David Stone

Steinhardt School of Culture, Education, and Human Development

New York University

Author Note

A few words of thanks: first to my parents, Tom and Rae, for keeping me on my toes through this whole process and keeping a light on for me when I needed to get out of town and really get some work done. Thanks also to Mary Farbood for her advisement and Juan Bello for his guidance on MIR topics large and small. Thanks to Shenglong Wang at the Mercer Street High Performance Computing cluster for heroically crushing code for 90 minutes straight so I could get lasagne and nolearn to run on a real GPU and giving me hope that I would ever see the results of an optimized convolutional network, and thanks to the staff at Avery Fisher Media Center who helped me check out what seemed at the time like 1,000,000 film soundtracks. And finally, a big thank you to my film score composition teachers Sonny Kompanek, who first brought up the idea of combining machine learning and film scoring in a lesson 3 years ago, and Ira Newborn, who drilled into me the primacy of “the sound” in capturing a mood onscreen, and who was appropriately nonplussed when I described my thesis to him. I don’t think I told him that Sonny gave me the idea.

Abstract

Search cost for music has fallen dramatically in recent years while commercial, television, and film 'syncs' have grown in importance as a revenue stream and marketing channel for recording artists. Toward providing a practical recommendation agent to better connect recording artists and music supervisors, the author proposes a bagged ensemble of deep convolutional networks for automatic classification of music by appropriate film genres. The ensemble was trained on more than 100 hours of labeled film underscoring and soundtracks and can accurately tag audio with multiple film genre labels ranked by confidence, as well as provide novel classifications that exceed the limitations of the given labels.

A Bagged Ensemble of Convolutional Deep Networks for Film Score Genre Labeling

Introduction

Music is a powerful tool for strengthening the emotional connection between the viewer and a moving image (Eckhardt and Bradshaw 2014). Neurological research shows that listeners react to changes in background music before anything else; for evolutionary reasons, the brain prioritizes auditory streams and the emotional reactions thereto, even if the listener is not consciously aware of it {(Näätänen and Winkler 1999) cited in (Fraser and Bradford 2013)}.

While this deep relationship between music and story has persisted for centuries, the circumstances of that relationship have changed dramatically. In the 21st century, digital technology has generated skyrocketing growth in the sheer quantity of media, as well as changing how we get it (Seabrook 2014). The increase in video content on cable and online combined with unprecedented access to licensing has led to the increasing popularity of synchronization ("sync") as a method for combining music and image. Sync, in contrast to under-scoring, is done by selecting a pre-recorded musical cue that fits the content and style of the image, frequently a pop song with vocals (Anderson 2013).

The recent prominence of sync over commissioned work has changed the workflow of scoring a project. The practice has gradually shifted from a matter of creating original compositions to a matter of selecting existing recordings that fit with the emotional and stylistic content of the image. In addition, the amount of music available for licensing has increased by many orders of magnitude, as sync gains a more important role as a marketing channel for emerging artists and a revenue stream for established artists. Year upon year, the pool of potential candidates for any given project grows.

As in many other domains of increased access and reduced search cost, "choice overload" leads to a heuristic decision process, at a cost to music supervisors, directors and emerging artists. This paper proposes a path to a content-based recommender system that will give music supervisors more efficient access to meaningful content, by modeling the relationship between cues and film genres. Drawing on new research in machine learning, music information retrieval, the perception musical emotion and style, as well as a dataset of labeled film score recordings, I will investigate the connection between raw recordings and film genre.

The first part of this paper covers the historical role of pop song synchronization in film, television, and TV commercials, tracing its fall during the height of rock 'n' roll, its re-emergence in past two decades, and its current creative and economic role. The next section covers the challenges inherent to aligning musical and visual meaning, metaphorical vs. technical descriptions of music, and the impact of "information overload" on the sync process. The third section explores the general solution of recommender systems and identifies what low-level features can be drawn from the audio signal that will correlate well with style and emotion, as well as review the MIR literature of genre and emotion estimation. Finally, the author proposes a multi-label tagging system based on an ensemble of deep convolutional neural networks.

Literature Review

History of Sync

Early Sync to the Rise of Commercial Antagonisms

During the first “Golden Age” of talkies, Hollywood, Broadway and record labels worked hand-in-glove. Features drove record sales and hit songs sent audiences to the box office (.Galdston, Carlin et al. 2013). Big-budget musicals were syncs writ large, entire films supported by pre-existing Broadway hits. The primacy of movie musicals lasted from the dawn of the talkie until well into the age of television. (Kessler 2013) A quick sampling of the Billboard archives reveals that, in the 1940s, 27 out of 94 #1 singles were featured in movies (29%). In the ‘70s, only 16[1] in 253 came from movie soundtracks (6%).

Exactly *what* happened in pop culture between 1950 and 1970 is the subject of reams of scholarship. Suffice to say it was a turbulent period technologically, socially and politically. By the end of the 1950s, the concept of “selling out” had taken hold in popular culture. Rock ‘n’ roll, which now formed the core of the pop record industry, set itself in an antagonistic pose to establishment media, and to TV advertisement especially. Rock musicians cherished and protected their authenticity and outsider status, often going out of their way to distance themselves from commercial culture. In one memorable episode from 1967, Jim Morrison threatened to destroy a Buick with a sledgehammer if one of the Doors’ songs was used in TV ad (Eckhardt and Bradshaw 2014).

Rock musicians were also wary of association with mainstream film and television. The most visible marriages of rock music and film from the period were "rock musicals" (*Rocky Horror Picture Show, Grease, Hair)*. Leaning more toward the "rock" side than the "musical" side, The Who released two classic films (*Tommy, Quadrophenia*), but these were adaptations of complete original works by the band, not syncs. The Who themselves were especially critical of advertising, titling their third album *The Who Sell Out* and filling the record with cutting parodies of ad culture (Eckhardt and Bradshaw 2014).

Today, however, the surviving members of The Who provide a curious contrast to the ironical and skeptical attitudes of their younger selves. In 2014, guitarist Pete Townsend approved the re-release of 15 of The Who's classic tracks, re-mixed expressly for the purpose of licensing (Adegoke 2014). It appears that the pendulum has swung back, but why?

The Re-Emergence of Sync In Film and TV

Many authors trace the rise of sync to the catastrophic erosion of record sales since 1999 (Seabrook 2014)(Adegoke 2014). According to the RIAA, record sales generated $20.4 billion (2012 dollars) in 1999. By 2013, revenues were down to $7 billion. (Lunny Jr 2014) The recording industry has consistently blamed these losses on the migration of listeners away from physical album purchases, driven by the proliferation of digital technology. It's logical that a quest to recoup lost revenue would lead to increased licensing. However, sync was on the upswing long before the turn of the millennium.

The mid-1980s marked the turning point for sync in films, TV, and commercials. Record companies looked to their back-catalog as a source of untapped revenue. Films began employ pop-music montages to such an extent that it became a hallmark of the decade's blockbusters: (*Footloose* [1984], *Rocky IV* [1985], *Top Gun* [1986]). And the press took notice of the influx of hit songs: Variety described *Flashdance* as "pretty much like looking at MTV for 96 minutes" (Calavita 2007).

MTV, which was just getting legs in 1983, supercharged the elements of musical montage and exposed the wider public to new ways of juxtaposing music and images. It's become somewhat of a cliché to note how comprehensive MTV's influence on sound and images has been (Calavita 2007). By the late '90s, a large proportion of major film releases were scored with syncs, especially films aimed at a younger audience (*Reality Bites* [1994], *Clueless* [1995], *Empire Records* [1995], *Spice World* [1997], *Godzilla* [1998]). Teen TV dramas like *My So-Called Life*, *Beverly Hills 90210*, *Melrose Place*, *Party of Five*, *Buffy the Vampire Slayer*, *Felicity*, *Dawson's Creek* and *The O.C.* were liberally seasoned with contemporary hits, paving the way for the use of sync in later shows aimed a slightly older audiences like *Grey's Anatomy* and *House* (Wakefield 2009).

The Internet and New Marketing Channels

By the turn of the millennium, sync was commonplace in mainstream film and television. TV commercials, however, remained a relatively limited market for cutting-edge pop. Some of the biggest pop stars in the world were allied with major ad campaigns (Michael Jackson and Pepsi, for example), but rock musicians generally still viewed advertisers with suspicion and disdain. Advertisers continually returned to the labels' back-catalog of classic soul and "oldies", but commercials were not seen as a place to "break" new acts (Anderson 2013).

The seeds for the change in commercial sync were sown in 1996, when the 104th Congress passed the Telecommunications Act, a dramatic deregulation of cable TV, telephone, Internet, and radio. It was the most comprehensive federal telecom legislation since 1934, the year the Federal Communications Commission was created. Among a raft of amendments to the law, limits on the number of radio stations in common ownership were lifted completely (Cummings 2007). This in turn led to a cycle of mergers, buyouts, and consolidation throughout the broadcast industry. In a short time, the majority of radio stations were owned by a handful of massive media conglomerates. This centralization demonstrably reduced the diversity of radio playlists. Taking place just as online, satellite, and cable outlets were multiplying, radio consolidation encouraged musicians and record companies to look elsewhere for effective exposure. Song "plugging", a direct-selling practice dating pack to broad-sheet ballads and Tin Pan Alley, made a comeback as artists and managers jockeyed for exposure (Anderson 2013).

A major turning point came in 1999, when Volkswagen ran an ad for the Cabriolet titled "Milky Way", scored by Nick Drake's "Pink Moon", from the 1972 album of the same name. When Drake died of an apparent suicide in 1975, Pink Moon had barely sold 5,000 copies. After "Milky Way" ran, he experienced a posthumous critical and commercial renaissance (Chadwick 2000). That same year, Moby's Play, which had experienced lackluster sales, exploded after being licensed in more than 300 advertisements. Well-placed syncs were a win-win for labels and agencies: mass-market exposure juiced the sales figures of previously unknown artists, and a measure of underground credibility rubbed off on the advertising client (Eckhardt and Bradshaw 2014).

Sync Today

Current total global revenue for the recording industry is approximately $15.1 billion (USD). The majority of this revenue continues to come from physical album sales, but a significant fraction is split between direct downloads and revenue from subscription and ad-supported streaming services. Revenues from synchronization deals can seem like a drop in the bucket in comparison: 2.1% of total global sales, totaling roughly $317 million (Seabrook 2014). This is partly because many sync deals are driven by cost-conscious clients and producers. “Lacey”, a recording exec interviewed anonymously by Eckhardt and Bradshaw, says: “Companies are looking for two things when they choose a song for their ad—they want to tap into some sort of cool cache and they also don’t want to spend a ton of money. Indie labels charge significantly less than the majors, and their product is inherently ‘cooler’” (Eckhardt and Bradshaw 2014).

However, a significant measure of the value of sync deals comes from the exposure generated for the artists. Up-and-coming artists look at sync deals as an essential part of promotion. “Lacey” again: “Initially licensing was more of a ‘gravy’ thing, but as print outlets have decreased and radio play has remained as difficult as ever, licensing has become more important to the marketing of a record.” Sync is not only a revenue generator for the label and the artist, but an important “marketing channel”. (Eckhardt and Bradshaw 2014) In this respect, sync bears similarity to product placement (Galdston, Carlin et al. 2013). Therefore, the ultimate goal of the label and the artist is not necessarily to garner earnings from the sync itself, but to inspire a song purchase. It's difficult to say how many purchases are caused by sync exposure, but the examples of Pink Moon and Play demonstrate that the true value of sync far outstrips 2.1% of global revenue.

Soundtrack for a Brand: Challenges in Synchronization

Although synchronization is clearly an effective and lucrative practice for musicians and clients, it doesn't come without challenges. Many of the aesthetic complications inherent to work for hire scoring apply equally to sync. In addition, there are significant workflow issues that come from picking pre-existing music from a large library.

Dancing About Architecture

Whether the music for a project is synchronization or an original score, the ultimate decision-maker is often an executive without professional musical training. (Anderson 2013) Music supervisors occupy the touchy role of translating the dramatic needs of executives into appropriate musical selections. In his autobiography No More Minor Chords, composer and conductor André Previn recounted an episode involving MGM boss Irving Thalberg, which illustrates how a non-musician, in an effort to verbalize some musical impression, can misstep with disastrous results. One day Thalberg heard some music from a score that he hated, and he said as much. Someone, no doubt trying to be helpful, said, "Well, that's a minor chord". The very next day a memo was circulated through the studio that said "No more minor chords in movie overtures" (Galdston 2013).

This story quite rightly reveals Thalberg as an absurd despot, but at the same time also reveals some very important limitations of music theory. Abstract, technical language is a useful tool in the right hands: a relatively compact and efficient way of communicating musical ideas between trained practitioners. However, it doesn't capture the complete picture.

The very nature of scoring and synchronization relies on the associative and suggestive powers of music, which arise from the timbre and affect of the piece and are best described with metaphorical language. Even the simplest musical terminology, e.g. "high/low" or "light/dark", is merely "frozen" metaphor to which we have become acculturated (Schroeder 2013). Laypersons are often surprisingly good at summing up the affective connotations of a track with complex poetic or visual language, e.g. "The sound of the open road," and describing different instrument timbres as different characters, e.g. "high-heeled sax" (Galdston, Carlin et al. 2013). The ultimate role of the music supervisor is to take the metaphors, references and I-know-it-when-I-hear-its of the director and translate them into a tangible sound (Anderson 2013).

Information Overload: The Paradox of Choice and the Limits of Attention

“As long as the centuries continue to unfold, the number of books will grow continually, and one can predict that a time will come when it will be almost as difficult to learn anything from books as from the direct study of the whole universe.”

—Denis Diderot, *Encyclopédie (1755)* (Diderot 1978)

"We get a lot of submissions; at this point I think about 9 million people have the address and phone number of my company.”

—Alexandra Patsavas, Music Supervisor: *Grey's Anatomy*, *The O.C.* (Wakefield 2009)

Each minute of the day, roughly 12 hours of music are uploaded to the music-streaming service Soundcloud (Dillet 2014). Spotify, an ad- and subscription-based service that streams the catalogs of major labels and indie artists, boasts 20 million unique tracks. It adds 20 thousand new tracks every day (Seabrook 2014). There are simply not enough years in a lifetime, let alone hours in a day, for a music supervisor or director to give a fair hearing to every possible track. As sync has become an important marketing channel for artists, supervisors are continually inundated with spec tracks.

What is the effect of this fire-hose of haphazardly sorted media? One of the benchmark principles of cognitive science, Hick’s Law, states that, in simple tasks, the number of alternative choices increases decision time log-linearly (Lidwell, Holden et al.

2003). In general, increased choice and increased information increase deliberation time because of the brain's physiological limits on information processing. Given limited time, people will change their decision-making processes, affecting the quality of their decisions, a phenomenon known as "information overload" or "choice overload". "Faced with elevated levels of information load, […] an individual's limited capacity to process information becomes overloaded, which results in undesirable consequences such as cognitive fatigue and confusion" (Eppler and Mengis 2004), cited in (Aljukhadar 2012).

Thus, pressed for time, music supervisors must inevitably make compromises when evaluating suitable tracks, relying on social or business considerations to make final decisions, rather than waiting for the perfect track to jump out of the haystack (Wakefield 2009). Artists and clients both lose, because the best fitting track can't be heard through the noise. So the question stands: how can the noise be reduced so the optimal sync becomes more likely?

Recommender Agents

Online retailers were among the first to come to grips with choice overload in the age of the Internet and cheap searches. Retailers that facilitate efficient shopping (i.e. matching up shoppers with what they want) encourage loyalty and gain a competitive edge. To that end, online retailers provide customers with "decision support systems" in the form of "recommendation agents", whose personalized recommendations alleviate cognitive load and maintain acceptable choice accuracy (Aljukhadar 2012).

From its genesis in the early 90s, recommender systems have expanded to many services beyond retail, including social networks, investment advice, dating, media streaming, and web search (Burke, Felfernig et al. 2011). Good recommendations can be the difference between a successful service and a failure, and effective recommendation algorithms are worth big money.

In 2007, Netflix famously put up $1 million for whoever could reduce recommendation errors by 10%. In order to do so, they had to release a massive amount of anonymous user data, as well as their proprietary algorithm, taking the gamble that an improved algorithm was worth the loss of trade secrets (Villarroel, Taylor et al. 2013), Early in 2014, Spotify purchased Boston-based Echonest for $100 million to gain control over its powerful "Truffle Pig" algorithm, which can extract more than 50 musical parameters from a track automatically (Lunden 2014). The CEO of Spotify, Paul Ek, is betting that his company's ability to analyze data and generate individualized playlists will give it an edge over emerging competition from Google and Apple (Seabrook 2014).

The foundation of a prototypical recommender system is filtering, either "collaborative" or "content-based" (Burke, Felfernig et al. 2011)(Kaminskas and Ricci 2012). Collaborative filtering, the most common method, works from the assumption that users with similar use habits, purchases, or ratings, are likely to continue to do so in the future.

An extremely transparent and simplified example of this practice can be seen in Amazon's "What Other Items Do Customers Buy After Viewing This Item?" feature. This is simply a ranked list of the purchases of other users who are close to the customer on the one-dimensional line of "did/did not view this item". The hope is that information about similar customers will be of some utility; Amazon collects other, finer-grained data, and encourages customers to rate items and thus create a more detailed and useful profile (2015). Last.fm, a music streaming service, uses a similar system, gathering implicit data (e.g. play time) and explicit data (e.g. song ratings) to generate a user taste profile (Haupt 2009) (Kaminskas and Ricci 2012).

Collaborative filtering is very useful only as long as there is enough user data, and it can be hindered by data sparsity, “cold starts”, and popularity bias. The rating predictions are heavily weighted toward items that have already been frequently rated, and new or novel selections don’t rise to the surface (Kaminskas and Ricci 2012).

Content-based filtering, on the other hand, suggests items that share content tags with items already rated highly by the user. Pandora is the most visible example of content-based filtering, relying on a relatively small library with very detailed and robust content tags (Burke, Felfernig et al. 2011). While Pandora makes excellent predictions about listener taste, there is a big trade-off in a content-based system that comes from the cost of acquiring accurately tagged data. Pandora relies on human experts to rate tracks on a long list of musical features, which is time-consuming, expensive, inflexible and non-scaling. These factors ultimately limit the size of the music library that Pandora can effectively manage (Kaminskas and Ricci 2012). One of the overarching goals of automated music classification is to dramatically reduce the cost of tagging massive data.

What makes a "good fit"?: Recommender Systems for Music Supervisors

No matter what your personal view is on minor chords, it isn't difficult to hear when a piece of music is "working" with an image or not. However, it can be exceedingly tricky to pin down why one piece of music works and another, possibly very similar track, does not. "Music fit" has been defined as "the consumer's subjective perception of the music's relevance or appropriateness to the central ad message" (MacInnis and Park 1991). In this case, the measurement of "fit" is derived from surveys about different musical selections. It draws a correlation between good "fit" and favorable listener response, but doesn't presume to indicate where exactly in the physical structure of a sound "fit" inheres, or give guidelines on how to find the best fitting track given an image and a message. MIR algorithms cannot "see" fitness directly, only patterns present in the signal. Much like the terminology gap between metaphor and music theory described earlier, "users are not able to define their needs in terms of low-level audio parameters (e.g., spectral shape features)" (Kaminskas and Ricci 2012).

(Oakes 2007) gives a good overview of the literature on "fit" in advertising background music, which he terms "congruity". Congruity is broken down into categories, including mood, valence, tempo, timbre, and genre. In general, higher congruity ratings in these categories correspond with increasingly favorable listener responses to the advertisements. If congruity is the ultimate goal of image/sound matching, these categories provide some important points of connection that will lead to a workable "reverse-engineering" of congruity/fitness.

Genre Perception

On the list of congruity categories provided in (Oakes 2007), timbre and tempo are the "lowest-level" features; measures that can be extracted from audio with minimal processing. Tempo has a clear definition: the rate of "beats" in time. (Percival 2014) In practice, this is usually identified as the rate of the predominant low-frequency periodicity present in a signal. It can be extracted by a number of methods, including auto-correlation (Percival and Tzanetakis 2014) (Grosche and Müller 2011), oscillating filters (Zhou, Mattavelli et al. 2008), or probabilistic methods (Ellis 2007).

Timbre is harder to pin down; it is defined by ANSI (1973) as the property of a sound that distinguishes it from another when duration, loudness and pitch are held constant. It is generally understood to include the spectral content of a sound at any given moment. However, perception of timbre is also contingent on how the spectral content changes over time, that is, the envelope of the sound (Houtsma 1997).

There is ample evidence that source identification is primarily driven by timbre cues, independent of pitch and loudness. In his groundbreaking work, (Gray 1942) determined that subjects could recognize phonemes with as little as 3 ms of exposure. At times, the duration was significantly shorter than the time necessary for a single cycle of the fundamental frequency. At 3 ms, no pitch information below 333 Hz can be retrieved, and the duration falls well below the temporal masking threshold of 20 – 200 ms (Pohlmann 2010). Testing for discrimination between sound groups, e.g. voices and instruments, (Suied, Agus et al. 2014) found performance exceeded chance at 4 ms, with some vowel sounds recognizable at 2 ms. While there was a robust voice advantage throughout the study, no advantage was observed for higher pitched sounds, suggesting that sound recognition is independent of pitch perception, and supporting the hypothesis that the foundation of sound evaluation lies in timbre features. (Patil, Pressnitzer et al. 2012) observed physiological evidence for this hypothesis, in the form of cortical responses to the same sound set. Patterns of excitation in the cortex of mammalian test subjects could be used to distinguish between sounds by using a standard machine learning classifier i.e. Support Vector Machines.

Further studies using short clips of complete, multi-instrument musical tracks have indicated the primacy of timbre and tempo for perceiving high-level features like genre and perceived emotion. Subjects can estimate the emotional content, musical style and decade of release of previously unheard recordings significantly better than chance, after exposure of only 400 ms (Krumhansl 2010). Even subjects with amusia, who are unable to identify any song by name, performed equally well on judgments of style and emotion. While this duration is longer than the impossibly thin slices sufficient for source identification, "if genre identification, as a contextualizing categorization, required the prior classifications of component features like melody, bass, harmony, and rhythm, then it is unlikely that such rapid identification would have been possible. That is, from a single tone one could not infer any reliable information about melody or rhythm. Rather, it seems highly probable that the rapid recognition of musical genre occurs concomitantly with the decoding of component features (Gjerdingen and Perrott 2008).

Therefore, there is ample perceptual justification for relying on the timbre of an audio signal as a proxy for the genre (however that may be defined) and emotional content of a piece of music. This is further supported by MIR research, discussed below, indicating that spectral content over a relatively long window is a good predictor of genre (Burred, Röbel et al. 2006).

Automatic Genre Recognition

Feature Extraction

The system proposed here is, at its core, a genre recognition system (MGR). Classification tasks form a sizable chunk of MIR research, dating back to research on phoneme classification in the 1960s, and music genre recognition is the "most widely studied area" (Fu, Lu et al. 2011)cited in (Sturm 2013). Great strides in the field have been made since the problem of automated genre tagging was first tackled in the early aughts, (Tzanetakis and Cook 2002), and classification accuracy on benchmark datasets has crept up persistently. While the particulars of MGR systems can vary widely, nearly all research has followed a standard model: a feature extraction stage followed by a classification stage.

During feature extraction, audio files are decomposed into smaller temporal chunks, transformed into the spectral domain through FFT, and then taken through further processing to extract relevant features. There is a sliding scale of feature engineering, ranging from unprocessed spectrograms to pseudo symbolic representations of rhythm, harmony, and pitch. Generally, "higher-level" features incorporate more rounds of signal processing and take up less memory. For example, raw spectrograms take up half the space of unprocessed PCM audio. MFCCs, depending on the number of coefficients retained, can be smaller by a factor of ten. Chroma-grams, tempo-grams, and other beat synchronous features are smaller still.

The primary advantage of more elaborate feature extraction and pre-processing schemes comes from projecting the audio data, which is rather high-dimensional, to a lower-dimensional subspace. Most classification algorithms perform poorly on high-dimensional datasets when the size of the feature vector exceeds the numbers of examples in the class, the so-called "curse of dimensionality" (Bishop 2006). The lower the dimensionality of the features, the more feasible small datasets are, and the classification algorithm will converge more quickly on a simple model. In addition, engineered features can filter out extraneous noise and provide a stronger training signal to the classification algorithm. Of course, there is no such thing as a free lunch, and more intensively engineered features are more alienated from the original audio samples, and by necessity discard information that may very well be useful to classification (Dieleman and Schrauwen 2014) (Humphrey, Bello et al. 2012). The best engineered features strike a balance between over-simplifying the data and drowning it in noise (Grzywczak and Gwardys 2014). In a survey of the literature since 2002, low-level timbre features rise to the top as the most successfully employed hand-crafted features.

(Tzanetakis and Cook 2002), in one of the founding experiments of MGR, used an ensemble of low-level and high-level features, combining MFCCs with simple timbre (spectral centroid, spectral flux, etc.) and tempo (beat histogram) features. The best individual features were the variance of the spectral centroid and the mean of the first MFCC over the analysis window. In addition, although they supported classification better than random chance, "the nontimbral texture features, pitch histogram features (PHF), and beat histogram features (BHF) perform worse than the timbral-texture features (STFT, MFCC) in all cases."

This initial observation has persisted to the present day; in the three state-of-the-art genre recognition systems identified and evaluated in (Sturm 2013), all three used MFCCs and some other timbral measures (i.e. centroid, spread, or rolloff). Because accurate automatic transcription is still an unresolved problem in MIR, and inaccurate transcription can severely impair classification, implementation of high-level pitch features is relatively rare. (Anglade, Benetos et al. 2010) proposed an ingenious system for generating prototypical harmonic rules and combining chord transcription with low level features. However, the harmonic rules were derived from a large corpus of pre-existing symbolic data, and inclusion of harmonic features in the classifier did not result in persuasive improvement over timbral features. The addition of harmonic features improved classification accuracy by 1.66%; the addition of total loudness as a feature improved accuracy by .66%. For a very slim improvement compared to an extremely simple feature like loudness, not to mention the requirement of a symbolic corpus, transcription appears to remain an inappropriate approach for classification from audio.

Although pitch transcription can't compete with low-level timbral features, the use of tempo-grams or extracted rhythm templates in concert with timbral features can add up to significant improvements (Tsunoo, Tzanetakis et al. 2009) (Schindler and Rauber 2014). Rhythm features are much more robust than automatic pitch or harmony transcription. Further, in many musical subsets the major distinguishing feature between genres is the presence of certain rhythmic clichés (Wright, Schloss et al. 2008). While the use of rhythmic features is promising, the implementation described here will only use timbre features.

In summation, timbre features have proven to be the most robust and practical features for automated genre recognition. In addition, there is significant support from psychoacoustic research that timbre in the predominant perceptual component in human genre recognition.

Classification Methods

Once the audio has been projected into a suitably low-dimensional and descriptive feature space, the data is ready to be classified. In the broadest terms, a classification algorithm will attempt to iteratively construct a function that partitions the high-dimensional feature space into class-matching subspaces (Tzanetakis and Cook 2000). In practical terms, the architecture of particular algorithms can vary widely.

Classification models fall under the two main branches in machine learning: supervised and unsupervised learning. Unsupervised learning, also referred to as "clustering", groups like-to-like feature vectors without any reinforcement from ground-truth labels. Rather, it seeks to describe the data by a set number of parameters, e.g. K-means or Expectation Maximization (Berenzweig, Logan et al. 2004). Clustering is useful for collaborative recommendation agents because it generates a similarity metric for the samples. Classes can be labeled after clustering by hand or by a measurement of their members.

Supervised learning, on the other hand, makes predictions and compares them to ground-truth labels, using that error signal to optimize a cost function. Supervised learning models, primarily decision trees (Foote 1997) and Support Vector Machines (Tzanetakis and Cook 2002), set the benchmark for MGR from the late 90s until the 2010's. Accuracy could be dramatically increased with ensembles of classifiers, e.g. random forests of decision trees, "bagging" (weighted parallel classifiers), or "boosting" (iteratively trained classifiers). A standard model for genre classification began to take shape: extract timbral features and classify them with an ensemble of SVMs.

By 2012, the MIR community had converged on such a model across a wide array of problems, including transcription and beat detection. This was for good reason: new features and schemes of classification were tested but could not reliably best the standard timbre/SVM model. Benchmark improvement had begun to slow dramatically, and was usually achieved by scaling up existing methods. MIR had reached a state of rapidly diminishing returns (Humphrey, Bello et al. 2012).

Deep Learning

In the past few years, research has begun to shift away from "hand-crafted" features, toward the use of deep neural networks (DNN) to extract relevant features automatically. DNNs can quickly converge on the optimal features for a given task, whereas hand-crafted features, whatever their sophistication, are ad hoc transformations that are evaluated and updated piecemeal. As (Humphrey, Bello et al. 2012) point out, many machine learning problems have converged toward a standard model of hierarchical transformations that is mirrored in the layered architecture of a DNN. Since 2010, features extracted by DNNs have broken records in countless machine learning tasks, from handwriting recognition to image tagging. MGR is no exception (Hamel and Eck 2010).

Deep neural networks are actually quite an old algorithm. The most famous, maybe infamous, early implementation was developed by Frank Rosenblatt at the Cornell Aeronautical Laboratory in 1957: the Perceptron. The early achievements of the Perceptron led to overheated hype of AI in the press; later research, notably Minsky and Papert's 1969 Perceptrons, burst the AI bubble and ushered in the first "AI winter". Minsky and Papert had pointed out that a single layer Perceptron can only converge on linear functions. Simple boolean functions like XOR can never be learned by a single layer Perceptron. Although additional layers make a neural network more than capable of modeling non-linear functions (and Minsky and Papert said as much), Perceptrons is nevertheless widely cited as a direct cause of decreased funding for neural network research during the 1970s (Marsland 2014). In the past 20 years, however, advances in network architecture (deeper networks and convolution), regularization (norm penalties, dropout, validation and early stopping), and optimization (SGD with momentum) have put neural networks back at the fore of automatic pattern recognition.

The simplest case of a neural network is a single "neuron". The neurons in neural networks are loosely based on the simplified biological model of McCullough and Pitt described in 1943: a vector of weights w multiplied by inputs x, plus a threshold (an additional weight b with a static input of 1) and an activation function (*Figure 1*).

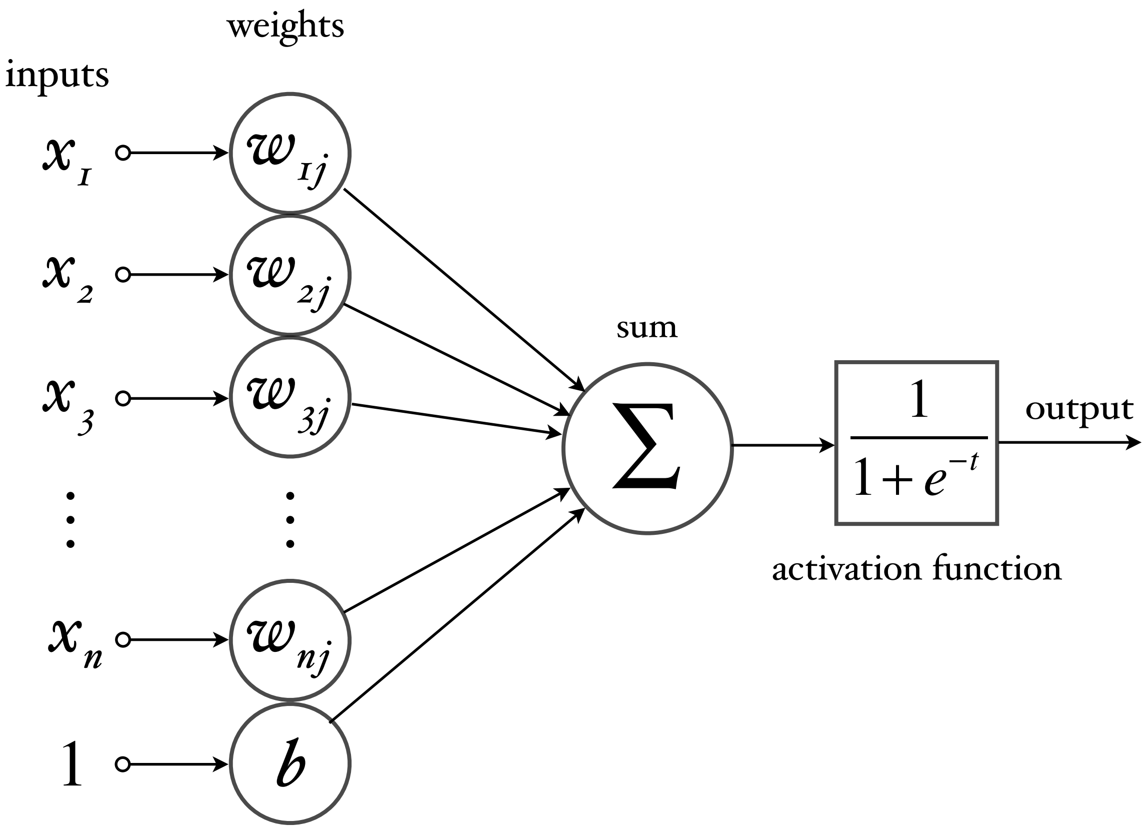

Figure 1. Artificial Neuron Architecture

The network is “trained” by updating the weight vector after running each training example: for each weight, the difference between the intended output and the generated output is scaled by the learning rate, and added to the weight. In the single-layer “Perceptron” case, the updating algorithm is obviously trivial. In order to deepen the network and allow for convergence on non-linear functions, the network must be updated by back-propagation of error. The weights of the inner, or “hidden” layers, can be updated by recursively differentiating the error of the cost function with respect to the output of each weight, moving backwards from the output to the input. The weight is then adjusted in the direction opposite the error gradient (again, scaled by the learning rate) (Marsland 2014).

Over the training epochs, the network output descends the error surface and settles on a loss minimum (gradient descent). Because back-propagation differentiates the output of each weight, the activation function for each neuron must be differentiable. Common differentiable activation functions are hyperbolic tangent, sigmoid (*Figure 1*), and rectified linear units or ReLU. ReLU is has become the standard for most networks because of its simple differentiation and ability to handle unscaled inputs (Dahl, Sainath et al. 2013).

Gradient descent without modification can be extremely slow and get trapped in local minima, so the updates are usually done on batches of training samples rather than the entire training set (Stochastic Gradient Descent or SGD), and scaled by a momentum value derived from previous step sizes (Nesterov Accelerated Gradient Descent, ADAgrad, rmsprop, etc.).

Deep networks are much more expressive than single layer networks, but without careful controls they can overfit. An overfit network will converge on a complex model of the training set that does not generalize well to new samples.

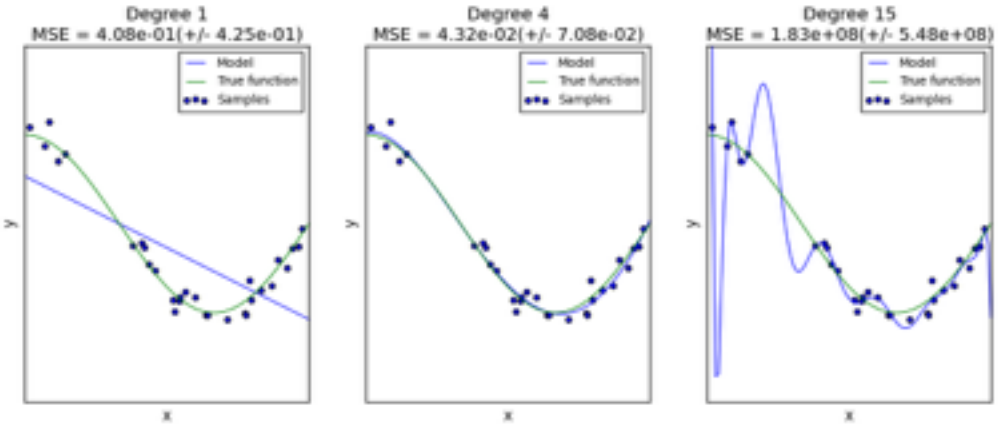

Fig. 2

Overfitting on Exponential + Noise (Scikit-Learn.org 2014)

In the figure above, we see the plot of an overfit estimator on a toy dataset generated by a simple polynomial + noise. The estimator has converged to a function that perfectly describes the training data; no point is left out. However, in doing so, it has allowed the noise signal to distort the model to a point far beyond any generalization. The complexity of neural networks often reaches a point of diminishing returns because a more complex model does not necessarily generalize better than a simple model. Overfitting can be postponed by directly adjusting updates and penalizing outliers (L1 and L2 norm regularization), and avoided altogether by periodically validating the model on a held-out dataset, and stopping training when performance on the validation set stops improving (Marsland 2014).

Although the neurons in a given layer are not connected to each other directly, they can develop implicit connection called "co-adaptation", in which a group of neurons divide up the description of a feature among themselves. The output of one neuron depends on the output of the other neurons in the layer. Co-adaptation is another source of overfitting, because it allows the network freedom to converge to a random "correct" solution, not the best generalizing solution. Dropout layers, simple randomized masks in between hidden layers, weaken the implicit connections of co-adapted neurons and force the neurons to be more individually descriptive. With dropout, the network essentially becomes an ensemble of smaller networks which are individually less susceptible to overfitting, while preserving the expressiveness of the overall model (Dahl, Sainath et al. 2013).

Although wide and deep networks with dropout and regularization take much longer to train than unregularized networks, optimizing a network on a Graphics Processing Unit (GPU) can take a fraction of the time needed on even the fastest CPUs. This is because, very much like the first Perceptron, GPUs can do high-dimensional linear algebra on hardware. The calculations for weight updates are extremely similar to the calculations on pixel shaders that GPUs do in their day jobs. Deep learning libraries like Theano or Caffe can use the CUDA platform developed by NVIDIA to cut training times of deep networks by a factor of 50. Since the hyper-parameters of a DNN must be derived experimentally, training times on the scale of hours or days, rather than months, are crucial to their usefulness.

Methodology

Data Collection, Labeling, and Class Balancing

The raw audio data was drawn from a comprehensive sampling of film soundtracks available at NYU's Bobst Library. The individual tracks were labeled by cue title, film title, composer, year of release, and most importantly, genre of the associated film. The genre assignments were taken from IMDB.com, the Internet Movie Database, for the sake of consistency and simplicity. IMDB is edited in wiki fashion, and the genre labels can be taken as the general consensus of a knowledgable and dedicated user base.

Each film in the online database is tagged with at least one of 22 genre tags. The majority of the films are tagged with more than one genre, and certain genre pairs had a high degree of overlap. If a track was tagged with more than one genre, it would occur once in each of those classes. Therefore, there is some repetition of sample data between the classes, but no sample recurs with the same tag. Table 1 shows the 10 most and least frequent genre bigrams.

Table 1

Ranked Genre Bi-Grams

Most Frequent Class Bi-Grams	Least Frequent Class Bi-Grams
935 : Drama+Romance	16 : Action+History
551 : Crime+Drama	16 : Adventure+History
544 : Comedy+Drama	16 : Horror+Romance
518 : Drama+Thriller	15 : Adventure+Horror
482 : Comedy+Musical	15: Crime+Fantasy
454 : Comedy+Romance	14 : Fantasy+Horror
406 : Adventure+Fantasy	13 : Fantasy+Western
379 : Action+Adventure	13 : Horror+Musical
370 : Musical+Romance	10 : Action+War
518 : Drama+Thriller	9 : Crime+Musical

Surprisingly, some tag pairs that would be expected did not occur at all, such as "Animation+Musical" and "History+Romance". Some of this can be explained by the very small sizes of the Animation and History classes. Although the tags appear on equal footing in the database, certain categories occur much more frequently. Furthermore, genres were not equally represented in the available soundtrack recordings. As the following breakdown of the total available tracks shows (*Fig. 3*), the "Drama" tag swamps all others, which matches the extensive overlap it shows in the bi-gram figure. While Drama and Comedy could be considered "super-genres", the two classes have significant overlap. A more promising bifurcation of the data would be between films tagged with Drama and those without. 33% of the total collected tracks, and 40% of tracks from the top ten classes have a Drama tag.

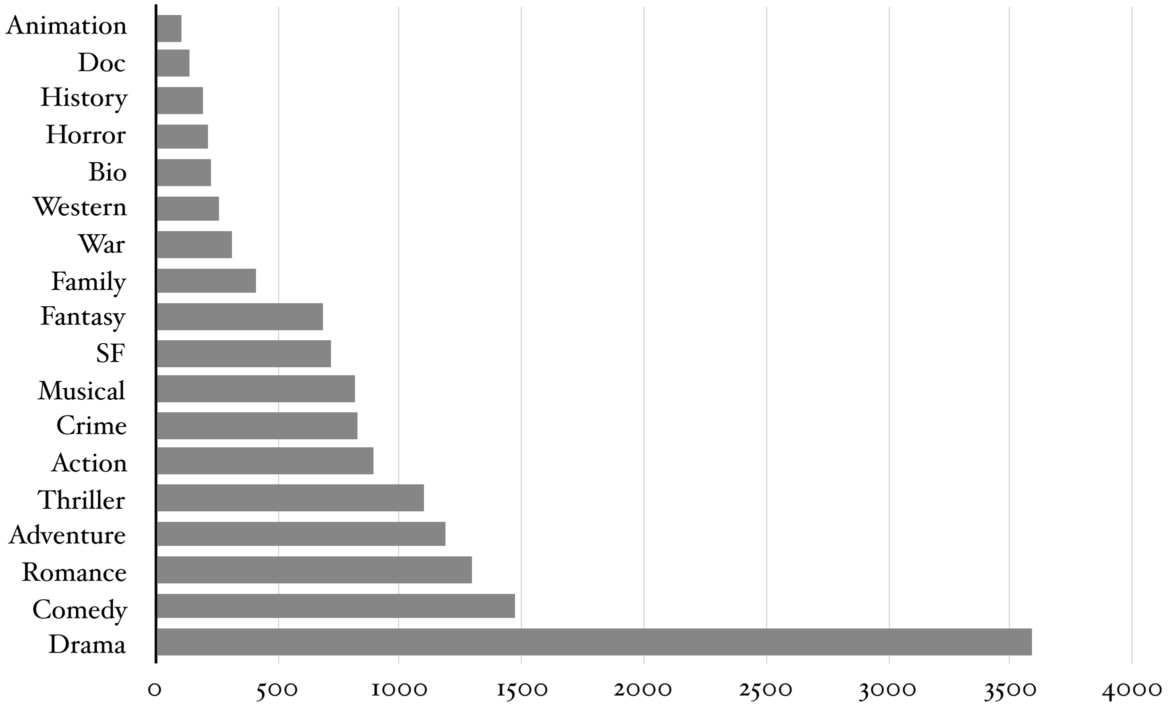

Figure 3. Genre Membership Ranked

The smallest five classes (Animation, Documentary, History, Horror, and Biography) were therefore discarded prior to feature extraction. Feature extraction was limited to 1000 randomly sampled tracks from each class. Because of varying track lengths between classes, the sample classes were balanced a final time after feature extraction, resulting in an equal amount of feature vectors for each class. Family, War, and Western were also discarded, as balancing to those classes would necessitate discarding more than half the data. Class balancing is so important prior to training because a neural network, in seeking to minimize loss, may overfit to a model that simply assigns each sample to the most overrepresented class. The 10 classes remaining after balancing were Action, Adventure, Comedy, Drama, Fantasy, Musical, Sci-Fi, Romance and Thriller.

Feature Extraction

The .mp3 files for each track were converted into ~9.2 second long tiles of the first 40 Mel-Frequency Cepstral Coefficients (MFCCs). While neural networks can be trained on raw spectrograms or even PCM, taking the first few dozen MFCCs allows compression by a factor of 10 or more, allowing faster training on a larger dataset with less dimensionality. In principal, MFCCs preserve the timbre of the signal while discarding information regarding pitch. As discussed above, timbre features are both perceptually important for human genre recognition and robust in machine genre classification.

Some version of the MFCC feature has been standard for nearly 50 years, because it is a relatively simple compression algorithm that preserves perceptually important information. By capturing the very low periodicities of the spectral envelope, the first few MFCCs represent the general shape of the entire spectrum extremely efficiently. The use of the Mel-frequency filter bank ensures a spectral resolution that is not perceptually redundant. In addition, MFCCs work well for classification tasks because taking a DCT of the spectrum de-correlates the spectral coefficients, which are highly covariant, i.e. statistically redundant (Batlle, Nadeu et al. 1998). The cepstral transformation is, in fact, an approximation of Principal Component Analysis (*Fig.* 4). The first MFCC, the DC component, is usually discarded. The final samples each consist of a 1 x 40 x 100 array of 32-bit floating point values.

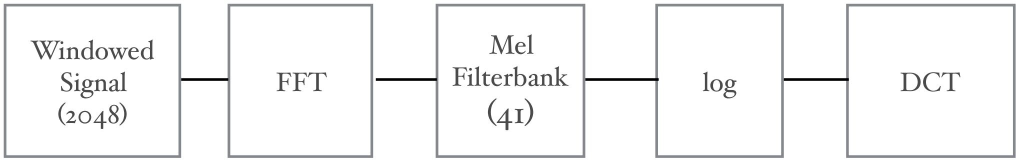

Figure 4. MFCC Signal Processing

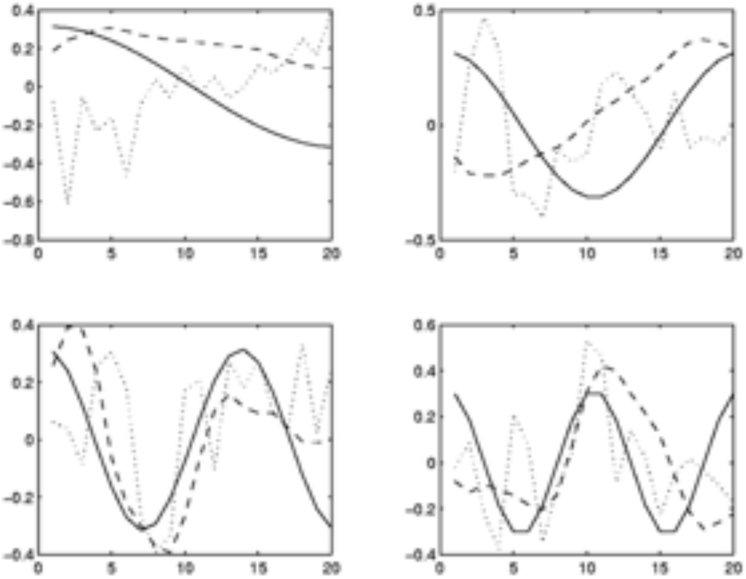

Figure 5. Components of DCT (—), PCA (- - -), and LDA (···)

(Batlle, Nadeu et al. 1998)

Test-Train Division and Final Inputs

After class balancing, the final dataset contained a 13,000 x 40 x 100 array for each class, totaling 130,000 samples in all. The average number of tags per track was between 2 and 3, so the total audio represented came to roughly 150 hours of unrepeated music. Out of the 130,000 samples, 7,534 were retained for testing, representing 196 tracks. None of the samples from any test sample's parent track occurred in the training set. Because of the multi-class (rather than multi-label) nature of the dataset, most samples recurred in multiple classes, and it would be possible for a test sample to be identical to a training sample but carry a different label, unless the test data were held out on the sample-parent level. Finally, the data were preprocessed by scaling to zero mean and unit variance for each feature (Stanford 2015). Dimensionality reduction was attempted with PCA, retaining 99% of the variance with 3400 of 4000 features, but the best results were obtained with unreduced vectors.

Two-Layer Neural Network

Hyper-Parameters and Results

The first feature extraction/classification model was a two layer neural network, with ten output neurons and a softmax classification layer. Hyper-parameters (learning rate, layer size, early-stopping patience, and batch size) were evaluated by experimentation and grid search. Model strength was evaluated by validation loss (on a random validation set of 20%), and by weighted-average F1-score (a ratio of precision to recall that captures over- and under-estimations).

The best model was surprisingly small and fast: 50 units in each hidden layer, trained in 49 epochs. Wider layers did not perform drastically worse, but trained more slowly and overfit more. Unsurprisingly for such a shallow network, it performed poorly on the test set, bottoming out at a validation loss of 2.163, with an F1 average, weighted by class representation, of 0.224 (*see Appendix for complete list of hyper-parameters and F1-scores*).

Discussion

Classification of Musicals was far and away the best, with an F1-score of 0.355, and Sci-Fi the most challenging class at 0.007. Sci-Fi and Thriller both have significant overlap with Action and Adventure, and both subsets classify very poorly compared with the more general parent classes. However Drama, being such a broad class that overlaps with nearly every other, cannot be modeled effectively at all, with a score nearly as bad as Sci-Fi: 0.014.

The Musicals class is an interesting case; obviously there is some audio feature present in the Musicals class that is rarely in any other. Below is a figure ranking the classes by year of film release. The Musical class is more than a standard deviation away from the rest of the classes, which are clustered around 1990.

Table 2

Ranked Mean and Std. Deviation of Classes

Class	Mean Year	Standard Deviation
Action	2000	15
Thriller	1997	17
Sci-Fi	1996	20
Crime	1995	14
Drama	1994	19
Adventure	1993	23
Fantasy	1988	14
Romance	1984	26
Comedy	1984	26
Musical	1965	24

It's possible that the network is picking up on some feature related to older recording techniques, such as attenuation in the low and high frequencies, tape saturation or limited dynamic range. In addition, tracks from Musicals are much less likely to be purely instrumental underscoring; that is, most of the Musical tracks are songs with voices. The strength of the predictions on the Musical class is probably a combination of the timbral features of mid-century recording techniques and the presence of voices in the track, two features which are common to the Musical samples and rare in the other classes.

Clearly, such a shallow network generates weak features for the softmax classifier, which, apart from one or two classes, does very poorly. Accuracy could be improved with additional fully connected layers, making a deeper network. A deepened network with two wider layers after in the input, gently stepping down the features, can do somewhat better on the validation set. A 1000 , 100 , 50 , 50 network achieved a validation loss of 2.13, but still overfit worse than the shallow net, logging an average F1 of 0.204 and even more inconsistent precision and recall on a class-by-class basis.

After 4 layers, DNNs simply hit a wall and a new approach needs to be taken, one that can take advantage of the high-dimensionality of the data and the time-dependent features encoded in the samples.

Convolutional Neural Network

Convolution neural networks can do two things better than simple deep networks. First, they are better at handling high dimensional data, because their neurons are not fully connected, but rather locally connected. Second, this local connectivity allows for the detection of translated, rotated, or scaled features. For these reasons, convolutional networks are the state of the art for image recognition.

A convolutional layer has 3 dimensions; conceptualizing the architecture can be a challenge as the term "layer" is now overdetermined. Here, "layer" refers to the three-dimensional convolutional layer, and "slice" refers to the neuron arrays along the depth of the layer. Each slice depth-wise represents the activations a different "kernel" or "filter" as it is convolved across an input array. The size of the slices is dictated by the filter size, often something like (3x3) or (2x2), and how they are "strided" across the input array. These filters downsample the input, and deliver an activation that describes the local features. The weights of each neuron are shared across all the neurons in the slice, so each slice shows the activation of the same filter in different localities on the input. This weight sharing greatly reduces the number of parameters in the network, replacing millions of weak neurons with a few thousand highly descriptive ones.

For example, a convolutional layer with 4 filters of size (2x2) and a stride of 1 pixel, looking at a 10x10 grayscale image, will be 9x9x4. The layer would be able to signal the presence of 4 different features anywhere on the image. Each neuron has 4 weights (plus a bias), and shares them with all other neighboring neurons in its filter slice of the 3D layer. Thus, the number of parameters to update for the entire convolutional layer is only 20 (filter size+1 x number of filters x the number of channels). A fully connected layer with the same number of neurons would have (81 x 4 x 100) + 100, or 32,500 weights!

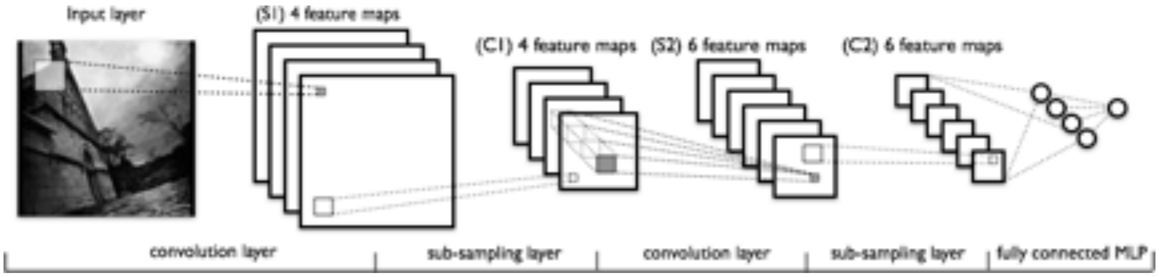

Figure 5. Convolutional Network Architecture

As a CNN is deepened, the neurons capture hierarchies of simple features like edges and construct filters that can detect complex shapes at any location. For musical signals in the form of spectrograms, this translates to the encoding of a position in time or frequency of a consistent feature.

Hyper-Parameters and Results

4 convolutional layers were interposed underneath the best scoring fully-connected model. Again, hyper-parameters were tested by exhaustive grid search and evaluated by validation loss and average F1-score. With the increased number of layers (now 16 including max-pool and dropout layers), there are a much greater number of hyper-parameters to test. The final hyper-parameters selected are therefore by no means the final word on the potential of such a network, but they provide a window onto the capabilities of the model. Exhaustive grid search over all parameters for such a large network performs equally well or worse than randomized search or hyper-parameters tuned by an additional regressor, e.g. an SVM.

The top-performing CNN significantly bested the fully-connected network, and was generally more responsive to increased depth, width, and training epochs. The best validation loss achieved was 1.791 in 655 epochs. Below is an F1-score comparison class-by-class between the fully-connected and convolutional network.

Table 3

F1-scores by Class for Fully-Connected and Convolutional Networks

Class	Fully-Connected Model	Convolutional Model	Change
Sci-Fi	0.008	0.220	+0.212
Romance	0.027	0.176	+0.149
Musical	0.356	0.483	+0.127
Thriller	0.042	0.137	+0.095
Crime	0.193	0.202	+0.09
Fantasy	0.113	0.202	+0.089
Drama	0.014	0.101	+0.085
Comedy	0.022	0.165	+.143
Action	0.233	0.219	-0.014
Adventure	0.204	0.196	-0.08

Discussion

While a single deep convolutional network demonstrated a significant improvement over the fully-connected network, it still made far too many misclassifications to be of any practical use. Some of these issues can be traced to the organization of the dataset. Part of the problem is that a simple softmax classification, which clamps to a single class, is inappropriate for a dataset in which each parent track has as many as three class labels. Secondly, since identical samples recur in disparate classes, it's possible that they knock the classifier back and forth during the training phase. or worse, distort the cost function when a sample from the training set recurs in the validation set with a different label. To an extent, this would prevent over-fitting, but with the average sample occurring in more than two classes, the distortion is very likely to be a significant hindrance to proper convergence, by inflating the validation loss and initiating early-stopping too soon.

Ensemble of Convolutional Networks

To increase the expressiveness of the model and compensate for conflicting training examples, an ensemble of ten deep convolutional binary classifiers (based on the best CNN hyper-parameters) was trained, each classifying one-vs-all. To allow for multi-label classification, the output neuron in each network was set to a sigmoid regression function instead of softmax, delivering a measure of the "confidence" of the binary classification. The collected estimations were then ranked and evaluated on multi-label metrics. This simple ensemble method is called "bagging" (**b**oostrap **agg**regat**ing**) (Breiman 1996). For each class, the corresponding model is trained of the full training set of the "positive" class and a random uniform sampling of the "negative" classes. No single classifier is trained on the full dataset, but the full ensemble sees every example at least once and the vast majority multiple times.

$$\mathrm{rank}_{ij} = \left|\{k : \hat{f}_{ik} \geq \hat{f}_{ij}\}\right|$$

$$\mathcal{L} = \{k : y_{ik} = 1, \hat{f}_{ik} \geq \hat{f}_{ij}\}$$

Results

Unsurprisingly, each binary classifier converged to a much lower validation loss than the 10-class models, averaging 0.22 over all models. While this metric cannot be directly compared with the 10-class metrics, it is indicative of the increased strength of the network on a simpler problem.

F1-scores for multi-label classification are not an especially useful metric, especially on a ranked list like this ensemble produces. However, there are alternative metrics that can provide some insight into the effectiveness of the ensemble.

Coverage error is a metric that measures the recall of a multi-label classifier. It calculates the average number of labels that must be selected from the predictions to fully recall all the relevant labels.

$$coverage\,(y,\hat{f}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} \max_{j:y_{ij}=1} \mathrm{rank}_{ij}$$

with

Label rank average precision measures the average, over each class, of the number of relevant labels retrieved as a fraction of the total retrievals.

$$LRAP\,(y,\hat{f}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} \frac{1}{|y_i|} \sum_{j:y_{ij}=1} \frac{|\mathcal{L}_{ij}|}{\mathrm{rank}_{ij}}$$

with

The ensemble, when tagging the 10-second samples individually, has a coverage error of 5.52. The best possible score for coverage error is equal to the average number of labels per track: 2.33. The classifier estimates slightly more than one incorrect label for every correct. The LRAP score is 0.597 out of 1.000.

There is still room for improvement, as the test set consists of complete families of track samples. Instead of tagging individual samples, it can be better to aggregate the predictions and calculate the average ranking across the entire track. Average predictions are much better than individual predictions, with coverage error falling to 5.02 and precision edging up to 0.618 out of 1.000. Taking the coverage error, we can threshold the predicted labels to the top five and calculate the recall, precision and F1-score (Table 4). Again, while these scores cannot be directly compared to the 10-class CNN—ranking and thresholding the ensemble predictions inflates recall by allowing five guesses per label instead of one—they are indicative of generally increased performance.

Table 4

Bagged, Pooled Ensemble Classification Report

	precision	recall	f1-score	support
Action	0.39	0.92	0.55	50
Adventure	0.40	0.85	0.54	62
Comedy	0.50	0.79	0.61	53
Crime	0.32	0.45	0.37	42
Drama	0.46	0.69	0.55	74
Fantasy	0.24	0.97	0.38	30
Musical	0.49	0.80	0.61	30
Romance	0.32	0.68	0.44	40
Sci-Fi	0.30	0.91	0.46	34
Thriller	0.33	0.89	0.48	44
avg / total	0.39	0.79	0.51	459

Discussion

The ensemble classifier easily outperformed the simpler classifiers, modeling the dataset both more accurately and expressively. Logistic regression better captures the multi-label, multi-class nature of the data, and bagging the classifiers increases f1-scores by a factor of 2 or more. However, accuracy is not the end of the discussion, especially as it relates to training a network as the foundation of a content-based recommender agent (McNee, Riedl et al. 2006). For some practical uses, it can be better for a recommender agent to have a bit of play; "good enough" can be more productive for discovery than "perfect" (Lopresti 2001) (Kaminskas and Ricci 2012). This ensemble's ability to select multiple labels generates some novel "mis-labelings"; depending on how one chooses to view them, these can be considered errors caused by weakness in the model, errors inherent to the coarse nature of the dataset labels, or fruitful and serendipitous descriptions that exceed the dataset labels.

(Sturm 2013) argues that the usual classification metrics used to evaluate MGR models are not sufficient to establish the model's success at modeling genre *per se*. A highly accurate classifier on an offline dataset may merely be modeling some other relationship between the examples that partly overlaps with genre, but does not describe genre itself. The Musical class in this dataset clearly carries distinguishing features (vocals, analog recording) that may or may not be distinguishing *genre* features. Vocals might be a generalizable genre feature, but recording technique is obviously not. An additional hurdle to true genre classification is the overdetermination of the term itself, which can refer to a limitless range of coarse or fine gradations of musical style (Tzanetakis and Cook 2002, Kaminskas and Ricci 2012). Thresholding the predictions of the ensemble to match the number of target labels, individual classifications can be investigated for possible productive mislabeling, indicating whether the model is pushing against the limitations of the dataset labels, or simply failing at the task. Furthermore, misclassifications can provide insight into the prototypical classes.

Table 5

Accurate Predictions (3 targets)

Film	Track	Targets	Predictions
The Wedding Banquet	Turkish March	Comedy, Drama, Romance	Comedy, Drama, Romance
Star Trek: Nemesis	A New Friend	Action, Adventure, Sci-Fi	Action, Adventure, Sci-Fi
Hello, Dolly!	Hello, Dolly!	Comedy, Musical, Romance	Comedy, Musical, Romance
Inside Man	Dr. Phil	Crime, Drama, Thriller	Crime, Drama, Thriller
Escape from New York	Over the Wall	Action, Sci-Fi, Thriller	Action, Sci-Fi, Thriller

Table 6

Accurate Predictions (2 targets)

Film	Track	Targets	Predictions
The Witches of Eastwick	The Dance of the Witches	Comedy, Fantasy	Comedy, Fantasy
La Tenta Rossa	The Rent Tent (Waltz)	Adventure, Fantasy	Adventure, Fantasy
Mysterious Island	The Giant Crab	Adventure, Sci-Fi	Adventure, Sci-Fi
Mary Poppins	The Bird Woman	Comedy, Fantasy	Comedy, Fantasy
The Reader	Go Back To Your Friends	Drama, Romance	Drama, Romance
Harry Potter & The Chamber of Secrets	The Chamber of Secrets	Adventure, Fantasy	Adventure, Fantasy

Table 7

Descriptive Mislabeling or Over-labeling

Film	Track	Targets	Predictions
The Grand Budapest Hotel	The Cold-Blooded Murder of Deputy Vilmos Kovacs	Adventure, Comedy	Crime, Thriller, Sci-Fi
The Wizard of Oz	The Jitterbug	Adventure, Fantasy	Musical, Comedy, Romance
The Dark Knight Rises	A Storm Is Coming	Action, Adventure	Action, Thriller, Sci-Fi
The Truman Show	Powaqqatsi - 5. Anthem	Comedy, Sci-Fi	Action, Sci-Fi, Thriller
Harry Potter & The Deathly Hallows, Part 1	Rescuing Hermione	Adventure, Fantasy	Adventure, Action, Fantasy
Mary Poppins	Let's Go Fly A Kite	Comedy, Fantasy	Musical, Fantasy, Comedy
War of the Worlds	Escape from the Basket	Adventure, Sci-Fi, Thriller	Action, Sci-Fi, Adventure
Superbad	Soul Finger	Comedy	Crime, Comedy, Drama
Natural Born Killers	The Future	Crime	Comedy, Crime, Drama
O Brother, Where Art Thou?	Down to the River to Pray	Crime	Drama, Fantasy, Crime
The 13th Warrior	Eaters of the Dead	Action, Adventure	Fantasy, Action, Sci-Fi
American Hustls	Jeep's Blues	Crime, Drama	Comedy, Romance, Crime

Table 8

Total Misclassification

Film	Track	Targets	Predictions
Sneakers	The Hand-Off	Comedy, Crime, Drama	Action, Sci-Fi, Adventure
Gangs of New York	Shimmy She Wobble	Crime	Comedy, Action, Drama
Taxi Driver	Diary of a Taxi Driver	Crime, Drama	Comedy, Musical, Fantasy
L.A. Confidential	Shootout	Crime, Drama	Action, Sci-Fi, Adventure
Once Upon a Time in Mexico	Malagueña	Action, Thriller	Drama, Comedy, Romance
The Theory of Everything	London, 1988	Drama, Romance	Action, Sci-Fi, Fantasy
Jaws	One Barrel Chase	Drama, Thriller	Comedy, Romance, Fantasy

Even the worst failures of the classifier can be parsed by a closer investigation of the individual tracks. A few examples are investigated here beginning with the first track in Table 7: “The Cold-Blooded Murder of Deputy Vilmos Kovacs” by Alexander Desplat, from Wes Anderson’s *The Grand Budapest Hotel*.

The two ground truth tags for the film are Comedy and Adventure, but the classifier tags the track with Crime, Thriller and Sci-Fi. Even without hearing the track, we can see from the title that the underscore accompanies an intense scene of suspense and violence. Sure enough, the track is rhythmically propulsive and sparsely instrumented, with timpani and snare drums in the foreground, blended with twangy plucked strings and a murmuring tone-wheel organ. The moment of the murder is highlighted by a cymbal crash and booming bass drum. The materials are harmonically repetitive and rhythmically syncopated. While not accurate to the film as a whole, the estimated tags fit the character of the track very well. However, the classifier was unable to detect the certain ironic melodrama common to the score's of Anderson's films, which would come across to a human expert and suggest the comedic element of the film's world.

Next, also from Table 7, we look at two tracks that are given nearly the same tags: the Bar-Kays 1967 hit "Soul Finger" from *Superbad*, and Leonard Cohen's "The Future", from Oliver Stone's *Natural Born Killers.* Each track is mislabeled with Crime and Comedy, respectively. Musically, the tracks have some surface similarities: electric instrumentation, steady back-beats, 60's pop. It's very interesting that both films are different sides of the pseudo-Tarantino coin—*Natural Born Killers* is a frenzy of stylized violence set to a classic pop soundtrack, and *Superbad* is a caper of teen-misdemeanors ironically glossed with the aesthetics of *Reservoir Dogs* or *Jackie Brown*. Each have an element of Comedy and Crime, but heavily biased toward one or the other.

O Brother, Where Art Thou?, from which the next track on Table 7 is sampled, has the exact same ground-truth label as *Natural Born Killers:* simply, Crime. The two films could hardly be more different, and the classifier captures the elements of Drama and Fantasy in the sampled track, "Down to the River to Pray". The cue is a stirring spiritual, sung by a choir of unaccompanied voices. The arrangement is lush and spacious, and accompanies a scene of the heroes stumbling upon a church congregation singing in the wilderness, tinged with the magical realism and mythical themes that run through the film. In each of these tracks, the classifier has failed to match the target labels, but the suggested labels exceed the ground-truth and capture details in the individual tracks.

None of the tracks in Table 8 were tagged with a single accurate label. At the top of the list is “The Hand-off”, from the comedic crime drama *Sneakers*, incorrectly tagged with the Action, Sci-Fi and Adventure labels. Again, the track title offers a clue to the misclassification; the cue is the underscoring for a tense and suspenseful scene. The 3:07 track is composed of two contrasting elements: one highly dissonant and rhythmic, centered around snare drum and deep staccato piano clusters; the other mysterious and ethereal, with legato melodies in female voice and soprano saxophone, accompanied by chiming woodwind chords and high violin tremolos. While not especially clever (Thriller is probably the most apt tag for the track), the Action and Sci-Fi elements are present in the audio.

“Shimmy She Wobble”, off of Martin Scorsese’s *Gangs of New York* is a very unusual track for the dataset. Apparently a field recording of a rural Southern fife-and-drum band, the cue is a blend of background voices, rollicking snare drums, bass drum and a single fife. In the film it partly accompanies a scene of the Five Points gangs assembling for battle in the streets of 19th century Manhattan. The film is tagged on IMDB as simply Crime; the classifier labeled the track with Comedy, Action and Drama. The comedic element here is a total mystery. The best guess is that the background voices and timbre of the fife correlate somewhat with the Comedy class. The Action tag, however, gives support to the idea that noisy percussion is a significant element of the Action prototype.

“Diary of a Taxi Driver” is, like “Shimmy She Wobble”, curiously mistagged as Comedy. Another unusual track from another Scorsese film, most of “Diary” is occupied by monologue from Robert DeNiro, underscored by Bernard Herrmann’s clashing brass and rolling cymbals. This provides a further clue that the classifier is somehow grappling with spoken work examples by labeling them as Comedy. How it came to do so is a mystery, but analysis of the dataset could confirm whether a significant proportion of Comedy examples contain speaking voices. The whole answer is certainly more complex, but such a connection would be a promising place to begin.

Conclusion

These experiments demonstrate that an ensemble of deep convolutional binary classifiers, trained on simple MFCC features, can successfully identify the film genres associated with film soundtracks. A comprehensive dataset of film audio features was collected and labeled with transparently sourced and consistent film genre labels. The 10 most numerous classes were selected for training and testing.

In an attempt to find the best model, three basic architectures were evaluated and built upon in sequence, up to the final bagging model of 10 binary classifiers. It was demonstrated, through the significantly improved accuracy of the deep convolutional model, that the best class features were invariant to scale and translation, and therefore timbre and time dependent.

Upon closer analysis of the dataset, it was determined that a fully-connected, softmax classifier was inappropriate for the data as organized, and would achieve suboptimal results even with ideal hyper-parameters. A final ensemble model was proposed to extend the power of the deep convolutional architecture and avoid clashing training examples and premature early stopping. Furthermore, by showing the flexibility, accuracy, and expressiveness of the bagged ensemble of CNNs, it demonstrates that such a model can make productive "mis-labelings" that describe the practical uses of the particular audio examples in finer detail than the ground-truth labels themselves. In the final aim of this thesis, which is to pave the way for a practical recommender agent for music supervisors and recording artists, such a system would not only aid search but also discovery and novelty.

The model described is only a snapshot of continuing research, as the model continues to respond favorably to expanded architectures and better tuned hyper-parameters. A re-evaluation and re-labeling of the dataset would certainly be promising, as it could remove outliers and present the examples in a multi-label format that could train a single fully connected logistic network to estimate multiple tags. In addition, more sophisticated ensembling, such as boosting, could increase classification accuracy. Finally, given enough memory space and time for training, the convolutional networks could train on raw spectrogram data, or at least an alternative compression feature with less of the approximation inherent to MFCCs. Research has shown that raw spectrograms can outperform MFCCs and similar features on speech recognition and music classification, and it's probable that there are pitch-dependent features that correlate strongly to film score genres.

Appendix 1

Neural Network Architectures

	2 Layer DNN	6 Layer CNN	5 Layer CNN Ensemble (x 10)
Input Dimension	1 x 4000	1 x 40 x 100	
Convolutional Layer	/	32 filters 5 x 5	/
Max Pool Layer	/	2 x 2	/
Dropout	/	0.1	/
Convolutional Layer	/	64 filters 3 x 3	32 filters 3 x 3
Max Pool Layer	/	2 x 2	2 x 2
Dropout	/	0.2	0.1
Convolutional Layer	/	128 filters 2 x 2	64 filters 2 x 2
Max Pool Layer	/	2 x 2	2 x 2
Dropout	/	0.3	0.2
Convolutional Layer	/	256 filters 2 x 2	128 filters 2 x 2
Max Pool Layer	/	2 x 2	2 x 2
Dropout	0.2	0.4	0.3
Dense Layer	50	50	50
Dropout	0.5	0.5	0.5
Dense Layer	50	50	50
Output Layer	10 (softmax)	10 (softmax)	1 (sigmoid)

References

(2014). "Underfitting vs. Overfitting." Retrieved December 1, 2015, 2015, from http://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html.

(2015). Data Preprocessing. Unsupervised Feature Learning and Deep Learning, Stanford University.

Adegoke, Y. (2014). The Who Gets Remixed to Reach New Generation in Ads.

Aljukhadar, M., Secal, S., Daoust, C.-E. (2012). (2012). "Using Recommendation Agents to Cope with Information Overload." International Journal of Electronic Commerce **17**(2): 41-70.

Anderson, T. (2013). "From Background Music to Above-the-Line Actor: The Rise of the Music Supervisor in Converging Televisual Environments." Journal of Popular Music Studies **25**(3): 371-388.

Anglade, A., et al. (2010). "Improving Music Genre Classification Using Automatically Induced Harmony Rules." Journal of New Music Research **39**(4): 349-361.

Batlle, E., et al. (1998). Feature decorrelation methods in speech recognition. a comparative study. ICSLP.

Berenzweig, A., et al. (2004). "A large-scale evaluation of acoustic and subjective music-similarity measures." Computer Music Journal **28**(2): 63-76.

Bishop, C. M. (2006). Pattern recognition and machine learning, springer.

Breiman, L. (1996). "Bagging predictors." Machine learning **24**(2): 123-140.

Burke, R., et al. (2011). "Recommender systems: An overview." AI Magazine **32**(3): 13-18.

Burred, J. J., et al. (2006). An accurate timbre model for musical instruments and its application to classification. Workshop on Learning the Semantics of Audio Signals, Athens, Greece.

Calavita, M. (2007). "" MTV Aesthetics" at the Movies: Interrogating a Film Criticism Fallacy." Journal of Film and Video: 15-31.

Chadwick, A. (2000). Profile: Work of musician Nick Drake and his rising posthumous stardom. NPR Morning Edition, National Public Radio.

Cummings, A. D. P. (2007). "Still Ain't No Glory in Pain: How the Telecommunications Act of 1996 and Other 1990s Deregulation Facilitated the Market Crash of 2002." Fordham J. Corp. & Fin. L. **12**: 467.

Dahl, G. E., et al. (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout. Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, IEEE.

Diderot, D. (1978). Diderot Encyclopédie: The Complete Illustrations, 1762-1777, Harry N. Abrams, Incorporated.

Dieleman, S. and B. Schrauwen (2014). End-to-end learning for music audio. Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on.

Dillet, R. (2014, January 25, 2014). "SoundCloud Raises $60 Million At A $700 Million Valuation." Retrieved November 26, 2015, from http://techcrunch.com/2014/01/25/soundcloud-raises-60-million-at-700-million-valuation/.

Eckhardt, G. M. and A. Bradshaw (2014). "The erasure of antagonisms between popular music and advertising." Marketing Theory: 1470593114521452.

Ellis, D. P. (2007). "Beat tracking by dynamic programming." Journal of New Music Research **36**(1): 51-60.

Eppler, M. J. and J. Mengis (2004). "The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines." The information society **20**(5): 325-344.

Foote, J. (1997). A similarity measure for automatic audio classification. Proc. AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora.

Fraser, C. and J. A. Bradford (2013). "Music to your brain: Background music changes are processed first, reducing ad message recall." Psychology & Marketing **30**(1): 62-75.

Fu, Z., et al. (2011). "A survey of audio-based music classification and annotation." Multimedia, IEEE Transactions on **13**(2): 303-319.

Galdston, P., et al. (2013). "Panel Discussion: Songs in Films Part Two." Music and the Moving Image **6**(3): 47-65.

Gjerdingen, R. O. and D. Perrott (2008). "Scanning the dial: The rapid recognition of music genres." Journal of New Music Research **37**(2): 93-100.

Grosche, P. and M. Müller (2011). "Extracting predominant local pulse information from music recordings." Audio, Speech, and Language Processing, IEEE Transactions on **19**(6): 1688-1701.

Grzywczak, D. and G. Gwardys (2014). Audio Features in Music Information Retrieval. Active Media Technology. D. Ślęzak, G. Schaefer, S. Vuong and Y.-S. Kim, Springer International Publishing. **8610:** 187-199.

Hamel, P. and D. Eck (2010). Learning Features from Music Audio with Deep Belief Networks. ISMIR, Utrecht, The Netherlands.

Haupt, J. (2009). "Last. fm: People−Powered Online Radio." Music Reference Services Quarterly **12**(1-2): 23-24.

Houtsma, A. J. (1997). "Pitch and timbre: Definition, meaning and use." Journal of New Music Research **26**(2): 104-115.

Humphrey, E. J., et al. (2012). Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics. ISMIR, Citeseer.

Kaminskas, M. and F. Ricci (2012). "Contextual music information retrieval and recommendation: State of the art and challenges." Computer Science Review **6**(2): 89-119.

Kessler, K. (2013). "Broadway in the Box: Television's Infancy and the Cultural Cachet of the Great White Way." **25**: 349-370.

Krumhansl, C. L. (2010). "Plink:" Thin slices" of music."

Lidwell, W., et al. (2003). Universal Principles of Design: 100 Ways to Enhance Usability, Influence Perception, Increase Appeal, Make Better Design Decisions, and Teach Through Design, Rockport Publishers.

Lopresti, D. (2001). Web document analysis and information retrieval. First International Workshop on Web Document Analysis (WDA 2001), Seattle, Washington, USA.

Lunden, I. (2014). "Spotify Acquired Music Tech Company The Echo Nest In A $100M Deal." Retrieved November 26, 2015, from http://techcrunch.com/2014/03/07/spotify-echo-nest-100m/.

Lunny Jr, G. S. (2014). "Copyright's Mercantilist Turn." Fla. St. UL Rev. **42**: 95.

MacInnis, D. J. and C. W. Park (1991). "The differential role of characteristics of music on high-and low-involvement consumers' processing of ads." Journal of consumer Research: 161-173.

Marsland, S. (2014). Machine Learning: An Algorithmic Perspective, Second Edition, CRC Press.

McNee, S. M., et al. (2006). Being accurate is not enough: how accuracy metrics have hurt recommender systems. CHI '06 Extended Abstracts on Human Factors in Computing Systems. Montr&#233;al, Qu&#233;bec, Canada, ACM**:** 1097-1101.

Näätänen, R. and I. Winkler (1999). "The concept of auditory stimulus representation in cognitive neuroscience." Psychological bulletin **125**(6): 826.

Oakes, S. (2007). "Evaluating empirical research into music in advertising: A congruity perspective."

Patil, K., et al. (2012). "Music in our ears: the biological bases of musical timbre perception."

Percival, G. and G. Tzanetakis (2014). "Streamlined tempo estimation based on autocorrelation and cross-correlation with pulses." Audio, Speech, and Language Processing, IEEE/ACM Transactions on **22**(12): 1765-1776.

Pohlmann, K. (2010). Principles of Digital Audio, Sixth Edition, McGraw-Hill Education.

Schindler, A. and A. Rauber (2014). Capturing the Temporal Domain in Echonest Features for Improved Classification Effectiveness. Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation. A. Nürnberger, S. Stober, B. Larsen and M. Detyniecki, Springer International Publishing. **8382:** 214-227.

Schroeder, S. (2013). "Music and Metaphor." The British Journal of Aesthetics **53**(1): 1-20.

Scikit-Learn.org (2014). "Underfitting vs. Overfitting." Retrieved December 1, 2015, 2015, from http://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html.

Seabrook, J. (2014). Revenue Streams: Is Spotify the Music Industry's Friend or Its Foe?

Seabrook, J. (2014). Revenue Streams: Is Spotify the Music Industry's Friend or Its Foe? New Yorker, Conde Nast.

Stanford (2015). Data Preprocessing. Unsupervised Feature Learning and Deep Learning, Stanford University.

Sturm, B. L. (2013). "Classification accuracy is not enough." Journal of Intelligent Information Systems **41**(3): 371-406.

Suied, C., et al. (2014). "Auditory gist: Recognition of very short sounds from timbre cues." The Journal of the Acoustical Society of America **135**(3): 1380-1391.

Tsunoo, E., et al. (2009). Audio genre classification using percussive pattern clustering combined with timbral features. Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on.

Tzanetakis, G. and P. Cook (2000). Audio information retrieval (AIR) tools. Proc. International Symposium on Music Information Retrieval.

Tzanetakis, G. and P. Cook (2002). "Musical genre classification of audio signals." Speech and Audio Processing, IEEE transactions on **10**(5): 293-302.

Villarroel, J. A., et al. (2013). "Innovation and learning performance implications of free revealing and knowledge brokering in competing communities: insights from the Netflix Prize challenge." Computational and Mathematical Organization Theory **19**(1): 42-77.

Wakefield, M. (2009). You Oughta Be In Pictures. Peforming Songwriter. **16:** 18-24.

Wright, M., et al. (2008). Analyzing Afro-Cuban Rhythms using Rotation-Aware Clave Template Matching with Dynamic Programming. ISMIR.

Zhou, R., et al. (2008). "Music onset detection based on resonator time frequency image." Audio, Speech, and Language Processing, IEEE Transactions on **16**(8): 1685-1695.