AUTOMATIC CLASSIFICATION OF ELECTRONIC MUSIC AND SPEECH/MUSIC AUDIO CONTENT

BY

AUSTIN C. CHEN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Professor Mark Hasegawa-Johnson

Abstract

Automatic audio categorization has great potential for application in the maintenance and usage of large and constantly growing media databases; accordingly, much research has been done to demonstrate the feasibility of such methods. A popular topic is that of automatic genre classification, accomplished by training machine learning algorithms. However, “electronic” or “techno” music is often misrepresented in prior work, especially given the recent rapid evolution of the genre and subsequent splintering into distinctive subgenres. As such, features are extracted from electronic music samples in an experiment to categorize song samples into three subgenres: deep house, dubstep, and progressive house. An overall classification performance of 80.67% accuracy is achieved, comparable to prior work.

Similarly, many past studies have been conducted on speech/music discrimination due to the potential applications for broadcast and other media, but it remains possible to expand the experimental scope to include samples of speech with varying amounts of background music. The development and evaluation of two measures of the ratio between speech energy and music energy are explored: a reference measure called speech-to-music ratio (SMR) and a feature which is an imprecise estimate of SMR called estimated voice-to-music ratio (eVMR). eVMR is an objective signal measure computed by taking advantage of broadcast mixing techniques in which vocals, unlike most instruments, are typically placed at stereo center. Conversely, SMR is a hidden variable defined by the relationship between the powers of portions of audio attributed to speech and music. It is shown that eVMR is predictive of SMR and can be combined with state-of-the-art features in order to improve performance. For evaluation, this new metric is applied in speech/music (binary) classification, speech/music/mixed (trinary) classification, and

a new speech-to-music ratio estimation problem. Promising results are achieved, including 93.06% accuracy for trinary classification and 3.86 dB RMSE estimation of the SMR.

Acknowledgments

First and foremost, I extend the utmost gratitude to my adviser, Professor Mark Hasegawa-Johnson, for guiding me through the challenging process that is graduate research work. Additionally, I would like to acknowledge Po-Sen Huang for his assistance regarding non-negative matrix factorization and blind source separation. Thanks also go out to Dr. Dan Ellis of Columbia University for allowing access to the Scheirer and Slaney audio corpus. Finally, I would not have made it this far without the love and support of my family, girlfriend, and numerous new friends whom I have met at Illinois and helped to make my time here enjoyable.

Table of Contents

Chapter 1. Introduction

Over the past decade, personal media consumption has quickly shifted from brick-and-mortar stores to online services. In fact, digital music sales surpassed physical sales for the first time in 2011. Internet-based music catalogs and streaming services such as iTunes and Pandora have gained significant popularity and now enjoy immense customer bases. As well, more and more individuals now utilize podcasts and internet-based radio as key sources of news and entertainment content. As consumers continue to look to the web to enjoy audio content, it becomes increasingly important to be able to search, sort, and otherwise manipulate large collections of media in order to enhance user experiences and facilitate database maintenance.

The categorizations known as genres are often the premier method of sorting musical content due to their capacity to separate tracks into stylistically similar subsets. Despite some musicologists' concerns of oversimplification or vagueness in definition, genres are critical labels for almost every type of media library [1]. Currently, most genre labeling is done manually, a tedious, time-consuming, and often inconsistent method; however, with the realization that songs within a genre usually share some set of distinctive features (e.g. rhythmic structure, composition), researchers have been able to analyze these qualities and utilize machine learning algorithms to predict the correct labeling for various types of songs. Automatic genre categorization has since evolved into an important facet of the intriguing music information retrieval (MIR) field.

Despite the success of these automatic classification methods, there is definite potential for improvement from a musicological standpoint. Most researchers chose broad categories (e.g. classical, pop, rock), which makes sense for the establishment of a widely applicable

classification methodology. It is possible, though, to delve deeper into the hierarchy of labels or more accurately process the individual genres. For instance, the representation of electronic music in most MIR papers is fairly inaccurate compared to the state of the genre today.

Over the past decade or so, “electronic” music has advanced from once was colloquially called “disco” or “techno” to a dynamic, thriving realm of music with styles such as “dubstep,” “electro house,” “progressive house,” and even new fusions such as “trap.” This movement, dubbed electronic dance music (EDM), has grown to the point where popular music artists have started to include quite obviously EDM-influenced synthesizers and bass drops in bridges and choruses of chart-topping singles. There is no doubt that electronic music has not only become mainstream, but has also evolved and splintered into varying subgenres as a result. For the most part, “electronic” by itself is no longer a sufficient descriptor, so it is logical to test established genre classification methods on these new styles of music.

A different variation of the audio classification problem involves differentiating between music and speech. The applications are obvious, for example allowing listeners to skip through mixed media intelligently. Particularly, speech/music segmentation has enormous potential for broadcast applications. A real-time implementation could greatly enhance TV viewers’ or radio listeners’ experiences by recognizing desired and undesired content, potentially allowing them to view only non-commercial portions or listen only to the speaking parts of a talk show. Determining key transitions in content type could allow audiences to navigate through continuous media streams in a more rewarding manner. As well, such classification adds value to the indexing of recorded content, making it simpler to sort and retrieve audio segments after they have been parsed and labeled as speech or music.

Automatic speech recognition (ASR) processes can also be assisted and improved using speech/music discrimination [2]. Captioning or transcription services may produce errors when fed music or other audio that is not pure spoken word. A discriminator could determine when it is and is not appropriate to analyze audio for speech recognition transcribing, not only reducing error but also reducing the amount of computation required and speeding up the process as a result.

Although human listeners can accurately and quickly discriminate music from speech, manual labeling of such content on an internet scale is infeasible due to the immense sizes of media databases and the wealth of amateur, user-created content being generated daily. Automation is desirable and, fortunately, easily attainable due to the intrinsic characteristics of these differing types of sound. It is precisely these analytically discernable differences in audio that will be taken advantage of for the following experiments in genre classification and speech/music discrimination.

Chapter 2. Background and Literature Review

2.1 Genre Classification

In the early 2000s, existing speech recognition research was expanded upon in hopes of finding practical music labeling applications. Commonly used metrics such as Mel-frequency cepstral coefficients (MFCCs), a succinct representation of a smoothed audio spectrum, were quickly adapted for music. These timbral features proved to be quite effective when applied to segmented audio samples.

However, given the complexities of music, additional features were necessary for sufficient genre analysis. Rhythmic and harmonic features (e.g. tempo, dominant pitches) were developed in response. Some went a step further to utilize quantizations of higher level descriptors like emotion and mood to assist in the categorization of audio. The effectiveness of these types of extracted features has led some researchers to ponder their applications for automatic playlist generation, music recommendation algorithms, or even assisting musicologists in determining how humans define genres and otherwise classify music [3].

Upon extracting features from music samples, relationships were mapped out among various genres. For example, rock music tends have a higher beat strength than classical music, whereas hip-hop might usually contain greater low energy (bass) than jazz. Statistical models were trained to process feature sets by genre before finally being applied to test sets of audio for classification.

One of the most well-known (and perhaps the earliest successful) genre classification papers is that of Tzanetakis and Cook [4]. Together, they spearheaded the initial usage of speech recognition feature sets for genre-related purposes. Timbre features utilized include spectral centroid, spectral rolloff, spectral flux, zero-crossings, MFCCs, and low-energy. Additional rhythmic and pitch-content features were also developed using beat and pitch histograms. Using Gaussian mixture model and k-nearest neighbor approaches, a 61% genre classification success rate was achieved. These encouraging results not only validated the feature-extraction and machine learning methodology, but provided a solid base for others to expand on and a corresponding benchmark to compare to.

Li et. al [5] utilized the exact same feature set as Tzanetakis and Cook with the addition of Daubechies wavelet coefficient histograms (DWCHs). Two further machine learning methods were compared, support vector machines (SVM) and linear discriminant analysis. Findings were positive, with an overall accuracy of 80%.

McKinney and Beebart [6] suggested the usage of psychoacoustic features, e.g. roughness, loudness, and sharpness. Additional low-level features such as RMS and pitch strength were incorporated as well. Features were computed for four different frequency bands before being applied to a Gaussian mixture model, resulting in a 74% success rate.

Lidy and Rauber [7] analyzed audio from a pyscho-acoustic standpoint as well, deciding to further transform the audio in the process of feature extraction. Using SVMs with pairwise classification, they were able to reach an accuracy of 75%.

Pohle et al. [8] supplemented Tzanetakis and Cook's feature set using MPEG-7 low level descriptors (LLDs); some of these include power, spectrum spread, and harmonicity. K-nearest neighbors, naïve Bayes, C4.5 (decision tree learner), and SVM were all utilized for machine learning. Although the original focus of the study was intended to classify music into perceptual categorization such as mood (happy vs. sad) and emotion (soft vs. aggressive), their algorithm was also adapted for genre classification with a 70% rate of success.

Burred and Lerch [9] chose to use MPEG-7 LLDs as well, but added features like beat strength and rhythmic regularity. Using an intriguing hierarchal decision-making approach, they were able to achieve a classification accuracy of around 58%.

As mentioned previously, these studies often failed to depict electronic music in a manner accurately representative of the state of the genre. If included at all, terms used ranged from "Techno" [5] and "Disco" [4] to "Techno/Dance" [9] and "Eurodance" [10]. Many of these labels are now antiquated or nonexistent, so audio features and machine learning strategies were selected from previous studies for application on samples from current electronic music subgenres that better embody the style of music.

2.2 Speech/Music Discrimination

Saunders [11] published one of the first studies on speech/music discrimination in hopes of isolating music portions of FM radio broadcasts. Based on analysis of the temporal zero-crossing rate (ZCR) of audio signals, a Gaussian classifier was developed from a training set of samples. A remarkable 98.4% segmentation accuracy was achieved with real-time performance, proving not only the viability of speech/music discriminators in general, but also the effectiveness of this

type of two-step feature extraction and machine learning process for signal classification purposes.

This work has been extended in several ways. The use of more complex features (e.g. spectral, rhythmic, harmonic) is beneficial for expanding the types of classes analyzed [4]. Scheirer and Slaney [2] proposed a low-energy metric, exploiting a characteristic energy peak that appears in speech signals around 4 Hz [12]. Spectral features such as spectral rolloff, spectral centroid, and spectral flux were also included in their feature vector, yielding a maximum accuracy of 98.6%. Mel-frequency cepstral coefficients proved to be valuable for speech/music discrimination applications as well, resulting in 98.8% [13] and 93% [14] accuracies in different experimental setups. Linear prediction coefficients have also been used with some success in several studies [15,16]. A more in-depth summary of prior speech/music discrimination studies can be found in table form in [17]. Noteworthy is the fact that a lack of a standard corpus makes classification rates hard to compare directly.

Throughout this wealth of experimentation, the focus has been on discriminating pure speech from pure music. In reality, much broadcast and internet audio cannot be classified as pure speech or pure music, with various mixtures of speech, singing, musical instruments, environmental noise, etc. occurring quite commonly. Even in radio applications, DJs speak over background music and news anchors talk over introduction themes or lead-in "bumpers." Being able to identify these types of mixtures could be beneficial, for example, to determine transitions between types of audio for segmentation purposes. Didiot et al. [18] addressed the problem of multi-class audio by classifying audio samples twice, first as music or non-music, then speech or non-speech. Using the resulting four final categories, maximum accuracies of 92%, 85%, and

74% were obtained for three different test sets. Lu, Zhang and Li [19] took a very similar approach, with the same final four categories (music, speech, speech + music, noise), and with comparable accuracy (92.5% after post-processing). A similar setup was implemented by Zhang and Kuo [20], likewise resulting in over 90% of the test audio being labeled correctly.

The authors of [21] chose seven categories: silence, single speaker speech, music, environmental noise, multiple speaker speech, simultaneous speech and music, and speech and noise. Using a wide array of features such as MFCCs and LPCs, over 90% of the samples were classified correctly. Razik et al. [22] mixed speech and music at ratios of 5, 10, and 15 dB to determine the effect on classification. Results indicated that lower mixing ratios caused a greater likelihood for the samples to be misclassified as instrumental music or music with vocals. Around 75% of the test set was classified correctly in a three-way (speech/music/mixed) application, using MFCCs and spectral features.

2.3 SMR

Despite the lack of directly relevant previous work, there is reason to believe that it is possible to classify mixed speech/music samples into categories based on their varying mixture ratios. By comparing the energy attributed to one versus the other, it becomes possible to relate such samples in quantitative manner. Define the speech-to-music ratio (SMR) in decibels r_n of the nth mixed audio signal $x_n[t]$ to be

$$r_n = 10 \log_{10} \frac{\sum_{t=1}^{T} s_n\,[t]^2}{\sum_{t=1}^{T} m_n[t]^2} \tag{2.3.1}$$

where $s_n[t]$ is the speech portion of the signal, $m_n[t]$ is the music portion of the signal (including sung vocals), and the complete audio signal $x_n[t] = s_n[t] + m\_n[t]$.

This speech-to-music ratio may be useful for a number of other settings also, as this is a hidden characteristic that is intrinsic to and descriptive of an audio clip. The property may be used accordingly to analyze or boost portions of an audio signal that are of particular interest. The metric is viable as a tool for signal enhancement, particularly speech enhancement for resynthesis or recognition. For example, Ephraim and Malah [23] rely on a prior probability distribution of speech and noise in order to improve the quality of samples degraded by additive noise. Even more elementarily, SMR could help identify transitions between speech and music content or simply define or label the genre of audio segments (e.g. scored dialogue has a high positive SMR, dialogue embedded in a party has a low positive SMR, a lead-in bumper has nearly zero SMR, and an accidentally recorded dialogue may have negative SMR). All of these applications require that SMR consider sung vocals to be part of the music rather than part of the speech; as we'll see, this definition complicates the SMR estimation problem.

2.4 eVMR

Though SMR is a hidden attribute of the audio mixture, it correlates strongly with features that can be measured in the signal. For example, it is common in audio engineering to pan musical instruments towards the left or right channels in order to avoid a "cluttered" or "muddy" sound. In fact, the creative ability to make use of sonic space in the stereo field is arguably what separates professional recording engineers from amateurs. During this process, it is commonplace to keep spoken and sung tracks in the middle of the mix as stereo center is the most prominent position available in relation to the listener. Well-documented in audio

engineering handbooks and other texts [24-26], this tendency is useful when attempting to isolate portions of a recording.

With this extra information about the presence of voice versus non-voice instruments contained within stereo signals, it is logical that a rough estimate of the SMR can be obtained by reversing the mixing process. The estimate is imperfect because it attributes sung vocals and other centered instruments (e.g. bass drum) to the estimated voice track rather than the estimated music track; nonetheless, it is sufficient. The development and evaluation of this simple but effective metric are explored. Both three-way speech/music/mixed and traditional two-way speech/music discrimination problems are addressed for assessment purposes. Further, the new feature is tested for its ability to estimate the hidden speech-to-music ratios of mixed audio samples within a classification framework. In all cases, the feature is supplemented with those utilized in the state of the art, with such mixtures demonstrating improved accuracy relative to the state of the art.

Chapter 3. Electronic Music Classification

3.1 Algorithms

3.1.1 Features for Classification

Seven commonly used features were selected based on their proven effectiveness for classifying music samples. Thought was also given as to how the features would specifically translate into the realm of electronic music. For example, the genre of dubstep emphasizes very low bass, potentially lending itself to greater low energy.

Low energy is the percentage of audio with less RMS energy than the average of the entire waveform. Each sample is divided into frames for the purpose of the computation. This feature is used as an effective measure of amplitude distribution and was implemented in [4].

The zero-crossing rate is calculated as the number of times a signal crosses the time axis. Zero-crossings can be used to measure the noisiness of audio signals.

Spectral rolloff is the frequency below which 85% of magnitude distribution lies; it is used as a measure of spectral shape.

Brightness (or high-frequency energy) is the amount of energy above a selected threshold, typically somewhere between 500 and 3000 Hz for speech and audio. Brightness is a primary measure of timbre (sharpness vs. softness).

Tempo is the speed at which a piece of music is played, measured in beats per minute (bpm). It is calculated here using the autocorrelation method developed by Eck [27]. Essentially, autocorrelation is computed using an optimized FFT approach, resulting in peaks at intervals of the tempo. This effect is sharpened for music samples by taking phase into account and using entropy as a measure of "peakiness." The strongest peaks are then selected from a window of the range of appropriate tempos.

Mel-frequency cepstral coefficients (MFCCs) are designed to succinctly represent a smoothed spectrum of an audio signal. Arguably the most widely used parameters in speech recognition, the steps for their computation are as follows:

1) The discrete Fourier transform (DFT) is taken on successive windowed segments of the signal, resulting in the short-time Fourier transform (STFT).

2) The log of each frame's spectrum is taken according to the calculation of the cepstrum; the log of the magnitude allows for extraction of the spectral envelope from the details, like an adaptation of the traditional spectrum for modified waveform manipulation.

3) Log-DFT values are grouped together in critical bands or bins and smoothed using triangular weighting functions for Mel-frequency scaling. This conversion to Mel non-uniformly scales the responses based on the perceptive response of the human ear, focusing on aspects important for our hearing system.

4) Coefficients are generated using the discrete cosine transform (DCT, very closely related to the DFT and Karhunen-Loève transform) to decorrelate the vectors from Mel-scaling. These

MFCCs compactly summarize the mel-spectrum response (the smoothed, modified spectral envelope).

Instead of the 13 coefficients that are typically used for speech recognition applications, the first 5 MFCCs were chosen for this feature for maximum effectiveness in music per Tzanetakis and Cook [4].

3.1.2 Machine Learning

Although many types of machine learning algorithms have been utilized in the literature, Gaussian mixture models (GMMs) were chosen for classification due to their efficiency and ease of implementation. Essentially a weighted summation of Gaussian distributions, each class (in this case, each subgenre) is taught to the algorithm using distinct parameters (here, audio features) of a training set of known audio samples. Then, for each sample in a separate test set, the model evaluates its corresponding features to determine which class's approximate distribution it most likely came from. By maximizing this conditional probability, an educated guess can be made as to which class the sample best belongs in.

3.2 Experiments

A number of music information retrieval tools were considered for usage in the automatic genre classification of electronic music. The University of Jyväskylä's MIR Toolbox [28] was eventually chosen over other applications such as Marsyas, jMIR, Yaafe, IMIRSEL M2K, and MA Toolbox. MIR Toolbox was favored due to its multitude of built-in feature extractors, inclusion of machine-learning algorithms, wide array of analysis tools, and clean implementation in MATLAB.

3.2.1 Genre selection

Deep house, dubstep, and progressive house were chosen as the subgenres of interest due to their fair representation of the broad spectrum of current electronic music. Deep house is a slower, sparser subgenre due to its soul and jazz influences; it is among the more minimalistic styles still listened to today. Dubstep is typically bass-heavy with a large amount of emphasis placed on the beat of the song. Typically around 70/140 bpm, many dubstep songs are notable for their inclusion of "wobbles," or quick timbre modulations of low basslines. By comparison, progressive house is a much more energetic and "happy" subgenre. Progressive tracks feature highly melodic synthesizer melodies, often preceded by sampled female pop vocals. These types of songs tend to be around 130 bpm.

3.2.2 Experimental Methodology

Ten songs were selected for each genre. Each of these thirty tracks was then clipped and rendered as a 30-second, 22050 Hz, 16-bit mono sample for analysis. A set of seven audio features was selected; most were timbral in nature due to their proven effectiveness in other genre classification applications. The exceptions were brightness, which was chosen to take into account the high synthesizer lines often found in progressive house compositions, and tempo, which follows very distinct distributions based on subgenre (around 70/140 bpm for dubstep, 130 bpm for progressive house). Descriptions and derivations of all of features analyzed can be found in the proceeding section.

Given the set of audio, a 3-fold cross-validation methodology was utilized for machine learning. This means that twenty tracks were taken as a training set for the algorithm, while the remaining

third was used as a test set for classification. The process was then repeated two more times such that every track was categorized once.

Gaussian mixture models were chosen for machine learning over other algorithms such as K-nearest neighbors due to efficiency and an improved classification accuracy from preliminary experimentation. Mixture models composed of a varying number of components (1-5 Gaussians per mixture) were tested for completeness.

3.3 Results

Across all mixture models, an 80.67% success rate was achieved. Deep house was correctly classified 84% of the time, with dubstep and progressive house rightly identified at rates of 68% and 90%, respectively. Such accuracy matches up quite well against the performances of prior studies: 80% [5], 61% [4], 74% [6], 70% [8], 58% [9], 79% [10], 75 % [7], and 79% [29].

A confusion matrix (Figure 3.3.2) reveals that of all of the errors, dubstep tracks were most often mislabeled as progressive house; this may be due to their similarly dense composition styles. As well, the pervasiveness of progressive house as the premier genre of EDM allows more room for the fusion and influence of other genres to occur; this may also help explain the confusion of some styles for progressive house. It is worth mentioning that aside from progressive house, deep house also enjoyed a low amount of miscategorization, likely due to its distinct and easily identifiable style of sparseness in composition.

3.3.1 Tables and Figures

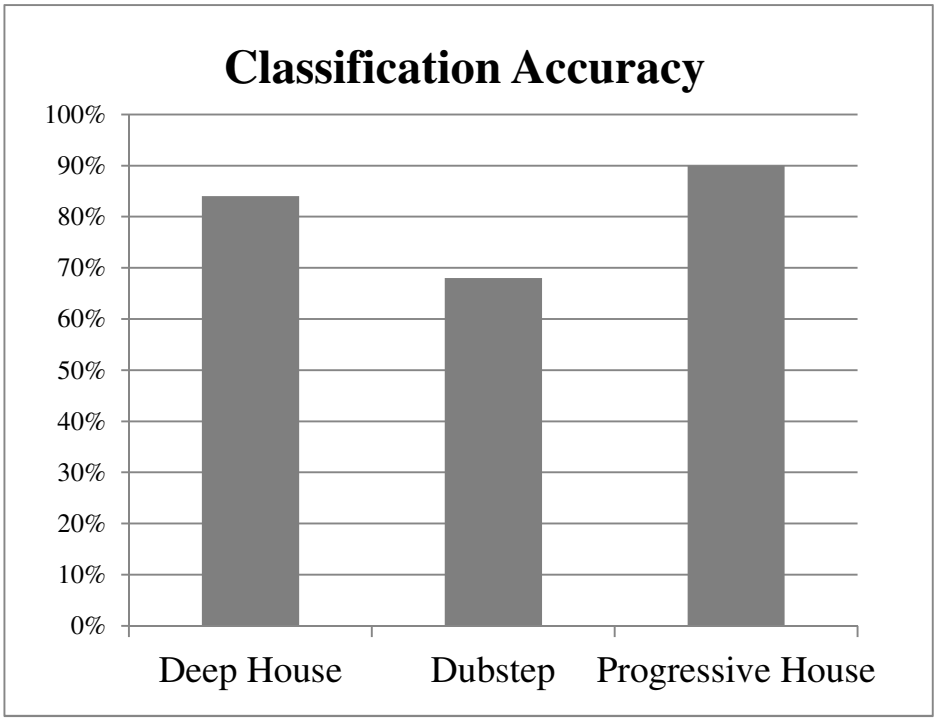

Figure 3.3.1 Classification accuracies for the three selected electronic subgenres

	Deep	Dubs	Prog
Deep	**84%**	0%	16%
Dubs	0%	**68%**	32%
Prog	10%	0%	**90%**

Figure 3.3.2 Confusion matrix for the three electronic subgenres analyzed. Predicted genres are represented by columns, with true genres in rows.

Chapter 4. Speech/Music Discrimination and Ratio Estimation

4.1 Algorithms

4.1.1 SMR

Mixed audio signals from broadcast and similar media sources can be broken down into two parts: speech and music. (Other components such as noise are ignored for the time being.) Based on the definition of speech-to-music ratio in Equation 2.3.1, the estimation of an audio file's intrinsic SMR can be expressed as a regression problem

$$\hat{r}_n = f\left(\vec{x}_n, \vec{\theta}\right) \tag{4.1.1}$$

where $\hat{r}_n$ is the estimated SMR, $\vec{x}_n$ is the evaluated featureset derived from signal $x_n[t]$, and $\vec{\theta}$ is a vector of parameters that comprise the regression model. The machine learning problem is thus to learn these parameters $\vec{\theta}$ in order to minimize the approximation error.

Studies of speech/music discrimination measure success based on rates of correct classification. However, since SMR is a regression problem, new metrics are necessary. Two metrics are introduced to better represent results for a set of N waveforms: mean absolute error

$$MAE = \frac{1}{N}\sum_{n=1}^{N} |r_n - \hat{r}_n| \tag{4.1.2}$$

and root mean squared error

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}|r_n - \hat{r}_n|^2} \tag{4.1.3}$$

In a number of other machine learning tasks (e.g. [30]), it has been demonstrated that real-valued regression models of the form of Equation 4.1.1 often can be developed most accurately by quantizing the output and training a polychotomous classifier instead. SMR can be binned for this purpose using histogram bins based on the SNR just noticeable difference (JND). For example, Allen and Neely [31] demonstrate a relationship between the loudness JND and the masking thresholds for pure tones and noise; from their data it is possible to show that the SNR JND varies between 1dB and 3dB for a signal presented to listeners at 40-120 dB SPL with energy concentrated in the frequency range 250-5000Hz. By binning the target SMR values into 3dB bins, the continuous-valued regression problem of Equation 4.1.1 is converted into a polychotomous classification problem.

A major benefit of this type of implementation is the ability to scale down to other types of classification problems in quick fashion. Recall that speech-to-music ratio is a hidden property implicit within a signal. Thus, if the problem is limited to audio clips with the SMR values of $r_n \in \{-\infty, \infty\}$, then the problem is reduced to classical speech/music discrimination. Likewise, a speech/music/mixed classification can be represented by $r_n \in \{-\infty, \infty, other\}$.

For either of these two problems, MAE and RMSE are inappropriate for measuring error. Instead, we can use a more traditional classification accuracy rate, or its arithmetic inverse, the rate of classification error, defined as

$$RCE = \frac{1}{N}\sum_{n=1}^{N}[r_n \neq \hat{r}_n] \tag{4.1.4}$$

These experiments pioneer a multi-SMR classification problem $\hat{r}_n \in \{-9 \ldots 21\}$, for which error may be measured appropriately using any of the three metrics proposed: MAE, RMSE, or RCE.

4.1.2 eVMR

On the feature extraction side, it is practical to attempt to predict the speech-to-music ratio of a mixed audio sample. However, given the spectral similarities between speech and some types of musical instruments, this task becomes a bit difficult. Traditional features such as zero-crossing rates and MFCCs may not be sufficient to determine defining differences in these types of samples. Instead, a crude stereo inversion trick that originated with recording engineers is adapted for a novel classification feature that requires minimal computation.

While reviewing the literature on speech/music discrimination, one finds that nearly every study chose to encode audio samples in a monaural format; this makes sense, as one channel is sufficient for spectral and other analysis, and an additional audio channel offers little benefit for speech recognition. However, the manner in which commercially recorded music tends to be mixed actually offers an interesting characteristic that can be exploited. Instruments are often panned slightly more towards the left or right channels in order to prevent a cluttered result. On

the other hand, vocal tracks are usually placed in the center of the stereo image. Similarly, raw speech is typically centered for simplicity or even recorded in mono. While the mix location of instrumental and voice tracks is arbitrary (the mixing engineer is free to position them at will), it is quite robust in practice because violations of the "centered voice" rule are jarring to listeners; violations cause listeners to pay attention to the location of the voice, rather than to the words the voice is producing. By isolating the normally center-mixed component of a stereo audio recording, therefore, it is possible to partially separate singing and speech from instrumentation.

This "inversion" or "phase cancellation" method has been used by recording engineers to create rough instrumental versions of songs from mostly uneditable formats. This trick has also been documented and utilized by music information retrieval and signal processing studies in the past [32-35]. Also referred to as the "karaoke effect", "center pan removal technique", or "left-right technique", a simple flip of one of the two stereo channels prominently reduces the volume of the vocals from a premixed track. Following, it becomes a quite easy matter to calculate the ratio of centered audio to panned audio. As a rough approximation of a signal's inherent speech-to-music ratio, we have termed this new metric the estimated voice-to-music ratio (eVMR) to account for the inclusion of both speech and singing.

In order to implement the eVMR in software, an arithmetic adaptation of the inversion process is used. By subtracting one channel from another (e.g. left $l[n]$ from right $r[n]$), sounds that are not in the center (not present in both channels) are isolated.

$$p[n] = l[n] - r[n] \tag{4.1.5}$$

This results in an approximation of the instrumental portion of the music $p[n]$, which includes any vocals or other effects that may be panned. It should be noted that $p[n]$ is not a discrete audio source originating at a fixed azimuth; rather, it is the set of all audio that is not perfectly centered in the recording, including typically three to thirty musical instruments, choral voices, and reverberation. The widely distributed azimuths of non-speech sources tend to defeat algorithms that rely on a fixed azimuth to identify interference signals, including filter-and-sum beamforming [36,37] and blind source separation algorithms [38-40].

The audio that has been removed after this first step roughly corresponds to the vocals and speech, though any instruments that have been placed in the center of the stereo image may be included as well. The RMS levels for the approximate music and voice separations are calculated following that of the total original source file.

$$RMS_{tot} = \sqrt{RMS_l^2 + RMS_r^2} \tag{4.1.6}$$

$$RMS_{rem} = \sqrt{RMS_{tot}^2 - RMS_p^2} \tag{4.1.7}$$

Finally, the estimated music-to-voice ratio is evaluated and expressed in decibels.

$$eVMR = 20\log_{10}\left(RMS_{rem}/RMS_p\right) \tag{4.1.8}$$

The final value tends to be higher than the original audio mixture's SMR due to the inclusion of singing parts.

It should be noted that this process can be thought of as a rudimentary version of blind source separation. Although a wealth of literature and techniques exist for source separation [36-38], the stereo trick is faster and, in this case, more accurate, because there is no requirement that centered and panned audio be uncorrelated and it requires no previous knowledge of the spectral characteristics of the individual audio sources.

In order to verify the preliminary validity and viability of eVMR for discrimination purposes, a sizeable selection of pure music, pure speech, and mixed audio samples were analyzed. The results, shown in Table 4.1.1, were quite positive, indicating a distinct difference between music and speech samples with mean eVMRs of 5.75 and 38.51 dB, respectively. At intermediate values of SMR, an affine relationship between eVMR and SMR was clearly evident: measured eVMR values are consistently greater than or equal to SMR values, likely because eVMR views speech and singing as indistinguishable sources. It becomes evident that eVMR measurements are predictive of, but not determinant of, SMR.

Standard deviations were somewhat high, perhaps due to the variety of samples and sources used. Different recording processes, mixing methods, musical content, and noise levels all have an impact on the SMR and eVMR. However, the strong affine relationship between eVMR and SMR and the extremely low computational complexity of the feature suggest the possibility of using eVMR in combination with other features for the purpose of estimating SMR; with appropriate design, the combined metric should retain the linear dependence, but with reduced variance.

Although the end goal is not to dissect signals, the same survey of power ratios was carried out based on the output from a state-of-the-art single-channel blind source separation algorithm simply for comparison purposes. Specifically, non-negative matrix factorization [41-43] was utilized in an attempt to isolate the speech and music portions of the mixed signals based on training sets of pure speech and pure music samples. The log-energy ratio of the results was then calculated (à la eVMR) using the two split outputs from the NMF algorithm [44]. As can be seen in Table 4.1.2, a non-linear relationship between the blind source separated power ratios (abbreviated "BSS") and SMR was the result. Futhermore, the computational complexity of NMF is significantly greater, verifying the worth of pursuing eVMR.

4.1.3 Supplementary features

The simple eVMR metric is supplemented by a number of features that have already been well proven in speech/music discrimination and other classification applications. Notably, zero-crossing rate and Mel-frequency cepstral coefficients present powerful options for differentiating between classes of audio. All of the following features except for RMS were utilized previously for the classification of electronic music subgenres; nevertheless, they will be reviewed here.

The zero-crossing rate is calculated from the number of times that a signal crosses the axis in the time-domain. It can be used as a measure of the noisiness of audio signals, and Saunders [11] first explored its potential for speech/music discrimination purposes. Statistics such as the mean, standard deviation, and skew of the ZCR are used to supplement the feature.

Mel-frequency cepstral coefficients [45] are the DCT of a modified spectrum of an audio signal. The mean and standard deviations of the coefficients, deltas, and delta-deltas are computed, providing a strong set of metrics.

Spectral rolloff is a measure of the skewness of spectral shape. Utilized in Tzanetakis and Cook [4], rolloff is determined by the frequency f_r below which 85% of the magnitude distribution lies.

Low-energy skew is another feature implemented in [4]. Representative of the amplitude distribution, it is defined by the percentage of audio frames with less RMS energy than average energy of the entire audio clip.

Finally, the root mean square of each signal is evaluated in hopes of exploiting the differences in energy levels between types of audio. Speech tends to be more sparse, whereas music is typically an amalgamation of numerous instruments and other sonic sources.

Unfortunately, some of these methods that were quite effective for the traditional speech/music discrimination problem ended up falling short for applications with mixed sources. Many features turned out to be not robust enough to handle the similarities between some of the synthesized samples. As a result, an assortment of feature combinations was tested in order to maximize the final classification accuracies of the various experiments. Classification results achieved using 13 of these featuresets can be found in the results section of the article.

4.1.4 Classification framework

On the machine learning side, audio discrimination and SMR evaluation can be performed using a wide variety of existing algorithms; as with electronic music classification, Gaussian mixture models (GMM) are utilized because they perform comparably to the best published algorithms on similar regression and classification problems (e.g. [46]) and can be implemented with standard speech recognition tools. GMMs were trained according the EM algorithm as described in [47,48]. A full covariance structure and a log likelihood stopping tolerance of 0.001 were used, although the maximum number of iterations was limited to 100. The final results were obtained by averaging the performance of classification runs using 1-5 Gaussians per mixture.

4.1.5 Tables

TABLE 4.1.1
EVALUATION OF EVMR

Sample	Mean (dB)	STD (dB)
Music	5.75	6.73
Speech	38.51	16.89
SMR = 3 dB	10.95	4.56
SMR = 6 dB	13.09	3.72
SMR = 9 dB	16.18	4.37
SMR =12 dB	17.46	3.77
SMR =15 dB	20.44	4.09

eVMR mean and standard deviation as a function of SMR

TABLE 4.1.2
EVALUATION OF BLIND SOURCE SEPARATION

Sample	Mean (dB)	STD (dB)
Music	-4.57	5.57
Speech	10.17	3.72
SMR = 3 dB	3.21	1.78
SMR = 6 dB	5.18	1.79
SMR = 9 dB	6.68	2.48
SMR =12 dB	7.31	2.91
SMR =15 dB	7.85	3.12

BSS mean and standard deviation as a function of SMR

4.2 Experiments

4.2.1. Training and Testing Data

Unfortunately, no large established competition or conference task exists for speech-music discrimination; as such, the majority of prior work settled for self-assembled sets of sound clips, making it difficult to compare results. Perhaps the closest thing to a standard corpus for this problem is a set of 15 second FM radio recordings assembled by Scheirer and Slaney [2] in 1996 which has been utilized by a few other researchers [18,49-51] and is used here for the purpose of establishing a baseline. Given that these samples are monaural in nature, though, they were insufficient for testing eVMR.

In order to test eVMR, 360 clips of pure speech and pure music (180 speech clips, 180 music clips) were extracted from personal music libraries and podcasts in an effort to create a contemporary corpus. Seven diverse music genres, eight languages, and an even distribution of male and female speakers were included to ensure a representative collection. A complete list of audio sources used can be found in Appendix B. All samples were encoded in 8-bit, 16 kHz stereo and trimmed to 4 seconds in length for classification. 50 millisecond frame lengths with 50% overlap were utilized for the computation of appropriate features, e.g. delta metrics.

Because past studies have not usually addressed mixed classification, it was necessary to generate new mixtures of pure speech and pure music in a controlled manner. The goal was to scale these raw samples appropriately in order to achieve a given speech-to-music ratio. It is imperative, though, that the mixed audio file maintain a similar RMS to the original audio clips, so that classifiers cannot find any revealing relationships between total energy and mixture

status. Thus, the total root mean square energy from the two pure source samples is calculated first:

$$RMS_{tot} = \sqrt{RMS_s^2 + RMS_m^2} \tag{4.2.1}$$

Given the desired speech-to-music ratio r_n (in decibels), it is then possible to determine the scale factors and synthesize the mixed signal $\tilde{x}[t]$:

$$M_s = \sqrt{RMS_{tot}^2 / (1 + 10^{-r_n/10}) RMS_s^2} \tag{4.2.2}$$

$$M_m = \sqrt{RMS_{tot}^2 / (1 + 10^{r_n/10}) RMS_m^2} \tag{4.2.3}$$

$$\tilde{x}[t] = M_s s[t] + M_m m[t] \tag{4.2.4}$$

Equations 4.2.1 through 4.2.4 vary the scaling of each audio file depending on the energy of the other, so changing the desired speech-to-music ratio will not have the same effect as adjusting gain sliders on an audio mixing console. Notably, speech clips have significantly greater sustained lengths of low energy (pauses and silences) than music, affecting the magnitude of the scale factors. Conversely, music is subject to compression in the mixing stage, the effects of which can make a track sound louder due to an increase in total energy.

Using the speech-to-music ratio equations, 5,580 mixed samples were synthesized and labeled accordingly by SMR in dB. Upon subjectively listening to a preliminary batch of output audio files, it was noted that an SMR of approximately +6 dB seemed appropriate for a typical application of one speaker talking over moderate background music. Accordingly, a -3 to +15 dB

SMR range was chosen to cover the various mixing situations that might be encountered in practical broadcast recording situations. This range was eventually expanded to -9 to +21 dB to provide meaningful endpoints for the estimation problem.

Feature extraction and classification algorithms described in Sec. II were implemented using the MIR Toolbox for MATLAB [28].

4.2.2 Procedures

The proposed eVMR feature was tested alone and in combination with four types of standard features, in three different tasks: speech/music discrimination, speech/music/mixed three-way classification, and SMR estimation.

In each of these tests, features are grouped into four large sets for ease of presentation. The eVMR feature is a scalar (one number per 4-second waveform), computed as shown in Equation 4.1.8. ZCR is a 3-dimensional feature vector including the mean, standard deviation, and skew values of zero-crossing rate during the 4-second waveform; this is a small subset of the features used in [20]. MFCC is a 65-dimensional feature vector including the 13 mel frequency cepstral coefficients as well as the means and variances of the delta coefficients and delta-delta coefficients, comparable to the features used in [13,14]. Finally, “features” is a 3-dimensional feature vector including the spectral rolloff, low-energy, and RMS energy measures; this vector is a subset of the features used in [2]. These three subset vectors were selected because preliminary experiments using a more complete selection of the features from [2,13,14,20] yielded no reduction in two-class error rate.

4.2.2.1 2-Way Classification

First, the time-domain and spectral features chosen were tested in the traditional two-way speech/music discrimination application in order to establish a baseline. Comparisons were made to prior work to ensure a suitable starting point for further extensions. In particular, the Scheirer and Slaney corpus was utilized to verify the performance of the baseline monaural classification system.

Data for this experiment consisted of 360 four-second audio waveform files (180 music files and 180 speech files). Audio files were divided into disjoint training and test sets in a 5-fold cross-validation paradigm. The training set in each fold, containing 72 files (36 music, 36 speech), was used to train GMMs. The test set in each fold, containing 288 files (144 music, 144 speech) was used to test classifiers.

Since performance of all classifiers was near 100% accuracy on this task, this task was also used to test the sufficiency of the proposed eVMR feature. eVMR by itself provides useful information for the speech/music discrimination, but it is not sufficient by itself for 100% accurate speech/music classification, therefore experiments were also performed evaluating feature combinations including eVMR.

4.2.2.2 3-Way Classification

Speech/music/mixed classification was examined to see if three-way classification accuracy could be improved using eVMR. Combinations of proven features such as zero-crossing rate and MFCCs were tried first in order to establish baseline results; previous publications including [20]

provide a baseline accuracy for this task, but on a different dataset. The feature eVMR was then tested alone and in combination with other features.

The training and test sets for three-way classification are a strict superset of those used for two-way classification. The training set in each fold includes 108 four-second audio files (54 music, 54 speech). The test set in each fold includes 432 files (144 music, 144 speech, 144 mixed).

4.2.2.3 SMR Estimation

Experiments to estimate a signal's elemental, real-valued speech-to-music ratio were performed on samples synthesized by mixing the speech and music files together at 31 discrete SMRs. 180 samples were constructed for each SMR value by choosing pairs uniformly at random from the original 180 raw speech and 180 raw audio files. Quantizing the ratios allow for the testing to be completed using a classification framework, similar in effect to binning values. Cross-validated experiments were conducted in which each fold contained 1,116 mixed-audio files and each test set contained 4464 mixed-audio files.

4.3 Results

4.3.1 2-Way Classification Results

Testing the Scheirer and Slaney corpus resulted in a 98.4% success rate using ZCR + MFCCs + features, validating the test framework against comparable studies of the past. Upon switching to stereo data, the baseline classifier achieved 99.3% classification accuracy for the labels "pure speech" and "pure music." For comparison, 98.6% accuracy was obtained using only ZCR, while only MFCC-related measures achieved 99.3%. These numbers very well parallel the success rates found in the literature.

Testing the eVMR feature resulted in 97.9% speech/music discrimination accuracy, a high classification rate. Although eVMR alone is less accurate than ZCR alone, their performances are comparable and, as will be shown in three-way classification, their error patterns complementary.

4.3.2 3-Way Classification Results

Classification rates from key featuresets, averaged across all folds of cross-validation, can be seen in Table 4.3.1. Using only eVMR, 81.9% of speech/music/mixed samples were labeled correctly; this accuracy is lower than that of ZCR, but comparable. ZCR and eVMR are partially complementary features; their combination achieved accuracy higher than that of either feature alone.

In order to evaluate feature combinations, the best previously published system (ZCR+MFCC+features) was chosen as baseline. Combinations including eVMR achieved

accuracy higher than the baseline, but the difference was not statistically significant for this corpus. Informally, spectral features and the eVMR were found to be complementary for this task: mixtures with an ambiguous eVMR were often spectrally unambiguous (e.g., because of the difference between singing and speaking voice), and conversely, spectrally ambiguous samples (e.g., experimental music and/or music with speech-like timbre) were often spatially unambiguous (characterized by significant off-center signal energy).

Figures 4.3.1 through 4.3.3 are confusion matrices for two of the cross-validation folds for the three-way classification experiment; predicted classes are represented by columns, with true labels in rows. As can be seen, a majority of the classification error arises from music samples being perceived as mixed audio, apparently because the singing components of music may be removed and interpreted as speech/voice in the eVMR algorithm. Correspondingly, a portion of mixed samples were misclassified as music due to these and other similarities.

Some error arose as well from speech samples being classified as mixed audio. These errors can be attributed to background or line-level microphone noise present in the source files, likely from improper recording techniques, lack of post-processing, or simply environmental sounds. Such noise, present in both stereo channels of the file, may end up being misinterpreted as part of the instrumental/background music separation in the eVMR method. Using a greater number of professionally mixed speech sources may reduce this error.

Confusion matrices for the state-of-the-art baseline (Figure 4.3.2) and the baseline + eVMR (Figure 4.3.3) are almost identical. The only difference is that eVMR enables the system to correctly identify a small portion of mixed samples (0.7%) that were previously misrecognized as speech.

4.3.3 SMR Estimation Results

A summary of SMR estimation results can be seen in Table 4.3.2. ZCR + features showed relatively low measures of error in comparison to other pre-eVMR featuresets. The inclusion of eVMR saw several metric combinations meet and exceed baseline results, with the highest performing combination being eVMR + ZCR + features. A scatter plot of an expanded classification for this featureset can be seen in Figure 4.3.4. The actual (ground truth) SMR is represented by the horizontal axis, with the predicted value graphed against the vertical axis. Slight amounts of jitter (0.2 dB) have been added for an enhanced visualization of the data distribution. A linear, diagonal band is clearly visible through the center of the plot, indicating the clear relationship between true SMRs and estimated values obtained through classification.

However, the width of the band implies that there may be a variance to the results. In order to further examine the classification precision, the mean average error (MAE) and root mean squared error (RMSE) were investigated. In all cases, the RMSE is noticeably greater than the MAE, possibly because of the small number of high-error outliers visible in Figure. 4.3.4; these outliers are further characterized by the column VAE (variance of amplitude error) in Table 4.3.2. Despite the presence of outliers, MAE and RMSE of the best-performing algorithm are only 2.91 and 3.86dB respectively, suggesting that it is indeed possible to estimate the proportions of speech and music in a mixed audio file via feature extraction and machine learning.

Several rounds of SMR estimation were also completed using the power ratio computed based on blind source separation results as a feature. eVMR performed similarly to BSS, and the addition of eVMR showed significant decreases in error.

4.3.4 Tables and Figures

TABLE 4.3.1
3-WAY CLASSIFICATION RESULTS

Featureset	Class. Rate	p
ZCR[a]	85.88%	
ZCR + features[b]	*92.13%*	baseline
MFCCs[a]	88.89%	
MFCCs + features	90.74%	
ZCR + MFCCs + features	90.97%	
eVMR	81.94%	
eVMR + features	92.59%	0.3991
eVMR + ZCR	87.73%	
eVMR + ZCR + features	**93.06%**	0.3020
eVMR + MFCCs	89.81%	
eVMR + MFCCs + features	91.20%	
eVMR + ZCR + MFCCs	90.05%	
eVMR + ZCR + MFCCs + features	91.20%	

p is the significance with which each better-performing result differs from the baseline of ZCR + features in a 1-sided t-test as in [27].
[a] "ZCR" and "MFCCs" represent the metric and corresponding statistics
[b] "Features" represents spectral rolloff, low-energy, and RMS.

	Music	Speech	Mixed
Music	**75.0%**	2.1%	22.9%
Speech	0.0%	**88.2%**	11.8%
Mixed	6.9%	10.4%	**82.6%**

Figure 4.3.1 Confusion matrix for three-way classification using eVMR as the only feature.

	Music	Speech	Mixed
Music	**89.6%**	0.0%	10.4%
Speech	0.0%	**93.8%**	6.3%
Mixed	6.3%	1.4 %	**92.4%**

Figure 4.3.2 Confusion matrix for three-way classification using ZCR, MFCCs, and additional features.

	Music	Speech	Mixed
Music	**89.6%**	0.0%	10.4%
Speech	0.0%	**93.8%**	6.3%
Mixed	6.3%	0.7%	**93.1%**

Figure 4.3.3 Confusion matrix for three-way classification using eVMR, ZCR, MFCCs, and additional features.

TABLE 4.3.2
SMR ESTIMATION RESULTS

Featureset	MAE	RMSE	VAE	p
ZCR[a]	5.41	7.07	20.67	
ZCR + features[b]	*3.36*	*4.45*	*8.52*	Baseline
MFCCs[a]	4.60	5.95	14.16	
MFCCs + features	4.18	5.46	12.31	
ZCR + MFCCs + features	3.99	5.18	10.94	
eVMR	4.50	6.07	16.66	
eVMR + features	3.01	4.06	7.49	1.41×10^{-5}
eVMR + ZCR	4.28	5.59	12.97	
eVMR + ZCR + features	**2.91**	**3.86**	**6.42**	<< .001
eVMR + MFCCs	4.42	5.65	12.42	
eVMR + MFCCs + features	4.04	5.23	11.06	
eVMR + ZCR + MFCCs	4.17	5.33	11.08	
eVMR + ZCR + MFCCs + features	3.83	4.93	9.67	
BSS	4.18	5.62	14.05	
BSS + features	4.03	5.54	14.52	
BSS + eVMR	2.99	4.04	7.44	
BSS + eVMR + features	2.98	4.11	7.98	

All results in dB, except p.

p is the significance with which each better-performing result differs from the baseline of ZCR + features in a 1-sided t-test as in [27].

[a] "ZCR" and "MFCCs" represent the metric and corresponding statistics

[b] "Features" represents spectral rolloff, low-energy, and RMS.

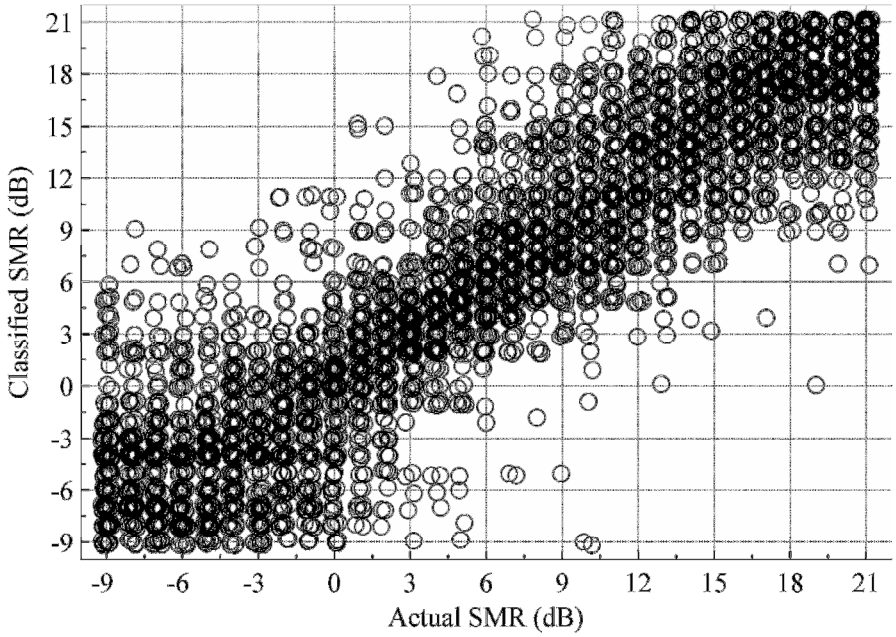

Figure 4.3.4 Scatterplot of SMR estimation results using eVMR, ZCR, and additional features (with jitter of 0.2 dB added).

Chapter 5. Discussion

5.1 Significance Considerations

For the evaluation conditions of 3-way speech/music/mixed audio classification and SMR estimation, one-sided t-tests [52] were conducted while reviewing the results in order to determine if the differences in performance are statistically significant. For each featureset, t-values were calculated as

$$t = \frac{MAE_{baseline} - MAE_{test}}{\sqrt{VAE_{baseline}/n + VAE_{test}/n}} \tag{5.1}$$

and corresponding probabilities computed using the CDF of the Student's t-distribution with the number of total samples minus two as the degrees of freedom. Results that fell below the baseline values were not analyzed.

In three-way classification, success rates for better performing featuresets do not appear to be statistically significant, even though some (such as eVMR + features and eVMR + ZCR + features) clearly show increases in accuracy. Adding eVMR thus does not necessarily improve the results of the baseline featureset in a manner that varies in the statistical sense, perhaps because the baseline performance is already so high. For SMR estimation, however, the t-tests showed higher degrees of significance. This is caused by a large effect size, combined with a large test corpus: the test corpus in Table 4.3.2 is more than ten times the size of that in Table 4.3.1. The statistical significance of SMR estimation differences suggests that the lack of statistical significance in three-way classification may be just an artifact of the small test corpus size; the utilization of simply a slightly larger test corpus may be sufficient to make the small difference in Table 4.3.1 statistically significant and thus resolve any issues.

In the binary speech/music discrimination task, the merit of eVMR becomes even more apparent. The pre-eVMR error rates were already so low (~1%) that it was impractical to evaluate the statistical significance of feature combinations. Instead, it can be observed that eVMR by itself is

capable of performing comparably against well-proven metrics such as ZCR, MFCCs, and their appropriate statistics while utilizing an incredibly basic algorithm and, accordingly, maintaining an incredibly low computational complexity. Regardless of the testing scenario, eVMR shows strength as a standalone feature, producing a very high classification accuracy on its own. Even while ignoring the application to the estimation of SMR, the findings indicate that eVMR remains a simple but viable metric for MIR.

5.2 Extensions

Larger sets of music would likely prove to be beneficial as well in the event of a follow-up study on electronic music classification, especially given the sensitivity of Gaussian mixture models to their corresponding training sets. As is evident from the analysis regarding statistical significance regarding eVMR, results would be expected to improve in true accuracy as larger and more representative collections were implemented for training and testing purposes. Additionally, the musicological scope could be expanded to other active electronic subgenres such electro house, tech house, drum & bass, and trance. For speech/music discrimination, additional languages of speech and other genres of music could be added to the dataset in order to make it more representative.

As utilized for speech/music experiments, the contributions of supplementary features (e.g. mean, standard deviation, and other statistics) could be investigated for their effectiveness in differentiating subgenres. Other genre-related extensions include a hierarchal classification technique per Burred and Lerch [9] since it is likely, for example, that harmonic descriptors may be more important for ultra-melodic styles such as progressive house than rougher styles like tech house. Further forms of machine-learning could be used to supplement or replace Gaussian mixture models for all experimental settings as well.

Chapter 6. Conclusion

Experimentation showed that current electronic music subgenres can be effectively classified despite the overlaps and vagueness of genre definitions in general. Audio features such as brightness, spectral rolloff, and Mel-frequency cepstral coefficients were computed for a set of thirty electronic music samples and used to develop Gaussian mixture models for categorization. A final classification accuracy of 80.67% was attained, comparable to previous genre classification results given the scope of both the dataset used and the study as a whole.

Additionally, work was done to develop a simple feature for audio classification that approximates the ratio of energies present in recorded vocal or speech and instrumental parts based on their typical locations in stereo mixes. eVMR proved to be a useful feature in speech/music discrimination applications, achieving success rates of 81.9% for three-way speech/music/mixed classification and 97.9% for two-way classification. Three-way classification was improved to 92.6% accuracy upon the inclusion of several spectral features, a performance which compares well with prior work.

eVMR was also examined in a novel application regarding mixed audio. Thousands of samples were synthesized by combining pure speech and pure music files in various proportions. GMMs were trained to perform an automatic estimation of the intrinsic speech-to-music ratio of these clips via a classification framework. A clear linear relationship between the actual and estimated SMR was evident upon plotting the results, verifying the feasibility of SMR estimation as a possible component algorithm in systems designed for the enhancement of broadcast media consumption and automatic speech recognition. SMR estimation carried out with the eVMR feature significantly outperformed the best available systems not including eVMR.

References

[1] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic Genre Classification of Music Content: A Survey," IEEE Signal Processing Mag., pp. 133-141, March 2006.

[2] E. Scheirer and M. Slaney, “Construction and evaluation of a robust multifeature speech/music discriminator,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing ’97, Munich, Germany, 1997.

[3] C. McKay and I. Fujinaga, “Automatic Genre Classification Using Large High-Level Musical Feature Sets,” in Proc. of the Int. Conf. on Music Information Retrieval 2004, Barcelona, Spain, Oct. 2004.

[4] G. Tzanetakis and P. Cook, “Musical Genre Classification of Audio Signals,” IEEE Trans. Speech Audio Process., vol. 10, no. 5, pp. 293-302, Jul. 2002.

[5] T. Li, M. Ogihara, and Q. Li, “A Comparative Study on Content-Based Music Genre Classification,” in Proc. of the Special Interest Group on Information Retrieval 2003, Toronto, Canada, Jul. 2003, pp. 282-289.

[6] M.F. McKinney and J. Breebaart, “Features for Audio and Music Classification,” in Symposium on Intelligent Algorithms, Eindhoven, the Netherlands, Dec. 2002.

[7] T. Lidy and A. Rauber, “Evaluation of Feature Extractors and Psycho-Acoustic Transformations for Music Genre Classification,” Vienna Univ. of Tech., Vienna, Austria, 2005.

[8] T. Pohle, E. Pampalk, and G. Widmer, “Evaluation of Frequently Used Audio Features for Classification of Music into Perceptual Categories,” in Proc. of the Fourth International Workshop on Content-Based Multimedia Indexing, Riga, Latvia, Jun. 2005

[9] J. J. Burred and A. Lerch, “A Hierarchal Approach to Automatic Musical Genre Classification,” in Proc. of the 6th Int. Conference on Digital Audio Effects, London, UK, Sept. 2003.

[10] E. Pampalk, A. Flexer, and G. Widmer, “Improvements of Audio-Based Music Similarity and Genre Classification,” in Proc. of the 6th Int. Conference on Digital Audio Effects, London, UK, Sept. 2003.

[11] J. Saunders, “Real-time discrimination of broadcast speech/music,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing ‘96, Atlanta, GA, 1996, pp. 993-996.

[12] J. Pinquier, J. Rouas, and R. André-Obrecht, “A fusion study in speech/music classification,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing ’03, Hong Kong, 2003, pp. 409-412.

[13] M. J. Carey, E. S. Parris, and H. Lloyd-Thomas, “A comparison of features for speech, music discrimination,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing ’99, Phoenix, AZ, 1999, pp. 149-152.

[14] O. M. Mubarak , E. Ambikairajah, and J. Epps, “Novel features for effective speech and music discrimination,” in Proc. Int. Conf. Engineering Intelligent Systems ’06, Islamabad, 2006, pp. 1-5.

[15] J. E. Munoz-Exposito, S. Garcia-Galán, and N. Ruiz-Reyes et al., “Speech/music discrimination using a single warped LPC-based feature,” in Proc. Int. Conf. Music Information Retrieval ’05, London, 2005, pp. 614-617.

[16] K. El-Maleh, M. Klein, and G. Petrucci et al., “Speech/music discrimination for multimedia applications,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing ’00, Istanbul, 2000, pp. 2245-2448.

[17] Y. Lavner and D. Ruinskiy, “A decision-tree based algorithm for speech/music classification and segmentation,” EURASIP J. Audio, Speech, Music Processing, vol. 2009, Jan. 2009.

[18] E. Didiot, I. Illina, and D. Fohr et al., “A wavelet-based parameterization for speech/music discrimination,” Comput. Speech Lang., vol. 24, no. 2, pp. 341-357, Apr. 2010.

[19] L. Lu, H. Zhang, and S. Z. Li, “Content-based audio classification and segmentation by using support vector machines,” Multimedia Sys., vol. 8, no. 6, pp. 482-492, 2003.

[20] T. Zhang and C. J. Kuo, “Audio content analysis for online audiovisual data segmentation and classification,” IEEE Trans. Speech Audio Processing, vol. 9, no. 4, pp. 441-457, May 2001.

[21] D. Li, I. K. Sethi, and N. Dimitrova et al., “Classification of general audio data for content-based retrieval,” Pattern Recog. Lett., vol. 22, no. 5, pp. 533-544, April 2001.

[22] J. Razik, C. Sénac, and D. Fohr et al., “Comparison of two speech/music segmentation systems for audio indexing on the web,” in Proc. Multi Conf. Systemics, Cybernetics, Informatics, Orlando, FL 2003.

[23] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” IEEE Trans. Acoust., Speech, Signal Processing, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[24] V. Breshears, “Mixing techniques for multi-channel (left/center/right) sound reinforcement systems,” in AES 111th Convention, New York, 2001.

[25] J. Eargle, “Mixing and Mastering” in Handbook of Recording Engineering, 4th ed. New York: Springer, 2003, ch. 22, pp.326-337.

[26] T. E. Rudolph and V. A. Leonard, “Mixing” in Recording in the Digital World: Complete Guide to Studio Gear and Software, 1st. ed. Boston: Berklee, 2001, ch. 18, pp. 213-222.

[27] D. Eck, “A Tempo-Extraction Algorithm Using an Autocorrelation Phase Matrix and Shannon Entropy,” Univ. of Montreal, Montreal, Canada, 2005.

[28] O. Lartillot and P. Toiviainen, “A Matlab toolbox for musical feature extraction from audio,” in Proc. Int. Conf. Digital Audio Effects, Bordeaux, 2007.

[29] G. Peeters, “A Generic System for Audio Indexing: Application to Speech/Music Segmentation and Music Genre Recognition,” in Proc. of the 10th Int. Conf. on Digital Audio Effects, Bordeaux, France, Sept. 2007.

[30] X. Zhuang, X. Zhou, and M. Hasegawa-Johnson et al. “Face age estimation using patch-based hidden Markov model supervectors,” in Proc Int. Conf. Pattern Recognition ’08, Tampa, FL, 2008, pp. 1-4.

[31] J. B. Allen and S. T. Neely, “Modeling the relation between the intensity just-noticeable difference and loudness for pure tones and wideband noise,” J. Acoust. Soc. Am., vol. 102, no. 6, pp. 3628-3646, 1997.

[32] B. Schuller, G. Rigoll, and M. Lang, “HMM-based music retrieval using stereophonic feature information and framelength adaptation,” in Proc. Int. Conf. Multimedia Expo ’03, Baltimore, 2003, pp. 713-716.

[33] A. Duda, A. Nürnberger, and Sebastian Stober, “Towards query by singing/humming on audio databases,” in Proc. Int. Conf. Music Information Retrieval ’07, Vienna, 2007.

[34] F. Weninger, B. Schuller, and C. Liem et al., “Music information retrieval: an inspirational guide to transfer from related disciplines,” Multimodal Music Processing, vol. 3, pp. 195-216, 2012.

[35] C. Avendano, “Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications,” in Proc. IEEE Workshop Applications Signal Processing Audio Acoustics, New Paltz, NY, 2003, pp. 55–58.

[36] O. L. Frost, “An algorithm for linearly constrained adaptive array processing,” in Proc. IEEE, vol. 60, no. 8, pp. 926-935, 1972.

[37] B. Lee, “Robust speech recognition in a car using a microphone array,” Ph.D. thesis, Dept. Elect. and Comp. Eng., Univ. of Illinois, Champaign, IL, 2006.

[38] F. Weninger, J. Feliu, and B. Schuller, “Supervised and semi-supervised suppression of background music in monaural speech recordings,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing ’12, Kyoto, 2012, pp. 61-64.

[39] T. Hughes and T. Kristjansson, “Music models for music-speech separation,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing ’12, Kyoto, 2012, pp. 4917-4920.

[40] K. Minami, A. Akutsu, and H. Hamada et al., “Video handling with music and speech detection,” MultiMedia IEEE, vol. 5, no. 3, pp. 17-25, Jul. 1998.

[41] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, pp. 788-791, Oct. 1999.

[42] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in Proc. Adv. Neural Information Processing Sys 13, Denver, CO, 2000, pp. 556-562.

[43] C. Févotte and J. Idier, “Algorithms for nonnegative matrix factorization with the β-divergence,” Neural Comput., vol. 23, no. 9, pp. 2421-2456, Sep. 2011.

[44] M. Kim and P. Smaragdis, “Single Channel Source Separation Using Smooth Nonnegative Matrix Factorization with Markov Random Fields," in Proc. IEEE Workshop Machine Learning Signal Processing, Southampton, UK, 2013.

[45] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust., Speech, Signal Process., vol. 28, no. 4, pp. 357–366, 1980.

[46] I. Y. Ozbek, M. Hasegawa-Johnson, and M. Demirekler, “Estimation of articulatory trajectories based on Gaussian mixture model (GMM) with audio-visual information fusion and dynamic Kalman smoothing,” IEEE Trans. Acoust., Speech, Signal Process., vol. 19, no. 5, pp. 1180–1195, 2010.

[47] B. H. Juang, S. E. Levinson, and M. M. Sondhi, “Maximum likelihood estimation for multivariate mixture observations of Markov chains,” IEEE Trans. Information Theory, vol. 32, no. 2, pp. 307-309, Mar. 1986.

[48] H. W. Sorenson and D. L. Alspach, “Recursive Bayesian estimation using Gaussian sums,” Automatica, vol. 7, no. 4, pp. 456-479, 1971.

[49] G. Williams and D. P. W. Ellis, “Speech/music discrimination based on posterior probability features,” in Proc. Euro. Conf. Speech Communication and Technology ’99, Budapest, 1999.

[50] A. L. Berenzeweig and D. P. W. Ellis, “Locating singing voice segments within music signals,” in Proc. IEEE Workshop Applications Signal Processing Audio Acoustics, Mohonk, NY, 2001, pp. 119–123.

[51] E. Alexandre-Cortizo, M. Rosa-Zurera, and F. López-Ferreras, “Application of fisher linear discriminant analysis to speech/music classification,” in Proc. Int. Conf. Comp. Tool (EUROCON), Belgrade, 2005, pp. 1666-1669.

[52] L. Gillick and S. J. Cox, “Some statistical issues in the comparison of speech recognition algorithms,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing ’89, Glasgow, 1989, pp. 532-535.

[53] B. Logan, “Mel Frequency Cepstral Coefficients for Music Modeling,” in Proc. of the Int. Conf. on Music Information Retrieval 2000, Plymouth, MA, Oct. 2000.

[54] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” IEEE Trans. Speech Audio Processing, vol. 33, no. 2, pp. 443-445, Apr. 1985.

[55] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” Proc. IEEE, vol.67, no. 12, pp. 1586-1604, Dec. 1979.

[56] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” IEEE Signal Process. Lett., vol. 6, no. 1, pp. 1-3, Nov. 2011.

[57] L. Kim, K. Kim, and M. Hasegawa-Johnson, “Robust automatic speech recognition with decoder oriented ideal binary mask estimation,” in Proc. Interspeech ’10, Makahari, 2010.

[58] J. Ajmera, I. McCowan, and H. Bourlard, “Speech/music segmentation using entropy and dynamism features in a HMM classification framework,” Speech Communication, vol. 40, no. 3, pp. 351-363, May 2003.

[59] M. Markaki and Y. Stylianou, “Discrimination of speech from nonspeech in broadcast news based on modulation frequency features,” Speech Communication, vol. 53, no. 5, pp. 726-735, 2011.

[60] A. Misra, “Speech/nonspeech segmentation in web videos,” in Proc. Interspeech ’12, New York, 2012.

Appendix A. Code Listing

A.1 Electronic Music Classification Code

A.1.1 Main Script

```
clc
clear

%% Check for correct directory
try
    cd AudioSamples
catch
    error('Please change the working directory to the ''EGC'' folder.')
end

%% Load the audio samples for training
cd train1
train1 = miraudio('Folder','Label',1:4);
cd ..
clc

cd train2
train2 = miraudio('Folder','Label',1:4);
cd ..
clc

cd train3
train3 = miraudio('Folder','Label',1:4);
cd ..
clc

%% Load the audio samples for testing
cd test1
test1 = miraudio('Folder','Label',1:4);
cd ..
clc

cd test2
test2 = miraudio('Folder','Label',1:4);
cd ..
clc

cd test3
```

```
test3 = miraudio('Folder','Label',1:4);
cd ..
clc

cd ..

%% Compute the features for each sample using MIR Toolbox
EGC_Features

%% Classify the test sets using Gaussian mixture modeling
EGC_Classify
```

A.1.2 Features Script

```
%% Calculate features for each training set
low_train1 = mirlowenergy(train1);
zero_train1 = mirzerocross(train1);
bright_train1 = mirbrightness(train1);
mfcc_train1 = mirmfcc(train1,'Rank',1:5);
roll_train1 = mirrolloff(train1);
tempo_train1 = mirtempo(train1,'Min',80);

low_train2 = mirlowenergy(train2);
zero_train2 = mirzerocross(train2);
bright_train2 = mirbrightness(train2);
mfcc_train2 = mirmfcc(train2,'Rank',1:5);
roll_train2 = mirrolloff(train2);
tempo_train2 = mirtempo(train2,'Min',80);

low_train3 = mirlowenergy(train3);
zero_train3 = mirzerocross(train3);
bright_train3 = mirbrightness(train3);
mfcc_train3 = mirmfcc(train3,'Rank',1:5);
roll_train3 = mirrolloff(train3);
tempo_train3 = mirtempo(train3,'Min',80);

%% Calculate features for each test set
low_test1 = mirlowenergy(test1);
zero_test1 = mirzerocross(test1);
bright_test1 = mirbrightness(test1);
mfcc_test1 = mirmfcc(test1,'Rank',1:5);
roll_test1 = mirrolloff(test1);
tempo_test1 = mirtempo(test1,'Min',80);

low_test2 = mirlowenergy(test2);
zero_test2 = mirzerocross(test2);
bright_test2 = mirbrightness(test2);
mfcc_test2 = mirmfcc(test2,'Rank',1:5);
```

```
roll_test2 = mirrolloff(test2);
tempo_test2 = mirtempo(test2,'Min',80);

low_test3 = mirlowenergy(test3);7
zero_test3 = mirzerocross(test3);
bright_test3 = mirbrightness(test3);
mfcc_test3 = mirmfcc(test3,'Rank',1:5);
roll_test3 = mirrolloff(test3);
tempo_test3 = mirtempo(test3,'Min',80);
```

A.1.3 Classification Script

```
%% Classify the test sets using Gaussian Mixture Models

classify1G1 = mirclassify(test1,{low_test1, zero_test1, bright_test1, ...
                                 mfcc_test1, roll_test1, tempo_test1}, ...
                          train1,{low_train1, zero_train1, bright_train1, ...
                                 mfcc_train1, roll_train1, tempo_train1}, ...
                                                          'GMM', 1);

classify1G2 = mirclassify(test1,{low_test1, zero_test1, bright_test1, ...
                                 mfcc_test1, roll_test1, tempo_test1}, ...
                          train1,{low_train1, zero_train1, bright_train1, ...
                                 mfcc_train1, roll_train1, tempo_train1}, ...
                                                          'GMM', 2);

classify1G3 = mirclassify(test1,{low_test1, zero_test1, bright_test1, ...
                                 mfcc_test1, roll_test1, tempo_test1}, ...
                          train1,{low_train1, zero_train1, bright_train1, ...
                                 mfcc_train1, roll_train1, tempo_train1}, ...
                                                          'GMM', 3);

classify1G4 = mirclassify(test1,{low_test1, zero_test1, bright_test1, ...
                                 mfcc_test1, roll_test1, tempo_test1}, ...
                          train1,{low_train1, zero_train1, bright_train1, ...
                                 mfcc_train1, roll_train1, tempo_train1}, ...
                                                          'GMM', 4);

classify1G5 = mirclassify(test1,{low_test1, zero_test1, bright_test1, ...
                                 mfcc_test1, roll_test1, tempo_test1}, ...
                          train1,{low_train1, zero_train1, bright_train1, ...
                                 mfcc_train1, roll_train1, tempo_train1}, ...
                                                          'GMM', 5);


classify2G1 = mirclassify(test2,{low_test2, zero_test2, bright_test2, ...
                                 mfcc_test2, roll_test2, tempo_test2}, ...
                          train2,{low_train2, zero_train2, bright_train2, ...
                                 mfcc_train2, roll_train2, tempo_train2}, ...
                                                          'GMM', 1)

classify2G2 = mirclassify(test2,{low_test2, zero_test2, bright_test2, ...
                                 mfcc_test2, roll_test2, tempo_test2}, ...
                          train2,{low_train2, zero_train2, bright_train2, ...
```

```
                                mfcc_train2, roll_train2, tempo_train2}, ...
                                                        'GMM', 2)

classify2G3 = mirclassify(test2,{low_test2, zero_test2, bright_test2, ...
                                mfcc_test2, roll_test2, tempo_test2}, ...
                          train2,{low_train2, zero_train2, bright_train2, ...
                                mfcc_train2, roll_train2, tempo_train2}, ...
                                                        'GMM', 3)

classify2G4 = mirclassify(test2,{low_test2, zero_test2, bright_test2, ...
                                mfcc_test2, roll_test2, tempo_test2}, ...
                          train2,{low_train2, zero_train2, bright_train2, ...
                                mfcc_train2, roll_train2, tempo_train2}, ...
                                                        'GMM', 4)

classify2G5 = mirclassify(test2,{low_test2, zero_test2, bright_test2, ...
                                mfcc_test2, roll_test2, tempo_test2}, ...
                          train2,{low_train2, zero_train2, bright_train2, ...
                                mfcc_train2, roll_train2, tempo_train2}, ...
                                                        'GMM', 5)


classify3G1 = mirclassify(test3,{low_test3, zero_test3, bright_test3, ...
                                mfcc_test3, roll_test3, tempo_test3}, ...
                          train3,{low_train3, zero_train3, bright_train3, ...
                                mfcc_train3, roll_train3, tempo_train3}, ...
                                                        'GMM', 1)

classify3G2 = mirclassify(test3,{low_test3, zero_test3, bright_test3, ...
                                mfcc_test3, roll_test3, tempo_test3}, ...
                          train3,{low_train3, zero_train3, bright_train3, ...
                                mfcc_train3, roll_train3, tempo_train3}, ...
                                                        'GMM', 2)

classify3G3 = mirclassify(test3,{low_test3, zero_test3, bright_test3, ...
                                mfcc_test3, roll_test3, tempo_test3}, ...
                          train3,{low_train3, zero_train3, bright_train3, ...
                                mfcc_train3, roll_train3, tempo_train3}, ...
                                                        'GMM', 3)

classify3G4 = mirclassify(test3,{low_test3, zero_test3, bright_test3, ...
                                mfcc_test3, roll_test3, tempo_test3}, ...
                          train3,{low_train3, zero_train3, bright_train3, ...
                                mfcc_train3, roll_train3, tempo_train3}, ...
                                                        'GMM', 4)

classify3G5 = mirclassify(test3,{low_test3, zero_test3, bright_test3, ...
                                mfcc_test3, roll_test3, tempo_test3}, ...
                          train3,{low_train3, zero_train3, bright_train3, ...
                                mfcc_train3, roll_train3, tempo_train3}, ...
                                                        'GMM', 5)

clc
fprintf('\n-------------------- Test Set 1 --------------------\n')
classify1G1
classify1G2
classify1G3
```

```
classify1G4
classify1G5

fprintf('\n-------------------- Test Set 2 --------------------\n')
classify2G1
classify2G2
classify2G3
classify2G4
classify2G5

fprintf('\n-------------------- Test Set 3 --------------------\n')
classify3G1
classify3G2
classify3G3
classify3G4
classify3G5
```

A.2 Speech/Music Code

A.2.1 Main Script

```
clc
clear

%% Check for correct directory
try
    cd AudioSamples
catch
    error('Please change the working directory to the ''SMD'' folder.')
end
cd ..

%% Define the length of the samples and windows for frame decomposition
samplength = 4;
framelength = 0.05;

%% Compute the features for each sample using MIR Toolbox
SMD_LoadFeat

%% Classify the test sets using Gaussian mixture modeling
SMD_Classify
```

A.2.2 Features Script

```
%% Loads audio samples and computes their features for SMR classification
%% Optimized for memory management
```

```
cd AudioSamples

%% Load the audio samples for testing
cd test1 %SMDtest
test1 = miraudio('Folder','Extract',0,samplength,'s','Start', 'Label',1:3);
    test1s = miraudio('Folder','Mono',0,'Extract',0,samplength,'s','Start',
'Label',1:3);
    evmr_test1 = eVMR(test1s);
    clear test1s
test1f = mirframe(test1,'Length',framelength,'s');
cd ..
clc

%BSSpow_test1 = BSSpow(test1);
zero_test1 = mirzerocross(test1);
zero_test1f = mirzerocross(test1f);
zero_std_test1 = mirstd(zero_test1f);
zero_mean_test1 = mean(mirgetdata(zero_test1f),2)';
zero_skew_test1 = skewness(mirgetdata(zero_test1f),2,2)';
lowen_test1 = mirlowenergy(test1);
silence_test1 = mirlowenergy(test1,'ASR');
rolloff_test1 = mirrolloff(test1);
rms_test1 = mirrms(test1);

mfcc0_test1 = mirmfcc(test1,'Rank',0:0);
mfccA_test1 = mirmfcc(test1,'Rank',1:5);
mfccB_test1 = mirmfcc(test1,'Rank',1:13);
mfcc0f_test1 = mirgetdata(mirmfcc(test1f,'Rank',0:0));
mfcc0_mean_test1 = mean(mfcc0f_test1,2)';
mfcc0_std_test1 = std(mfcc0f_test1,0,2)';
[mfcc_d1m_test1, mfcc_d1s_test1, mfcc_d2m_test1, mfcc_d2s_test1] = ...
    MFCC_Deltas(test1,test1f);

clear test1 test1f

%% Load the audio samples for training
if exist('train1') ~= 1 % ... if they're not already loaded

    cd train1 %SMDtrain
    train1 = miraudio('Folder','Extract',0,samplength,'s','Start',
'Label',1:3);
        train1s =
miraudio('Folder','Mono',0,'Extract',0,samplength,'s','Start', 'Label',1:3);
        evmr_train1 = eVMR(train1s);
        clear train1s
    train1f = mirframe(train1,'Length',framelength,'s');
    cd ..
    clc

    %BSSpow_train1 = BSSpow(train1);
    zero_train1 = mirzerocross(train1);
    zero_train1f = mirzerocross(train1f);
    zero_std_train1 = mirstd(zero_train1f);
    zero_mean_train1 = mean(mirgetdata(zero_train1f),2)';
    zero_skew_train1 = skewness(mirgetdata(zero_train1f),2,2)';
    lowen_train1 = mirlowenergy(train1);
    silence_train1 = mirlowenergy(train1,'ASR');
```

```
    rolloff_train1 = mirrolloff(train1);
    rms_train1 = mirrms(train1);

    mfcc0_train1 = mirmfcc(train1,'Rank',0:0);
    mfccA_train1 = mirmfcc(train1,'Rank',1:5);
    mfccB_train1 = mirmfcc(train1,'Rank',1:13);
    mfcc0f_train1 = mirgetdata(mirmfcc(train1f,'Rank',0:0));
    mfcc0_mean_train1 = mean(mfcc0f_train1,2)';
    mfcc0_std_train1 = std(mfcc0f_train1,0,2)';
    [mfcc_d1m_train1, mfcc_d1s_train1, mfcc_d2m_train1, mfcc_d2s_train1] =
...
        MFCC_Deltas(train1,train1f);

    clear train1 train1f

end

%% Reload test and training sets for classification
cd test1 %SMDtest
test1 = miraudio('Folder','Extract',0,samplength,'s','Start', 'Label',1:3);
cd ..
cd train1 %SMDtrain
train1 = miraudio('Folder','Extract',0,samplength,'s','Start', 'Label',1:3);
cd ..
clc

%% Return to the parent directory
cd ..
```

A.2.3 Classification Script

```
%% Classify the test sets using Gaussian Mixture Models

GMM = [1:5];
N = length(GMM);

for g = 1:N

classify1(g) = mirclassify(test1,{zero_test1, zero_mean_test1,...
                                  zero_std_test1, zero_skew_test1, ...
evmr_test1, ...
BSSpow_test1, ...
                                  lowen_test1, ...
                                  rms_test1, rolloff_test1, ...
                                  mfcc0_test1, mfcc0_mean_test1,
mfcc0_std_test1, ...
mfcc_d1m_test1, mfcc_d1s_test1, mfcc_d2m_test1, mfcc_d2s_test1, ...
                                  mfccA_test1, mfccB_test1}, ...
                           train1,{zero_train1, zero_mean_train1, ...
                                  zero_std_train1, zero_skew_train1, ...
evmr_train1, ...
BSSpow_train1, ...
                                  lowen_train1, ...
                                  rms_train1, rolloff_train1, ...
```

```
                                mfcc0_train1, mfcc0_mean_train1,
mfcc0_std_train1, ...
mfcc_d1m_train1, mfcc_d1s_train1, mfcc_d2m_train1, mfcc_d2s_train1, ...
                                mfccA_train1, mfccB_train1}, ...
                                                        'GMM', g);
end

clc
fprintf('\n-------------------- Classification Results --------------------
\n')

sum = 0;
for g = 1:N
    classify1(g)
    sum = sum + mirgetdata(classify1(g));
end

average = sum/N
```

Appendix B. Audio Sources

This appendix contains information regarding the original sources for all audio samples utilized for experimental purposes. Precise extraction points for each clip in the database have been included to ensure the reproducibility of results.

B.1 Electronic Music Classification Sources

Avicii - "Levels"
http://www.amazon.com/Levels-Avicii/dp/B006ZZAMNO/
02:30;00-03:00;00

Avicii ft. Salem Al Fakir - "Silhouettes"
http://www.amazon.com/Silhouettes/dp/B0084IZFY6/
02:30;00-03:00;00

Bassnectar - "Bass Head" from Timestretch
http://www.amazon.com/Timestretch-Bassnectar/dp/B003BOJAAS/
01:30;00-2:00;00

Basto - "Again and Again (Extended Mix)"
http://www.beatport.com/track/again-and-again-extended-mix/2996293
02:30;00-03:00;00

Calvin Harris - "Feel So Close (Nero Remix)"
http://www.amazon.com/Feel-So-Close-Nero-Remix/dp/B00B9ER9FA/
01:30;00-2:00;00

Coldplay - "Paradise (Mostafa Negm Dubstep Remix)"
http://www.last.fm/music/Coldplay/_/Paradise+(Mostafa+Negm+Dubstep+remix)
01:30;00-2:00;00

David Guetta and Nicky Romero - "Metropolis"
http://www.beatport.com/track/metropolis-original-mix/3413298
02:30;00-03:00;00

Deep Unit & Don Uli - "Erosion (Stefano Libelle & Der Pender Remix)"
http://www.amazon.com/gp/product/B007X07SD2/ref=dm_ws_sp_ps_dp
02:30;00-03:00;00

Ellie Goulding - "Lights (Bassnectar Remix)" from Lights (The Remixes Part 1)
http://www.amazon.com/gp/product/B0050TITXG/
01:30;00-2:00;00

Flux Pavilion ft. Example - "Daydreamer"
http://www.amazon.com/Daydreamer-feat-Example-Extended-Version/dp/B007USZVWM/
01:30;00-2:00;00

Fuseboxers - "La Tina (Mio Martini Remix)"
http://www.amazon.com/gp/product/B007X07SD2/ref=dm_ws_sp_ps_dp
02:30;00-03:00;00

Goose Bumps - "Truba (Tiff & Trashkid Remix)"
http://www.amazon.com/gp/product/B007X07SD2/ref=dm_ws_sp_ps_dp
02:30;00-03:00;00

Hardwell - "Spaceman (Original Mix)"
http://www.beatport.com/track/spaceman-original-mix/3245917
02:30;00-03:00;00

Hardwell & Dannic - "Kontiki (Original Mix)"
http://www.beatport.com/track/kontiki-original-mix/3414486
02:30;00-03:00;00

Jack Back ft. David Guetta, Nicky Romero, and Sia - "Wild One Two"
http://www.amazon.com/Feat-David-Guetta-Nicky-Romero/dp/B007M45S18/
02:30;00-03:00;00

Knife Party - "Centipede (original_mix)" from Rage Valley EP
http://www.amazon.com/Rage-Valley-EP-Knife-Party/dp/B0086EYCF6/
01:30;00-2:00;00

Knobs & Wires - "Desire"
http://www.amazon.com/gp/product/B007X07SD2/ref=dm_ws_sp_ps_dp
02:30;00-03:00;00

The Naked And Famous - "Young Blood (Tiësto & Hardwell Remix)" from Club Life - Volume 2 Miami
http://www.amazon.com/Club-Life-2-Miami-Ti%C3%ABsto/dp/B007NCTFJ0/
02:30;00-03:00;00

Nero - "Promises (Skrillex Remix)"
http://www.amazon.com/Promises-Skrillex-Nero-Remix/dp/B009PHMFXO/
01:30;00-2:00;00

Nervo ft. Hook N Sling - "Reason (Original Mix)"
http://www.beatport.com/track/reason-original-mix/4014561
02:30;00-03:00;00

Nick Waters and David Hopperman - "Feel the Beat (Original Mix)"
http://www.amazon.com/gp/product/B007X07SD2/ref=dm_ws_sp_ps_dp
02:30;00-03:00;00

Plusculaar - "Dusty Book"
http://www.amazon.com/gp/product/B007X07SD2/ref=dm_ws_sp_ps_dp
02:30;00-03:00;00

Rihanna - "You Da One (Fonik Remix)"
https://soundcloud.com/fonik-uk/rihanna-you-da-one-fonik-remix
01:30;00-2:00;00

Roma Alkhimov - "Gloss"
http://www.amazon.com/gp/product/B007X07SD2/ref=dm_ws_sp_ps_dp
02:30;00-03:00;00

Ron Ractive - "Flora und Fauna"
http://www.amazon.com/gp/product/B007X07SD2/ref=dm_ws_sp_ps_dp
02:30;00-03:00;00

Shakes ft. Kneon - "8Iron Driver (Our Mix)"
http://www.amazon.com/gp/product/B007X07SD2/ref=dm_ws_sp_ps_dp
02:30;00-03:00;00

Skrillex - "Scary Monsters And Nice Sprites"
http://www.amazon.com/Scary-Monsters-And-Nice-Sprites/dp/B004KRQFI0/
01:30;00-2:00;00

Skrillex ft. Sirah - "Bangarang"
http://www.amazon.com/Bangarang-Skrillex/dp/B005FLX1HS/
01:30;00-2:00;00

The Sun Warrios - "Milk"
http://www.amazon.com/gp/product/B007X07SD2/ref=dm_ws_sp_ps_dp
02:30;00-03:00;00

Swedish House Mafia ft. John Martin - "Don't You Worry Child (Extended Mix)"
http://www.amazon.com/Dont-Worry-Child-feat-Martin/dp/B009FRCWAU/
02:30;00-03:00;00

B.2 Speech/Music Sources

B.2.1 Music Sources

Atomic Tom - "Take Me Out" from the Moment
http://www.amazon.com/Moment-Atomic-Tom/dp/B003W8QODY/
00:31;03-00:35;03
00:57;01-01:01;01
02:29;28-02:33;28

Avicii - "Levels"
http://www.amazon.com/Levels-Avicii/dp/B006ZZAMNO/
00:06:04-00:10:04
00:18;26-00:22;26
01:55;19-01:59;19

Avicii ft. Salem Al Fakir - "Silhouettes"
http://www.amazon.com/Silhouettes/dp/B0084IZFY6/
00:29;20-00:33;20
01:00;02-01:04;02
02:11;07-02:15;07

B.o.B ft. Lil Wayne - "Strange Clouds" from Strange Clouds
http://www.amazon.com/Strange-Clouds-B-B/dp/B0071BXZGM/
00:54;02-00:58;02
01:52;07-01:56;07

02:26;20-02:30;20

The Band Perry - "If I Die Young" from The Band Perry
http://www.amazon.com/The-Band-Perry/dp/B003ZDZ1WG/
00:33;16-00:37;16
01:10;19-01:14;19
02:02;17-02:06;17

Basto - "Again and Again (Extended Mix)"
http://www.beatport.com/track/again-and-again-extended-mix/2996293
00:52;28-00:56;28
01:12;05-01:16;05
02:03;02-02:07;02

Beethoven - "Fur Elise" from 100 Masterpieces, Vol. 4 (1788-1810)
http://www.amazon.com/100-Masterpieces-Vol-4-Classical-Music/dp/B002QPQF8I
01:07;00-01:11;00

Big Sean ft. Chris Brown - "My Last (Clean)" from Finally Famous
http://www.amazon.com/Finally-Famous-Big-Sean/dp/B004NDVKGI/
00:45;25-00:49;25
01:02;28-01:06;28
02:28;14-02:32;14

Brahms - "Waltz" from 100 Masterpieces, Vol. 7 (1854-1866)
http://www.amazon.com/100-Masterpieces-Vol-7-Classical-Music/dp/B002QPQBV4/
00:44;06-00:48;06
01:02;11-01:06;11
01:38;28-01:42;28

Brett Eldredge - "Signs" from Bring You Back
http://www.amazon.com/Bring-You-Back-Brett-Eldredge/dp/B0053G7A18/
00:35;07-00:39;07
01:05;15-01:09;15
01:47;07-01:51;07

Bruno Mars - "Just the Way You Are" from Doo-Wops & Hooligans
http://www.amazon.com/Doo-Wops-Hooligans-Bruno-Mars/dp/B003ZJ0ZX0/
00:31;16-00:35;16
00:54;23-00:59;23

03:07;25-03:11;25

Chopin - "Polonaise in A, Op. 40 No. 3" from 100 Masterpieces, Vol. 5 (1811-1841)
http://www.amazon.com/100-Masterpieces-Vol-5-Classical-Music/dp/B002QPMDZ2/
01:35;02-01:39;02
01:50;09-01:54;09

David Guetta and Nicky Romero - "Metropolis"
http://www.beatport.com/track/metropolis-original-mix/3413298
01:02;24-01:06;24
02:25;20-02:29;20
03:03;07-03:07;07

Foster the People - "Helena Beat" from Torches
http://www.amazon.com/Torches-Foster-People/dp/B004UUKDNA/
00:34;10-00:38;10

The Fratellis - "Chelsea Dagger" from Costello Music
http://www.amazon.com/Costello-Music-Fratellis/dp/B000MXPE74/
00:09;09-00:13;09
00:26;01-00:30;01
00:57;24-01:01;24

Good Charlotte - "The Anthem" from The Young and the Hopeless
http://www.amazon.com/The-Young-And-Hopeless/dp/B008DVJOMI/
00:29;23-00:33;23
00:49;27-00:53;27
01:01;29-01:05;29

Handel - "Largo (from 'Xerxes')" from 100 Masterpieces, Vol. 2 (1731-1775)
http://www.amazon.com/100-Masterpieces-Vol-2-Classical-Music/dp/B002QPIRAW/
00:19;19-00:23;19

Hardwell - "Spaceman (Original Mix)"
http://www.beatport.com/track/spaceman-original-mix/3245917
01:38;27-01:42;27
02:01;28-02:05;28
02:51;07-02:55;07

Hardwell & Dannic - "Kontiki (Original Mix)"

http://www.beatport.com/track/kontiki-original-mix/3414486
00:35;02-00:39;02
02:45;17-02:49;17
03:51;00-03:55;00

Jack Back ft. David Guetta, Nicky Romero, and Sia - "Wild One Two"
http://www.amazon.com/Feat-David-Guetta-Nicky-Romero/dp/B007M45S18/
00:13;29-00:17;29
01:07;09-01:11;09
02:10;23-02:14;23

Jay-Z - "Izzo (H.O.V.A)" from The Blueprint
http://www.amazon.com/The-Blueprint-Jay-Z/dp/B00005O54T/
01:17;20-01:21;20
02:02;14-02:06;14
02:53;02-02:57;02

Jessie James - "Boys In The Summer"
http://www.amazon.com/Boys-In-The-Summer/dp/B003T5KRJW/
00:14;27-00:16;27
00:37;15-00:41;15
00:52;06-00:56;06

Justin Bieber ft. Ludacris - "Baby" from My World 2.0
http://www.amazon.com/My-World-2-0-Justin-Bieber/dp/B0037AGASG/
00:27;08-00:31;08
01:02;11-01:06;11
01:30;29-01:34;29

Katy Perry - "Hot N Cold" from One of the Boys (820a6b0c)
http://www.amazon.com/One-Boys-Katy-Perry/dp/B0017ZB8M6/
00:11;14-00:15;14
00:32;22-00:36;22
01:16;16-01:20;16

Ke$ha - "Your Love Is My Drug" from Animal
http://www.amazon.com/Animal-Ke-ha/dp/B002XNEII2/
00:30;22-00:34;22
01:02;15-01:06;15
01:53;12-01:57;12

Lady Antebellum - "I Run To You" Lady Antebellum
http://www.amazon.com/Lady-Antebellum/dp/B0014CBXOK/
00:25;11-00:29;11
01:07;23-01:11;23
03:33;26-03:37;26

Lady Gaga - "Bad Romance" from The Fame Monster
http://www.amazon.com/Fame-Monster-Explicit-Lady-Gaga/dp/B0034KHFM4/
00:33;29-00:36;29
01:41;04-01:45;04
02:10;05-02:14;05

Leona Lewis - "Bleeding Love" from Spirit
http://www.amazon.com/Spirit-Leona-Lewis/dp/B0012TBGYC/
00:45;01-00:49;01
01:27;13-01:31;13
02:07;17-02:11;17

Lil Wayne ft. Drake - "Right Above It" from I Am Not a Human Being
http://www.amazon.com/Am-Not-Human-Being/dp/B00437DYLI/
01:06;15-01:11;15
02:06;25-02:11;25
02:56;26-03:01;26

Lloyd Banks ft. Kaney West, Swizz Beats, Ryan Leslie, and Fabolous, "Start It Up" from H.F.M. 2 (The Hunger for More 2)
http://www.amazon.com/H-F-M-2-Hunger-For-More-2/dp/B0046JLTBU/
01:42;24-01:46;24
02:32;19-02:36;19
03:18;07-03:22;07

Ludacris ft. Sean Garrett and Chris Brown "What Them Girls Like" from Theater of the Mind
http://www.amazon.com/Theater-Mind-Ludacris/dp/B001EUSYE4/
01:17;24-01:21;24
01:36;15-01:40;15
02:24;04-02:28;04

Lupe Fiasco - "Kick, Push" from Lupe Fiasco's Food & Liquor
http://www.amazon.com/Lupe-Fiascos-Food-Liquor-Fiasco/dp/B000FS9MTW/

01:16;21-01:20;21
01:49;09-01:53;09
02:58;07-03:02;07

Maroon 5 - "One More Night" from Overexposed
http://www.amazon.com/Overexposed-Explicit-Version-Maroon-5/dp/B008LC8TJI/
00:14;28-00:18;28
02:09;09-02:13;09
02:45;12-02:49;12

Matt Nathanson - "Detroit Waves" from Some Mad Hope
http://www.amazon.com/Some-Mad-Hope-Matt-Nathanson/dp/B000RHRFVI/
00:25;27-00:29;27
00:47;17-00:51;17
02:03;03-02:07;03

Mozart - "Clarinet Concerto in A, 2nd movement" from 100 Masterpieces, Vol. 4 (1788-1810)
http://www.amazon.com/100-Masterpieces-Vol-4-Classical-Music/dp/B002QPQF8I/
02:27;05-02:31;05
06:12;18-06:16;18

The Naked And Famous - "Young Blood (Tiësto & Hardwell Remix)" from Club Life - Volume 2 Miami
http://www.amazon.com/Club-Life-2-Miami-Ti%C3%ABsto/dp/B007NCTFJ0/
01:01;08-01:05;08
01:37;20-01:41;20
04:35;05-04:39;05

Natasha Bedingfield - "Pocketful of Sunshine" from Pocketful of Sunshine
http://www.amazon.com/Pocketful-Sunshine-Natasha-Bedingfield/dp/B000Y14U4M/
00:14;22-00:18;22
01:18;06-01:22;06
02:52;18-02:56;18

Nervo ft. Hook N Sling - "Reason (Original Mix)"
http://www.beatport.com/track/reason-original-mix/4014561
01:01;01-01:05;01
01:31;13-01:35;13
02:00;02-02:04;02

Nicki Minaj ft. Drake - "Moment 4 Life" from Pink Friday
http://www.amazon.com/Pink-Friday-Nicki-Minaj/dp/B0042RUMEQ/
01:38;05-01:42;05
02:31;28-02:35;28
03:22;22-03:26;22

O.A.R. - "Try Me" from All Sides
http://www.amazon.com/All-Sides-O-A-R/dp/B001AUKUVI/
00:21;23-00:25;23
01:42;12-01:46;12
02:57;07-03:01;07

Oasis - "Live Forever" from Definitely Maybe
http://www.amazon.com/Definitely-Maybe-Remastered-Deluxe-Edition/dp/B00IN5KX06/
00:41;11-00:45;11
01:08;03-01:12;03
01:52;01-01:56;01

Offenbach - "Barcarolle" from 100 Masterpieces, Vol. 7 (1854-1866)
http://www.amazon.com/100-Masterpieces-Vol-7-Classical-Music/dp/B002QPQBV4/
00:56;09-01:00;09
01:34;23-01:38;23
01:57;23-02:01;23

OutKast - "So Fresh, So Clean" from Stankonia
http://www.amazon.com/Stankonia-Outkast/dp/B00002R0MA/
01:00;08-01:04;08
02:02;14-02:06;14
02:52;28-02:56;28

Phoenix - "Lisztomania" from Wolfgang Amadeus Phoenix
http://www.amazon.com/Wolfgang-Amadeus-Phoenix/dp/B0021X515S/
00:09;20-00:13;20
01:12;24-01:16;24
02:54;21-02:58;21

Rascal Flatts - "What Hurts The Most" from Me and My Gang
http://www.amazon.com/Me-My-Gang-Rascal-Flatts/dp/B000JBXOC6/
00:47;28-00:51;28
01:19;24-01:23;24

02:20;04-02:24;04

Red Hot Chili Peppers - "Californication" from Californication
http://www.amazon.com/Californication-Red-Hot-Chili-Peppers/dp/B00000J7JO/
00:31;06-00:35;06
01:04;12-01:08;12
02:59;18-03:03;18

Santana ft. Alex Band "Why Don't You & I" from Why Don't You & I
http://www.amazon.com/Why-Dont-You-Alt-Version/dp/B00137OJ0Q/
00:15;01-00:19;01
00:28;07-00:32;07
02:03;24-02:07;24

Selena Gomez & the Scene - "Round & Round" from A Year Without Rain
http://www.amazon.com/Year-Without-Selena-Gomez-Scene/dp/B003YCI1PM/
00:12;24-00:16;24
00:50;23-00:54;23
01:47;00-01:51;00

Smetana - "The Moldau" from 100 Masterpieces, Vol. 8 (1867-1876)
http://www.amazon.com/100-Masterpieces-Vol-8-Classical-Music/dp/B002QPQEF2/
02:45;23-02:49;23
03:44;00-03:48;00
09:50;09-09:54;09

Something Corporate - "Hurricane" from Leaving Through the Window
http://www.amazon.com/Leaving-Through-Window-Something-Corporate/dp/B000066B4V/
00:28;29-00:32;29
00:37;25-00:41;25
01:00;12-01:04;12

Sugarland - "Stuck Like Glue" from The Incredible Machine
http://www.amazon.com/The-Incredible-Machine-Sugarland/dp/B003OUXEMY/
00:34;06-00:38;06
01:15;29-01:19;29
01:54;24-01:58;24

Swedish House Mafia ft. John Martin - "Don't You Worry Child (Extended Mix)"
http://www.amazon.com/Dont-Worry-Child-feat-Martin/dp/B009FRCWAU/

00:08;10-00:12;10
00:18;26-00:22;26
01:20;12-01:24;12

Switchfoot - "Dare You To Move" from The Beautiful Letdown
http://www.amazon.com/The-Beautiful-Letdown-Switchfoot/dp/B000189DW2/
00:44;03-00:48;03
00:57;21-01:01;21
02:44;02-02:48;02

Taio Cruz - "Dynamite" from Rokstarr
http://www.amazon.com/Rokstarr-Taio-Cruz/dp/B003FP0Y12/
00:32;22-00:36;22
01:19;23-01:23;23
02:04;12-02:08;12

Taylor Swift - "Our Song" from Taylor Swift
http://www.amazon.com/Taylor-Swift/dp/B0014I4KH6/
00:24;13-00:28;13
00:40;06-00:44;06
02:46;18-02:50;18

Taylor Swift - "Ours" from Speak Now
http://www.amazon.com/Speak-Now-Taylor-Swift/dp/B003WTE886/
00:37;10-00:41;10
01:02;05-01:06;05
01:13;15-01:17;15

Tchaikovsky - "The Sleeping Beauty - Introduction" from 100 Masterpieces, Vol. 9 (1877-1893)
http://www.amazon.com/100-Masterpieces-Vol-9-Classical-Music/dp/B002QPSFLS/
00:09;20-00:13;20
00:31;05-00:35;05
02:01;12-02:05;12

Vivaldi - "Flute Concerto in G minor 'La Notte'" from 100 Masterpieces, Vol. 2 (1731-1775)
http://www.amazon.com/100-Masterpieces-Vol-2-Classical-Music/dp/B002QPIRAW/
00:53;12-00:57;12
01:21;06-01:25;06

Wagner - "Sigfried's Death and Funeral March" from 100 Masterpieces, Vol. 9 (1877-1893)

http://www.amazon.com/100-Masterpieces-Vol-9-Classical-Music/dp/B002QPSFLS/
01:20;28-01:24;28
03:32;27-03:36;27
05:00;17-05:04;17

Wiz Khalifa & Snoop Dogg ft. Bruno Mars - "Young, Wild, & Free" from Mac & Devin Go To High School
http://www.amazon.com/Mac-Devin-Go-High-School/dp/B005PV1THC/
00:31;25-00:35;25
01:21;07-01:25;07
02:07;24-02:11;24

The Wreckers - "Leave The Pieces" from Stand Still, Look Pretty
http://www.amazon.com/Stand-Still-Look-Pretty-Wreckers/dp/B0009F43V8/
00:37;27-00:41;27
01:12;21-01:16;21
01:27;02-01:31;02

Yellowcard - "Ocean Avenue" from Ocean Avenue
http://www.amazon.com/Ocean-Avenue-Yellowcard/dp/B0000A0WKG/
00:13;11-00:17;11
00:57;04-01:01;04
02:08;17-02:12;17

Zac Brown Band - "Chicken Fried" from The Foundation
http://www.amazon.com/The-Foundation-Zac-Brown-Band/dp/B001I10AAA/
00:26;20-00:30;20
00:51;22-00:55;22
03:06;02-03:10;02

B.2.2 Speech Sources

딴지일보 나는 꼼수다, "마지막회"
https://itunes.apple.com/us/podcast/naneun-kkomsuda/id438624412?mt=2
04:12;01-04:16;01
04:56;15-05:00;15
05:03;28-05:07;28
05:09;00-05:13;00
05:18;29-05:22;29
05:30;08-05:34;08

05:46;10-05:50;10
06:10;23-06:14;23
07:46;20-07:50;20
09:05;16-09:09;16
10:11;19-10:15;19
10:31;12-10:35;12
10:40;23-10:44;23
10:55;09-10:59;09
11:14;28-11:18;28

Arsenal Podcast, 05/20/2011
http://www.podbean.com/podcast-detail?pid=17819
05:01;29-05:05;29
05:08;19-05:12;19
05:13;19-05:17;19
05:18;19-05:22;19
05:27;15-05:31;15
05:32;22-05:36;22
05:37;22-05:41;22
05:42;22-05:46;22
05:47;22-05:51;22
05:52;22-05:56;22
06:06;26-06:10;26
06:11;26-06:15;26
06:16;26-06:20;26
06:21;26-06:25;26
06:26;26-06:30;26
10:07;20-10:11;20
10:12;20-10:16;20
10:17;20-10:21;20
10:22;20-10:26;20
10:27;20-10:31;20

Coffee Break French Podcast, Word Of The Day Review 02/20/2013
http://radiolingua.com/shows/french/coffee-break-french/
00:17;06-00:21;06
00:22;05-00:26;05
00:27;08-00:31;08
00:32;15-00:36;15
00:37;17-00:41;17

01:02;04-01:06;04
01:50;09-01:54;09
02:03;17-02:07;17
02:14;19-02:18;19
02:28;19-02:32;19
03:34;20-03:38;20
03:50;03-03:54;03
03:58;05-04:02;05
04:08;08-04:12;08
04:15;02-04:19;02
04:22;01-04:26;01
05:38;19-05:42;19
06:11;12-06:15;12
06:21;02-06:25;02
06:29;02-06:33;02

Deutsche Welle Deutsch - Warum nicht? Podcast, "Serie 1 - Lektion 06"
http://www.dw.de/learn-german/deutsch-warum-nicht/s-2548
00:33;00-00:37;00
00:38;04-00:42;04
00:43;08-00:47;08
00:49;03-00:53;03
00:56;26-01:00;26
01:03;17-01:07;17
01:13;28-01:17;28
01:24;20-01:28;20
01:33;15-01:37;15
01:38;15-01:42;15
03:21;24-03:25;24
03:29;29-03:33;29
03:35;05-03:39;05
03:51;16-03:54;16
04:06;09-04:10;09
04:12;07-04:16;07
04:20;15-04:24;15
04:25;27-04:29;27
04:31;00-04:35;00
04:42;10-04:47;10

El Café de Nadie Podcast, "Juan Sebastián Gatti"

http://www.elcafedenadie.com/
01:05;26-01:09;26
01:10;26-01:14;26
01:15;27-01:19;27
01:20;27-01:24;27
01:25;27-01:29;27
01:30;28-01:34;28
01:48;26-01:52;26
01:53;26-01:57;26
02:14;01-02:18;01
02:20;09-02:24;09
02:36;27-02:40;27
02:41;29-02:45;29
02:53;17-02:57;17
03:07;04-03:11;04
03:12;04-03:16;04
03:34;04-03:38;04
03:55;07-03:59;07
04:00;07-04:04;07
13:32;02-13:36;02
13:46;11-13:50;11
13:54;29-13:58;29
14:05;24-14:09;24
14:20;15-14:24;15
16:56;25-17:00;25
17:27;02-17:31;02

Grammar Girl Quick and Dirty Tips for Better Writing Podcast, Episode 58 - "'Only': The Most Indisious Misplaced Modifier"
http://www.quickanddirtytips.com/grammar-girl
00:44;08-00:48;08
00:50;05-00:54;05
00:58;01-01:02;01
01:04;23-01:08;23
01:11;01-01:15;01
01:16;01-01:21;01
01:22;22-01:26;22
01:29;11-01:33;11
01:40;16-01:44;16
02:12;23-02:16;23

02:42;05-02:46;05
03:07;04-03:11;04
04:45;23-04:49;23
06:18;22-06:22;22
06:59;05-07:03;05
07:08;11-07:12;11
07:13;11-07:17;11
07:18;11-07:22;11
07:23;11-07:27;11
07:28;11-07:28;11

ItalianPod101.com Culture Class: Essential Italian Vocabulary, #8 - "Food Souvenirs"
http://www.italianpod101.com/category/culture-class/?order=asc
00:31;16-00:35;16
00:37;07-00:41;07
00:42;18-00:46;18
00:47;14-00:51;14
01:14;05-01:18;05
01:19;05-01:23;05
02:04;04-02:08;04
02:09;04-02:13;04
02:14;04-02:18;04
02:19;08-02:23;08
03:00;09-03:04;09
03:05;13-03:09;13
03:54;22-03:59;22
03:59;22-04:03;22
04:04;22-04:08;22
04:07;12-04:11;12
04:11;04-04:15;04
04:54;00-04:59;00
04:59;00-05:03;00
05:04;02-05:08;02

JapanesePod101.com Onomatopoeia, #18 - "Cast Your Gaze Upon This Japanese Lesson!"
http://www.japanesepod101.com/
00:22;05-00:26;05
00:27;05-00:31;05
00:32;05-00:36;05
00:37;05-00:41;05

00:42;05-00:46;05
00:47;05-00:51;05
00:52;05-00:56;05
00:57;05-01:01;05
01:02;05-01:06;05
01:07;05-01:11;05
01:12;05-01:16;05
01:17;05-01:21;05
01:22;05-01:26;05
01:27;05-01:31;05
01:32;05-01:36;05
01:37;05-01:41;05
01:42;05-01:46;05
01:47;05-01:51;05
01:52;05-01:56;05
01:57;05-02:01;05

Voice of America Chinese News Podcast, "6-7 AM Program 11/30/2012"
https://itunes.apple.com/us/podcast/zao6-7dian-guang-bo-jie-mu/id297399054?mt=2
00:22;08-00:26;08
00:27;08-00:31;08
00:32;08-00:36;08
00:37;08-00:41;08
00:42;08-00:46;08
00:47;08-00:51;08
00:52;08-00:56;08
00:57;08-01:01;08
01:02;08-01:06;08
01:07;08-01:11;08
01:12;08-01:16;08
01:17;08-01:21;08
01:22;08-01:26;08
01:27;08-01:31;08
01:32;08-01:36;08
01:37;08-01:41;08
01:42;08-01:46;08
01:47;08-01:51;08
01:52;08-01:56;08
01:57;08-02:01;08