ON THE IMPORTANCE OF MODELING TEMPORAL INFORMATION IN MUSIC TAG ANNOTATION

Jeremy Reed and Chin-Hui Lee

School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA, 30332 USA
jeremy.reed@gatech.edu,chl@ece.gatech.edu

ABSTRACT

Music is an art form in which sounds are organized in time; however, current approaches for determining similarity and classification largely ignore temporal information. This paper presents an approach to automatic tagging which incorporates temporal aspects of music directly into the statistical models, unlike the typical bag-of-frames paradigm in traditional music information retrieval techniques. Vector quantization on song segments leads to a vocabulary of acoustic segment models. An unsupervised, iterative process that cycles between Viterbi decoding and Baum-Welch estimation builds transcripts of this vocabulary. Latent semantic analysis converts the song transcriptions into a vector for subsequent classification using a support vector machine for each tag. Experimental results demonstrate that the proposed approach performs better in 15 of the 18 tags. Further analysis demonstrates an ability to capture local timbral characteristics as well as sequential arrangements of acoustic segment models.

Index Terms— Music, Hidden Markov models, Information retrieval, Vector quantization, Speech processing

1. INTRODUCTION

Recently, semantic tags have become a popular means to organizing, retrieving, and discovering multimedia content. In the case of music, tags are replacing traditional music taxonomies such as genre and style. Tags, which are short, descriptive keywords, are assigned by two groups: experts or regular users. An example of an expert-oriented music discovery site is Pandora[1], where musicians categorically rate music along several musical dimensions. A popular website that takes advantage of the collective knowledge is Last.fm[2], where any user can assign any tag to any work at any time. Regardless, both types suffer from the *cold-start problem*, where new or untagged songs cannot be retrieved because no tags exist for the song. In the case of user-assigned tags, other problems include uninformative tags to the general population (e.g., *albums I own*) and vandalism, where users assign tags inappropriately.

A solution to the cold-start problem is automatic tagging, which provides an initial set of tags and allows suggested playlists to incorporate new songs. In addition, automatic tagging algorithms can flag tags that do not describe the content of the work and arose due to vandalism. Previous techniques for providing tags automatically use a bag-of-frames framework [1], which ignores the temporal structure of music by treating small frames of audio as independent and identically distributed. The algorithm detailed in this paper uses speech recognition technology to build a vocabulary of acoustic tokens, which incorporate temporal structure probabilistically. Specifically, an initial set of songs is tokenized into a small set of 128 representative models, called acoustic segment models (ASMs). Each ASM is modeled with a hidden Markov model (HMM) to incorporate temporal information. Latent semantic analysis (LSA) [2] converts each song into a weighted vector of ASM symbol counts and their co-occurrences, which then train vector-based classifiers; i.e., support vector machines (SVMs).

[1] www.pandora.com
[2] www.last.fm

2. PREVIOUS WORK

First approaches to relate music to semantic content utilized data-mining techniques on webpages about the musical works. In [3], semantic basis functions maximize the semantic meaning of words based on musical features. Slaney [4] models the connection between anchor points in the acoustic space and semantic audio descriptions in a hierarchical multinomial clustering model. However, tags have some differences than freely-flowing text. Tags are short, descriptive keywords and generally refer to a particular audio quality; hence, it is possible to directly model the tags acoustically. For example, in [1] a Gaussian mixture model (GMM) is estimated for each song and builds a tag-level GMM using a mixture-of-hierarchies algorithm.

However, these approaches rely on a bag-of-frames model or features derived from the entire audio file; e.g., rhythm features. Examples of past approaches to model temporal aspects include derivatives of features [1] and modulation cepstra [3]. In [5], Casey and Slaney argue that sequences are important in determining musical similarity; however, the authors examined only repeated sections within a song and did not investigate inter-song similarity. Further, the modeling approach relied on uniform segmentation; i.e., texture windows. This paper presents an approach that incorporates temporal information into a probabilistic framework by using HMMs in a similar fashion as automatic speech recognition, and first proposed in [6] for genre recognition. While this approach compares favorably to other approaches, determining the advantages in modeling temporal structures proved difficult to assess because genre recognition is ill-defined [7]. By investigating the proposed approach on tags, the authors demonstrate the advantages of incorporating temporal information directly into the statistical models.

3. ASM-BASED AUDIO TAGGING SYSTEM

This section describes the proposed algorithm in two sub-sections: the front-end tokenizer and the back-end vector-modeling with tag classifiers, as diagrammed in Figure 1.

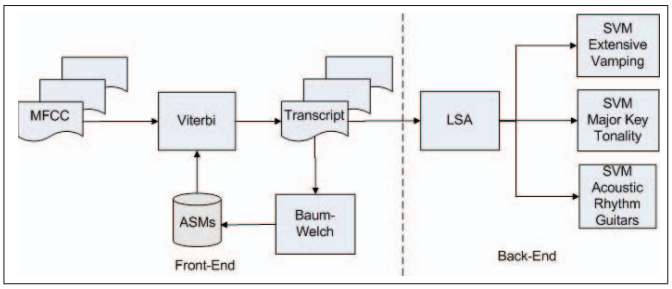

Fig. 1. System diagram.

3.1. Tokenizer

The front-end tokenizes a song into a string of ASM indexes parallel to the way automatic speech recognition tokenizes an utterance into a string of words or phonemes. First, a maximum-likelihood segmentation algorithm [8], which was originally designed to segment speech into subword units (phonemes), tokenizes the training corpus into acoustic segments. Next, vector quantization of the segment centroids builds a vocabulary of ASMs, which are small acoustic tokens. Each song can then be temporally represented as a string of symbols (i.e., a transcript) from the ASM index that best represents the segment. The transcripts and acoustic models are refined through an iterative process between Baum-Welch estimation and Viterbi decoding. First, each ASM in the vector quantized codebook is modeled with a 3-state HMM, which has a 16-mixture GMM with a diagonal covariance matrix in each state. Baum-Welch estimation updates the HMMs using the transcriptions as a reference. Using the updated HMMs, Viterbi decoding creates a new transcription of ASM indexes. The algorithm then uses the new transcripts in the next iteration to refine the HMMs further using Baum-Welch estimation and create new transcripts from these updated HMMs using Viterbi decoding. Generally, only two or three iterations of Baum-Welch estimation followed by Viterbi decoding are needed for convergence.

It should be noted that the segmentation algorithm in [8] is expensive in terms of computation time. Therefore, a small set of songs is used to bootstrap the initial HMMs. The entire training set is added after the first round of Baum-Welch updates. Further, to ensure that the segmentation is based on the slowly changing spectral shape, only the first 8 MFCCs are used for the segmentation procedure [9]. After the vector quantization step that produces the original transcriptions, the 8 MFCCs are replaced with the more traditional 39-dimensional vector consisting of MFCCs 0 through 12, plus their derivatives and accelerations, to incorporate more spectral information.

3.2. Vector Modeling

The front-end tokenizer outputs a transcription of ASM indexes for each song in the training set, which the back-end classifier converts into a vector, using LSA. The final classification uses an SVM for each tag. First, the unigram and bigram counts in each song are obtained, where a unigram is the occurrence of an individual ASM and a bigram is the occurrence of an ordered ASM pair. For the purposes of this paper, a *term* refers to either a unigram or a bigram. For example, if the output transcription of a song is $(3, 44, 3)$, then the output vector would have a 2 in the location for ASM term 3, and a 1 in the location for ASM terms 44, (3,44), and (44,3), but zeros everywhere else. This results in a vector of size $M = J + J * J$, where J is the number of ASMs in the vocabulary. For example, for a vocabulary of size 128, the resulting vector has a dimension of $128 + 128 * 128 = 16512$.

Next, the entropy [2] of each term is calculated as

$$\epsilon_i = -\frac{1}{\log N} \sum_{j=1}^{N} \frac{c_{i,j}}{t_i} \log t_i \quad (1)$$

where N is the number of training songs, $c_{i,j}$ is the count for term i in song j, and t_i is the number of times term i appears in the entire training database. A term entropy close to zero indicates that the term appears in few songs and a value close to one indicates the term occurs in almost every song. The matrix W contains the entropy weighted counts for term i in song j,

$$w_{i,j} = (1 - \epsilon_i) \frac{c_{i,j}}{n_j} \quad (2)$$

where n_j is the number of ASM tokens in song j. To reduce the sparsity in W, singular value decomposition reduces the dimension to 250, which experimentally resulted in a good performance.

Finally, a SVM is trained with SVM^{light} [10] for each tag such that the positive class refers to *tag present* and the negative class refers to *tag missing*. For each test point SVM^{light} returns the distance between a sample and the separating hyperplane, which is compared to a threshold for the classification decision.

3.3. Baseline Classifier

The classifier in [1] provides the baseline comparison of performance. Essentially, each song is first modeled with a 16-mixture GMM. Next, a tag GMM with 8 mixtures is estimated by a mixture-of-hierarchies algorithm. The contribution of a song to a tag is weighted by the salience of the tag within the song. However, since Pandora provides binary labels, weights are either zero or one. In addition, [1] assumes a closed tag set; therefore, a background model built from songs missing the tag in question provides a comparison using a thresholded log-likelihood ratio (LLR) test. The authors caution the use of the phrase *anti-model* because Pandora does not list all the attributes relevant to a song, but only a few of the most salient. A better phrase is *background* or *universal model*.

4. EXPERIMENTS

4.1. Experimental Setup

A ten-fold cross validation is performed on a subset of the USPop2002 dataset [3]. Only songs which Pandora has tagged are considered. An artist filter ensures that no artist overlaps between the training and testing set for any particular fold. A total of 18 tags, shown in Table 1, were chosen because each occurs at least 500 times in the USPop2002 dataset; i.e., at least 50 examples in the test set for each evaluation fold. Tags from Pandora have either a temporal or global aspect, or both. A temporal aspect means the ordering of sounds is important for the tag in question; whereas a global characteristic deals with qualities that describe the overall sound of the music.

4.2. Performance Measures

Performance was measured in terms of equal error rate (EER) for annotation and mean average precision (MAP) for retrieval. EER is

[3] http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html

Tag/Attribute	Prop	Base
major key tonality (TG)	**34.54**	42.76
electric guitar riffs (T)	**40.78**	54.30
minor key tonality (TG)	**40.20**	48.87
acoustic & electric instrumentation (G)	**39.23**	48.24
acoustic rhythm guitars (TG)	**28.74**	33.25
vocal harmonies (TG)	**41.74**	47.77
extensive vamping (T)	**42.24**	45.56
focus on studio production (TG)	**39.24**	43.53
subtle use of vocal harmony (TG)	41.15	43.68
mild rhythmic syncopation (T)	46.64	49.16
a vocal-centric aesthetic (G)	43.65	44.69
a dynamic male vocalist (G)	**43.65**	44.69
hard rock roots (TG)	**19.13**	19.44
melodic songwriting (T)	**46.35**	46.67
electric rock instrumentation (G)	35.20	34.70
acoustic rhythm piano (TG)	37.58	**35.79**
repetitive melodic phrasing (T)	44.59	**41.92**

Table 1. Results for each tag in terms of EER for proposed (Prop) and baseline (Base). The letters in the parenthesis indicate whether the tag is temporally (T) and/or globally (G) based. Bold face indicates McNemar statistical significance.

the point at which the false acceptance rate and false rejection rate are equal. MAP gives the precision at each recalled document. For example, if the system returns the ordered results of [hit, miss, hit], then the MAP is [1,0.5,0.67] and the MAP at level $K = 2$ is 0.5. The McNemar's test is a non-parametric statistical test to determine whether two classifiers are significantly different and is shown to have a low Type I error [11].

4.3. Annotation Results

The proposed approach performs better for annotation in terms of equal error rate (EER) for 15 of the 18 tags, as shown in Table 1. However, by comparing the temporal characteristics, more interesting results are obtained. The table is organized by ranking the differences in EER through the *t*-statistic so that the proposed approach worked the best for the top tags when compared to the baseline approach. With the exception of *mixed acoustic and electric instrumentation* all tags appearing in the top half of the table contain temporal aspects, showing the ability to capture temporal information by the proposed approach. In addition, three of the four global tags appear in the bottom half of the table. One possible, and encouraging, reason as to why *mixed acoustic and electric instrumentation* performed well with the proposed approach is because some of the ASMs had a strong preference for songs that contained acoustic instruments and vice verse. The authors wish to exploit this property in the future by having a separate vocabulary set for each tag, similar to the parallel phone recognition and language modeling (PPRLM) approach to language recognition [12]. Only four tags failed the McNemar's test ($\alpha = 0.05$), which shows that for most tags, the best performer is not due to randomness in the training and testing sets.

Another interesting result is that the best performing tags under the proposed approach largely dealt with aspects of timbre, whether global or temporal. Examples include *electric rock instrumentation* (EER = 35.20) and *acoustic rhythm guitars* (EER = 28.74). The worst performing tags dealt with melody (*repetitive melodic phrasing* (EER = 44.59) and *melodic songwriting* (46.35)) and rhythm

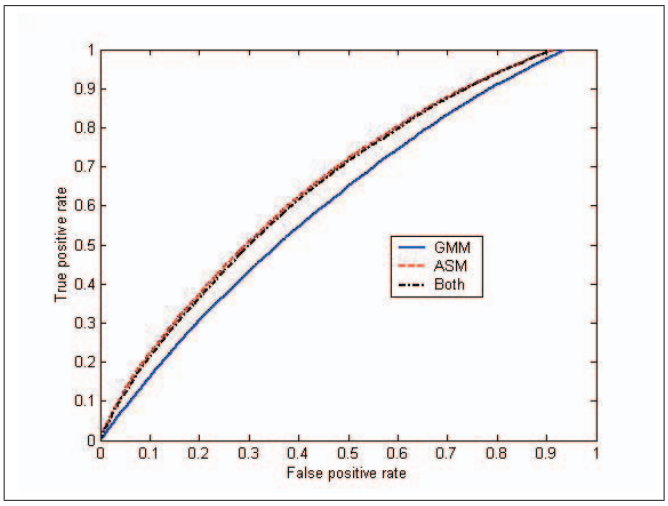

Fig. 2. Average ROC curves for the proposed approach (ASM), baseline approaches (GMM), and the combined approach (Both), which takes the best performing algorithm for each tag.

	Proposed	Baseline
Mean AP ($K = 5$)	0.4477	0.3313
Mean AP ($K = 10$)	0.4249	0.3321
Mean AP ($K = 15$)	0.409	0.3277

Table 2. Retrieval mean average precision at level *K*.

(*mild rhythmic syncopation* (EER = 46.64)). The authors conjecture the poor performance in melody attributes is due to the choice in features; i.e., MFCCs. Other features such as chromagrams [13] should lead to superior performance. In addition, rhythm is largely affected by the granularity of the segmentation algorithm. The maximum likelihood segmentation algorithm in [8] is designed for segmenting speech into subword units, such as phonemes. Therefore, the ASMs tended to be very short and on the order of note onset, sustain, and release [6]. In the future, the authors wish to investigate the use of beat tracking and onset detection for the segmentation step.

The average ROC curves, taken as the average across the 18 tags, are shown in Figure 2 and compared to the case where the best algorithm is chosen for each tag. The advantages of incorporating the temporal aspects in the proposed approach leads to better performance. Further, there is no noticeable advantage in using the baseline bag-of-frames approach for any tag. This demonstrates that global characteristics are modeled well in the proposed approach.

4.4. Retrieval Results

An important application of semantic tags is retrieval, where one searches by semantic tags and is returned a list of relevant songs. To measure retrieval for both the proposed and baseline approaches, the results for each tag are sorted by the score of the LLR test in the case of the baseline approach and SVM scores in the case of the proposed approach. Table 2 demonstrates that the proposed approach performs better in terms of MAP at the levels considered.

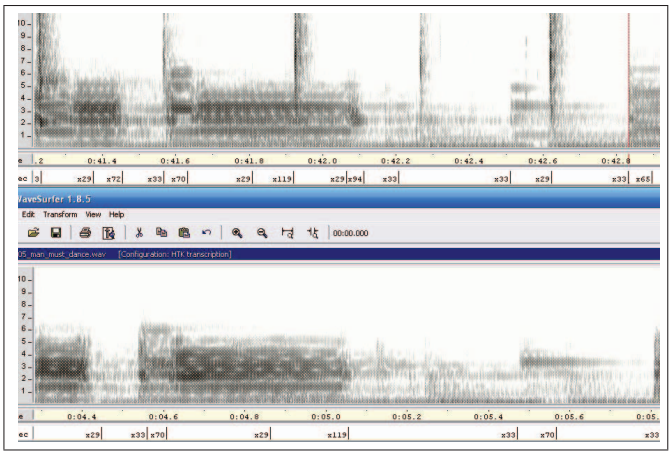

Fig. 3. Example of two similar melodies as a solo (lower waveform) and with polyphonic ornamentation (upper waveform).

4.5. Temporal Analysis

The previous two sections demonstrates the proposed approach performs better than the baseline bag-of-frames classifier for tags or musical attributes that contain temporal aspects. This section investigates how the proposed approach captures local timbral information and identifies temporal structure.

Figure 3 shows two parts of similar songs with similar electric guitar riffs. Specifically, the lower waveform is a solo with a single, clean electric guitar and the upper waveform has an additional high-hat and finger snap. The tokenization procedure finds an underlying *timbral melody* with the sequence (x33, x70, x29, x119, x33). However, the upper waveform also has *timbral embellishments*. Specifically, the finger snap at 41.9 seconds causes the insertion of the sequence (x29, x94) between x119 and x33. Further, the high-hat hit at 42.3 seconds causes the last x33 to repeat. By having a shared acoustic vocabulary, the proposed approach is able to identify how two musical pieces may have locally similar characteristics, even when additional instruments are added. Note that *timbral melody* refers to a sequential realization of timbral units and does not refer to the usual notion of *tonal melody*. However, the authors wish to investigate tonal melody by incorporating pitch-based features; i.e., chromagram.

Most importantly, unlike previous approaches to modeling the temporal structure [5], the approach presented here is able to model durations of sounds in a probabilistic sense, rather than a fixed length. The advantage is best seen in the finger snap at 41.9 seconds, which shortens the duration of x119 in the top waveform when compared to the solo part in the bottom waveform. Under a fixed length approach, one of two possibilities occur. Either the quantization is too course (i.e., a small codebook is used) and the finger snap would be smoothed over, or a large codebook is used and the entire sound would map to a different ASM symbol, which changes the underlying timbral melody. By incorporating temporal segmentation probabilistically, the underlying melody still exists, but an inserted ASM models the added finger snap.

5. CONCLUSION

An automatic tag annotation algorithm based on speech recognition technology demonstrates the importance of incorporating temporal information. The authors are especially encouraged by the fact that tags which contain both temporal and global aspects perform better with the proposed approach. In the future, the authors wish to build a vocabulary for each tag to incorporate global timbre characteristics as is done for the PPRLM approach to language recognition. In addition, the authors want to improve the initialization of the ASM models by incorporating beat-tracking software. These improvements may increase the ability for rhythm and melody attribute detection, possibly by incorporating research on MIDI melody retrieval.

6. REFERENCES

[1] D. Turnbull, L. Barrignton, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 467–476, 2008.

[2] J. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proc. IEEE*, vol. 88, no. 8, pp. 1279–1296, 2000.

[3] B. Whitman, *Learning the meaning of music*, Ph.D. thesis, Mass. Inst. Technol., 2005.

[4] M. Slaney, "Semantic-audio retrieval," in *Proceedings of ICASSP*, August 2002, pp. 4108–4111.

[5] M. Casey and M. Slaney, "The importance of sequences in musical similarity," in *Proceedings of ICASSP*, May 2006, pp. V5–V8.

[6] J. Reed and C.-H. Lee, "A study on music genre classification based on universal acoustic models," in *Proceedings of Intern. Symp. of Music Info. Retrieval (ISMIR)*, October 2006, pp. 89–94.

[7] J.-J. Aucouturier and F. Pachet, "Representing music genre: A state of the art," *J. New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.

[8] T. Svendsen and F. Soong, "On the automatic segmentation of speech signals," in *Proceedings of ICASSP*, April 1987, pp. 77–80.

[9] A. Meng and J. Shawe-Taylor, "An investigation of feature models for music genre classification using support vector classifier," in *Proceedings of Intern. Symp. of Music Info. Retrieval (ISMIR)*, September 2005, pp. 604–609.

[10] T. Joachims, *Making large-scale SVM learning practical*, pp. 10–119, Advances in Kernel Methods - Support Vector Learning. MIT-Press, 1999.

[11] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, pp. 1895–1923, 1998.

[12] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEE Trans. Speech, and Audio Process.*, vol. 4, no. 1, pp. 31–44, 1996.

[13] M. A. Bartsh and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *Proc, IEEE Workshop on Appl. of Signal Process. to Audio and Acoustics*, 2001, pp. 15–18.