IEICE TRANS. INF. & SYST., VOL.E91–D, NO.6 JUNE 2008

LETTER

Melody Track Selection Using Discriminative Language Model*

Xiao WU[†a)], **Ming LI**[†], *Nonmembers*, **Hongbin SUO**[†], *Student Member*, *and* **Yonghong YAN**[†], *Nonmember*

SUMMARY In this letter we focus on the task of selecting the melody track from a polyphonic MIDI file. Based on the intuition that music and language are similar in many aspects, we solve the selection problem by introducing an n-gram language model to learn the melody co-occurrence patterns in a statistical manner and determine the melodic degree of a given MIDI track. Furthermore, we propose the idea of using background model and posterior probability criteria to make modeling more discriminative. In the evaluation, the achieved 81.6% correct rate indicates the feasibility of our approach.

key words: *melody style, melody track selection, melody extraction*

1. Introduction

Determining the melody track of a given polyphonic MIDI file is a critical task for many music information retrieval applications (MIR), especially for a query-by-humming system [1]. Existing approaches are usually based on a set of empirically obtained features, such as average note duration and occupation rate, and rely on certain classifiers to select the melody tracks [2], [3].

In this letter, we propose a rather different method for the track selection task, which treats melody as natural language and models the probability of melody occurrences with n-gram grammar [4]. Since music and language share many similar characteristics, the n-gram model which is proved to be effective in language processing should also be effective in music processing. Although there are several previous research attempting to model musical melodies with N-gram for various purposes [5]–[7], this idea has been never introduced into the melody selection task. Furthermore, in contrast to these works which base their N-gram score on the likelihood probability, we propose the posterior probability-based classification by estimating the probability space of accompaniment, to make the modeling more discriminative. Apart from melody track selection, we believe that the proposed method can also be applied to other MIR tasks, such as melody style abstracting [8], [9] and composer recognition [10].

Manuscript received September 18, 2007.
Manuscript revised January 25, 2008.

[†]The authors are with the ThinkIT Speech Lab., Institute of Acoustics, Chinese Academy of Sciences, China.

*This work is partially supported by MOST (973 program · 2004CB318106), National Natural Science Foundation of China (10574140, 60535030), The National High Technology Research and Development Program of China (863 program, 2006AA010102 · 2006AA01Z195).

a) E-mail: xwu@hccl.ioa.ac.cn

DOI: 10.1093/ietisy/e91–d.6.1838

2. Statistical Melody Modeling

Music and language are similar in many aspects. Generally, both of them are constructed from basic units (note and word) using rules which are not absolute (musicology and linguistics), and both of them have semantic structures and sub-structures [11]. Therefore, approaches of melody processing should parallel with those of language processing. In melody selection task, the melody track is intuitively more melodic in perception than the accompaniment track. If such a character could be described with the probability space of statistical language model (LM), the problem can be solved. In this study, we attempt to model the melodic degree of a given MIDI track using n-gram grammar, and further find out which track is most probable to be the melody part.

2.1 The Alphabet

We choose the note difference rather than note itself to form the alphabet for three reasons. First, previous research has shown that pitch intervals are more musically meaningful than absolute pitch values [12]. Second, note-difference representation can significantly reduce the alphabet size and make the model estimation easier. Last, a differential n-gram actually gets $n + 1$ notes considered, which makes the melody model more expressive without increasing the order n.

Assume $n_i = (p_i, d_i)$ to be the ith note whose pitch p and duration d are counted in semitones and seconds respectively. We obtain the alphabet unit $u_i = (p_i^{'}, d_i^{'})$ by

$$p_i^{'} = integer(p_{i+1} - p_i), \text{and subject to } |p_i^{'}| \leq 8 \tag{1}$$

and

$$d_i^{'} = \begin{cases} -1 & \frac{d_{i+1}}{d_i} \leq 0.6 \\ 0 & \frac{d_{i+1}}{d_i} \in (0.6, 1.7] \\ +1 & \frac{d_{i+1}}{d_i} > 1.7 \end{cases} \tag{2}$$

Here pitch intervals are limited to 8 semitones because most melodies are stepwise and contain few leaps which are larger than perfect fifth. Besides, duration difference is modeled as one of "much shorter", "roughly the same" and "much longer". Applying above mapping, we get the a lexicon with 17×3 alphabets.

2.2 N-Gram Modeling

A statistical melody model is a probability distribution $P(seq)$ over all possible sequences $seq = (u_i, \ldots, u_n)$ [4]. This probability is always rewritten as

$$P(seq) = P\big((u_1, \ldots, u_n)\big) = \prod_{i=1}^{n} P(u_i|h_i) \tag{3}$$

With $n-1$ order Markov assumption of history h_i, $P(u_i|h_i)$ in above equation can be approximated by

$$P(u_i|h_i) \approx P_{ngram}(u_i|h_i) = P(u_i|u_{i-n+1}, \ldots, u_{i-1}) \tag{4}$$

Then the probability can be derived by counting n-gram occurrences in the training corpus.

To balance between the model capability and the model trainability, we choose n to be 3 and thus results in 132651 different trigrams. In addition, the Katz backoff smoothing is used to battle the sparseness. Furthermore, with the purpose of reducing the impact by singular trigrams, we set a probability floor to each trigram.

2.3 Introducing the Posterior Criteria

With N-gram melody model we are able to estimate the melodic likelihood of a given note sequence. This likelihood can be used in our application. However, posterior criterion is a better choice for classification related tasks. Maximizing the posterior probability not only attempts to make the model close to targets, but also make it far from non targets. Based on this idea, we train a background model to estimate the distribution of accompaniment. Therefore, instead of using $P(seq|M)$ as the melodic score, we develop a more discriminative one based on posterior probability criteria.

To deduce the new score we bring forward the "sequence picking problem". Imagine there are n sequences seq_i and their labels about melody or non-melody l_i are unknown. Besides, we have two models trained with large corpus, the target model M_{tar} and the background model M_{bk}. The melody score of seq_k is the probability of the event that seq_k is selected as the **only** melody among all sequences. With independence assumption, such a probability $P_{pick}(k)$ can be obtained:

$$P\big(l_k = M_{tar}, l_i = M_{bk} \text{ for all } i \neq k \big| seq_1, \ldots seq_n\big) \tag{5}$$

$$= \frac{P(M_{tar}|seq_k)}{P(M_{bk}|seq_k)} \cdot \prod_{i=1}^{n} P(M_{bk}|seq_i) \tag{6}$$

If we further assume that the prior probabilities $P(M_{tar})$ and $P(M_{bk})$ are identical, the equation above can be rewritten as

$$P_{pick}(k) = \frac{P(seq_k|M_{tar})}{P(seq_k|M_{bk})} \cdot \prod_{i=1}^{n} P(M_{bk}|seq_i) \tag{7}$$

Intuitively, $P_{pick}(k)$ indicates how probable a note sequence can survive in the "picking problem", and thus is supposed to be more discriminative. Further more, as the first part of Eq. (7) only relates to the likelihood of seq_k on M_{tar} and M_{bk} while the second part is identical for all sequences, we refine the final form of the discriminative score by omitting the second part of Eq. (7)

$$Score(seq) = \frac{P(seq|M_{tar})}{P(seq|M_{bk})} \tag{8}$$

The form of $Score(seq)$ is quite straightforward: a highly melodic note sequence should not only have good likelihood on the target model, but also should have poor likelihood on the background model. Using this equation, we get the melodic degree of each MIDI track, and finally select the one with the highest score to be the melody part.

3. Evaluations

3.1 Data Description

Generally, our MIDI collection can be divided into two parts, which are pop songs MIDI files and classic music MIDI files. All of them are crawled from internet. The components of the database are itemized as follows: 1000 runs For each pop MIDI file, the most clear melody track is manually labeled. Since $P3$ is ensured to have no crossing with other sets, we use it as evaluation set in our experiments. In addition, classic MIDI files are only used as accompaniment data, because they are in different genre with pop songs, thus may deteriorate the probability estimation if added for training the target model.

3.2 Results

Table 1 presents results with different training data and criteria. The statistics supports our analysis that posterior probability criteria outperforms the maximum likelihood criteria in track selection task. Furthermore, the results also demonstrate that increasing the size of training corpus is beneficial to the performance. The best result, which is the combination of posterior criteria and training with all available corpus (exclude test set, of course), achieves 81.6% correct rate. The above result proves the validity of the proposed idea, that is, language processing method can be applied to the melody selection task.

4. Conclusions

In this letter we apply n-gram model to the melody track

Table 1 Melody selection results.

Training Corpus		Criteria	
Target	Background	Likelihood	Posterior
$P1$	$P1$	24.2%	67.7%
$P1$	$P1 + C$	24.2%	67.9%
$P1 + P2$	$P1 + P2$	28.2%	80.2%
$P1 + P2$	$P1 + P2 + C$	28.2%	81.6%

selection task. Such a model is able to learn the melody occurrence patterns from the training corpus, and determine the melodic degree of a given note sequence. The proposed method gives convincing results, indicating the feasibility of the our idea which attempts to introduce language processing approaches to music information retrieval. Furthermore, the posterior criterion also shows a significant advantage over likelihood criterion in our experiments. In the future, we hope the proposed method could be applied to other MIR tasks, such as melody style abstracting and composer recognition.

References

[1] N. Kosugi, Y. Nishihara, T. Sakata, M. Yamamuro, and K. Kushima, "A practical query-by-humming system for a large music database," Proc. Eighth ACM International Conference on Multimedia, pp.333–342, 2000.

[2] M. Tang, C.Y. Lap, and B. Kao, "Selection of melody lines for music databases," Proc. Computer Software and Applications Conference, 2006.

[3] D. Rizo, P. de Leon, C. Perez-Sancho, A. Pertusa, and J. Inesta, "A pattern recognition approach for melody track selection in MIDI files," Proc. International Conference on Music Information Retrieval, 2006.

[4] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go fromhere?," Proc. IEEE, vol.88, no.8, pp.1270–1278, 2000.

[5] S. Doraisamy, "Polyphonic music retrieval: The N-gram approach," ACM SIGIR Forum, vol.39, no.1, p.58, 2005.

[6] K. Ishida, T. Kitahara, and M. Takeda, "ism: Improvisation supporting system based on melody correction," Proc. 2004 Conference on New Interfaces for Musical Expression, pp.177–180, 2004.

[7] J. Downie, "Music retrieval as text retrieval (poster abstract): Simple yet effective," Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.297–298, 1999.

[8] M.K. Shan and F.F. Kuo, "Music style mining and classification by melody," IEICE Trans. Inf. & Syst., vol.E86-D, no.3, pp.655–659, March 2003.

[9] D. Cope, Computers and musical style, AR Editions, Inc. Madison, WI, USA, 1991.

[10] E. Pollastri and G. Simoncelli, "Classification of melodies by composer with hidden Markov models," Web Delivering of Music, 2001. Proceedings. First International Conference on, pp.88–95, 2001.

[11] R. Bod, "A unified model of structural organization in language and music," Journal of Artificial Intelligence Research, vol.17, pp.289–308, 2002.

[12] C. Krumhansl, Cognitive Foundations of Musical Pitch, Oxford Univ. Press, 1990.