MODELING GENRE WITH THE MUSIC GENOME PROJECT: COMPARING HUMAN-LABELED ATTRIBUTES AND AUDIO FEATURES

Matthew Prockup[+*], **Andreas F. Ehmann**[*], **Fabien Gouyon**[*]
Erik M. Schmidt[*], **Oscar Celma**[*], **and Youngmoo E. Kim**[+]
Electrical and Computer Engineering, Drexel University[+] and Pandora Media Inc.[*]
{mprockup,ykim}@drexel.edu {aehmann,fgouyon,eschmidt,ocelma}@pandora.com

ABSTRACT

Genre provides one of the most convenient categorizations of music, but it is often regarded as a poorly defined or largely subjective musical construct. In this work, we provide evidence that musical genres can to a large extent be objectively modeled via a combination of musical attributes. We employ a data-driven approach utilizing a subset of 48 hand-labeled musical attributes comprising instrumentation, timbre, and rhythm across more than one million examples from Pandora® Internet Radio's *Music Genome Project*®. A set of audio features motivated by timbre and rhythm are then implemented to model genre both directly and through audio-driven models derived from the hand-labeled musical attributes. In most cases, machine learning models built directly from hand-labeled attributes outperform models based on audio features. Among the audio-based models, those that combine audio features and learned musical attributes perform better than those derived from audio features alone.

1. INTRODUCTION

Musical *genre* is a high-level label given to a piece of music (e.g., Rock, Jazz) to both associate it with similar music pieces and distinguish it from others. Genre is a very popular way to organize music as it is being used by virtually all actors in the music industry, from record labels and music retailers, to music consumers and musicians via radio and music streaming services on the internet.

Just because genres are widely used does not necessarily mean that they are easy to categorize, or easy to recognize. In fact, previous research shows that the music industry uses inconsistent genre taxonomies [21], and there is debate over whether genre is the product of objective or subjective categorizations [28]. Furthermore, it is debated whether individual musical properties (e.g. tempo, rhythm, instrumentation), which are not always exclusive to a single genre, represent defining components [1, 10]. For example, an Afro-Latin clave pattern occurs many places, both in Antonio Carlos Jobim's *The Girl from Ipanema* (Jazz) and in The Beatles' *And I Love Her* (Rock). It can even be heard in the recently popular song, *All About that Bass*, by Meghan Trainor. However, when discriminating the more specific subgenres of 'Bebop' Jazz (fast swing) and 'Brazilian' Jazz (Afro-Latin rhythms), this clave property becomes much more salient. Despite these intriguing relationships, a large-scale analysis of the association of musical properties to genre, to the knowledge of the authors, has yet to be performed.

If it were possible to define a categorization of music genres that is useful, meaningful, consensual and consistent *at some level*, then an automated categorization of music pieces into genres would be both achievable and highly desirable. Since early research in Music Information Retrieval (MIR), and still to date, the automatic genre recognition from music pieces has precisely been an important topic [1, 28, 30].

In this work, we explore the intriguing relationship of genre and musical attributes. In Section 3, we will overview the expertly-curated data used. In Section 4, we detail an applied musicology experiment that uses expertly-labeled musical attributes to model genre. We then report in Section 5 on a series of experiments regarding automated categorization of music pieces into genres using audio signal analysis. In the following section, we will briefly outline each of these approaches.

2. APPROACH

In this work we explore four approaches to modeling musical genre, investigating both expert human annotations as well as audio representations (Figure 1). We explore a subset of 12 'Basic' musical genres (e.g. Jazz) as well as a selected subset of 47 subgenres (e.g. Bebop). In the first approach, we address via data-driven experiments whether objective musical attributes of music pieces carry sufficient information to categorize their genre. The next set of approaches uses audio features to model genre automatically. In the second approach, we use audio features directly. The third approach uses audio features to model each of the musical attributes individually, which are then used to model genre. In the fourth approach, the estimated attributes are used in conjunction with raw audio features.

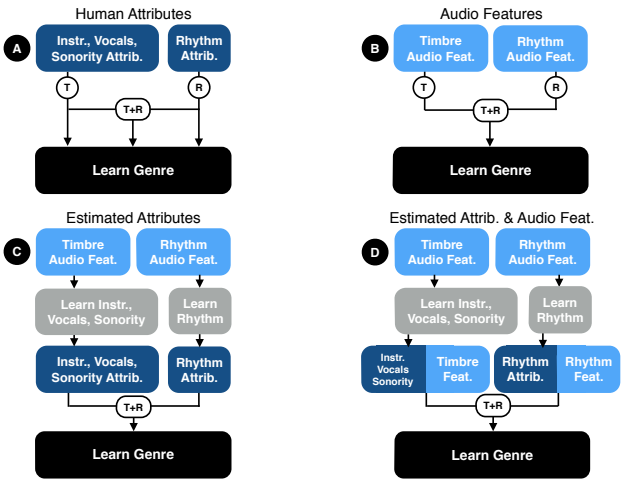

Figure 1. An overview of the experiments performed.

By injecting human-inspired context, we hope to automatically capture elements of genre in a manner similar to that of models derived from attributes labeled by music experts.

3. DATA - THE MUSIC GENOME PROJECT®

Both the musical attribute and genre labels used were defined and collected by musical experts on a corpus of over one million music pieces from Pandora® Internet Radio's *Music Genome Project*® (MGP) [1]. The labels were collected over a period of nearly 15 years and great care was placed in defining them and analyzing each song with that consistent set of criteria.

3.1 Musical Attributes

The musical attributes refer to specific musical components comprising elements of the vocals, instrumentation, sonority, and rhythm. They are designed to have a generalized meaning across all genres (in western music) and map to specific and deterministic musical qualities. In this work, we choose subset of 48 attributes (10 rhythm, 38 timbre). An overview of the attributes is shown in Table 1.

Meter attributes denote musical meters separate from simple duple (e.g, cut-time, compound-duple, odd)
Rhythmic Feel attributes denote rhythmic interpretation (e.g., swing, shuffle, back-beat strength) and elements of rhythmic perception (e.g., syncopation, danceability)
Vocal attributes denote the presence of vocals and timbral characteristics of voice (e.g., male, female, vocal grittiness).
Instrumentation attributes denote the presence of instruments (e.g., piano) and their timbre (e.g., guitar distortion)
Sonority attributes describe production techniques (e.g., studio, live) and the overall sound (e.g., acoustic, synthesized)

Table 1. Explanations of rhythm and timbre attributes.

[1] "Pandora" and "Music Genome Project" are registered trademarks of Pandora Media, Inc. `http://www.pandora.com/about/mgp`

Each of the attributes is rated on continuous scale from 0-1. In some contexts, it is helpful to convert them to binary labels if they show only low (absence) or high (presence) ratings with little in between [25].

3.2 Genre and Subgenre

In this work we will explore a selected subset of 12 'Basic' genres and 47 additional sub-genres. 'Basic' genre is assembled as a mix of very expansive genres (e.g., Rock, Jazz) as well as some more focused ones (e.g., Disco and Bluegrass), serving as an analog to many previous genre experiments in MIR. The presence of a genre is notated independently for each song by a binary label. A selection of genre labels and a simplistic high-level organization for discussion purposes is shown in Table 2.

Basic Genre: Rock, Jazz, Rap, Latin, Disco, Bluegrass, etc.
Jazz Subgenre: Cool, Fusion, Hard Bop, Afro-Cuban, etc.
Rock Subgenre: Light, Hard, Punk, etc.
Rap Subgenre: Party, Old School, Hardcore, etc.
Dance Subgenre: Trance, House, etc.
World Subgenre: Cajun, North African, Indian, Celtic, etc.

Table 2. Some of the musical genres and subgenres used.

4. MUSICAL ATTRIBUTE MODELS OF GENRE

In order to see the extent to which genre can be modeled by musical attributes, we first perform an applied musicology experiment using the set of expertly-labeled attributes from Section 3.1 and relate them to labels of genre. A model for each induvidual genre is trained on each of the musical attributes alone and in rhythm- and timbre-based aggregations. This will show the role that each attribute or collection of attributes plays and how they interact with one another in order to create joint representations of genre. Each model employs logistic regression trained using stochastic gradient decent (SGD) [25]. The training data was separated on a randomly shuffled 70%:30% (train:test) split with no shared artists between training and testing. Due to the size of the dataset, a single trial for each attribute is both tractable and sufficient. The learning rate for each genre model is tuned adaptively.

4.1 Evaluating the Role of Musical Attributes

In order to evaluate each of the models, the area under the receiver operating characteristic (ROC) curve will be used. Each genre has large and varying class imbalance, so this is first corrected for by weighting training examples appropriately in the cost function. However, accuracy alone still does not tell the whole story. High accuracy can be achieved by predicting only the negative class (genre absence). Area under the ROC curve allows for a more comparable difference between each of the models than raw accuracy alone. It gives insight into the trade-off between true positive and false positive rates. Alternatively

we could have used precision and recall (PR) curves for evaluation, but it is shown that if one model dominates in the ROC domain, it will also dominate in the PR domain and vice-versa [5]. In this work, the area under the ROC curve will be referred to as AUC.

The results for each of the attribute-based genre models are shown in Tables 3 and 4. The tables outline the AUC values for classifying genre using timbre attributes, rhythm attributes, and their combination. Table 3 summarizes all results, showing the mean of all AUC values for each genre model contained in the subgroups defined in Section 3.2. Using attributes of rhythm and timbre together show better performance than using each alone. Secondly, timbre tends to perform better than rhythm. This suggests that the timbre attributes in this context are better descriptors. However in some cases, the rhythm attributes, even though there are less of them (10 rhythm, 38 timbre), are not that far behind. They are especially important in defining Jazz and Rap, where rhythms such as swing in Jazz or syncopated vocal cadences over back-beat heavy drums in Rap play defining roles.

Genre Group	Timbre	Rhythm	Both
Basic	0.905	0.841	0.918
Rock Sub	0.910	0.819	0.919
Jazz Sub	0.925	0.856	0.945
Rap Sub	0.901	0.891	0.940
Dance Sub	0.961	0.881	0.965
World Sub	0.885	0.833	0.904
Mean	0.913	0.848	0.931

Table 3. An overview of all models using musical attributes.

In Table 4 we show the individual AUC results for the set of 'Basic' genres and subgenres of Jazz. Within these individual groups, rhythm and timbre attributes together are once again able to better represent genre than when used individually. Each of the 'Basic' genres can be represented reasonably well with just timbre, as each has slightly differing instrumentation. However, we again see the importance of rhythm, describing what instrumentation and timbre cannot capture alone. Genres heavily reliant on specific rhythms (e.g., Funk, Rap, Latin, Disco, Jazz) are all able to be represented rather well with only rhythm attributes. In the Jazz subgenre this emphasis on rhythm in certain cases is even more clear. In the next subsection, we will dive deeper into the attributes that best describe the Jazz subgenres.

4.2 The Influence of Rhythm and Timbre in Jazz

In order to more deeply explore the defining relationships of rhythm and instrumentation within a subgenre, we will look further into Jazz. Table 5 shows a subset of the important musical attributes for the Jazz subgenres. The AUC accuracy of classifying each subgenre based on individual musical attributes is shown.

The presence of solo brass (e.g, trumpet), piano, reeds (e.g., saxophone) and auxiliary percussion (e.g., congas) are important defining characteristics of instrumentation.

Basic				**Jazz**			
Genre	**Timbre**	**Rhythm**	**Both**	**Subgenre**	**Timbre**	**Rhythm**	**Both**
Rock	0.843	0.759	0.856	New Orleans	0.970	0.957	0.989
Blues	0.913	0.783	0.915	Boogie	0.943	0.893	0.978
Gospel	0.810	0.664	0.843	Swing	0.970	0.933	0.984
Soul	0.869	0.793	0.887	Bebop	0.976	0.965	0.988
Funk	0.937	0.862	0.937	Cool	0.964	0.928	0.975
Rap	0.926	0.890	0.951	Hard Bop	0.944	0.905	0.967
Folk	0.943	0.760	0.952	Fusion	0.843	0.750	0.886
Country	0.952	0.794	0.955	Free	0.906	0.855	0.936
Reggae	0.893	0.819	0.905	Afro-Cuban	0.961	0.910	0.972
Latin	0.940	0.904	0.945	Brazilian	0.871	0.847	0.905
Disco	0.899	0.891	0.902	Acid	0.886	0.660	0.891
Jazz	0.937	0.850	0.963	Smooth	0.862	0.667	0.871
Mean	0.905	0.814	0.918	**Mean**	0.925	0.856	0.945

Table 4. Experimental results for 'Basic' genre and Jazz subgenre models using musical attributes.

Jazz	**Timbre**			Aux.	**Rhythm**				
Subgenre	Solo Brass	Piano	Reeds	Perc.	BackBeat	Dance	Swing	Shuffle	Syncop.
New Orleans	**0.808**	**0.786**	**0.790**	0.680	0.652	0.564	**0.936***	0.513	0.515
Boogie	0.510	**0.924***	0.544	**0.714**	0.592	**0.712**	**0.737**	0.505	0.676
Swing	**0.721**	**0.784**	**0.748**	0.679	0.624	0.578	**0.923***	0.511	0.508
Bebop	**0.725**	**0.850**	**0.862**	**0.703**	0.662	0.525	**0.946***	0.509	0.602
Cool	0.639	**0.750**	**0.836**	**0.701**	0.697	0.424	**0.890***	0.504	0.568
HardBop	0.606	**0.774**	**0.737**	0.669	**0.726**	0.555	**0.808***	0.684	0.606
Fusion	0.604	0.497	0.669	0.507	0.574	0.577	0.507	0.500	0.693*
Free	0.606	0.538	**0.784**	0.615	**0.809***	**0.765**	0.577	0.515	0.558
Afro-Cuban	0.696	**0.822**	**0.706**	**0.832***	**0.782**	0.648	0.512	0.501	**0.790**
Brazilian	0.560	**0.736**	0.568	0.572	**0.761***	0.555	0.532	0.504	0.635
Acid	0.591	0.513	0.658*	0.507	0.585	0.622	0.509	0.515	0.635
Smooth	0.530	0.577	**0.748***	0.590	0.559	0.614	0.513	0.509	0.573

Table 5. Attributes important to the Jazz subgenres are shown. AUC values greater than 0.70 are bold. The highest performing attribute for each genre is denoted with a *.

Boogie and Afro-Cuban styles, even though different, place heavy emphasis on the piano, which is shown here as well. Bebop, Hard-bob, and Afro-Cuban Jazz show emphasis placed on solo brass, piano, and reeds, as they rely heavily on solo artists of these instruments (e.g., "Dizzy" Gillespie, Miles Davis, Thelonious Monk, John Coltrane). The presence of auxiliary percussion is also a good descriptor of Afro-Cuban Jazz, where the use of hand drums (e.g., bongos, congas) is very prevalent.

Rhythm is also important in Jazz subgenres. The danceability, back-beat, and presence of swing and syncopation are defining characteristics of certain Jazz rhythms. It is important to note that a high AUC does not necessarily denote the presence of that attribute, only its consistent relationship. For example, back-beat is a good predictor of Free Jazz possibly due to its absolute absence. Alternatively, one may think that the presence of swing is important in all Jazz. Bebop, Hard Bop, New Orleans, and Swing Jazz do have a heavy dependence on swing being present. However, Afro-Cuban Jazz relies on straight time, clave-based rhythms, so syncopation is actually a better predictor. It is also important to note that while the attributes of swing and shuffle are musically related, there is a clear distinction in their application. In this case, swing is very important, while shuffle is only slightly useful (e.g., Boogie). However, outside of the Jazz genre, the opposite case may be true, where shuffle is the more important attribute (e.g. Blues, Country). This suggests that it is important to make a clear distinction between swing and shuffle.

5. PREDICTING GENRE FROM AUDIO

There is a large body of work on musical genre recognition from and audio signals [28,30]. However, most known prior work in this area focuses on discriminating a discrete set of basic genre labels with little emphasis on what defines genre. In response, researchers have tried to develop datasets that focus on style or subgenre labels (e.g., ballroom dance [7, 13, 24], latin [19], electronic dance [23], Indian [17]) that have clear relations to the presence of specific musical attributes. However, because models are designed for these specific sets, the methods used do not adapt to larger more generalized music collections. For example, tempo alone is a good descriptor for the ballroom dance style dataset, which is not true for more general collections [12].

Other work in genre recognition avoids the problem of strict genre class separations. Audio feature similarity, self organizing maps, and nearest-neighbor approaches can be used estimate genre of an unknown example [22]. Similarly, auto-tagging approaches use audio features to learn the presence of both musical attributes and genre tags curated by the public [2, 8] or by experts [29].

In this work, we compare modeling genre both with audio features directly and with stacked approaches that exploit the relationships of audio features and musical attributes.

5.1 Timbre Related Features

In order to capture timbral components and model vocal, instrumentation, and sonority attributes, block-based Mel-Frequency Cepstral Coefficients (MFCC) are implemented. Means and covariances of 20 MFCCs are calculated across non-overlapping 3-second blocks. These block-covariances are further summarized over the piece by calculating their means and variances [27]. This yields a 460 dimensional timbre based feature set.

5.2 Rhythm Related Features

In order to capture aspects of each rhythm attribute, a set of rhythm-specific features was implemented. All rhythm features described in this section rely on global estimates of an accent signal [3]

The *beat profile* quantizes the accent signal between consecutive beats to 36 subdivisions. The beat profile features are statistics of those 36 bins over all beats. The feature relies on estimates of both beats [9] and tempo.

The *tempogram ratio* feature (TGR) uses the tempo estimate to remove the tempo dependence in a tempogram. By normalizing the tempo axis of the tempogram by the tempo estimate, a fractional relationship to the tempo is gained. A compact, tempo-invariant feature is created by capturing the weights of the tempogram at musically related ratios relative to the tempo estimate.

The *Mellin scale transform* is a scale invariant transform of a time domain signal. Similar musical patterns at different tempos are scaled relative to the tempo. The Mellin scale transform is invariant to that tempo scaling. It was first introduced in the context of rhythmic similarity by Holzapfel [16], around which our implementation is based. In order to exploit the natural periodicity in the transform, the discrete cosine transform (DCT) is computed. Median removal (by subtracting the local median) and half-wave rectifying the DCT creates a new feature that emphasizes transform periodicities.

The previous rhythm features are also extended to multiple-band versions by using accent signals that are constrained to be within a set of specific sub-bands. This affords the ability to capture the rhythmic function of instruments in different frequency ranges. The rhythm feature set used in this work is an aggregation of the median removed Mellin Transform DCT and multi-band representations of the beat profile and the tempogram ratio features. This yields a 372 dimensional rhythm based feature set that was shown in previous work to be relatively effective at capturing musical attributes related to rhythm (see [25] for more details).

5.3 Genre Recognition Experiments

In addition to the experiment from Section 4, we present three additional methods for modeling genre, each based on audio signal analysis. The second method (Figure 1b) performs the task of genre recognition with rhythm and timbre inspired audio features directly. The third method (Figure 1c) is motivated similar to the first experiment, which employs the expertly-labeled musical attributes. However, inspired by work in transfer learning [4], audio features are used to develop models for the humanly-defined attributes which in turn are used to model genre. Through this supervised pre-training of musical attributes, models of genre can be learned from attributes' estimated presence. In the fourth approach (Figure 1d), inspired by [6] and [18], the learned attributes are combined with the audio features directly in a shared middle layer to train models of genre.

Similar to Section 4, genre is modeled with logistic regression fit using stochastic gradient decent (SGD). The data was separated on the same 70%:30% (train:test) split. Once again, there were no shared artists between training and testing. Due to the size of the dataset, a single trial for each genre, as well as for each learned musical attribute, is both tractable and sufficient. The learning rate for each model is tuned adaptively.

5.3.1 Using Audio Features Directly

Of the four presented approaches, the second uses audio features directly to model genre. The features from Sections 5.1 and 5.2 are used in aggregation and a model is trained and tested for each individual genre. This provides a baseline for what audio features are able to capture without any added context. However, this lack of context makes it hard to interpret what about genre they are capturing.

5.3.2 Stacked Methods

The third and fourth approaches are also driven by audio features. However instead of targeting genre directly,

models are learned for each of the vocal, instrumentation, sonority, and rhythm attributes. Inspired by approaches in transfer learning [4], and similar in structure to previous methods in the MIR community [20], the learned attributes are then used to predict genre. This approach is formulated similar to a basic neural network with a supervised pre-trained (and no longer hidden) musical attributes layer.

The rhythm-based attributes are modeled with a feature aggregation of the Mellin DCT, multi-band beat profile, and multi-band tempogram ratio features. The vocals, instrumentation, and sonority attributes are modeled with the block-based MFCC features. Each attribute is modeled using logistic regression for binary labels (categorical) and linear regression for continuous labels (scale-based). If an individual attribute is formulated as a binary classification task (see Section 3.1), the probability of the positive class (its presence) is used as the feature value.

The first version of the stacked methods (third approach) uses audio features to estimate musical attributes and employs only those estimated attributes to model genre. The second version (fourth approach) concatenates the audio features and the learned attributes in a shared middle layer to model genre [6, 18].

5.4 Results

In this section, we will give an overview of all of the results from the audio-based methods, and compare them to the models learned from the expertly-labeled attributes. In order to show the overall performance of each method in a compact way, only combined rhythm and timbre approaches will be compared. Once again each genre model will be evaluated using area under the ROC curve (AUC). In order to better evaluate the stacked models, we will finish with a brief evaluation of the learned attributes.

5.4.1 Learning Genre

A summary of the results for the audio experiments using rhythm and timbre features is shown in Table 6. The human attribute model results are also included for comparison. Similar to Table 3, the mean AUC of each genre grouping is shown.

Genre Group	Human Attrib.	Audio Feat.	Learned Attrib.	Audio + Learned
Basic	0.918	0.892	0.878	0.899
Rock Sub	0.919	0.902	0.903	0.911
Jazz Sub	0.945	0.910	0.893	0.923
Rap Sub	0.940	0.916	0.914	0.927
Dance Sub	0.965	0.963	0.955	0.966
World Sub	0.904	0.850	0.846	0.865
Mean	0.931	0.905	0.897	0.915

Table 6. An overview of experimental results using audio-based models that utilize timbre and rhythm features.

Compared to the human attributes approach, using audio features alone to model genre performs relatively well. This is especially true for the 'Basic', Rock, and Dance groups, where the audio feature AUC results are very close to human attribute performance. Across the other groups, the differences between the audio feature models and the musical attribute models suggest that the audio features lose some important, genre-defining information. Furthermore, performance that was close to musical attributes when using only audio features alone is also close when musical attributes learned from audio features. This suggests that, in these cases, the audio features may be capturing similarly salient components related to the musical attributes that describe these genre groups.

Overall, the learned attributes perform just as good as or worse than the audio features alone. This suggests that they are at most as powerful as the audio features used to train them. However, combining audio features and learned attributes shows significant improvement (paired t-test $p < 0.01$ across all genres) over using audio features or learned attributes alone. This evidence suggests that audio features and learned attribute models each contain slightly different information. The added human context of the learned attributes is helpful to achieve results that approach those of the expertly-labeled attributes. This also suggests that the decisions made from learned labels are possibly more similar to the decisions made from human attribute labels, and the errors in estimating the musical attributes are possibly to blame for the performance decrease when used alone.

Basic Genre	Human Attrib.	Audio Feat.	Learned Attrib.	Audio + Learned	Jazz Subgenre	Human Attrib.	Audio Feat.	Learned Attrib.	Audio + Learned
Rock	0.856	0.831	0.835	0.839	New Orleans	0.989	0.947	0.951	0.956
Blues	0.915	0.892	0.883	0.899	Boogie	0.978	0.962	0.939	0.962
Gospel	0.843	0.798	0.794	0.805	Swing	0.984	0.929	0.929	0.940
Soul	0.887	0.833	0.818	0.842	Bebop	0.988	0.951	0.943	0.957
Funk	0.937	0.911	0.886	0.918	Cool	0.975	0.900	0.901	0.916
Rap	0.951	0.963	0.951	0.969	HardBop	0.967	0.946	0.930	0.952
Folk	0.952	0.905	0.903	0.916	Fusion	0.886	0.844	0.812	0.867
Country	0.955	0.885	0.880	0.897	Free	0.936	0.920	0.923	0.931
Reggae	0.905	0.926	0.885	0.929	AfroCuban	0.972	0.934	0.912	0.946
Latin	0.945	0.921	0.905	0.923	Brazilian	0.905	0.879	0.858	0.904
Disco	0.902	0.936	0.893	0.938	Acid	0.891	0.841	0.763	0.846
Jazz	0.963	0.907	0.906	0.916	Smooth	0.871	0.868	0.853	0.894
Mean	0.918	0.892	0.878	0.899	**Mean**	0.945	0.910	0.893	0.923

Table 7. Experimental results for the 'Basic' genres and Jazz subgenres using audio-based models.

The left half of Table 7 shows the results for predicting the 'Basic' genre labels. Within this set, we see some interesting patterns start to emerge. In the case of Rap, Reggae, and Disco, audio features are actually able to outperform the musical attributes. This suggests that our small selected subset of 48 human attribute labels do not always tell the whole story, and that the audio features, which are much larger in dimensionality, possibly contain additional and/or different information. As in previous results, the learned attribute models perform similarly to methods that use audio features directly, but with a few exceptions. In the cases that the audio feature models do better than the human-labeled musical attribute models, the learned attribute models are able to perform *at most* as good as the human-labeled musical attribute models. This once again suggests that the learned attribute approach may be better approximating the decisions the human-labeled attribute approach is making. When adding audio and learned attributes together, the added context is once again beneficial, with combined methods outperforming models that use audio or learned attributes alone. Audio feature models that perform better than the human attributes models

are additionally improved, showing again that the audio features and human attribute labels contain complementary information.

The right half of Table 7 shows the results for predicting the Jazz subgenre labels. The Jazz genre shows more expected relationships between the human attribute, audio feature, and learned attribute methods. The combined method outperforms each of the audio feature and learned attribute methods. The human attribute method performs better than all audio-based methods.

5.4.2 Learning Attributes

In order to further explore the stacked audio-based models, we performed a small evaluation of how well the audio features are able to learn each of the expertly-labeled musical attributes. Sticking with a common theme, we will explore the results of modeling attributes that are important to Jazz (from Table 5). Table 8 shows the ability to directly predict these attributes from audio features. AUC accuracies are reported for the binary attributes; R^2 values are reported for continuous attributes. The results of evaluating the model for the training and testing sets is shown.

Musical Attributes	**Audio Features**	**Training AUC/R^2**	**Testing AUC/R^2**	**Label Type**
Solo Brass	Timbre	0.796	0.798	binary
Piano	Timbre	0.721	0.716	binary
Reeds	Timbre	0.790	0.789	binary
Aux Percussion	Timbre	0.750	0.750	binary
FeelSwing	Rhythm	0.907	0.902	binary
FeelShuffle	Rhythm	0.919	0.920	binary
FeelSyncopation	Rhythm	0.772	0.770	binary
FeelBackBeat	Rhythm	0.400	0.393	continuous
FeelDance	Rhythm	0.527	0.515	continuous

Table 8. The results for learning binary (AUC) and continuous (R^2) attributes important to Jazz are shown.

First of all, we see that testing and training AUC is almost identical. Because of this, a single trial (fold) is appropriate when learning attribute models. The learned models should generalize over all music without over fitting. This justifies using the the same 70%:30% (train:test) split for each layer in the stacked models. We see that MFCC's do somewhat well for brass and reeds, but the lower AUC overall shows that these timbre features are not doing enough to capture these attributes, which may be a source of error in genre models that rely heavily on timbre. However, the rhythm results are much better, especially for swing and shuffle, which was argued in Section 4 and Table 5 as an important distinction to make when predicting Jazz subgenres.

Attribute Type	**Num**	**Mean**	**Median**	**Maximum**
Continous Rhythm (R^2)	3	0.432 ± 0.077	0.393	0.515
Continous Timbre (R^2)	12	0.266 ± 0.192	0.194	0.514
All Continuous	15	0.299 ± 0.186	0.389	0.515
Binary Rhythm (AUC)	7	0.889 ± 0.059	0.902	0.946
Binary Timbre (AUC)	26	0.794 ± 0.074	0.794	0.925
All Binary	15	0.814 ± 0.080	0.806	0.946

Table 9. Overall summary of learned attributes.

Table 9 shows a summary of learning the all of the selected 48 attributes from audio features. It shows similar trends to Table 8, with rhythmic attributes better described by audio features than timbral attributes. Furthermore, the continuous timbral attributes, which are sometimes complicated perceptually (e.g., vocal grittiness), are not modeled very well at all. This suggests that MFCC's, and possibly other spectral approximations, do not provide the full picture into what we perceive as the components of timbre. This is especially true in the context of instrument identification in mixtures, which is a main utility of the timbre features in this context. While these models as a whole can be improved, the problems of instrument identification and rhythm analysis are separate, large, and active research areas [14, 15, 25, 26].

6. CONCLUSION

In this work, we demonstrated that there is potential to demystify the constructs of musical genre into distinct musicological components. The attributes we selected from music experts are able to provide a great deal of genre distinguishing information, but this is only an initial investigation into these questions. We were also able to discover and outline the importance of certain attributes in specific contexts. This strongly suggests that the expression of musical attributes are necessary additions to definitions of genre.

It was also shown here (and in previous work [25]) that audio features motivated by timbre and rhythm are, with some success, able to model musical attributes. Audio features are also able to describe musical genre directly and through stacked approaches that exploit the learned models of musical attributes. This is strong evidence suggesting that audio-based approaches are learning the presence of the musical attributes, to some degree, when distinguishing genre. In some cases, the audio-based models were more powerful than the human musical attribute models. This suggests that there is more to genre than our chosen subset of rhythm and orchestration attributes, and it makes us contemplate that there is more about the definition of genre yet to be discovered.

In seeking to improve on this work, we next look to investigate replacing the feature concatenation with late fusion of context-dependent classifiers (e.g., rhythm, timbre), which has shown improved results for genre classification [11]. It may also be helpful to use a greater number of the available attributes than the chosen 48, as well as additional attribute types (e.g., melody, harmony). Furthermore, perhaps the most interesting direction is to treat each musical attribute model as a hidden layer in a neural network. In these cases, the models that are trained to predict musicological attributes will serve as a form of domain-specific pre-training. These models would perform full back propagation across an additional layer which connects our attributes to genres. This will potentially help to learn better models of genre as well as adjust the models of musical attributes in order better capture their genre relationships.

7. REFERENCES

[1] Jean-Julien Aucouturier and Francois Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.

[2] Thierry Bertin-Mahieux, Douglas Eck, and Michael Mandel. Automatic tagging of audio: The state-of-the-art. *Machine audition: Principles, algorithms and systems*, pages 334–352, 2010.

[3] Sebastian Böck and Gerhard Widmer. Maximum filter vibrato suppression for onset detection. In *Proc. of the 16th International Conference on Digital Audio Effects (DAFx-13)*, 2013.

[4] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[5] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proc. of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[6] Li Deng and Dong Yu. Deep convex net: A scalable architecture for speech pattern classification. In *Proc. of Interspeech*, 2011.

[7] Simon Dixon, Elias Pampalk, and Gerhard Widmer. Classification of dance music by periodicity patterns. In *Proc. of the International Society for Music Information Retrieval Conference*, 2003.

[8] Douglas Eck, Paul Lamere, Thierry Bertin-Mahieux, and Stephen Green. Automatic generation of social tags for music recommendation. In *Advances in neural information processing systems*, pages 385–392, 2008.

[9] Daniel PW Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007.

[10] Franco Fabbri. A theory of musical genres: Two applications. *Popular music perspectives*, 1:52–81, 1982.

[11] Arthur Flexer, Fabien Gouyon, Simon Dixon, and Gerhard Widmer. Probabilistic combination of features for music classification. In *Proc. of the International Society for Music Information Retrieval Conference*, pages 111–114, 2006.

[12] Fabien Gouyon and Simon Dixon. Dance music classification: A tempo-based approach. In *Proc. of the International Society for Music Information Retrieval Conference*, 2004.

[13] Fabien Gouyon, Simon Dixon, Elias Pampalk, and Gerhard Widmer. Evaluating rhythmic descriptors for musical genre classification. In *Proc. of the AES 25th International Conference*, pages 196–204, 2004.

[14] Philippe Hamel, Sean Wood, and Douglas Eck. Automatic identification of instrument classes in polyphonic and poly-instrument audio. In *Proc. of the International Society for Music Information Retrieval Conference*, pages 399–404, 2009.

[15] Perfecto Herrera-Boyer, Geoffroy Peeters, and Shlomo Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1):3–21, 2003.

[16] André Holzapfel and Yannis Stylianou. Scale transform in rhythmic similarity of music. *IEEE Trans. on Audio, Speech and Language Processing*, 19(1):176–185, 2011.

[17] S Jothilakshmi and N Kathiresan. Automatic music genre classification for indian music. In *Proc. Int. Conf. Software Computer App*, 2012.

[18] Peter Knees, Tim Pohle, Markus Schedl, and Gerhard Widmer. Combining audio-based similarity with web-based data to accelerate automatic music playlist generation. In *Proc. of the 8th ACM international workshop on Multimedia information retrieval*, pages 147–154. ACM, 2006.

[19] Miguel Lopes, Fabien Gouyon, Alessandro L Koerich, and Luiz ES Oliveira. Selection of training instances for music genre classification. In *Proc. of the International Conference on Pattern Recognition*, pages 4569–4572. IEEE, 2010.

[20] F. Pachet and P. Roy. Improving multilabel analysis of music titles: A large-scale validation of the correction approach. *IEEE Trans. on Audio, Speech and Language Processing*, 17(2):335 –343, 2009.

[21] François Pachet and Daniel Cazaly. A taxonomy of musical genres. In *Content-Based Multimedia Information Access Conference*, pages 1238–1245, 2000.

[22] Elias Pampalk, Arthur Flexer, and Gerhard Widmer. Improvements of audio-based music similarity and genre classificaton. In *Proc. of the International Society for Music Information Retrieval Conference*, volume 5, pages 634–637, 2005.

[23] Maria Panteli, Niels Bogaards, and Aline Honingh. Modeling rhythm similarity for electronic dance music. *Proc. of the International Society for Music Information Retrieval Conference*, 2014.

[24] Geoffroy Peeters. Rhythm classification using spectral rhythm patterns. In *Proc. of the International Society for Music Information Retrieval Conference*, pages 644–647, 2005.

[25] Matthew Prockup, Andreas F. Ehmann, Fabien Gouyon, Erik M. Schmidt, and Youngmoo E. Kim. Modeling musical rhythm at scale using the music genome project. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015.

[26] Jeffrey Scott and Youngmoo E Kim. Instrument identification informed multi-track mixing. In *Proc. of the International Society for Music Information Retrieval Conference*, pages 305–310, 2013.

[27] Klaus Seyerlehner, Markus Schedl, Peter Knees, and Reinhard Sonnleitner. A refined block-level feature set for classification, similarity and tag prediction. *Extended Abstract to MIREX*, 2011.

[28] Bob L Sturm. The state of the art ten years after a state of the art: Future research in music information retrieval. *Journal of New Music Research*, 43(2):147–172, 2014.

[29] Derek Tingle, Youngmoo E Kim, and Douglas Turnbull. Exploring automatic music annotation with acoustically-objective tags. In *Proc. of the international conference on Multimedia information retrieval*, pages 55–62. ACM, 2010.

[30] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Trans. on Audio, Speech and Language Processing*, 10(5):293–302, 2002.