PAPER

Robust and Fast Phonetic String Matching Method for Lyric Searching Based on Acoustic Distance

Xin XU[†a)], ***Member and*** **Tsuneo KATO**[†], ***Senior Member***

SUMMARY This paper proposes a robust and fast lyric search method for music information retrieval (MIR). The effectiveness of lyric search systems based on full-text retrieval engines or web search engines is highly compromised when the queries of lyric phrases contain incorrect parts due to mishearing. To improve the robustness of the system, the authors introduce acoustic distance, which is computed based on a confusion matrix of an automatic speech recognition experiment, into Dynamic-Programming (DP)-based phonetic string matching to identify the songs that the misheard lyric phrases refer to. An evaluation experiment verified that the search accuracy is increased by 4.4% compared with the conventional method. Furthermore, in this paper a two-pass search algorithm is proposed to realize real-time execution. The algorithm pre-selects the probable candidates using a rapid index-based search in the first pass and executes a DP-based search process with an adaptive termination strategy in the second pass. Experimental results show that the proposed search method reduced processing time by more than 86.2% compared with the conventional methods for the same search accuracy.

key words: *lyric search, phonetic confusion matrix, two-pass search, dynamic programming*

1. Introduction

Current commercial music information retrieval (MIR) systems accept queries in a range of forms by text, humming, singing, and acoustic music signals. Among these, text queries of lyric phrases are commonly used [1]. As many MIR systems apply full text search engines to lyric search, it has been widely regarded that the issue of lyric search has been solved by state-of-the-art text retrieval techniques. However, there is a problem. The investigations on real world queries, as conducted in this paper, suggested that users are likely to input incorrect lyric phrases into MIR systems, resulting in a failure. The investigation found that incorrect queries that replace a word with another word of a similar pronunciation can occur at rates as high as 19.3% of total collected lyric queries. These incorrect lyric queries are due to mishearing or the unreliability of human memory, as users only memorize the lyric phrases when they enjoy hearing a part of a song and usually do not use the aid of its lyric sheet. The investigation also verified that major commercial web search engines implemented with fuzzy matching algorithms were not helpful. A search method is expected to be able to identify the lyric containing the part that is most acoustically similar to the query. Phonetic string matching, which is used in such applications as name retrieval [2], was considered to be an appropriate method to solve this problem of low search accuracy.

However, another important requirement for lyric search is that it must satisfy a real-time response. As the search algorithm of phonetic string matching methods is based on exhaustive dynamic programming (DP), the computational complexity results are in the order of $m * n * I_t$ per query. Here m is the length of the query, n is the average length of a lyric and I_t is the number of lyrics to search. Since commercial MIR systems usually provide hundreds of thousands of lyrics, the computational complexity is too high to realize a real-time search.

In order to solve these two problems peculiar to lyric search, in this paper the authors propose a novel method to make lyric search simultaneously robust and fast.

First, in order to improve lyric search accuracy, the proposed method applied "acoustic distance" to measure the acoustic confusability between phonetic strings due to mishearing. The acoustic distance values are derived from a DP-based phonetic string matching calculation using a phoneme confusion matrix to quantize acoustic similarity between phonemes. The values in the confusion matrix are obtained by an automatic speech recognition (ASR) experiment. Although the confusion matrix should be based on singing voice, a huge amount of singing data is not available. In this paper, telephone speech data of Japanese phonetically balanced sentences, which are more easily-obtainable data in the field of speech recognition, were used as the training data for the acoustic models of ASR. This is inspired by concepts in Spoken Document Retrieval (SDR) and Spoken Utterance Retrieval (SUR) [3], [4].

To solve the second problem in real-time search, a two-pass search algorithm is applied in the proposed lyric search method. It uses a fast inverted-index-based search in the first pass and DP-based search with an adaptive termination strategy in the second pass. In the first pass, the proposed method pre-selects the probable lyric candidates by means of a rapid approximate search based on the accumulation of pre-computed and indexed partial acoustic distances. Then, in the second pass, the lyric candidates are sorted by the approximate acoustic distances and evenly divided into groups. The exhaustive DP matching between the query and the lyrics is carried out group by group. During the DP matching, a cut-off function for the adaptive termination is calculated by the DP distances to make the matching process more efficient. Once the function value exceeds a predeter-

Manuscript received November 18, 2013.
Manuscript revised May 7, 2014.

[†]The authors are with the KDDI R&D Laboratories Inc., Fujimino-shi, 356–8502 Japan.
a) E-mail: sh-jo@kddilabs.jp

DOI: 10.1587/transinf.2013EDP7418

mined threshold for some group, which means the correct lyric has been found, the search is terminated. The experimental results show that the processing time is greatly reduced by using the proposed two-pass search strategy, without loss of search accuracy. The group size that contributed to the best performance was proved to be 100 lyrics, which is about one fifteenth of the candidates.

The remainder of this paper is organized as follows: Section 2 presents related works on lyric search. The analysis of mistaken queries is described in Sect. 3. Section 4 introduces how to calculate acoustic distance during phonetic string matching. All of the search processes of the proposed lyric search method are described in detail in Sect. 5. In Sect. 6, the experiments are carried out to evaluate the proposed method in terms of search accuracy and processing time. The paper is summarized in Sect. 7.

2. Related Works

Several related studies attempted to use phonetic string matching methods to solve the search problem caused by misheard lyrics. They were verified to be more robust than the text retrieval methods. Ring and Uitenbogerd [5] tried to find the correct lyric by minimizing the edit distances between phoneme strings of queries and the lyrics. However, edit distance does not present the degree of confusability between phonemes. To model the similarity of misheard lyrics to their correct versions statistically, Hussein [6] introduced a probabilistic model of mishearing that is trained using examples of actual misheard lyrics from a user-submitted misheard lyrics website "kissthisguy" [7], and developed a phoneme similarity scoring matrix based on the model. The performance of this method depends on the size of the training database. As described in [6], a total number of 20788 pairs of the misheard lyrics and the correct lyrics are used. However, such a big database like "kissthisguy" is not available in other languages. For example, in order to search lyircs in Japanese, it is impractical to collect sufficient misheard lyrics to build a practical probabilistic model.

On the other hand, in order to reduce the processing time, conventional high-speed DP matching processors use index or tree-structured data to pre-select the hypothetical candidates [8], [9]. As an example, [8] used a suffix array as the data structure and applied phoneme-based DP matching to detect keywords quickly from a very large speech database. In order to avoid an exponential increase in the processing time caused by increasing keyword length, it divided the original keyword into short sub-keywords. Then, it searched the sub-keywords on the suffix array by DP matching. If the DP distance between a sub-keyword and a path of the suffix array is not more than a predetermined threshold value, these pathes remained as the candidates of search results. By repeating the DP matching process between the original keyword and the candidates, the final result is detected. As well as other high-speed DP methods, the predetermined threshold for sub-keywords is proportional to the length of the queries.

However, lyric search has a distinctive characteristic: it is too difficult to determine an absolute threshold to decide whether a lyric is the exactly correct one for the incorrect query or not, since it is related to the individual variations of mishearing. Therefore, the previous studies on lyric search used the common criterion of looking up the entire lyric search space and estimating the lyric at minimum distance from the query to be the user's target. Based on the investigation of real world queries in Sect. 3.3 of this paper, the DP distances between the queries and the correct lyrics have no statistical relationship with the lengths of the queries. The conventional high-speed DP processors are not able to keep high search accuracy for the lyric search case.

3. Analysis of Real World Lyric Queries

Several investigations were carried out on collected real world queries. The statistical features of queries and some issues peculiar to lyric search are presented in this section.

3.1 Statistical Features of Real World Queries

To analyze the queries of lyric phrases for MIR in the real world, the authors investigated major Japanese question & answer community websites, "okwave" [10] and "oshiete goo" [11]. It was found that many questions used lyric phrases to request the names of songs and singers. As 1140 queries of lyric phrases asked by various questioners were collected, the authors compared each query with its corresponding lyric to categorize whether lyric phrases in the query are correct or not (correct query or incorrect query) and how they were mistaken. The lyrics and queries are written in Japanese or English, or a mixture of both.

Figure 1 shows the distribution of incorrect queries in the different types and correct queries within the collected data. The incorrect queries, which make up around 79%, are classified into the following types:

- Confusion of notations: Chinese characters in the queries are substituted for reading symbols (kana), and

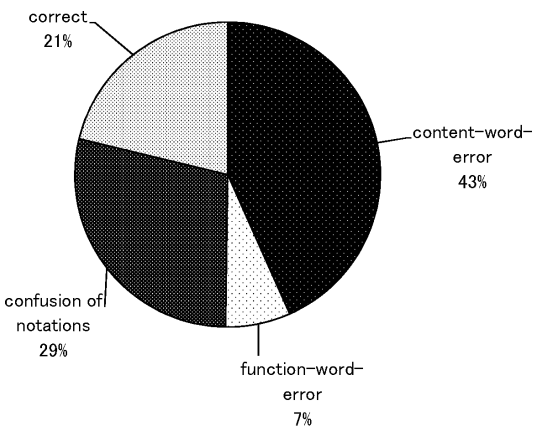

Fig. 1 The distribution of mistaken queries in the different types and correct queries within the collected queries.