8-4-2016

# A Physiologically Motivated Approach to the Classification of Natural Sounds using High Order Sound Statistics

Brian B. Bishop
*University of Connecticut - Storrs*, brian.b.bishop@gmail.com

# A Physiologically Motivated Approach to the Classification of Natural Sounds using High Order Sound Statistics

Brian Benjamin Bishop

B.S., University of Connecticut, Storrs, 2006

A Thesis

Submitted in Partial Fulfillment of the

Requirements of the Degree of

Master of Science

at the

University of Connecticut

2016

# APPROVAL PAGE

Master of Science Thesis

## A Physiologically Motivated Approach to the Classification of Natural Sounds using High Order Sound Statistics

Presented by

Brian Benjamin Bishop, B.S.

Major Advisor _____

Dr. Monty Escabi

Associate Advisor _____

Dr. Heather Read

Associate Advisor _____

Dr. Bahram Javidi

University of Connecticut

2016

# Acknowledgements

I would like to thank my Advisor, Dr. Monty Escabi, for his guidance, knowledge, and friendship, without which this work would not have been possible. I would also like to thank the other members of my advisory committee, Dr. Heather Read and Dr. Javidi Bahram, for their help and participation in this project.

Additionally, I would like to thank Dr. Shigeyuki Kuwada for his mentorship in my early career, without which I would probably never have studied audition and would probably not be in the field of engineering. I would also like to thank Dr. Duck Kim, who contributed to my early development as a scientist and engineer as well.

Lastly, I would like to thank anyone else: friends, family, coworkers, classmates, UConn faculty and administrative staff, etc. who have had a hand in helping me with this achievement, either directly or indirectly.

# Table of Contents:

# Abstract

The ability of humans and animals to classify sounds into behaviorally relevant categories is an ongoing field of study in auditory neuroscience and psychophysics. We employed a physiologically motivated signal processing scheme to extract time-varying statistics from a large database of real sounds. Our hypothesis is that two sound statistics, intensity and contrast, will be sufficient to classify broad categories of sounds using a naïve Bayesian classifier. We seek to evaluate this hypothesis, determine to what extent these two time-varying statistics can be used to classify sounds and to evaluate the limitations of the technique. The sounds themselves were organized into hierarchical categories based on the species or physical phenomenon generating the sound. Results of the naive Bayesian classifier suggest that classification using these statistics is better than chance for 13 different categories and in some cases has accuracy well above 50%. Performance generally increases with increasing duration of the validation sounds. Sounds from similar sources, such as flowing water and rain or two types of birds are often confused for each other. The classifier has high performance when comparing sounds at a much higher level of hierarchical classification, such as animal vocalizations compared to non-animal environmental sounds. Alternately, categories composed of more disparate sound sources, such as new world primates, have comparatively poor performance. This suggests that contrast and intensity statistics provide critical information that can contribute to sound categorization and that the hierarchical approach to classification is appropriate for many, but not all, types of sounds.

# 1 Introduction:

A fundamental question about the auditory system is “how are sounds classified into behaviorally relevant categories?”. Along with the ability to localize sound, sound classification enables organisms to locate mates and food and avoid predators and thus survive. It also enables higher level behaviors such as speech and music perception.

Sound, which is just pressure waves in a medium, is readily measured and quantified by instruments such as microphones and hydrophones. Indeed, the sensory organs of the ear operate on similar and relatively well understood principals similar to those of microphones: the transduction of pressure waves in a medium into electrical signals. As such, it stands to reason that the physical inputs to the auditory system can be modeled computationally and that whatever features are relevant to the classification task within the auditory system can be extracted computationally from digital sound waveforms recorded with microphones.

Ethological theories of the biological processing of natural stimuli first emerged with Barlow's 1953 study of the frog retina [1]. Barlow hypothesized that the on/off behavior of adjacent ganglion cells in the frog retina were optimally suited to identifying a dark fly on a bright background. Later work by Barlow [2], Field [3] and Olshausen and Field [4] took this hypothesis further, suggesting that biological systems are optimized for the natural stimuli most relevant to their survival. Barlow proposed, in information-theoretic terms, a minimum-entropy code, i.e. one which preserves the most possible information while minimizing the metabolic costs. Olshausen and Field demonstrated that for natural images, the receptive fields of neurons in the mammalian primary visual cortex could be predicted under this optimal-coding

hypothesis. In particular, they showed that natural images are "sparse", that they can be "represented in terms of a small number of descriptors out of a large set".

The motivation for this study is a response to this early work in the visual system. We assume, under Barlow's hypothesis, that biological processing is optimized for natural stimuli and that the classification task is vital for survival as it is intuitively easy to do. It then stands to reason that the best way to design a computational classifier with performance similar to a human is to identify the sound features that are used behaviorally and physiologically for the classification task and to incorporate those into the computational classification system.

This task is complex, as there are many candidate sound features of behavioral importance and that are represented neurally (see, e.g. [5]). The goal of this study is then to determine the relevance of two candidate sound features, intensity and contrast, to the sound classification problem, rather than attempt to use the entire space of candidate sound features. There is good evidence that these two sound features are encoded by neurons at multiple levels of the auditory system [6] [7] [8] [9] [10] [11] [12] [13] [14].

McDermott and Simoncelli [5] have shown that for a particular class of sounds known as sound textures, synthetic sounds that are perceptually similar to a real sound can be created from white noise through an iterative approach that replicates the statistical properties of the real sound. Sound textures are stationary processes, meaning that their statistical properties do not vary as a function of time. Their results suggest that intensity and contrast statistics are required for the synthetic sound to have a realistic percept. That study also suggests that there are other statistics required for generating perceptually accurate synthetic sounds, particularly correlations between frequency channels. Sound classification, however, is not exactly the same task, and so the goal

of this thesis work is to focus on intensity and contrast in particular to quantify the extent to which they are sufficient for sound category classification. Furthermore, the work of McDemermott and Simoncelli explored the role of sound statistics for sounds that are approximately stationary, whereas many sounds such as animal vocalizations have intricate dynamics and cannot be modeled as a stationary process. Intensity and contrast statistics have low computational cost to compute relative to other candidate sound features such as correlations between frequency channels, so they are a good candidate starting point for determining what sound features are sufficient for the classification task. This will hopefully inform further studies in both computational and neurophysiological sound classification.

Extraction of the candidate sound features was done with a physiologically motivated model of the auditory system, modeling the cochlea and elements of the early auditory system. This model of the auditory system is a digital signal processing chain that was applied to lossless digital audio waveforms. A database of sound segments was created and organized by listening to albums of sounds and grouping them into hierarchical categorizations and then organizing the result using the relational database software MySQL. Classification is done using a naïve Bayesian classifier and the maximum a posteriori (MAP) rule. Best practices from data mining and machine learning applications were applied to the data to ensure reliable results.

## 1.1 Applications:

Classification systems based on machine learning algorithms are an emerging field and could potentially have widespread applications. Silicon Valley based corporations like Google, Facebook and Microsoft are interested in classifying multimedia content within their servers to

improve their core search algorithms, improve "suggested" content, tag specific events, detect copyright infringement and other illegal content and to ultimately improve the targeting of their core advertising products. Researchers in Google's labs, for example, have been attempting to classify content and events within YouTube videos (e.g. [15], [16]) using video content. The visual imagery is just one element of video, however, and this could be supplemented with an audio classifier as well. Classifying based on the higher order statistical properties of sounds could ultimately be an efficient way to classify sounds for such purposes.

Military and law enforcement have a similar interest in classifying the content of visual and audio media posted publicly online, as evidenced by the publicly touted "Holodeck" project currently being developed by The MITRE Corporation [17], which is able to classify the content of public videos for military and law enforcement purposes. Key objects, such as tanks and other military hardware, Islamic State flags, and other images associated with terrorism or crime can be identified in videos and flagged for a human analyst to review. Once again, this could easily be extended to classify the audio content of videos to gather better intelligence. An example might be classifying explosions or sustained gunfire based on the higher order statistics of those sounds and comparing the result to that of the visual classifier to reduce false positives.

Medical devices provide another potential application for a statistical sound classifier. Modern medical devices are not currently equipped to modify the statistical properties of sounds. Many sources of noise that present problems for the hard of hearing, such as "cocktail party" noise (many people speaking at once, as in a cocktail party), busy streets or industrial noise such as HVAC noise are well described by their higher order statistics [5]. Hearing aids and cochlear implants could have a sound classification process which triggers additional signal processing to help reduce the noise from these types of sources when they are present.

Industrial and academic users may also have use for the classifier detailed within. Sieve Analytics, for example, offers a sound recording product, the Automated Remote Biodiversity Monitoring Network (ARBIMON) that is used for recording and analyzing environmental acoustic information. This type of product has uses, for example, in monitoring the habits of animal species, which is useful for basic science research as well as environmental impact studies that are of use to industry.

This work builds on prior work by Rahul Narayan and Monty Escabi. In particular, the database of sound segments was organized and expanded substantially, the classifier algorithm was optimized for speed and numerical stability, the signal processing chain was optimized with MATLAB best practices and additional experiments and validations were performed.

# 2 Methods:

The primary goal is to cultivate a database of various natural and man-made sounds which have been organized hierarchically and to use statistics derived from those sounds to generate a classifier using machine learning algorithms. Ideally, the statistics should be physiologically relevant, so they are obtained after processing sounds with a signal processing scheme that is representative of physiological auditory processing.

## 2.1 Database:

A preliminary database has been set up using the open source relational database management system, MySQL. The current database schema, the basic layout of how the database tables are related, is shown in Figure 1. Relational databases, and MySQL in particular, were chosen because they have a number of useful properties that make them ideal for this project. They are an established standard with over 40 years of use and development ( [18]). There is a wealth of knowledge and support along with a well-developed and easy to learn scripting language, SQL, which facilitates querying the database for entries with a particular property. Additionally, relational databases based on the SQL language are common in web-based applications, so the MySQL database developed for this project will help with the implementation of a publicly accessible database of sounds available over the internet.

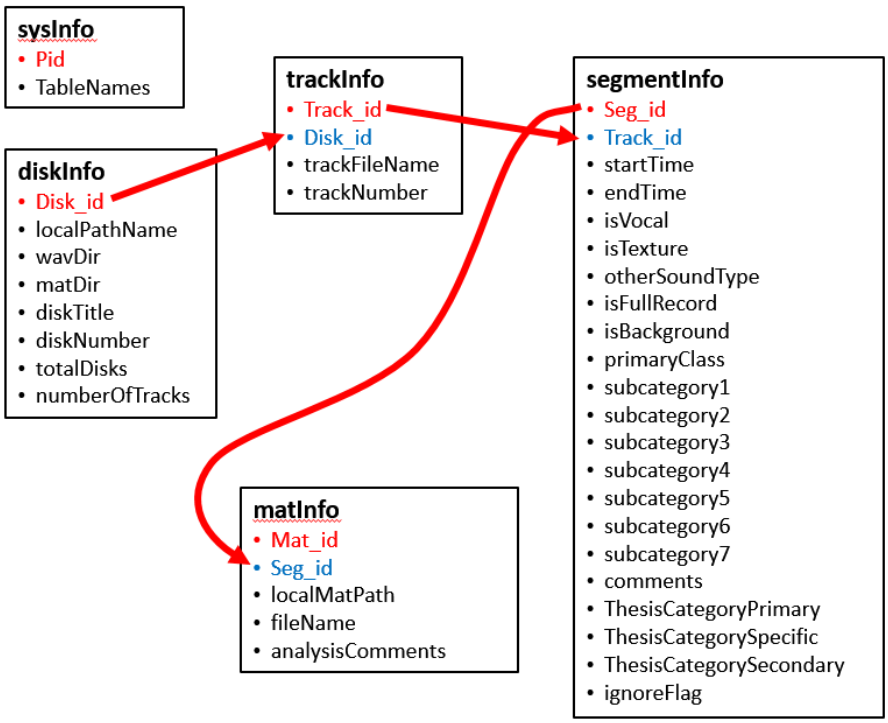


Figure 1: Dendrogram showing a sample of the ad hoc hierarchical sound categorizations. A). Non-animal sounds are grouped by the phenomenon that makes them while animal sounds (B) are grouped in a rough biological taxonomy. A direct mirroring of the actual biological taxonomy is avoided to facilitate usage by non-experts in taxonomy (which includes the author).

Relational databases reduce the amount of storage required by avoiding the storage of redundant data through the use of primary and foreign keys, which relate data stored on different tables. Relational databases can be set up on a public or private server and multiple users can access them simultaneously, with permission levels given to each user that allow read and/or write access, allowing the data to be shared with other researchers while maintaining security. MySQL, in particular, was chosen because it is open source and relatively compatible with three commonly used desktop operating systems, Microsoft's Windows product line, Apple's OsX and most versions of Linux [18].

The tables in this particular relational database contain information at the level of the album, track, segment and post-processed data. The album table contains information such as the names of each album, the disk number, total number of disks and number of tracks. The tracks table contains information such as track names and numbers. The segments table contains timestamps pointing into the tracks that contain subjectively "clean" sound segments (in that they are composed of predominantly one classification) and the hierarchical classification system. Currently, the hierarchical classifications have eight levels of classification. The processed data table contains information about the particular analysis that was done, such as the types of statistics that were extracted. Because of size constraints, the actual raw and processed data is stored outside of the database in a directory structure. The database instead contains pointers to the disk location of the data. Additionally, a MATLAB utility was written to allow users to import the database structure into MATLAB to be manipulated there, so that SQL knowledge is not a prerequisite for interacting with the database.

A sample of how the classification system itself is currently organized is shown in the dendrogram in Figure 2. Sounds include natural sounds, such as running water and wind, man-made sounds, such as engine noise, and biological sounds, such as canine barking and howling, bird songs and human speech. The database contains over 3000 tracks from 16 collections (see Appendix I) from which over 3500 sound segments have been identified and classified.

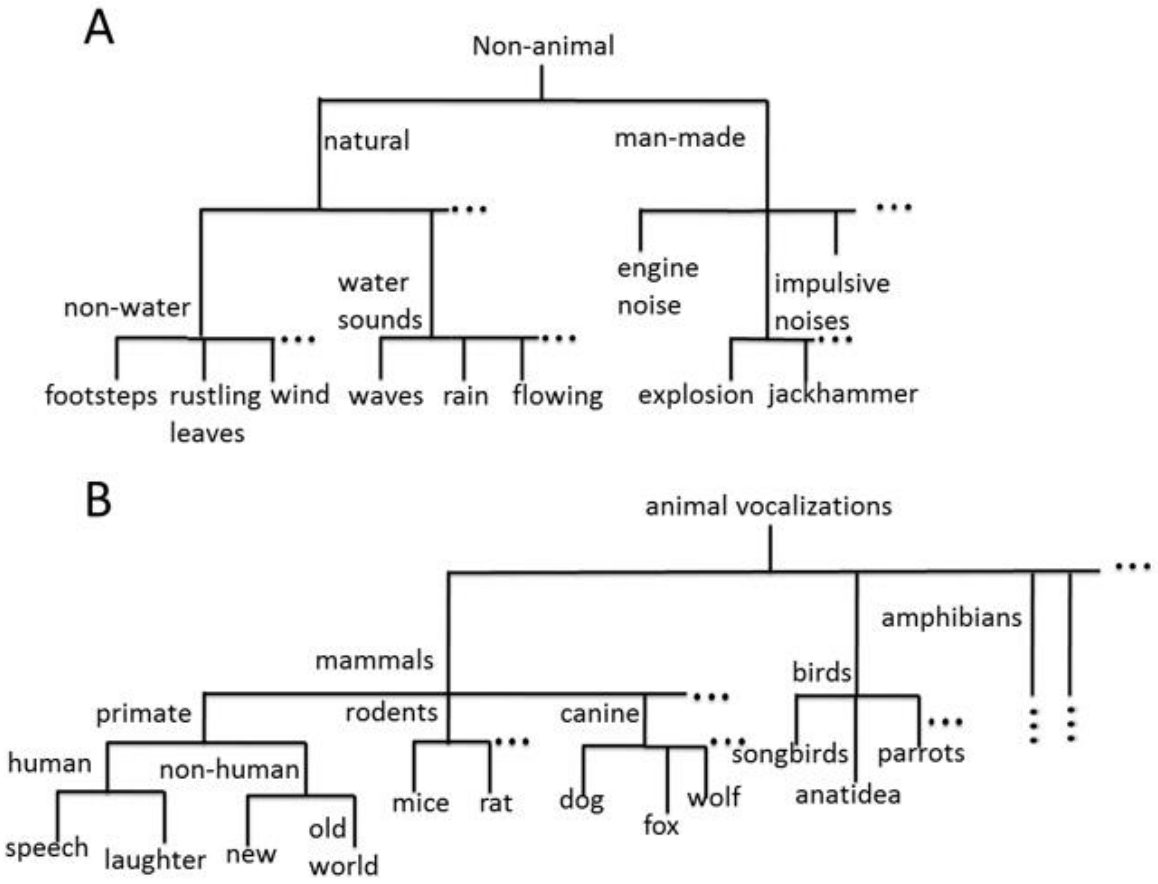


Figure 2: Dendrogram showing a sample of the ad hoc hierarchical sound categorizations. A). Non-animal sounds are grouped by the phenomenon that makes them while animal sounds (B) are grouped in a rough biological taxonomy. A direct mirroring of the actual biological taxonomy is avoided to facilitate usage by non-experts in taxonomy (which includes the author).

## 2.2 Auditory signal processing model:

Sound segments were processed with a physiological model of the inner ear followed by feature extraction that attempts to mimic auditory processing. A schematic of the auditory processing chain is shown in Figure 3 and is similar to the procedure described in Rodriguez et al. 2010. The

digital audio waveforms are all uncompressed CD quality sounds sampled at 44.1 kHz. Audio waveforms are first normalized to unit variance:

$$x(t) = \frac{x(t)}{\sigma_x} \quad (1)$$

where $x(t)$ is the digital audio waveform and $\sigma_x^2$ is the variance. This normalization accounts for different mastering levels on each track and the differences in overall loudness during individual segments. The waveforms are next passed through a filter bank that attempts to mimic the frequency selective processing of the cochlea, which will be referred to as the auditory filter bank.

The auditory filter bank models the transformations of the cochlea, which is part of the inner ear and the peripheral auditory system [19]. The peripheral auditory system is composed of the outer, middle and inner ear. The outer ear, which is composed of the cartilaginous pinna, amplifies sounds and shapes the spectrum of sounds to aid in localization. The effective spectral filter applied by the pinna is highly variable depending on sound location ( [20] [21] [22]). Due to this, and because no one would argue that behavioral sound classification is impaired when using devices which bypass the outer ear (e.g. in ear headphones), this stage of auditory processing is ignored.

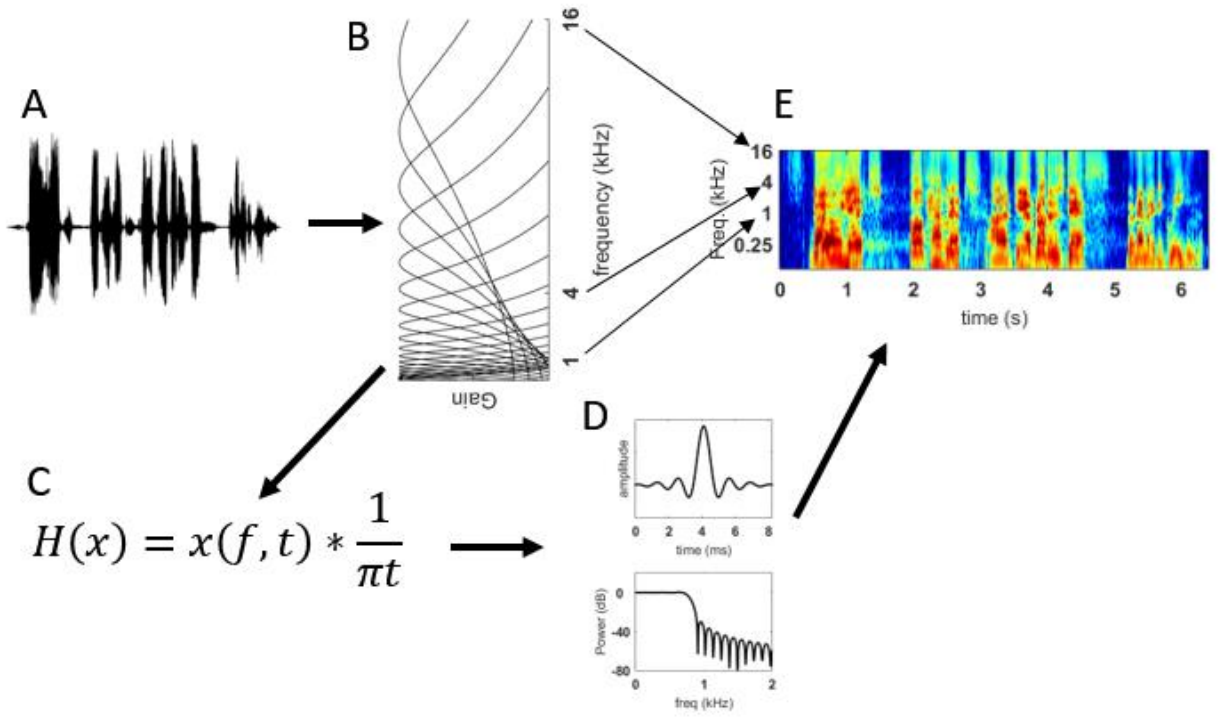


Figure 3: The signal processing chain for the auditory processing model. Unmodified digital waveforms (A) are passed through a filter bank (B) which mimics the frequency selectivity of the cochlea. The Hilbert transform (C) is taken to obtain the amplitude modulations with respect to frequency. An optimal low pass filter as described by Roark and Escabi with an 800 Hz cutoff (D) is applied to remove higher modulation frequencies which are not physiologically represented in the brain. The end result is the modulation spectrogram (E). Thin arrows from (B) to (E) show how the frequency channels of the cochlear filter bank map to the frequencies of the modulation spectrogram (ignoring the steps in C and D).

The next stage of the peripheral auditory system is the middle ear. The primary structures of interest in the middle ear are three membranes, the tympanic membrane, the oval window and the round window and the three small bones known as the ossicles, the malleus, incus and stapes. The tympanic membrane is connected to the malleus, which then connects to the incus followed by the stapes, which is connected to the oval window. The middle ear converts pressure waves in low density air into pressure waves in a high density fluid, known as the perilymph, with minimal power loss due to reflection that would normally be associated with such a phase

change boundary. In other words, it effectively performs an impedance matching between the two mediums. In particular, pressure waves in air contact the tympanic membrane, causing vibrations in the ossicles, causing the oval window to vibrate and transduces the vibrations into the fluid of the cochlea. The round window sits inferior to the oval window and vibrates out of phase to the oval window. It is required because the cochlea is a closed system and the fluid in the cochlea is incompressible, so without such a mechanism, fluid vibration in the cochlea would not be possible. The effect of the middle ear is primarily the efficient transduction of pressure waves between two mediums rather than something that imposes significant changes on the acoustic signal so it is also ignored in this model.

The primary structure of the inner ear responsible for audition is the cochlea. The cochlea is comprised of a fluid-filled spiral-shaped tube separated into three sub-tubes by two thin membranes. The superior-most tube is the scala vestibuli, which is connected to the oval window. It is filled with perilymph and connects to the middle sub-tube, the scala media, at the helicotrema, which is the apex of the cochlea's spiral. The scala vestibule and scala media are separated by Reissner's membrane. The scala media contains the Organ of Corti and is filled with a potassium rich fluid called endolymph and is separated from the inferior-most sub-tube, the scala tympani, which also contains perilymph and abuts the round window, by the basilar membrane. The Organ of Corti contains hair cells along its surface that cause action potentials (rapid changes in electrical potential caused by an electrochemical reaction) in response to vibrations of the basilar membrane and the endolymph that surrounds them. These action potentials are then transduced along the auditory nerve to the rest of the brain. In a simplistic sense, conversion of pressure waves to action potentials functions as a half-wave rectification, which, along with other neural processing, is generalized in the model as the extraction of

amplitude modulation (AM), or envelope. There are two types of hair cells, inner and outer. The function of the inner hair cells has already been described, while the outer hair cells act as a non-linear amplifier, helping to amplify small vibrations to a greater degree than larger ones [19].

The basilar membrane is relatively stiff, with its stiffness varying along its length. This variability in stiffness causes different portions of the membrane to vibrate more or less effectively depending on the frequency of vibration within the perilymph. Higher frequency sounds are better transduced at the base of the cochlea, near the oval and round windows, while low frequency sounds are better transduced at the apex, near the helicotrema. The particulars of this frequency selective property are well understood [19] [23] and are the primary part of the cochlear processing that will be replicated by the auditory filter bank.

The auditory filter bank is composed of third order gammatone filters [23] with impulse response function of the form:

$$h_k(t) = at^{n-1}e^{-2\pi b_k t}\,cos(2\pi f_k t + \phi) \qquad (2)$$

where $k$ is the frequency channel, $a$ is an amplitude coefficient ($a = 1$ here), $t$ is the time, $n$ is the filter order ($n = 3$), $b_k$ is the filter bandwidth in the $k^{th}$ frequency channel, $f_k$ is the center frequency of the $k^{th}$ frequency channel and $\phi$ is a fixed phase offset ($\phi = 0$ here). The bandwidth, $b_k$ was chosen to follow perceptual critical bandwidths [24] [25], $b_k = 25 + 75(1 + 1.4f_k^2)^{0.69}$. A sample impulse response function for a gammatone filter with center frequency of 1.025 kHz is shown in Figure 4B while the filter bank is shown at a larger scale in Figure 4A. Gammatone filters are chosen because they have a sharp high frequency transition and gradual low frequency tail that resemble auditory nerve fiber tuning functions [26].

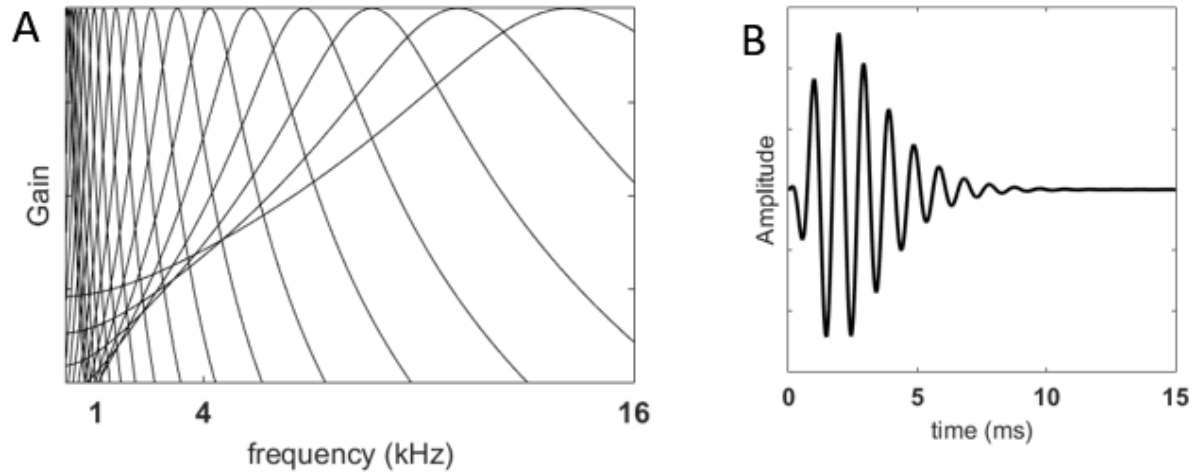


Figure 4: A). The filter gains vs frequency for the model cochlear filter bank. Each curve is a different band pass gammatone filter with bandwidth $b_k = 25 + 75\left(1 + 1.4f_k^2\right)^{0.69}$. Note that the filters have equal energy per octave. B). Filter impulse response for a sample gammatone filter with $f_c = 1025$ Hz and $b_k = 166$ Hz.

The auditory filter bank uses logarithmically spaced frequency channels between 0.25 kHz and 16 kHz with a 1/10 octave spacing. The result after filtering a digital audio waveform through this filter bank is:

$$X(t, f_k) = X_k(t) = h_k(t) * x(t) \tag{3}$$

Where $*$ is the convolution operator.

After the auditory filter bank, the envelope is extracted for each frequency channel $k$. Envelope extraction is an important function of the auditory system and is considered to be the primary information bearing feature of sounds. Neurons sensitive to the sound envelope are found throughout the auditory system, so this is an important stage of the model (see [27] for a comprehensive review). The envelope is mathematically defined as the absolute value of the analytic signal:

$$s(t, f_k) = s_k(t) = \left| X_k(t) + jH\big(X_k(t)\big) \right| * h_{lp}(t) \quad (4)$$

where $j$ is the imaginary unit, $H(\cdot)$ is the Hilbert transform and $h_{lp}(t)$ is an optimal low pass filter as described in [28] with 800 Hz cutoff, 200 Hz transition bandwidth and 60 dB stopband attenuation. This low pass filter is applied because auditory nerve fibers do not phase lock to modulation frequencies beyond this range, regardless of the neuron's best frequency [27]. The result, $s_k(t)$, is the envelope extracted from the spectrogram.

## 2.3 Time Varying Amplitude Statistics and Distributions:

Two time varying statistics are extracted from the spectrogram of each sound segment, intensity and contrast. Numerous studies have demonstrated that there are neurons sensitive to these statistics at all levels of the auditory system [6] [7] [8] [9] [10] [11] [12] [13] [14].

Prior to extracting contrast and intensity, the spectrogram was normalized for the mean value over the entire spectrogram, in particular:

$$s_{k,norm}(t) = s_k(t) - \frac{1}{k_{tot} t_{max}} \int_0^{t_{max}} \sum_{\forall k} s_k(\tau) d\tau \quad (5)$$

Where $\mathrm{t}_{max}$ is the total duration of the segment, and $k_{\mathrm{tot}}$ is the total number of frequency channels. This was intended to remove any overall trends in the processed spectrograms so that they could be directly compared between sounds. Other normalization schemes were considered, but ultimately abandoned because they did not provide an overall improvement in the correct

classification rate. For simplicity of notation, $s_{k,norm}(t)$ will be replaced with $s_k(t)$ for the remainder of the text.

Next, the probability of a particular intensity value was computed across all frequency channels of the spectrogram, conditioned on some range of times, that is:

$$P(s_k(t+\tau) = I \mid 0 < \tau < \tau_{max}, \forall k) \tag{6}$$

The results of this operation, computed at each time sample of the sampled spectrogram ($f_s = 3000\ Hz$), are shown in Figure 5C for a 6.4 second segment of speech. In Figure 5C, the conditional probability is computed at the sampling rate of the spectrogram, so $\tau_{max} = \frac{1}{f_s} = 0.3333\ ms$. This intermediate quantity shows the probability of an intensity value occurring at a given time point and is used to compute the intensity and contrast in probabilistic terms.

Intensity is essentially the loudness of a sound over some short time window, either within a frequency channel or across channels. The total amount of data was reduced prior to the machine learning task by specifically defining intensity as the expected value of the spectrogram over all frequency channels and the time interval $(0, \tau_{\max})$,

$$I(t) = E[s_k(t+\tau) \mid 0 < \tau < \tau_{max}, \forall k] \tag{7}$$

Here, the expected value formulation would use the conditional probabilities computed in equation (6), where a value of $\tau_{\max} = 50$ ms is used for the computation. This compares to auditory neurons in the inferior colliculus that are sensitive to intensity and which have an integration time typically between 50-200 ms [7]. Furthermore, this falls within the established range for intensity integration within the same range [7] [10]. Unless stated otherwise, intensity

will refer to the mean intensity over all frequency channels. Figure 5D (blue curve) shows an example of the intensity over time for a 6.4 second segment of speech.

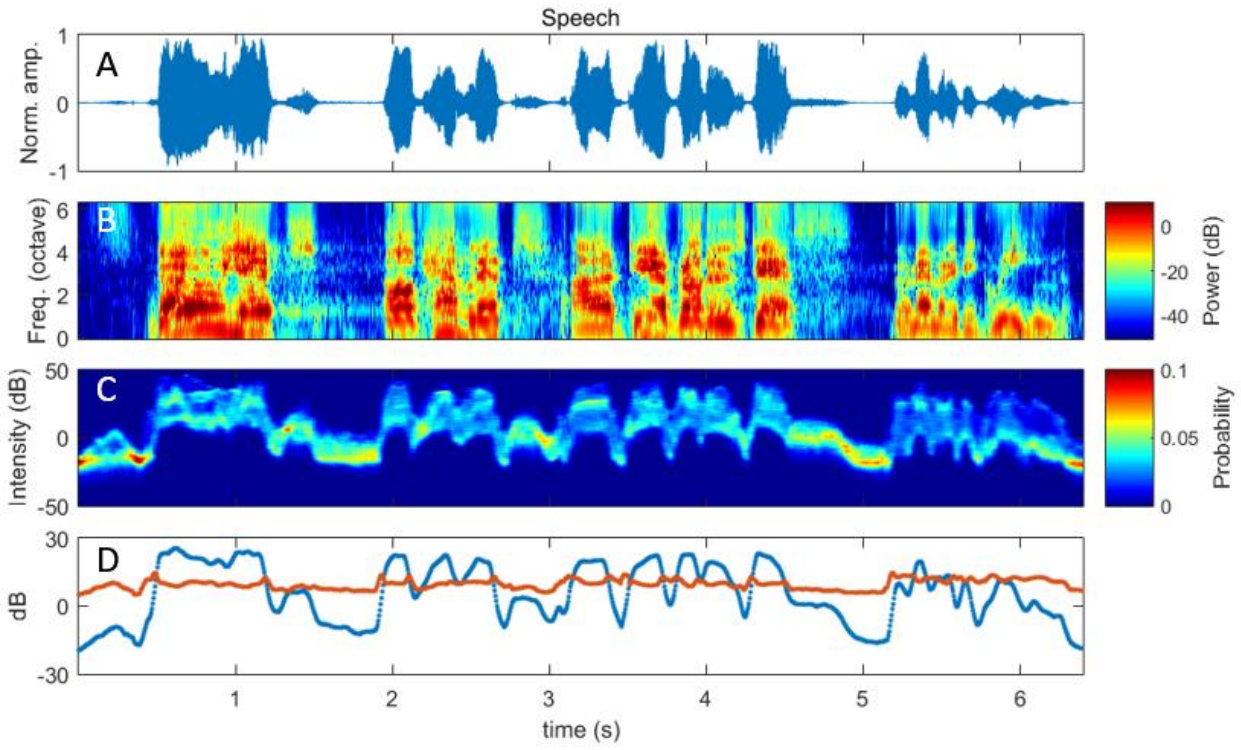


Figure 5: A). Normalized amplitude vs time for a 6.4 second sample of a speech waveform used for subplots B-D. B). The spectrogram, frequency channel (in octaves, starting at 200 Hz) is shown on the ordinate. C). The instantaneous probability of a given intensity value vs. time for the modulation spectrogram. D). The intensity (blue) and contrast (red) values obtained from averaging across frequency and 50 ms time windows.

Contrast is defined as the standard deviation of the spectrogram over all frequency channels and the same time interval, $(0, \tau_{\max})$:

$$C(t) = \sqrt{E[(s_k(t) - I(t))^2 |\, 0 < t < \tau_{max}, \forall k]} \tag{8}$$

Contrast could also be defined within each frequency channel over some time interval, although this is not considered further. Figure 5D (red curve) shows an example of the contrast over time for the same 6.4 second segment of speech.

After computing these values, they are converted to decibels (dB) by taking, e.g. $20\log_{10}(I(t))$. Escabi (2003) showed that the auditory system is most sensitive to these quantities on a logarithmic scale. Indeed, perception of intensity of light and sound is typically measured over a logarithmic range, which has been captured in the Weber-Fechner law. Joris et al. (2004) offer an extensive review of the amplitude modulation literature further demonstrating that the auditory system encodes amplitude modulation on a logarithmic scale.

Acoustic intensity and contrast have important analogs in visual science, where neurons sensitive to such quantities were first discovered [29]. Visual intensity, also known as luminance, is the total amount of light power incident on a unit area. Similarly, sound intensity is the sound power per unit area. Visual contrast has multiple definitions, but in general it describes the total deviation in visual intensity over some dimension relative to some absolute measure of the intensity such as the mean. An early definition due to Michelson is $\frac{I_{\max}-I_{\min}}{I_{\max}+I_{\min}}$, where $I_{\max}$ and $I_{\min}$ are the highest and lowest luminance [30]. This is analogous (although not identical) to the definition of modulation depth for audio signals.

Joint and marginal distributions of contrast and intensity were created for each category of sounds from the statistics extracted from each sound segment. The joint distributions, $Pr_n(s_i, c_j)$ were created by taking a two-dimensional histogram of contrast and intensity statistics from all sound segments within a particular category and normalizing each by the total number of samples so that it obeys the law of total probability. Here $s_i$ is the intensity of the $i^{th}$ intensity

bin, $c_j$ is the contrast of the $j^{th}$ contrast bin and $n$ is the $n^{th}$ category. As is suggested by the notation, this is a discrete distribution which is used to approximate the underlying continuous distribution, which is more difficult to estimate given a finite database of sound segments. Marginal distributions are generated as

$$Pr_n(s) = \sum_c Pr_n(s|C = c)Pr_n(C = c) \tag{9}$$

And similarly for the marginal distribution of contrast.

An additional stage of manipulation was applied to these distributions in order to improve their generalizability in the naïve Bayesian classifier paradigm. If any sample from a sound segment maps to a point in a distribution with a zero value, the conditional probability associated with that entire segment becomes zero, regardless of how well the distribution fits otherwise. In an effort to prevent the most egregious examples of this from affecting the classifier, the distributions are smoothed with a narrow 2D Gaussian filter (standard deviation 0.25 dB in both the contrast and intensity dimensions). Any samples that are still zero valued are replaced with the minimum machine precision of MATLAB. The general equation for a Gaussian filter that is uncorrelated in the two dimensions is:

$$G(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right) \tag{10}$$

In this case, the filter is discretized using a 1 dB sample spacing and with $\sigma_x = \sigma_y = 0.25$ dB. This filter is applied to the distributions using a two-dimensional convolution:

$$Pr_n^*(s_i, c_j) = G(s_i, c_j) * Pr_n(s_i, c_j) \tag{11}$$

Because of the tails of the Gaussian filter, the resulting distributions, $\Pr_n^*(s_i, c_j)$, have some amount of probability mass in every intensity-contrast bin, with bins closer to the distribution generated directly from the real data having higher mass.

## 2.4 Naïve Bayesian Classifier:

Classification is carried out using a naïve Bayesian classifier. Statistical classification schemes such as the naïve Bayesian classifier typically attempt to determine the most probable classification from an assortment of categories given the evidence. This is typically aided by prior knowledge of the underlying probabilities associated with the evidence given each category. The procedure is known as the maximum a posteriori (MAP) rule [31]:

$$\hat{C} = \underset{\forall m}{argmax}\, P(C_m|\boldsymbol{x}) \tag{12}$$

Here, $\hat{C}$ represents the best estimate of the classification based on maximizing the conditional probability $P(C_m|\boldsymbol{x})$ over all of the categories, $C_m$, given the evidence $\boldsymbol{x} = \{x_1, x_2, \dots, x_N\}$, also known as the feature vector. $argmax$ is a function of a function where the output is the value of some parameter, in this case, all values of $m$, the category number, that maximizes the function. The argument of the $argmax$ in equation (12) can be expanded using Bayes theorem, shown below for two events A and B

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{13}$$

Here, $P(A|B)$ is known as the posterior distribution and $P(B|A)$ is known as the prior distribution, which is typically generated from past knowledge. An additional assumption, the so called "naïve" assumption, assumes independence between the individual features, the $x_n$. Without the naïve assumption, the chain rule of conditional probabilities would have to be applied N times, with the resulting requirement that there be an estimate for all of those conditional probabilities (which is not feasible for all but the most massive data sets). After all simplifications, the argument of the $argmax$ in equation (12) becomes:

$$\hat{y} = P(C_m) \prod_{n=1}^{N} p(x_n|C_m) \tag{14}$$

Note that the probability $1/p(\boldsymbol{x})$ that would be required by Bayes theorem is dropped because it does not depend on $m$ and is therefore constant for all categories and does not affect the output of the $argmax$ function. Additionally, the $P(C_m)$ term is dropped, in this case, by assuming the categories are equiprobable and log probability is used to avoid numerical precision issues that result from taking the product of N values that are usually much less than 1. This yields the final form:

$$\hat{y} = \sum_{n=1}^{N} log\big(p(x_n|C_m)\big) \tag{15}$$

In the particular case used here, the feature vector, $\boldsymbol{x} = \{(s_1, c_1), (s_2, c_2), \dots, (s_N, c_N)\}$, is an ordered set of time-varying contrast and intensity pairs for a given sound segment under test. The $C_m$ are the different sound categories, such as speech, flowing water or white noise.

Figure 6 demonstrates the basic procedure using an example set of intensity and contrast pairs from a speech segment. Figure 6A shows the intensity and contrast statistics for a short sample of speech. These pairs are mapped to the associated points on the joint intensity-contrast distribution for each of the three example categories, speech (Figure 6B), white noise (Figure 6C) and flowing water (Figure 6D). This is done for all of the intensity-contrast pairs (in this case 128 such pairs) and equation 14 is applied, followed by the decision rule in equation (12). In this example, the decision rule clearly decides for speech, and the categorization is correct.

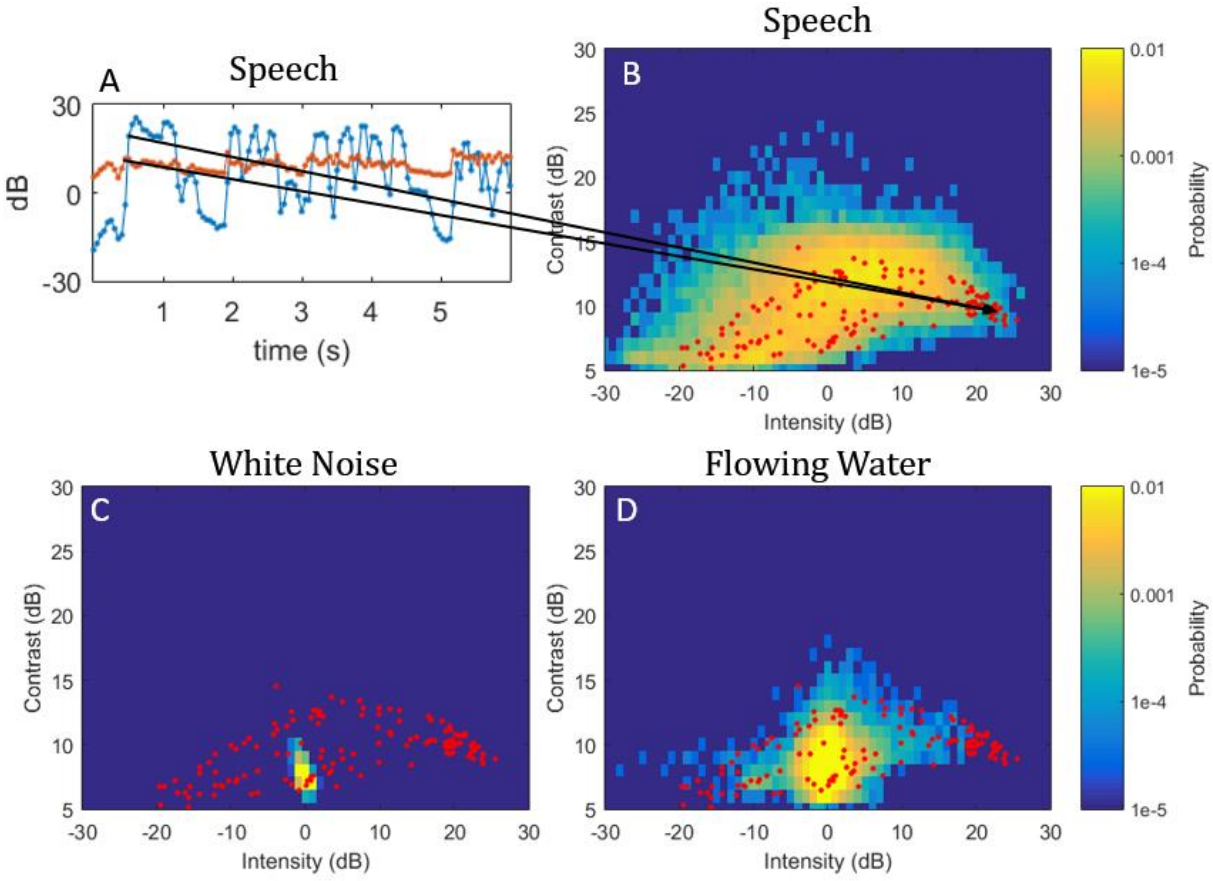


Figure 6: A). A 6.4 second segment of contrast and intensity statistics for a speech signal. Intensity is blue, contrast is red. B). The probability mass function for the speech category with the individual intensity-contrast pairs from A. overlaid (red dots). C). Same as B. except the pmf is for the white noise category. D). Same as B. except the pmf is for the flowing water category. The decision rule works as follows: take the product of the underlying probability mass

corresponding to each red dot and do this for each category. Decide for whichever category has the highest product, in this case a correct decision is clearly made for speech.

## 2.5 Cross validation:

An important component of ensuring that the results of machine learning and classification algorithms are correct is the choice of cross-validation scheme [32]. Cross-validation is the method by which a classification algorithm is verified to ensure that it will generalize to new data. Cross-validation ensures that positive results are not statistical artifacts or otherwise anomalous when compared to new data.

Two common cross-validation techniques were used, with slight modifications to compensate for the particulars of this experiment. The first cross-validation method is a variant of twofold cross-validation [32]. In this paradigm, each sound segment was divided in half, with one half used for building the model (prior) distributions and the other half used for validation. The validation dataset for each category is then a concatenation of every sound segment in that category. For each category, N consecutive samples were taken from the validation dataset and the naïve Bayesian classifier was applied to determine the classification. This was repeated 1000 times and a confusion matrix was formed. Rows of the confusion matrix show the validation class and columns show the result of the naïve Bayesian classifier. Results on the diagonal indicate a correct classification while results off the diagonal indicate that the classifier has "confused" the validation sample with a different category.

The second cross-validation technique that was tried was the leave one out cross-validation, again, adapted to the particulars of this experiment. In the leave one out paradigm, every sound

segment except for one, the current validation segment, is used to generate the model distributions [32]. Using a random starting point, N consecutive samples are then drawn from the sound segment to be validated and validation proceeds as shown in Figure 6. This is repeated for each sound segment and repeated 50 times within each sound segment (which may result in redundant validations for short sound segments or large N). The results are assembled into a confusion matrix.

For both cross-validation schemes, values of N from 1 to 128 in doubling steps are used, corresponding to a minimum sound duration of 50 milliseconds up to 6.4 seconds. The cross-validation scheme is applied to both the joint distribution of contrast and intensity as well as the marginal distributions.

# 3 Results:

## 3.1 Intensity and contrast statistics for example sounds

Sound segments (819, see Appendix II for a complete list of albums and track numbers used) from 13 categories were used to build ensemble distributions of the time-varying intensity and contrast statistics. Many more sound segments were available in the database, but only categories with at least 20 sound segments of subjectively determined “good” quality were used.

Before discussing these ensemble distributions, it is instructive to consider the time-varying statistics of sounds. Sounds can be broadly divided into two categories, those with relatively stationary statistics over short time scales and those without. Background sounds and ensemble

sounds, such as running water or a chorus of insects are a typical example of the former category, while a single speaker or animal vocalizing is typical of the later. It's generally instructive to consider sounds from each category, stationary and non-stationary, when looking at examples. While this is a useful distinction, for the naïve Bayesian classifier proposed here to work, there is an assumption that sounds from the same category have similar statistical properties and so must exhibit some degree of stationary over some sufficiently long time scale.

For each sound segment in the database, the intensity and contrast statistics were extracted over a 50 ms interval as described in the methods section. An example of these statistics are shown in Figure 5, above. Figure 5A shows the original digital waveform for a 6.4 second segment of speech from the Renaissance Theatre Company's 1992 production of William Shakespeare's Hamlet. Figure 5B shows the spectrogram of this speech segment and Figure 5C shows the distribution of amplitudes conditioned on the time sample and as a function of time. Figure 5D is obtained by taking the mean and standard deviation over 50 ms intervals (yielding 128 total samples over the 6.4 seconds) of the spectrogram in Figure 5B. Speech typically has a wide dynamic range in both intensity and contrast as compared to many other categories of sound. Figure 7 compares these results for the same speech segment (Figure 7A-D) against sounds from two other categories, running water (Figure 7E-H) and white noise (Figure 7I-L). The intensity and contrast statistics of flowing water appear to be stationary over a much shorter interval with a much smaller dynamic range, while white noise has an even smaller dynamic range. This is expected as flowing water is a type of sound texture, which are ensembles of many acoustic events and are usually stationary over a short time scale (cite McDermott and possibly others). White noise is an even more extreme (and unnatural) stationary process [33] [34].

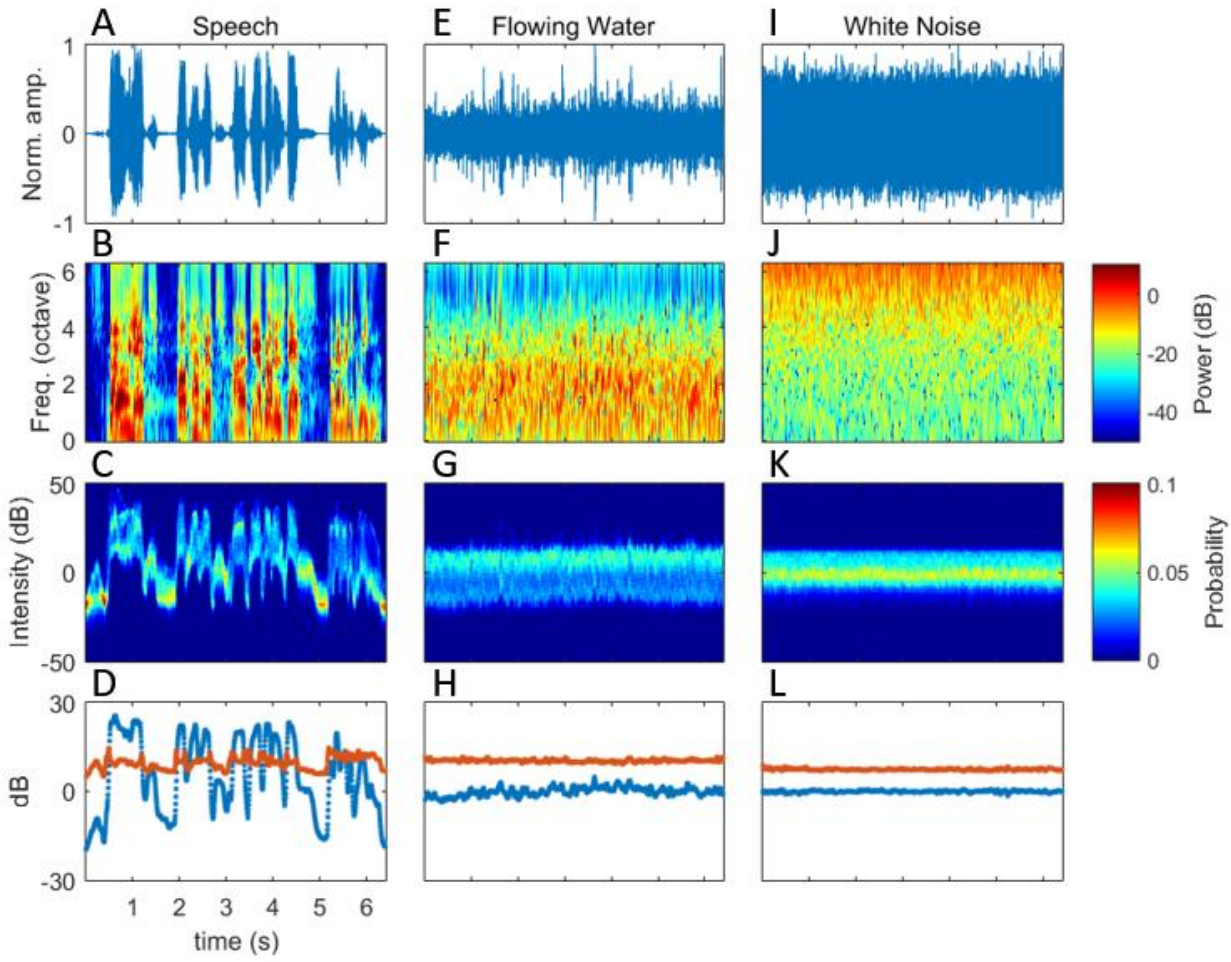


Figure 7: Comparison of the speech sample from figure 6 (A-D same quantities as figure 6 in the same order) to a relatively stationary sound texture, flowing water (E-H), to the control sound, white noise (I-L). All sounds are 6.4 seconds long. Note that speech has the most variability in both intensity and contrast, while white noise has almost no variability. This is also reflected in the spectrogram: speech has the most complex spectrotemporal features, while white noise has the least complex.

Figure 8 shows a comparison of three different animal vocalizations, the same segment of human speech from before (Figure 8a-d), a parrot, the scarlet macaw (Figure 8e-h) and a large domestic dog barking (Figure 8i-l). Some of the confounds of this experiment are evident in this example. In particular, the scarlet macaw's vocalizations are quite spread out, with a fairly high level of background sound in between (which sound like insect and/or frog choruses). This can be seen by examining, for example, the scarlet macaw's spectrogram (Figure 8f), and noting the

relatively high power level between the two main calls. Compare this to the power level between the dog barks in Figure 8j, which comes from a compilation of studio quality sound effects [35], where the sound recordings are much cleaner and more free of conflicting background sound. Comparing the mean intensity during vocalizations to the mean intensity when there is no vocalization to obtain a signal to noise ratio (SNR), there is almost a 25 dB difference in SNR when comparing the scarlet macaw calls to dog barks.

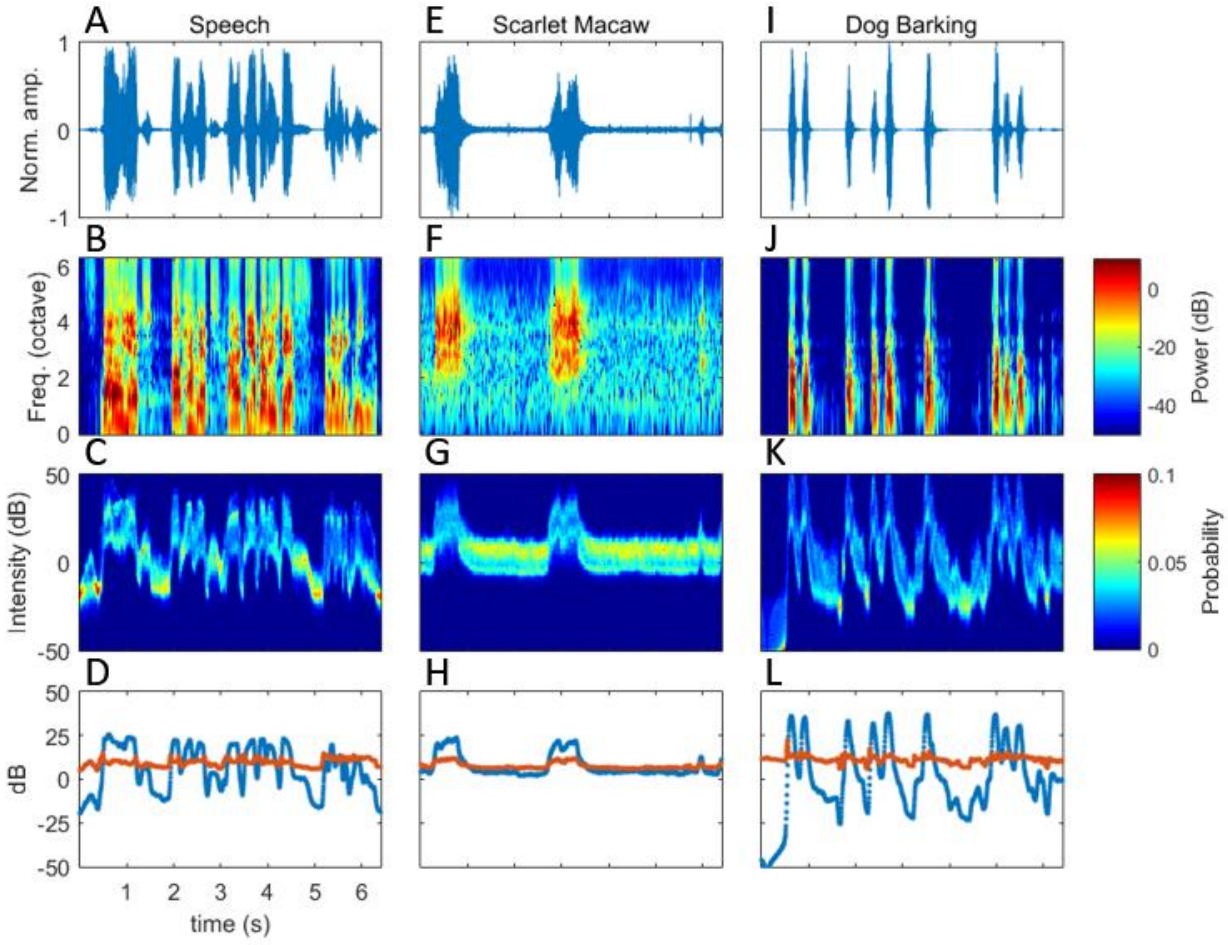


Figure 8: Comparison of three types of animal vocalizations using the same quantities as figure 6: the speech segment from figure 6 (A-D), a parrot, the scarlet macaw (E-H) and a studio quality track of a medium sized dog barking (I-L). Note the difference in both temporal and spectral modulations between the three, which manifest as slightly different contrast and intensity statistics vs time. Speech has complex, broadband spectral modulations, the scarlet

macaw vocalization is relatively band pass while the dog bark is slightly less broad band than speech, but more impulsive (less complex harmonic behavior).

Categorization was carried out at multiple hierarchical levels on the data, going from most general to most specific and the results are compared to determine if perceptually similar categories have objectively similar joint intensity-contrast statistics as measured by the classifier. Prior to examining the classifier's results, it is instructive to examine both the joint and marginal distributions of intensity and contrast for a handful of categories.

Figure 9 shows the joint (Figure 9B) and marginal distributions (Figure 9A is the marginal distribution of intensity and Figure 9C is contrast) for white noise, which serves as a control with known statistical properties to compare to empirical categories. White noise was generated using MATLAB's randn function with a sampling rate of 44.1 kHz, giving a total bandwidth of 22.05 kHz according to the Nyquist criteria. No initial filtering was done to this white noise, however, it is passed through the auditory filter bank in order to extract contrast and intensity statistics. 50 sequences of white noise were generated each 20 seconds long, giving a total of $4.41 \times 10^7$ samples at the original 44.1 kHz sampling rate and 20000 samples of intensity and contrast statistics, which should be sufficient to estimate the joint statistics. The joint intensity-contrast statistics of white noise are narrow in both dimensions. The mean intensity is 0 dB, which is due to the normalizations in equation 1 and 7. The minimum contrast is 5 dB, which is again a consequence of those normalizations. The marginal statistics are also quite narrow.

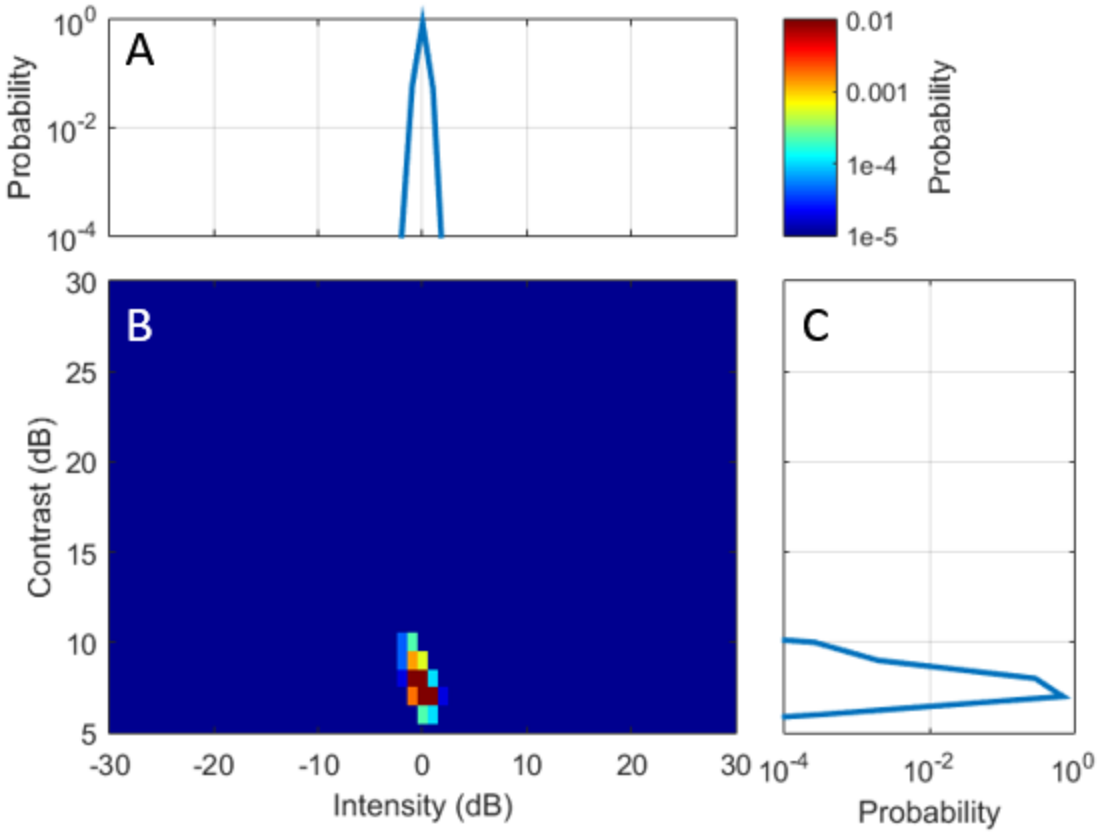


Figure 9: Joint (B) and marginal intensity (A) and contrast (C) distributions for white noise. White noise has a narrow distribution in both dimensions with a strong negative correlation.

Figure 10 shows the joint and marginal distributions for the speech category. The speech category is composed of 51 samples of speech from Shakespearian actors. Each sample is approximately 15-40 seconds in duration, with 10 total speakers, 6 of which are male and 4 of which are female. The total number of intensity-contrast samples is 19745. The distribution of speech is much broader in both the intensity and contrast dimension with a clear positive correlation coefficient.

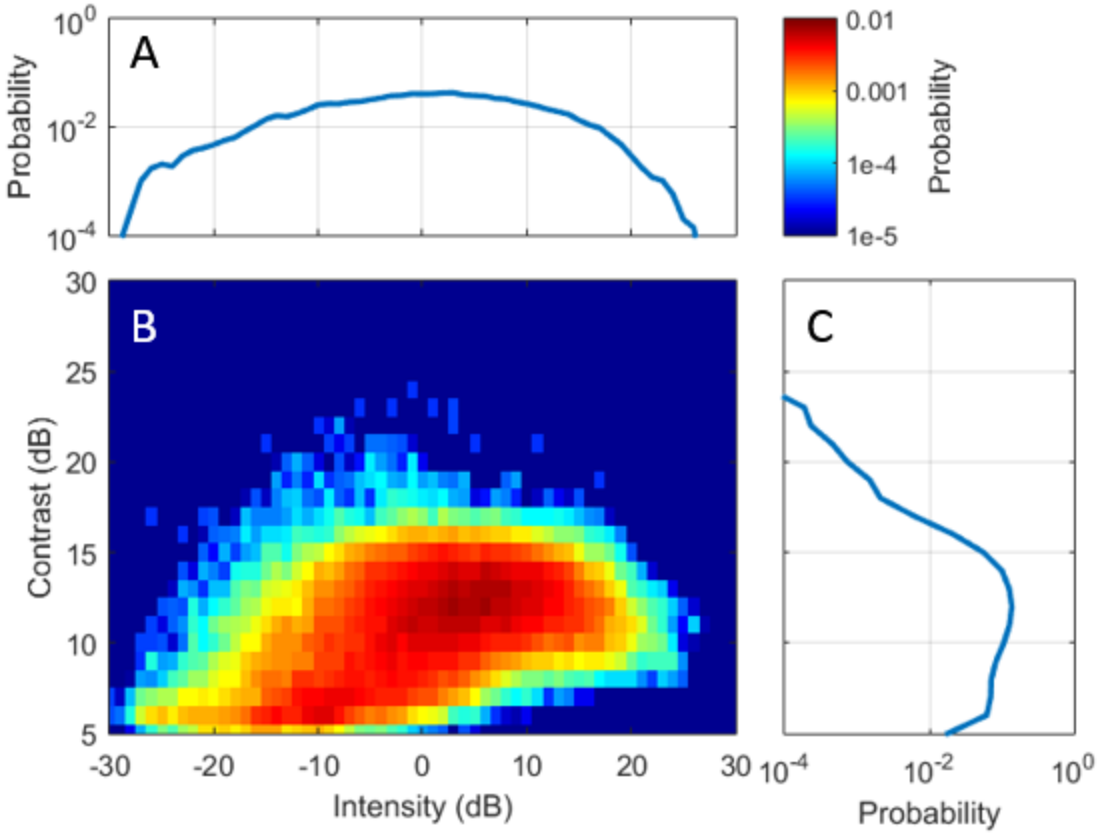


*Figure 10:* Joint (B) and marginal intensity (A) and contrast (C) distributions for speech. Speech has a large standard deviation in both the contrast and intensity dimensions with a strong positive correlation. The distributions are non-Gaussian, requiring additional parameters to fully describe.

Figure 11 shows the joint and marginal distributions for flowing water, a sound texture. The flowing water category is composed of 55 segments with a total of 38358 samples from various sources encompassing a wide range of water environments from babbling brooks to rivers, faucets and waterfalls. Flowing water has a much narrower distribution in both contrast and intensity than speech or other vocalizations.

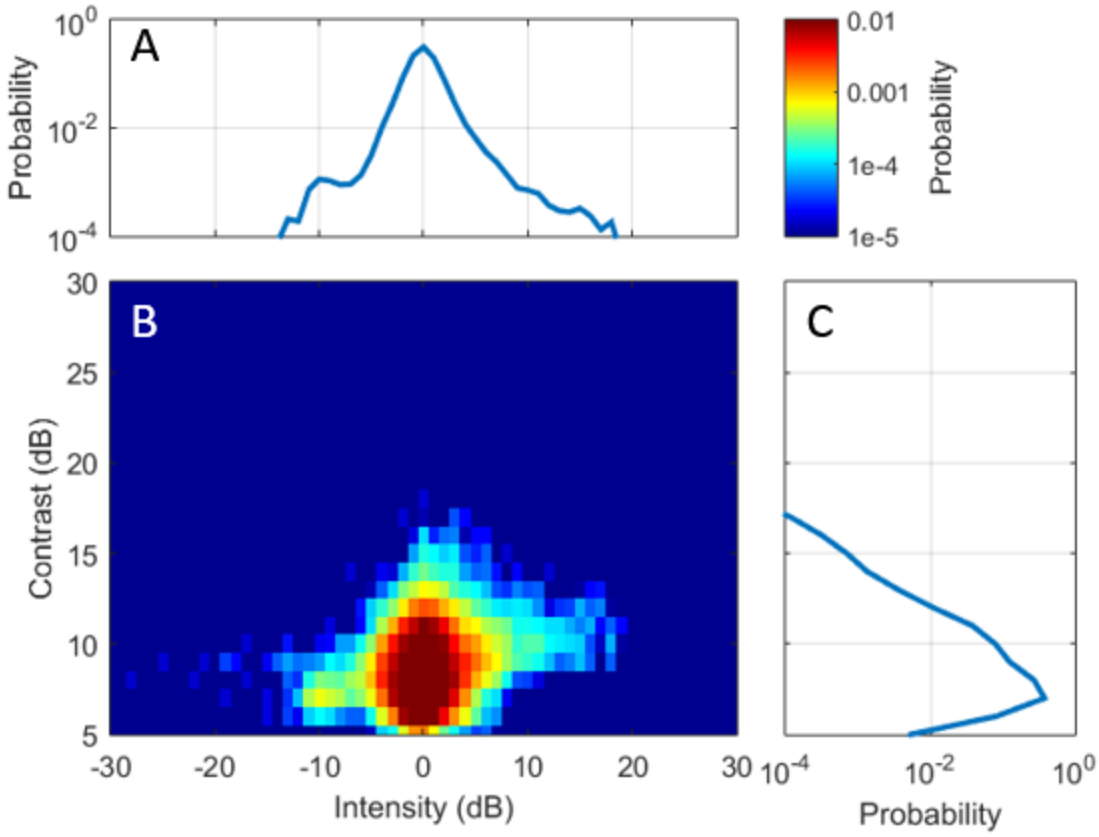


Figure 11: Joint (B) and marginal intensity (A) and contrast (C) distributions for flowing water. Flowing water has a narrower distribution in both dimensions than speech but substantially wider than white noise. The two dimensions are mostly uncorrelated.

Figure 12 shows the joint and marginal distributions for parrot vocalizations. This distribution is one of the largest categories in the database and contains 587 segments from 160 different species of parrots. An important question to resolve in this thesis is the degree to which different species can be grouped together and still have reasonable performance of the classifier. In the case of parrots, it appears that they are fairly homogeneous and can be grouped and still have reasonable classifier performance even against other categories of birds. The distribution of intensity-contrast statistics for the parrot category is similar to that of speech, with a broad distribution in both the contrast and intensity dimension and a positive correlation between

intensity and contrast. The parrot category's distribution is still perceptually quite different from the speech distribution, with a higher variance in both the intensity and contrast dimension.

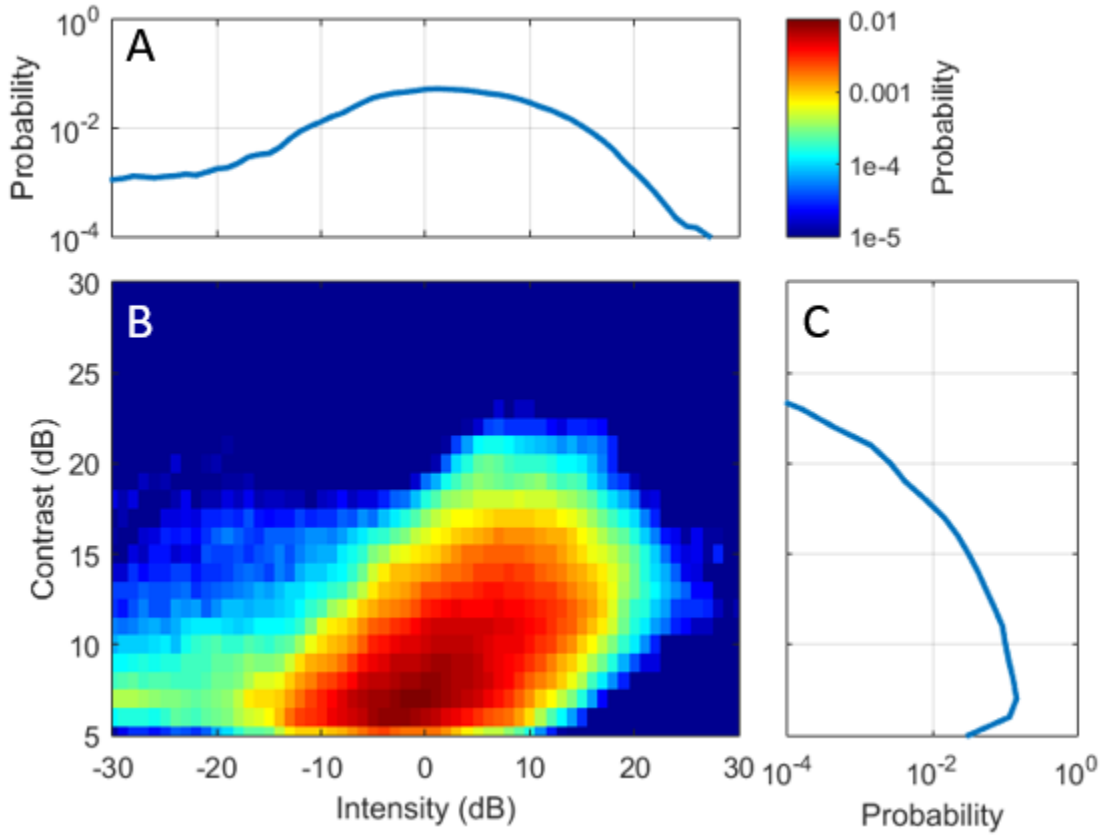


Figure 12: Joint (B) and marginal intensity (A) and contrast (C) distributions for parrot calls. Parrot calls have a wide distribution in both contrast and intensity, similar to speech, but with substantially more probability mass above 15 dB contrast.

Examination of the 2nd order moments of the individual distributions yields an interesting result. Vocalizations appear to have a higher variance in both the intensity and contrast dimensions then ensemble or background sounds (Figure 13) as well as a stronger correlation between the two dimensions. The background sound categories, flowing water, rain, waves, wind and internal combustion engine (I.C.E.) have near zero correlation, with the maximum absolute correlation of any of these categories equal to 0.16 and an absolute mean of 0.10. The vocalization

categories (canine, chirping insects, frogs, new world primates, parrots, ducks and geese, and speech) all have a positive correlation with a maximum of 0.47 (speech) and a mean of 0.33. The minimum value is for the frogs (0.14) where many of the samples are themselves ensembles of many frogs rather than a single frog vocalization.

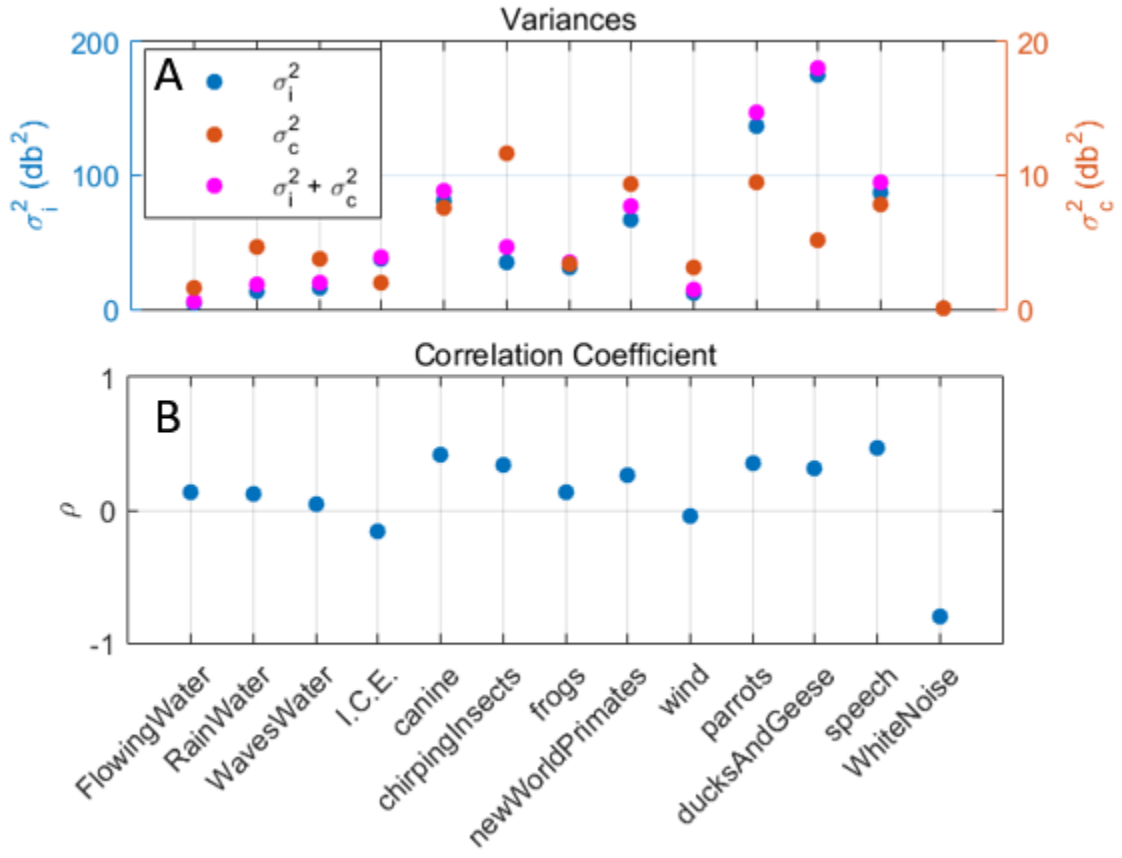


*Figure 13:* Variance (A) and correlation (B) coefficients for the distributions of intensity and contrast for all 13 categories used in the analysis. Note the different scales used for the variance of intensity ($\sigma_i$, blue dots, left ordinate) and contrast ($\sigma_c$, red dots, right ordinate). The sum of the two variances (magenta dots) use the left ordinate. Background sounds have the smallest absolute correlation and variance, while vocalization sounds have much larger correlation coefficients and variance. Bird calls (parrots and ducks and geese) have the highest variability, while speech has the highest correlation between the intensity and contrast dimensions.

## 3.2 Classifier performance:

Classification was done across a variety of conditions which varied the following: the number of categories to be compared, the level of hierarchy compared within the categories themselves (for example, comparing vocalizations to non-vocalizations and then comparing subsets of each such as human speech vs all other vocalizations), the sound duration used to generate the statistics, whether the joint distribution of intensity-contrast was used or just one marginal distribution. All tests were repeated and compared with the two outlined cross validation schemes.

Figure 14 shows the confusion matrix for two cross validation schemes for the classifier output across all categories with more than 20 available sound segments of "good" quality. This yields 13 total categories including groups that are hierarchically distinct, for example speech and flowing water as well as subgroups with a common higher level category, such as different water sounds: flowing water, rain and waves. The confusion matrices show performance when 128 consecutive intensity-contrast pairs are used for the classification. Performance is statistically significant (chance performance = 7.69%) for all categories and for both cross validation methods. Because of the large number of validation samples used in this analysis, almost every result is statistically significant at a high level and so significance will not typically be reported. This is common with data mining applications using large data sources [32]. There is confusion above chance for a number of categories, most of which involve the three worst categories: chirping insects, frogs and new world primates. Additional confusions occur between flowing water and rain water, flowing water and waves, waves and wind, canine vocalizations and speech and the two bird categories, parrots and anatidae, which is a phylogenic categorization that encompasses all species of ducks and geese.

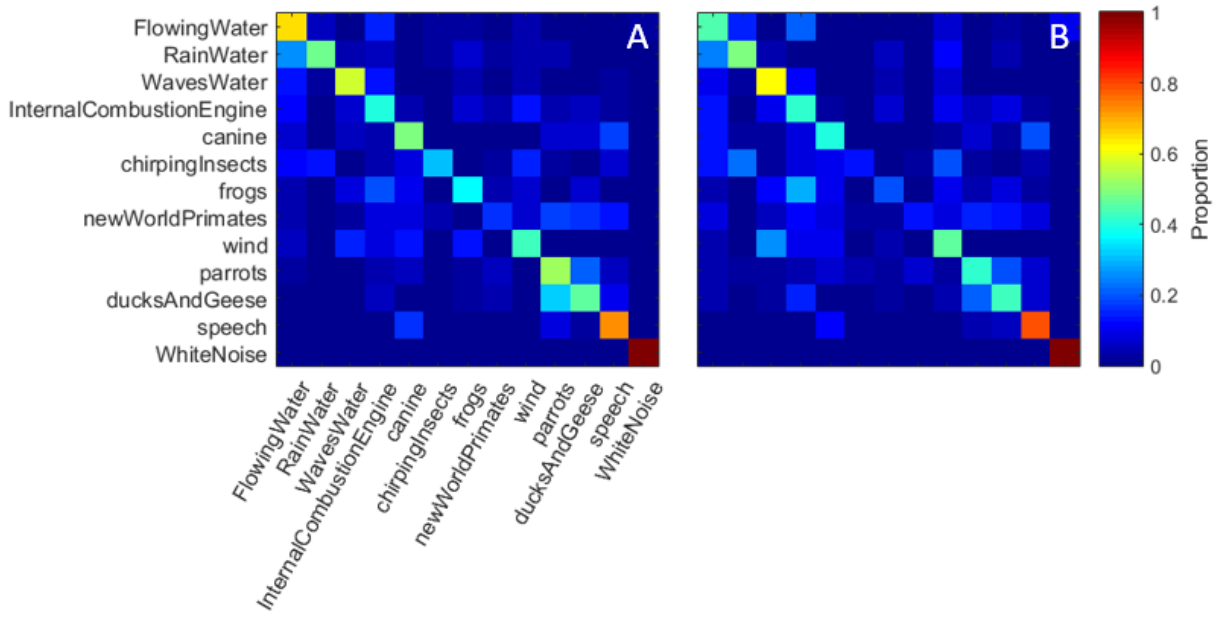


*Figure 14:* Confusion matrices showing the performance for the two cross-validation techniques, two-fold (A) and leave-one-out (B) for 128 validation points. Performance is similar, although the two-fold cross validation has better performance on average.

Figure 15 shows the proportion correct vs. the number of samples for the two different cross validation methods. In terms of mean performance of the non-white noise sounds (black line) two-fold cross validation (Figure 15A) yields superior results to the leave-one-out cross validation (Figure 15B), with mean performance about 5% better for the two-fold cross validation (45.5% for leave-one-out vs 50.2% for two-fold). The performance of individual categories shows more variability between the two cross validation schemes, which can be seen by examining performance for the flowing water or frogs categories. In general, however, the performance of the two cross-validation schemes show qualitatively similar patterns with increasing sound duration for most of the categories. White noise is excluded from the calculation of mean performance as it is correctly classified 100% of the time for all but the shortest durations.

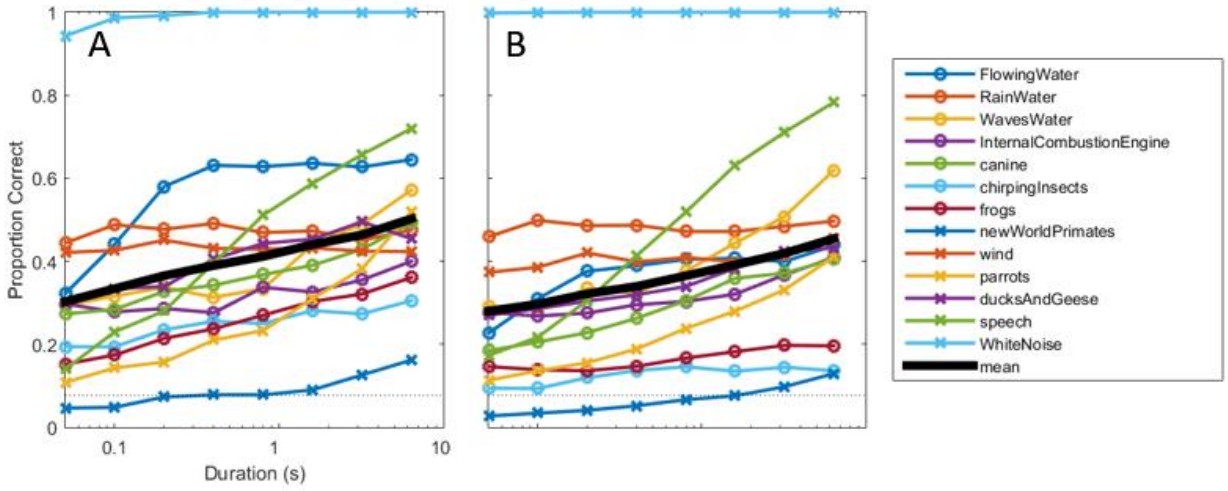


*Figure 15:* Proportion correct vs duration of the validation segments for the two cross-validation methods, two-fold (A) and leave-one-out (B). Performance increase, on average (thick black line), by about 15% as the number of samples increases. For the longest duration (6.4 seconds), classification performance is above chance for all categories.

Figure 16 shows confusion matrices for 4 sound durations, 0.05, 0.2, 0.8 and 3.2 seconds, corresponding to 1, 4, 16, and 64 intensity-contrast samples using the two-fold cross validation method. Background type sounds seem to have the best performance when there is a low number of samples, likely due to the more concentrated probability mass functions (pmf) (as shown in Figure 13), while a larger number of samples is required for performance above chance with most of the vocalizations.

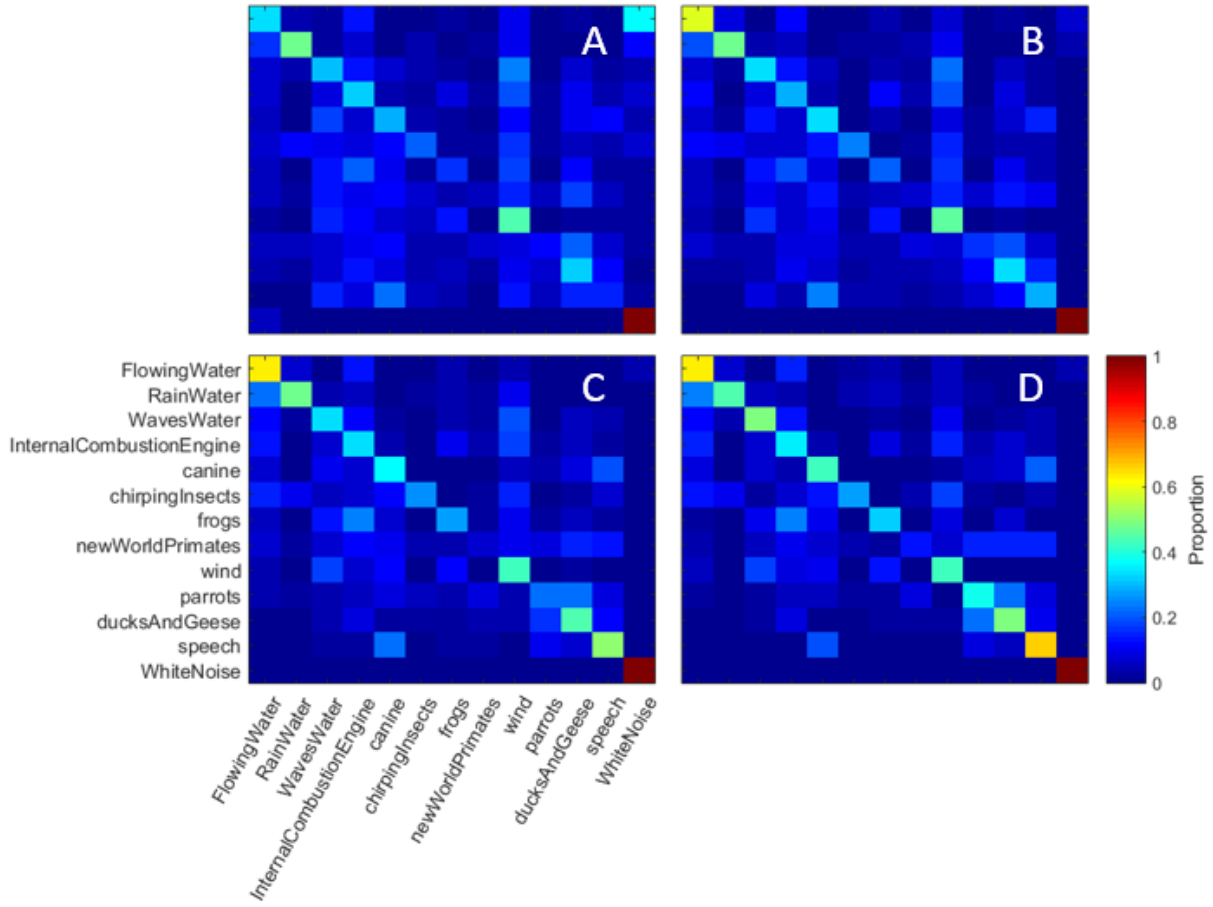


*Figure 16:* Confusion matrices for four different validation segment lengths, A). 1 sample (0.05 s), B). 4 samples (0.2 s), C). 16 samples (0.8 s) and D). 64 samples (3.2 s) using two-fold cross-validation. At lower segment lengths, performance is worse and confusions are more randomly distributed, while at longer lengths, confusions appear to become more systematic, for example, between frogs and I.C.E. or the two types of birds.

Three of the categories, chirping insects, frogs and new world primates, have substantially worse performance than the other categories. These categories were removed to assemble a "best subset" of categories for the remainder of the analysis. Those three categories, particularly new world primates, encompass some subjectively rather disparate calls from multiple species and it may not be appropriate to group them together at such a high level. Similarly, frog and insect calls are difficult to tell apart on many tracks, often co-occurring and usually have other

background noise on the tracks (such as flowing water that is part of frog habitats) that may conflict with other categories.

Figure 17A&B show the confusion matrices using both cross validation methods for this best subset (analogous to Figure 14) and Figure 17C&D show the performance vs. sound duration for the two cross validation methods (analogous to Figure 15). This figure illustrates that certain subjective or even taxonomic classifications may not be sufficient to group certain types of sounds. Alternately, there may be other confounds within a particular category that require substantially more sophisticated analysis to isolate and control. By comparison to Figure 15, the maximum mean performance for this best subset is improved by 9.4% and 10.7% relative to the full set of 13 categories for the two-fold and leave-one-out cross-validation, respectively.

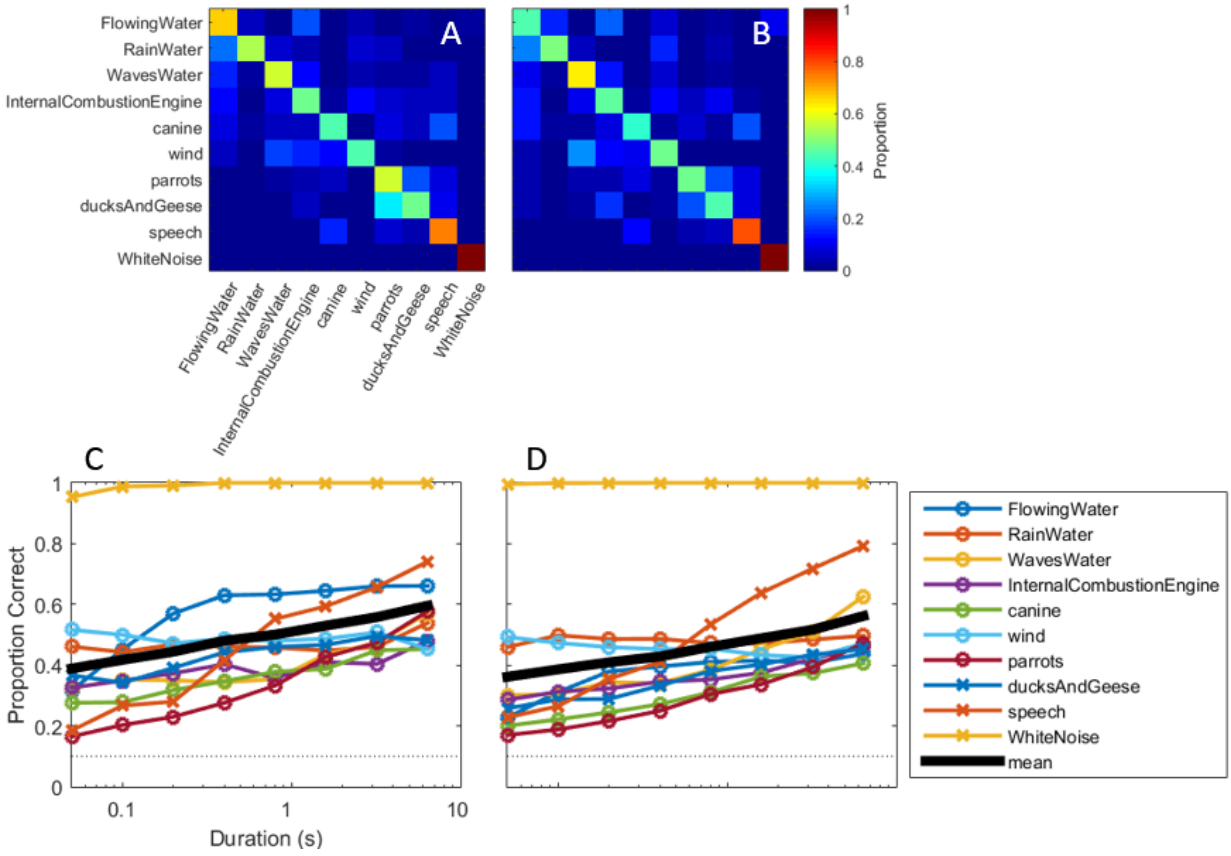


*Figure 17:* Confusion matrices and proportion correct vs duration for the best subset of 10 categories using the two cross validation methods, two-fold (A and C) and leave-one-out (B and D). These categories all perform near or above 40% correct at the longest validation segment duration.

It is also interesting to evaluate the classifier's performance when using the marginal distributions of intensity and contrast like those shown in Figure 9-Figure 12. The classifier was run on the best subset of categories using only the two-fold cross validation scheme (as the leave-one-out cross validation scheme seems to consistently perform at a rate about 5% worse than the two-fold scheme). Figure 18 shows the proportion correct vs duration for intensity only (Figure 18A) and contrast only (Figure 18B). Mean performance overall performance is similar between the two at all durations, with a peak performance of about 45% in both cases, however the particular categories that are best classified is quite different between the two statistics.

Intensity is best for classifying speech, water sounds and anatidae, while contrast performs best for the I.C.E., canine and wind, all of which have a correct classification rate at least 20% greater using contrast statistics rather than intensity statistics. Comparing the results using marginal distributions (both have equal mean percent correct performance of 45%, to within 0.01%) to the joint distribution, there is a 14.5% increase in performance by using the joint distributions.

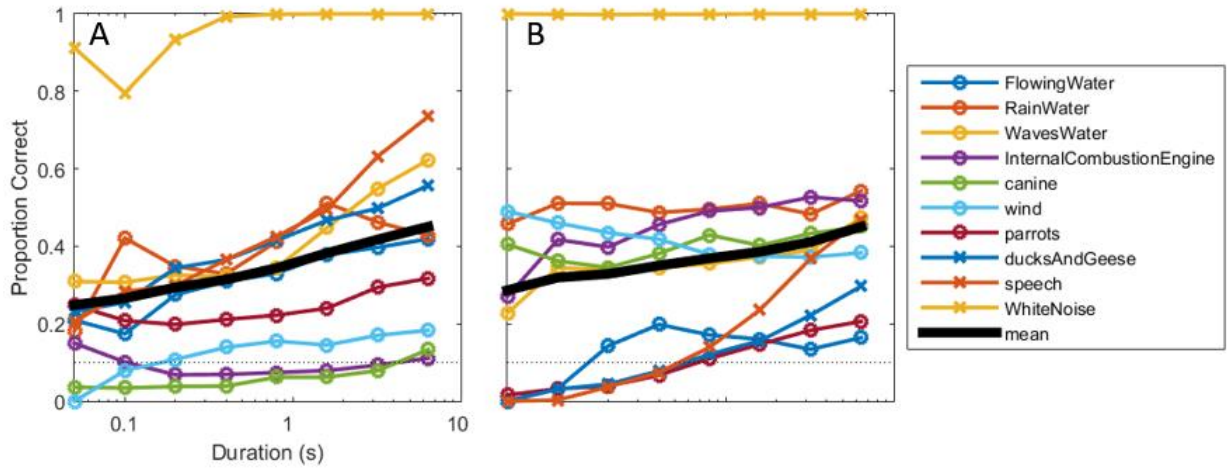


*Figure 18:* Proportion correct for the best subset using only the marginal distributions of intensity (A) and contrast (B) using only the two-fold cross-validation scheme. Performance is substantially degraded for some categories (e.g. wind, canine and I.C.E.) when using the marginal distribution of intensity while other categories are less affected (e.g. speech, waves or anatidae). A different set of categories is detrimentally affected when using the marginal distribution of contrast (e.g. parrots, anatidae and flowing water). Surprisingly, joint distributions only seem to improve the average performance by about 14%.

Figure 19 shows a scatter plot of proportion correct performance when using contrast or intensity only for each category in the best subset using 6.4 second validation segments. Categories which are above the dotted line perform better when using just contrast statistics (I.C.E., canine, wind and rain), while those below the line perform better when using just intensity statistics (flowing water, waves, parrots, anatidae and speech). The mean performance (black dot) lies almost

exactly on the dotted line, indicating that neither statistic is more valuable to the overall performance of the classifier than the other.

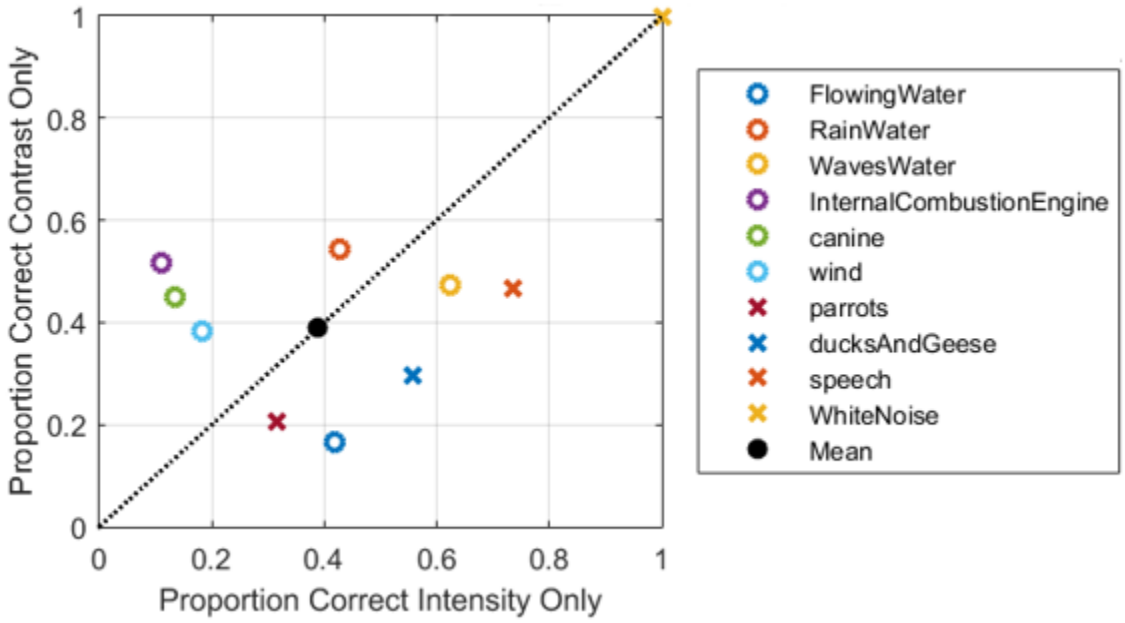


*Figure 19:* Scatter plot of the proportion correct for each category of the best subset using the marginal distributions of intensity and contrast to perform the classification. Validation segment length is 6.4 seconds. Dotted line indicates that the classifier performs equally when using either marginal distribution. Categories above the dotted line (I.C.E., canine, wind and rain) have better classification performance when using contrast only, while those below (flowing water, waves, parrots, anatidae and speech) perform better when using intensity only. Mean performance (excluding white noise) is almost exactly on the dotted line, indicating that neither statistic is more favorable for classification.

A final set of results compares higher levels within the class hierarchy to test if animal vocalizations and the non-animal sounds in the database are fundamentally different in terms of contrast and intensity and to test the same for speech compared to other animal vocalizations. Figure 20 shows the proportion correct for animal vocalizations vs non-animal sounds using the two-fold cross validation method. For each category, 25 segments are selected randomly (with replacement if there's not enough segments in a particular category) and then the vocalization classes (insects, frogs, new world primates, canine, parrots, anatidae and speech) are grouped and

the same is done for the non-animal sounds (flowing water, rain, waves, wind and ICE). The model distribution for each category is constructed as normal. The classifier has an 80% correct classification rate at the longest sound segment duration, with the number of correct classifications roughly constant at about 80% correct for the non-animal sounds and steadily increasing from 63% to 85% correct with duration for the vocalization sounds.

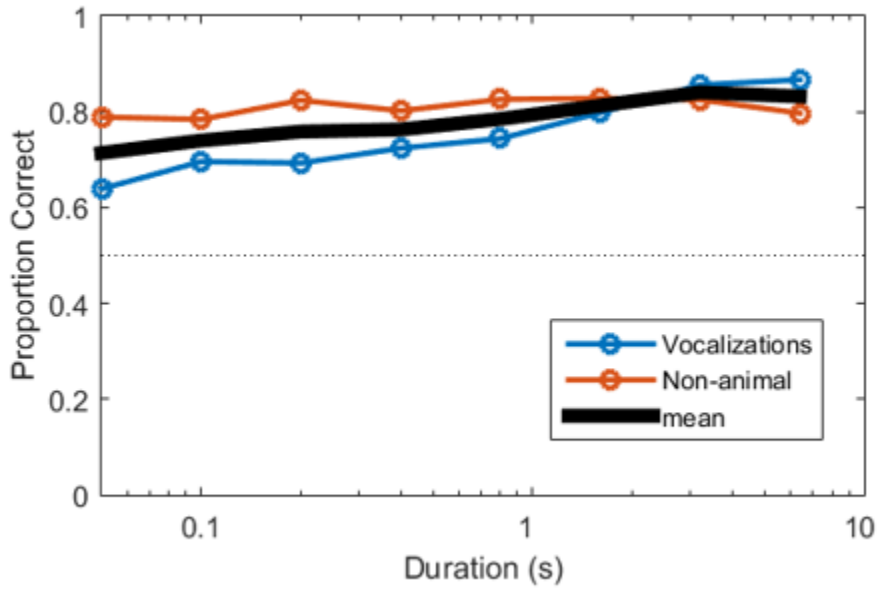


*Figure 20:* Proportion correct vs duration for animal vocalizations vs. non-animal sounds (water sounds, I.C.E. and wind). Maximum performance is around 80%. Vocalizations have a higher misclassification rate at low duration, likely because of the periods between vocalizations, which often have background sounds similar to the categories used for the non-animal category.

Figure 21 shows the same experiment comparing speech to all other forms of animal vocalization (insects, frogs, new world primates, canine, parrots and anatidae). The classifier has an overall performance of 66% at the shortest duration, increasing to 80% at the longest duration. There is a substantial and systematic difference in misclassification rate between the two categories, which

likely emerges because speech is a relatively homogeneous category in comparison to all the other vocalizations.

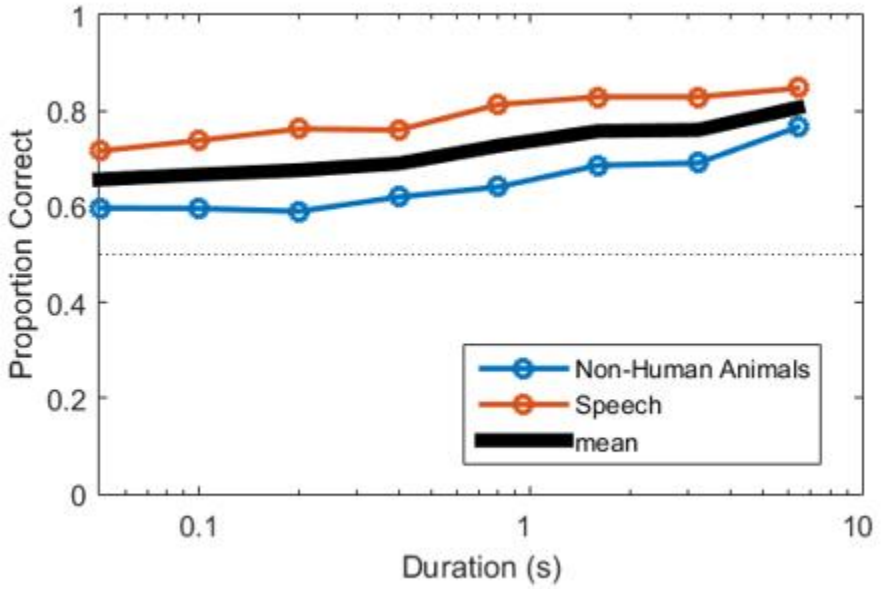


*Figure 21:* Proportion correct vs duration for speech vs. non-human vocalizations (canine, frogs, insects, both types of birds and non-human primates). Maximum average performance is around 80% with a large difference at all durations in misclassification rate between the two categories.

# 4 Discussion

One of the primary confounds in this experiment is in isolating the correct portion of sounds from natural sound tracks. Consider the example vocalizations from Figure 8. The macaw calls are temporally spread out, with well over a second in between each call. A first instinct might be to specifically identify and isolate individual calls and then concatenate them to get a longer sound sample and build a model that way. This could be achieved by using an amplitude

threshold to identify the calls and then applying a smooth window, such as a flattop window, to remove discontinuities due to the concatenation. This may not be the best approach as the natural cadence of calls may be of fundamental importance to this classifier and any windowing will directly affect the amplitude statistics. Compare this to the well-studied example of speech. The spectrotemporal characteristics of speech have been reported [34] [36] and it would be expected that these characteristics would directly impact the performance of a classifier based on amplitude statistics since the modulation power spectrum is obtained from the Fourier transform of the spectrogram used to calculate the amplitude statistics. The findings in this report, that speech is the most correctly classified sound category using just contrast and intensity statistics, suggests that there are fundamental spectrotemporal features of speech that make it sufficiently different from other signals. These likely include the temporal modulations, especially those associated with syllables, which have been reported to be fluctuate at rates between 2-5 Hz [37] and based on the results of [34], contain a large amount of the speech signal energy. Using the syllabus rate as an analog for the call rate of animal vocalizations, it can be argued that the call rate may also be important for a classifier based strictly on amplitude statistics.

If instead each syllable or word was isolated and concatenated, removing the natural 2-5 Hz cadence, then it stands to reason that it would be harder to classify natural speech against the speech that has been artificially truncated and pieced together to form the model. Returning to the scarlet macaw vocalization, it now becomes a challenge to determine when the pause between vocalizations is a fundamental feature that is necessary for classification versus a nuisance caused by an “uncooperative” animal subject. An intermediate example is the dog bark from Figure 8, the closely spaced barks around the 5 second mark (along with basic common sense) suggest that a dog has some maximum rate of barking and that there is a natural pause

between each bark that could be quite relevant for classification, much like the natural cadence of speech.

One method that was used to prevent misclassification due to pauses in more dynamic signals was the narrow Gaussian filter applied to the model distributions. Without any method of generalizing the model distributions, there would effectively be a bias towards categories with enough data to ensure that the entire model distribution was non-zero. Simpler methods of avoiding the problem of zero-valued bins in the model distribution, such as using a constant minimum probability mass, bias against animal vocalizations where there might be a relatively long duration between calls. Most of the probability mass would be associated with the background sounds on the track, which are typically centered around 0 dB intensity with low contrast. The tails of the Gaussian filter drop off rapidly, ensuring that there is low probability mass in bins that are not adjacent to any of the model data. This ensures that the a stray sample appearing outside of the appropriate model distribution does not cause a zero valued posterior distribution, while ensuring that there is sufficiently low probability mass in bins that are far from the unfiltered model. Fitting a Gaussian mixture model to the model distributions was attempted with similar, but inferior, results.

Another method that was used to prevent misclassification due to pauses in more dynamic signals was the use of consecutive samples in both validation and modeling. Under the independence assumption of the naïve Bayesian classifier, it should be possible to draw samples randomly from the sound segments. Instead, using consecutive samples ensures that the natural cadence of sounds has additional influence on the classification. Consider the scarlet macaw vocalization, which has short, highly dynamic periods of contrast and intensity followed by near constant periods of contrast and intensity. Compare this to a less dynamic sound, like white

noise, which has near constant contrast and intensity throughout. The large number of non-dynamic samples might cause a misclassification to the less dynamic sound, but even just one of the dynamic samples will favor the correct posterior distribution heavily. Using consecutive samples helps ensure that the MAP rule will correctly classify sounds with fundamental temporal modulations.

## 4.1 Discussion of classifier

Most confusions seem to occur between categories that have some similarities. For example, there are the obvious confusions between different types of water sounds. These could be due to the similar physics that are hypothesized to be involved in the creation all water sounds [38] [39] [40] [41], namely the vibrations associated with trapped gas bubbles that occur when water droplets collide. Geffen et al. (2011) has also shown that water sounds exhibit "scale-invariance not only within spectral channels, but also across the full spectral bandwidth". Geffen et. al. were able to create perceptually realistic artificial water sounds from a sum of variable frequency- and time-scale random chirp stimuli. The number of summed chirps determined the quality of the water sound, with a range of qualities from a dripping tap to a roaring stream as the number of summed chirps increased. Their approach was similar to more direct physical models [39] [40], namely in the usage of damped oscillator type stimuli to model the effect of the gas bubble and water droplet interactions, but with more perceptual salience due to a focus on mirroring natural sound statistics rather than strictly physics-based modeling.

Water sound confusions could also be due to the fact that the different types of water sounds are inseparable to some extent. A recording of falling rain will almost certainly have the noise of

flowing water on the track as the rain pools and flows to the point with the lowest potential energy. Similarly, waves are just an atypical form of flowing water governed by tidal forces and wind rather than primarily the earth's gravity as would be the case for a river or stream. Based on a similar observation, waves likely confuse with wind noise because waves are driven by wind, and so almost all wave tracks have some amount of wind on them, even if the track quality is good and the wave noise is well isolated from loud gusts of wind.

Another interesting confusion is between canine vocalizations and human speech. The canine category includes dog barks from multiple breeds of dog, wolf howls and barks and miscellaneous canine calls such as the yelps of foxes. The confusion might be explained because many canine sounds are impulsive, i.e. short duration, broadband sounds produced at a rate that is typically less than 5 Hz (Figure 8, 3$^{rd}$ column). Similarly, speech is also a broadband sound with complex harmonic structure and a syllable rate less than 5 Hz [36] [34] [37]. While the harmonic structure of the two is quite different, they may become similar after the data reduction involved in producing the intensity-contrast statistics.

The final confusion of note occurs between two rather disparate families (in the biological taxonomy sense) of birds: parrots and anatidae. Parrots primarily reside in tropical climates, with most species native to Central and South America. Ducks and geese, on the other hand, are native to North America. Their calls are also quite distinct perceptually. Regardless of these perceived differences, one of the most prominent confusions between classes is between ducks and parrots. A hypothesis is that the calls are not quite as distinct as one might naively think. Duck quacking and goose honking are short, burst like calls with fairly low rate. Similarly, most parrots make short chirping, squeaking or screeching calls which have similar rates to quacking and honking.

For animal vocalizations, the rate and duration of the calls may be one of the more important factors influencing the naïve Bayesian classifier used in this report, which suggests that additional metrics may be required to fully separate different types of animal vocalizations based on a more detailed representation of the spectrogram. For example, channel correlations improve the percept of synthetic sounds [5], and so may also improve the performance of a sound category classifier.

## 4.2 Comparison to behavior and physiology

Numerous studies have shown sensitivity to intensity and contrast statistics at most levels of the auditory system (auditory nerve: [8], midbrain: [6] [7] [9] [10], auditory cortex: [11] [12] [13] [14]). One of the chief response mechanisms seems to be adaptation to the sound statistics [9] [10] [8]. Sensitivity to sound level statistics has also been shown psychophysically [5] [42]. The auditory system has two basic functions, localization and sound identification (the so called what? and where? questions [19]). Localization has been well studied and the primary sound features used for localization appear to be primarily binaural cues, interaural level and timing differences, not monaural sound level statistics [19]. Based on these facts, it stands to reason that neuronal and behavioral sensitivity to sound level statistics are likely utilized for sound classification tasks. This research suggests that there is a substantial amount of information just in the channel averaged, time-varying, mean and standard deviation of the intensity envelope (intensity and contrast statistics) for simple machine learning algorithms to do substantially better than chance classification of sounds. There are many real world limitations of this study in terms of number of sound segments available, quality of sound segments, and confounds

associated with the realistic properties of sounds, such that there is significant room for refinement, not only of the classifier and database but also recursive refinement that could result from a better understanding of the behavior and neuroscience. Additional study will be required to determine if there are neurons specifically associated with sound classification and if those neurons receive inputs from neurons sensitive to sound level statistics.

Psychophysical sound recognition tasks of comparable difficulty to the ones presented in this paper are lacking in the literature. In particular, the general classification of arbitrary sound classes of variable duration has not been addressed. In contrast, much of the literature has focused on simple sounds such as tones and musical notes [43] [44] and vowels [45]. However, specific sound recognition tasks of more complex categories, such as identifying instruments [44] [46], human speakers [47] [48] and popular songs [49] do exist in the literature. Once again, however, there are limitations to the comparison, as most of the literature is conducted with substantially shorter sound segments than the 6.4 second maximum duration considered here.

Categorization performance has been shown to improve with stimulus duration, especially for more complex sound categories such as identifying a human speaker, instrument or song. Schellenberg et. al. (1999) showed that for listeners attempting to identify popular songs, identification performance above chance improved by 8% when using a 200 ms sound segment compared to a 100 ms sound segment. Schweinberger et. al. showed that for listeners attempting to identify famous voices with stimulus lengths from 0.25 seconds to 2 seconds, “Voice recognition improvements with stimulus duration were with a growth function. Gains were most rapid within the first second and less pronounced thereafter.” Martin (1999), in his PhD thesis, showed that human listeners were able to correctly identify an instrument from a single tone at a rate of45.9%, which improved to 66.9% correct identification with a 10 second musical segment.

Similarly, listeners' performance when trying to identify the correct class of instruments (strings, woodwinds, brass, etc.) improved from 91.7% with the single note to 96.9% with a 10 second musical segment.

### 4.3 Next steps

An important next step in this research would be to compare the performance of this classifier when using lossy compressed audio files, such as those compressed with the mp3 standard. In the age of the internet, lossless sound files are much harder to locate than compressed audio which is often freely available on the internet. Websites such as YouTube have expansive databases of user tagged multimedia content. Many of these videos are also covered by fair use which would allow a human researcher to add segments of the content to the sound database. The next step would first involve comparing the outputs of the auditory model using a reconstruction of an audio file compressed with a lossy coding scheme (similar to what the ear would hear when playing and mp3 file through speakers) and then compare that to the auditory model output for the lossless version of the same audio file. If they are sufficiently similar, then the sound database and classification task could be furthered using sound files with lossy compression as well as lossless files.

## 5 References

[1] H. B. Barlow, "Summation and Inhibition in the Frog's Retina," *J. Physiol.,* vol. 119, pp. 69-88, 1953.

[2] H. B. Barlow, "Single units and sensation: A neuron doctrine for perceptual psychology?," *Perception,* vol. 1, pp. 371-394, 1972.

[3] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am.,* vol. 4, no. 12, pp. 2379-2394, 1987.

[4] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature,* vol. 381, no. 13, pp. 607-609, 1996.

[5] J. H. McDermott and E. P. Simoncelli, "Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis," *Neuron,* pp. 71, 926-940, 2011.

[6] F. A. Rodriguez, C. Chen, H. L. Read and M. A. Escabi, "Neural Modulation Tuning Characteristics Scale to Efficiently Encode Natural Sound Statistics," *J. Neuroscience,* pp. 15969-15980, 2010.

[7] M. A. Escabi, L. M. Miller, H. L. Read and C. E. Schreiner, "Naturalistic Auditory Contrast Improves Spectrotemporal Coding in the Cat Inferior Colliculus," *J. Neuroscience,* pp. 11489-11504, 2003.

[8] B. Wen, G. I. Wang, D. Isabel and B. Delgutte, "Dynamic Range Adaptation to Sound Level Statistics in the Auditory Nerve," *J. Neuroscience,* pp. 13797-13808, 2009.

[9] I. Dean, N. S. Harper and D. McAlpine, "Neural population coding of sound level adapts to stimulus statistics," *Nature Neuroscience,* vol. 8, pp. 1684-1689, 2005.

[10] I. Dean, B. L. Robinson, N. S. Harper and D. McAlpine, "Rapid Neural Adaptation to Sound Level Statistics," *J. Neuroscience,* pp. 6430-6438, 2008.

[11] D. L. Barbour and X. Wang, "Contrast Tuning in Auditory Cortex," *Science,* vol. 299, pp. 1073-1075, 2003.

[12] I. Nelken, Y. Rotman and O. B. Yosef, "Responses of auditory-cortex neurons to structural features of natural sounds," *Nature,* vol. 397, pp. 154-157, 1999.

[13] N. C. Rabinowitz, B. D. Willmore, J. W. Schnupp and A. J. King, "Contrast Gain Control in Auditory Cortex," *Neuron,* vol. 70, pp. 1178-1191, 2011.

[14] N. C. Rabinowitz, B. D. Willmore, J. W. Schnupp and A. J. King, "Spectrotemporal Contrast Kernels for Neurons in Primary Auditory Cortex," *J. Neuroscience,* vol. 32, pp. 11271-11284, 2012.

[15] B. Ni, Y. Song and M. Zhao, "YouTubeEvent: On large-scale video event classification," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, Barcelona, 2011.

[16] Z. Wang, M. Zhao, Y. Song, S. Kumar and B. Li, "Youtubecat: Learning to categorize wild web videos," in *CVPR*, 2010.

[17] The MITRE Corporation, "Computer Vision Offers New Tools for Searching the Video Explosion," January 2015. [Online]. Available: https://www.mitre.org/publications/project-stories/computer-vision-offers-new-tools-for-searching-the-video-explosion.

[18] Oracle Corporation, "Oracle Database SQL Reference 10.1," December 2003. [Online]. Available: https://docs.oracle.com/cd/B12037_01/server.101/b10759.pdf.

[19] K. L. Tremblay and R. F. Burkard, Translational Perspectives in Auditory Neuroscience Normal Aspects of Hearing, San Diego, CA: Plural Publishing, 2012.

[20] C. A. Kuwada, B. Bishop, S. Kuwada and D. O. Kim, "Acoustic recordings in human ear canals to sounds at different locations," *Otolaryngology -- Head and Neck Surgery,* pp. 615-617, 2010.

[21] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. I: Stimulus synthesis," *J. Acoust. Soc. Am.,* vol. 85, pp. 858-867, 1989.

[22] J. Blauert, Spatial Hearing, Revised Edition, Cambridge, MA: The MIT Press, 1997.

[23] T. Irino and R. D. Patterson, "Temporal asymmetry in the auditory system," *J. Acoust. Soc. Am.,* vol. 99, pp. 2316-2331, 1996.

[24] H. Fletcher, "Auditory Patterns," *Rev. Mod. Phys.,* vol. 12, pp. 47-65, 1940.

[25] E. Zwicker, G. Flottorp and S. S. Stevens, "Critical Band Width in Loudness Summation," *J. Acoust. Soc. Am.,* vol. 29, pp. 548-557, 1957.

[26] R. D. Patterson, M. Allerhand and C. Giguere, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Am.,* vol. 98, pp. 1890-1894, 1995.

[27] P. X. Joris, C. E. Schreiner and A. Rees, "Neural Processing of Amplitude-Modulated Sounds," *Physiol Rev,* pp. 541-577, 2004.

[28] R. M. Roark and M. A. Escabi, "B-Spline Design of Maximally Flat and Prolate Spheroidal-Type FIR Filters," *IEEE Transactions On Signal Processing,* vol. 47, pp. 701-716, 1999.

[29] B. A. Wandell, Foundations of Vision, Sunderland, MA: Sinauer Associates Inc., 1995.

[30] A. Michelson, Studies in Optics, Chicago, IL: U. of Chicago Press, 1927.

[31] Y. Bar-Shalom, X. Rong Li and T. Kirubarajan, Estimation with Applications to Tracking and Navigation, Hoboken, NJ: Wiley-Interscience, 2001.

[32] D. Larose and C. Larose, Data Mining and Predictive Analytics, Hoboken, New Jersey: John Wiley & Sons, Inc., 2015.

[33] A. Papoulis and S. U. Pillai, Probability, Random Variables and Stochastic Processes, New York, NY: McGraw-Hill, 2002.

[34] N. C. Singh and F. E. Theunissen, "Modulation spectra of natural sounds and ethological theories of auditory processing," *J. Acoust. Soc. Am.,* vol. 114, pp. 3394-3411, 2003.

[35] Sony Corporation, Performer, *Sony Pictures Sound Effects Series Volumes 1-10.* [Sound Recording]. 2003.

[36] T. M. Elliott and F. E. Theunissen, "The Modulation Transfer Function for Speech Intelligibility," *PLOS Computational Biology,* 2009.

[37] J. M. Pickett, The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology, Boston, MA: Pearson, 1998.

[38] P. Guyot, J. Pinquier and R. Andre-Obrecht, "Water Sound Recognition Based On Physical Models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[39] K. van Den Doel, "Physicallybased models for liquid sounds," in *Proceedings of the ICAD 04 - Tenth Meeting of the Intervational Conference on Auditory Display*, Sydney, 2004.

[40] T. G. Leighton and A. J. Walton, "An experimental study of the sound emitted from gas bubbles in a liquid," *Eur. J. Phys.,* vol. 8, pp. 98-104, 1987.

[41] T. G. Leighton, M. Wilkinson, A. J. Walton and J. E. Field, "Studies of non-linear bubble oscillations in a simulated acoustic field," *Eur. J. Phys.,* vol. 11, pp. 352-358, 1990.

[42] J. H. McDermott, M. Schemitsch and E. P. Simoncelli, "Summary statistics in auditory perception," *Nature Neuroscience,* vol. 16, pp. 493-498, 2013.

[43] R. D. Patterson, R. W. Peters and R. Milroy, "Threshold duration for melodic pitch," in *Hearing - Physiological Bases and Psychophysics, Proceedings of the 6th International Symposium on Hearing*, Berlin, 1983.

[44] K. Robinson and R. D. Patterson, "The Duration Required To Identify the Instrument, the Octave, or the Pitch Chroma of a Musical Note," *Music Perception,* vol. 13, no. 1, pp. 1-15, 1995.

[45] K. Robinson and R. D. Patterson, "The stimulus duration required to identify vowels, their octave, and their pitch chroma," *J. Acous. Soc. Am.,* vol. 98, no. 4, pp. 1858-1865, 1995.

[46] K. D. Martin, "Sound-Source Recognition: A Theory and Computational Model (thesis)," Massachusetts Institute of Technology, Cambridge, MA, 1999.

[47] A. J. Compton, "Effects of Filtering and Vocal Duration upon the Identification of Speakers, Aurally," *J. Acous. Soc. Am.,* vol. 1748, 1963.

[48] S. R. Schweinberger, A. Herholz and W. Sommer, "Recognizing famous voices: influence of stimulus duration and different types of retrieval cues.," *J Speech Lang Hear Res. ,* vol. 40, no. 2, pp. 453-463, 1997.

[49] E. G. Schellenberg, P. Iverson and M. C. McKinnon, "Name that tune: Identifying popular recordings from brief excerpts," *Psychonomic Bulletin & Review,* vol. 6, no. 4, pp. 641-646, 1999.

# I. Appendix I - Audio Collections:

[1] Sony Corporation, Performer, *Sony Pictures Sound Effects Series Volumes 1-10.* [Sound Recording]. Sony Corporation. 2003.

[2] J. Storm, Composer, *Great Smoky Mountains National Park : Storms in the Smokies.* [Sound Recording]. Cornell Laboratory of Ornithology. 1994.

[3] J. Storm, Composer, *Great Smoky Mountains National Park : Summer and Fall.* [Sound Recording]. Cornell Laboratory of Ornithology. 1994.

[4] J. Storm, Composer, *Great Smoky Mountains National Park : winter & spring.*. [Sound Recording]. Cornell Laboratory of Ornithology. 1994.

[5] *Sounds of Nature & The Great Outdoors.* [Sound Recording]. Madacy Records. 1994.

[6] L. H. Emmons, B. M. Whitney and D. L. Ross Jr., Composers, *Sounds of Neotropical Rainforest Mammals: An Audio Field Guide.* [Sound Recording]. Cornell Laboratory of Ornithology. 1998.

[7] *Sounds of the Fascinating Animal World.* [Sound Recording]. Madacy Records. 1994.

[8] *The Diversity of Animal Sounds.* [Sound Recording]. Cornell Laboratory Of Ornithology. 2009.

[9] D. Stokes and L. Stokes, Composers, *Stokes Field Guide to Bird Songs: Eastern Region.* [Sound Recording]. Little, Brown & Company. 2010.

[10] T. S. Schulenberg, Composer, *Voices of Amazonian Birds, Vol. 1: Tinamous Through Barbets.* [Sound Recording]. Cornell Laboratory Of Ornithology. 2000.

[11] B. M. Whitney, Composer, *Voices of New World Parrots.* [Sound Recording]. Cornell Laboratory of Ornithology. 2002.

[12] W. Shakespeare, Composer, *Hamlet.* [Sound Recording]. The Renaissance Theatre Company. 1992.

[13] T. S. Schulenberg, Composer, *Voices of Amazonian Birds, Vol. 2: Toucans Through Antbirds.* [Sound Recording]. Cornell Laboratory Of Ornithology. 2000.

[14] C. Davidson, Composer, *Frog and Toad Calls of the Rocky Mountains: Vanishing Voices.* [Sound Recording]. Cornell Laboratory Of Ornithology. 1996.

[15] *A Field Guide to Western Bird Songs: Western North America.* [Sound Recording]. Houghton Mifflin Harcourt. 1999.

[16] *Walk in the Forest.* [Sound Recording]. Special Music. 1994.

# II. Appendix II – Track listing:

This appendix lists the tracks used in the results section, sorted by compilation/album. The Sony Pictures Sound Effects Series uses track names because this is a digital compilation rather than an audio disk, so the publisher did not number the tracks. All other sources use track numbers for space considerations.

[1] **Sony Pictures Sound Effects Series:**

**Volume 1:** River Current Medium, River Current Medium, Rain Drops From Roof, Rain Drops From Roof, Rain Heavy 01, Rain Into Puddle, Rain Into Puddle, Lake Waves Lapping, Lake Waves Lapping, Dog Large Barking, Dog Medium Barking 01, Dog Small Barking, Dog Small Growling Licking, Dog Small Growling Licking, Dog Small Growling Licking, Dog Whining Begging, Dog Whining Begging, Dogs Large Barking 01, Wind Light, Wind Low Rumble, Wind Medium, Wind Medium, Wind Medium, Wind Strong, Wind Strong, Wind Strong

Volume 10: 1920s American Car Steady Out, 1920s American Car Steady Out, 1924 Ford Stake Bed Truck Ext Crank Start Idle Revs Off 01, 1924 Ford Stake Bed Truck Ext Crank Start Idle Revs Off 02, 1924 Ford Stake Bed Truck Ext Crank Start Idle Revs Off 03, 1924 Ford Stake Bed Truck Ext In Medium Idle Off, 1924 Ford Stake Bed Truck Ext In Slow Idle Off, 1924 Ford Stake Bed Truck Ext Medium By, 1924 Ford Stake Bed Truck Ext Start Idle Away 01, 1924 Ford Stake Bed Truck Ext Start Idle Away 02, 1924 Ford Stake Bed Truck Ext Start Idle Medium Away, 1924 Ford Stake Bed Truck Ext Start Idle Reverse Away Idle O, 1924 Ford Stake Bed Truck Ext Start Idle Slow Away, 1924 Ford Stake Bed Truck Ext Start Idle Slow Engine Revs, 1924 Ford Stake Bed Truck Int Idle With Maneuvers Away, 1924 Ford Stake Bed Truck Int Medium To Slow Idle Shift Forw, 1924 Ford Stake Bed Truck Int Medium To Slow Idle Shift Forw, 1924 Ford Stake Bed Truck Int Medium To Slow Idle Shift Forw, 1924 Ford Stake Bed Truck Int Start Idle Away Long, 1924 Ford Stake Bed Truck Int Start Idle Medium To Slow Away, 1924 Ford Stake Bed Truck Int Start Idle Medium To Slow Away, 1927 Sedan In Idle Away, 1929 Sedan Ext Slow By,  Away, 1929 Sedan Int Start Idle Revs Off, 1929 Sedan Int Start Slow Shift Off, 1929 Sedan Int Start Slow Shift Off, 1929 Sedan Int Start Slow Shift Off,  Stop,  Shifts 01,  Shifts 02, 1937 Plymouth Ext Medium By 01, 1937 Plymouth Ext Medium By 02, 1937 Plymouth Ext Slow By, 1937 Plymouth Ext Start Idle Off, 1937 Plymouth Ext Steady Away,  Decelerate,  Decelerate, 1937 Plymouth Int Idle Away Constant Shift, 1937 Plymouth Int Idle Away Constant Shift, 1937 Plymouth Int Idle Away Constant Shift, 1937 Plymouth Int Start Away Idle Away Idle Stop, 1937 Plymouth Int Start Away Idle Away Idle Stop, 1937 Plymouth Int Start Awy Shift Idle, 1937 Plymouth Int Start Idle Away Stop 01, 1937 Plymouth Int Start Idle Away Stop 02, 1937 Plymouth Int Start Idle Away Stop 02, 1937 Plymouth Int Start Idle Away Stop 02, 1937 Plymouth Int Start Idle Away, 1937 Plymouth Int Steady With Maneuvers Off, 1947 Ford V8 Truck Ext In Medium Idle Off, 1947 Ford V8 Truck Ext In Slow Idle Off, 1947 Ford V8 Truck Ext In Slow Off, 1947 Ford V8 Truck Ext Start Fast Away, 1947 Ford V8 Truck Ext Start Idle Away Slow 01, 1947 Ford V8 Truck Ext Start Idle Away Slow 02, 1947 Ford V8 Truck Ext Start Idle Revs Off, 1947 Ford V8 Truck Ext Start Medium Away, 1948 Diesel Truck Ext Fast By 01, 1948 Diesel

Truck Ext Fast By 02, 1948 Diesel Truck Ext In Medium Stop With Brakes Idle Off, 1948 Diesel Truck Ext In Slow Stop With Brakes Idle Off, 1948 Diesel Truck Ext Slow By, 1948 Diesel Truck Ext Start Idle Fast Away, 1948 Diesel Truck Ext Start Idle Reverse Idle Off, 1948 Diesel Truck Ext Start Idle Slow Away, 1948 Diesel Truck Ext Start Medium Away, 1948 Diesel Truck Int Start Away Stop Idle Off, 1948 Diesel Truck Int Start Idle Revs Slow Forward Off, 1948 Diesel Truck Int Start Idle Revs, 1948 Diesel Truck Int Start Idle Revs, 1954 Big Rig Ext Fast By 01, 1954 Big Rig Ext Fast By 03, 1954 Big Rig Ext In Idle Off, 1954 Big Rig Ext In Slow Idle Stop, 1954 Big Rig Ext Slow By 01, 1954 Big Rig Ext Slow By 02, 1954 Big Rig Ext Slow By 03, 1954 Big Rig Ext Start Away, 1954 Big Rig Ext Start Idle Away Medium Shifts, 1954 Big Rig Ext Start Idle Away Slow Revs, 1954 Big Rig Ext Start Idle Away, 1954 Big Rig Ext Start Idle Fast Away With Shifts, 1954 Big Rig Ext Start Idle Reverse Off, 1954 Big Rig Int Start Away Various Shifts Idle, 1954 Big Rig Int Start Away Various Shifts Idle, 1954 Big Rig Int Start Away Various Shifts Idle, 1954 Big Rig Int Start Idle Away Idle Off, 1954 Big Rig Int Start Idle Away Idle Off, 1954 Big Rig Int Start Idle Away Off, 1954 Big Rig Int Start Idle Away Off, 1954 Big Rig Int Start Idle Revs Off, 1954 Big Rig Int Start Idle Revs Off, 1954 Big Rig Int Start Out Slow Idle Stop, 1954 Big Rig Int Start Out Slow Idle Stop, 1954 Big Rig Int Start Rev Idle Off, 1954 Big Rig Int Start Rev Idle, 1954 Big Rig Int Start Steady Slow Idle Off, 1954 Big Rig Int Start Steady Slow Idle Off, Bus Vintage Start Idle, Ford Model T Away By Out, Ford Model T Start Maneuvers Stop

**Volume 2:** Kennel Interior, Kennel Interior, Kennel Interior, Jungle Birds And Insects

**Volume 3:** Bath Water Movement

[2] **Earthtunes Storms in the Smokies:** 4, 5, 6, 7, 10, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 34, 35, 36, 39, 46, 47, 48, 49, 50, 51, 52, 53, 54, 56

[3] **Earthtunes Summer and Fall:** 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 15, 16, 17, 19, 20, 21, 22, 23, 24, 25, 28, 29, 30, 44, 45, 46, 48, 51, 52, 53, 75, 83, 84, 85, 86, 88, 90

[4] **Earthtunes Winter and Spring:** 1, 4, 11, 12, 13, 14, 15, 16, 20, 21, 22, 23, 24, 28, 29, 30, 31, 45, 46, 47, 86, 87, 89, 91, 92, 96, 97

[5] **Sounds of Nature and the Great Outdoors:** 1, 3, 4, 5, 6, 7, 9, 13, 14, 15, 17, 20, 21, 22, 23, 24, 25, 27, 35, 36, 37, 38, 41, 47, 48, 49, 50, 52, 56, 58, 59, 60

[6] **Sounds of Neotropical Rainforest Mammals:**

Disk 1: 9, 10

Disk 2: 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54

[7] **Sounds of the Fascinating Animal World:** 7, 11, 16, 17, 26, 27, 28, 29, 30, 37, 38, 39, 40, 53, 57, 61, 65, 70

[8] **The Diversity of Animal Sounds:** 5, 6, 11, 12, 13, 27, 42, 49, 52

[9] **The Stokes Field Guide To Bird Songs Eastern Region:** 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34

[10] **Voices of Amazonian Birds:** 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57

[11] **Voices of New World Parrots:**

**Disk 1:** 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50

**Disk 2:** 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59

**Disk 3:** 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31

[12] **William Shakespeare Hamlet:**

**Disk 1:** 2, 6, 7

**Disk 2:** 2, 3, 4, 6

**Disk 3:** 1, 2, 3