.* MIT Press, Cambridge, 1983.

[40] Daniel J. Levitin. Absolute pitch: Self-reference and human memory. *International Journal of Computing and Anticipatory Systems*, 4:255–266, 1999.

[41] George List. The boundaries of speech and song. In D.P. McAllester, editor, *Readings in Ethnomusicology*, pages 253–268. Johnson Reprint Co., 1971.

[42] Philip Loizou. Colea: A matlab software tool for speech analysis. [Online] Retrieved March 18, 2003, from `http://www.utdallas.edu/~loizou/speech/colea.htm`

[43] Esther Ho Shun Mang. *Speech, Song and Intermediate Vocalizations: A Longitudinal Study of Preschool Children's Vocal Development.* PhD thesis, University of British Columbia, 1999.

[44] Marion Mast, Ralf Kompe, Stefan Harbeck, Andreas Kießling, Heinrich Niemann, Elmar Nöth, Ernst Günter Schukat-Talamazzini, and Volker Warnke. Dialog act classification with the help of prosody. In *Fourth Internalional Conference on Spoken Language Processing*, volume 3, pages 1732–1735, 1996.

[45] Daniel McNeil and Paul Freiberger. *Fuzzy Logic.* Simon & Schuster, New York, 1983.

[46] Brian C. M. Moore, editor. *Hearing.* Academic Press, Toronto, 1995.

[47] James A. Moorer. On the transcription of musical sound by computer. *Computer Music Journal*, pages 32–38, November 1977.

[48] Yasuyuki Nakajima, Yang Lu, Masaru Sugano, Akio Yoneyama, Hiromasa Yanagihara, and Akira Kurematsu. A fast audio classification from MPEG coded data. In *International Conference on Acoustics, Speech and Signal Processing*, volume VI, pages 3005–3008. IEEE, 1999.

[49] Numerical Recipes. Numerical Recipes home page, containing the text of Numerical Recipes in C books on-line. [Online] Retrieved March 18, 2003, from `http://www.nr.com/nronline_switcher.html`

[50] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing.* Prentice Hall, Nwe Jersey, 1999.

[51] Jonathan G. Secora Pearl. Song, speech and brain: a survey of thge literature regarding brain processing of human vocal sounds, 2001. Depth Paper, University of California, Santa Barbara.

[52] Martin Piszczalski. *A Computational Model for Music Transcription.* PhD thesis, University of Stanford, 1986.

[53] Martin Piszczalski and Bernard A. Galler. Predicting musical pitch from component frequency ratios. *Journal of the Acoustical Society of America*, 66(3):710–720, September 1979.

[54] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipies in C.* Cambridge University Press, 1992.

[55] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition.* Prentice Hall, New Jersey, 1993.

[56] G.V. Ramana Rao and J. Srichand. Word boundary detection using pitch variations. In *Fourth International Conference on Spoken Language Processing*, volume 2, pages 813–816, 1996.

[57] Curtis Roads. *The Computer Music Tutorial.* MIT Press, Cambridge, 1996.

[58] Juan G. Roederer. *The Physics and Psychophysics of Music.* Springer-Verlag, New York, 1995.

[59] Stéphane Rossignol, Philippe Depalle, Joel Soumagne, Xavier Rodet, and Jean-Luc Colette. Vibrato: Detection, estimation, extraction, modification. In *Proceedings of the COST-G6 Workshop on Digital Audio Effects (DAFx-99)*, December 9–11 1999.

[60] Stéphane Rossignol, Xavier Rodet, Jöel Soumagne, Jean-Luc Collette, and Philippe Depalle. Features extraction and temporal segmentation of acoustic signals. In *International Computer Music Conference*, pages 199–202, 1998.

[61] Hajime Sano and B. Keith Jenkins. A neural network model for pitch perception. *Computer Music Journal*, 13(3):41–48, Fall 1989.

[62] John Saunders. Real-time discrimination of broadcast speech/music. In *International Conference on Acoustics, Speech and Signal Processing*, pages 993–996. IEEE, 1996.

[63] Eric Scheirer and Malcolm Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 1331–1334. IEEE, 1997.

[64] Eric D. Scheirer, Richard B. Watson, and Barry L. Vercoe. On the perceived complexity of short musical segments. In *Proceedings of the 6th International Conference on Music Perception and Cognition*, August 2000.

[65] Manfred R. Schroeder. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise.* W.H.Freeman, New York, 1991.

[66] Jerome M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, 1993.

[67] Bernard W. Silverman. Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability*, 1986.

[68] Mark D. Skowronski and John G. Harris. Human factor cepstral coefficients. *Journal of the Acoustical Society of America*, 112(5):2279, November 2002.

[69] V.S. Subrahmanian. *Multimedia Database Systems.* Morgan Kaufmann Publishers, Inc., San Francisco, 1998.

[70] Marc Swerts and Raymond Veldhuis. The effect of speech melody on voice quality. *Speech Communication*, 33:297–303, 2001.

[71] Dmitry Terez. Fundamental frequency estimation using signal embedding in state space. *Journal of the Acoustical Society of America*, 112(5):2279, November 2002.

[72] Dmitry Terez. Robust pitch determination using nonlinear state-space embedding. In *International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 345–348, 2002.

[73] Barry Truax, editor. *Handbook for Acoustic Ecology.* A.R.C. Publications, Vancouver, 1978.

[74] Windy H. Valerio, Alison M. Reynolds, Beth M. Bolton, Cynthia C. Taggart, and Edwin E. Gordon. *Music Play: The early childhood music curriculum guide for teachers, parents and caregivers.* GIA Publications Inc., Chicago IL, 1998.

[75] Luc M. Van Immerseel and Jean-Pierre Martens. Pitch and voiced/unvoiced determination with auditory model. *Journal of the Acoustical Society of America*, 91(6):3511–3526, June 1992.

[76] Rivarol Vergin, Azarshid Farhat, and Douglas O'Shaughnessy. Robust gender-dependant acoustic-phonetic modelling in continuous speech recogntion based on a new automatic male/female classification. In *Fourth Internalional Conference on Spoken Language Processing*, volume 2, pages 1081–1084, 1996.

[77] Linguistic Data Consortium website. [Online] Retrieved March 18, 2003, from `http://www.ldc.upenn.edu`

[78] Gethin Williams and Dan Ellis. Speech/music discrimination based on posterior probability features. In *Proc. Eurospeech99, Budapest*, September 1999.

[79] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton. Content-based classification, search and retrieval of audio. *IEEE MultiMedia*, pages 27–37, Fall 1996.

[80] Bin Yang. Content based search in multimedia databases. Master's thesis, Simon Fraser University, June 1995.

[81] William A. Yost, Arthur N. Popper, and Richard R. Fay. *Human Psychophysics.* Springer-Verlag, New York, 1993.

[82] Tong Zhang and C.-C. Jay Kuo. Heuristic approach for generic audio data segmentation and annotation. In *Proceedings of ACM International Multimedia Conference*, November 1999.

[83] Tong Zhang and Jay C.-C. Kuo. Hierarchical classification of audio data for acrhiving and retrieving. In *International Conference on Acoustics, Speech and Signal Processing*, volume VI, pages 3001–3004. IEEE, 1999.