Mobile Melody Recognition System with Voice-Only User Interface

Timo Sorsa
Nokia Research Center
P.O.Box 407, FIN-00180 Helsinki, Nokia Group,
Finland
timo.sorsa@nokia.com

Katriina Halonen
Nokia Research Center
P.O.Box 100, FIN-33720 Tampere, Nokia Group,
Finland
katriina.halonen@nokia.com

ABSTRACT

A melody recognition system with a voice-only user interface is presented in this paper. By integrating speech recognition and melody recognition technology we have built an end-to-end melody retrieval system that allows a users to do voice controlled melodic queries and melody generation using a dial-in service with a mobile phone.

1. INTRODUCTION

As a result of new, widely accepted music storage and transfer formats, the amount of musical information accessible by users has risen rapidly during the past years. Content-based, automated tools for data management and retrieval have become an attractive goal from the research as well as from the commercial point of view. Indications of this increasing interest are, for example, standardization efforts such as MPEG-7 [1].

One approach for content-based data management in music domain is query-by-humming type of solutions. A number of query-by-humming applications have been presented during the recent years (see for example [2]-[5]). All of the proposed applications are web or PC based implementations.

Recently, voice markup languages have appeared that enable developers to build web-based speech applications [6]. Users control the dialogue with speech commands, which are automatically recognized at the system end, and speech synthesis is used to provide the user with instructions and prompts. Speech telephony technology is used to develop applications into dial-in end-to-end services, accessible by a mobile or fixed phone.

In this paper we consider one possible use of query-by-humming technology in commercial mobile applications.

The paper is organized as follows. First we present an overview of the implemented melody recognition service. Then, in Section 3 we report some user evaluation results. We end with Section 4 discussing some conclusions and considerations.

2. SYSTEM OVERVIEW

The implemented query-by-humming system is intended for mobile melody retrieval and generation based on user given acoustic inputs.

The system is controlled with voice commands given by the user. A voice user interface (UI) is well suited for this type of applications since the melody input is most often given by voice as well. The voice UI also enables the usage of the service in situations where a keypad is hard to use.

2.1 System architecture

A general overview of the mobile melody retrieval system is presented in Figure 1. The main functional blocks are the voice-only UI, melody transcription module and the database engine.

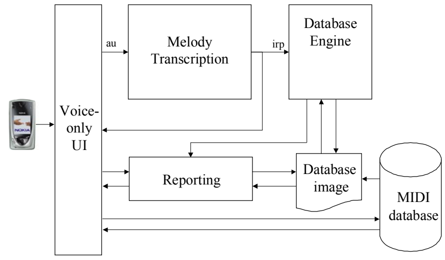

Figure 1. General overview of the melody recognition system with a voice-only user interface.

2.1.1 Melody Transcription

The melody transcription engine used in this system is the same as reported in [7]. Some parameters, such as analysis thresholds, have been tuned in order to get better average performance in the mobile context taking the GSM and channel coding into account.

Constant amplitude thresholds are used for detecting note boundaries. An autocorrelation pitch tracker [8] with center clipping [9] is used for estimating the pitch within each 20 ms analysis frame. On a note level the pitch is estimated as the median of the pitch values of the frames within a note.

2.1.2 Database Engine

The database engine used in the system is based on an engine made by Lemström and Perttu [10]. A fast bit-parallel dynamic programming algorithm by Myers is used for approximate string matching [11].

2.1.3 Voice-Only User Interface

The voice UI was implemented using a standard voice markup language VoiceXML[6]. Speech recognition is used for user input, and speech synthesis for system output. On calling the service, the user hears short instructions for selecting between melody search and generation modes. In both cases, the user is prompted to produce a melody sample by humming, whistling or playing. The sample is played back to the user, and the user can record a new sample or send the sample for transcription.

In melody search, five best matches are retrieved from the database and played back to the user on request. The user can browse the results by going forward or backward in the list.

In melody generation, the transcription output is transformed into AU format and played back. The user can then save or reject the generated melody.

3. SYSTEM PERFORMANCE

The implemented system was evaluated with user tests. Nine users tested the system by calling the dial-in service. The user feedback was collected with a questionnaire and discussions.

The evaluation tests indicated positive user feedback and acceptance for the concept. The users felt the user interface and the back-end audio analysis as well as the database engine to be well integrated into an entertaining and easy-to-use service. The concept proved to be well suited for mobile use and the users considered the service to be a good additional service for mobile music distribution and generation.

3.1 Transcription and Retrieval Accuracy

The melody transcription and retrieval perform essentially in a similar manner as with the PC application reported in [7], although the mobile context introduces some additional errors in the transcription, mostly in the detection of note boundaries.

The users rated the melody transcription accuracy a bit below 3 on the scale 1-5 (5 = the best). Even though this rating is not tied to any reference it indicates that the transcription should be further tuned and alternative ways for improvement should be considered.

The small decrease in transcription accuracy affects the retrieval somewhat. However, the retrieval performance is still roughly at the same level with the PC implementation. That is, the recall percentage is at the level of 65-85% like reported in [7], depending on the chosen inputs.

3.2 User Interface

Voice commands are a quick and fairly natural way to interact. The difficulties of voice interfaces come from the serial nature of speech. Only one instruction can be offered at a time, compared to a graphical user interface (GUI) with several active areas and buttons.

Most users considered the instructions given by the UI clear and informative but also thought there were too many of them. During the database search, clear indication of the system status was hoped.

4. CONCLUSIONS AND FUTURE WORK

The system reported in this paper is an end-to-end system that offers an easy-to-use music retrieval and generation application for mobile phone users. The evaluation tests indicate the voice-only UI to be a good choice for this kind of applications. The user tests also indicate the concept to be well accepted by the users.

The presented system allows the user to generate or search for a melody. These two tasks set different requirements for the transcription. For search, it is often beneficial to clean the transcription from small inaccuracies present in the input signals. This, however, often makes the transcription sound clumsy in comparison to the original melody. It would be beneficial to tune the transcription process differently for music retrieval and generation tasks.

One major advantage of the presented system over PC-based solutions is the well-controlled signal path. The properties of the mobile terminal and network are well known and the recording and transcription parameters can be tuned in the design phase.

Even with well-controlled signal path two major weaknesses remain in query-by-humming applications: the optimal parameter settings differ from user to user and in a general case the input is relatively noisy or even erroneous.

The retrieval accuracy could be significantly improved if the transcription parameters were tuned separately for each user. Some sort of learning scheme could be used in which the application would learn to adjust parameters by the user's input.

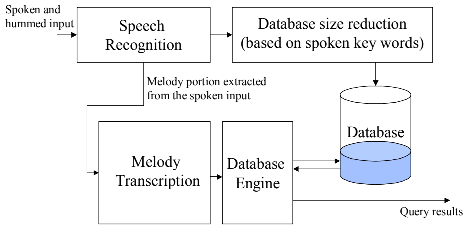
Spoken and hummed input
Speech Recognition
Database size reduction (based on spoken key words)
Melody portion extracted from the spoken input
Database
Melody Transcription
Database Engine
Query results

Figure 2. A block diagram of a speech enhanced melody recognition system.

A way for increasing the retrieval accuracy of a noisy signal is to enhance the retrieval process with speech (Figure 2). Spoken key words (e.g., music style or artist name) can be used to limit the database search to relevant parts of the database.

Some of the technologies described in this paper are subject to pending patents.

5. REFERENCES

[1] MPEG Requirements Group, "Mpeg-7 Overview", Doc. ISO/MPEG N4674, MPEG Jeju Meeting, March 2002, edited by Martinez, J. M.

[2] McNab, R.J., Smith, L.A., Bainbridge, D., and Witten, I.H., "The New Zealand digital library MELody inDEX", *D-Lib Magazine, May 1997.*

[3] Ghias, A., Logan, J., Chamberlin, D., and Smith, B.C., "Query by humming – musical information retrieval in an audio database", ACM Multimedia '95, San Francisco, USA, 1995.

[4] http://name-this-tune.com

[5] Chai, Wei, "Melody Retrieval on the Web", M.Sc Thesis, Massachusetts Institute of Technology, August 2001.

[6] http://www.w3.org/Voice/

[7] Sorsa, T., "Development and Evaluation of a Melody Recognition System", M.Sc Thesis, Helsinki University of Technology, October 2000.

[8] Rabiner, L.R., Cheng, M.J., Rosenberg, A.E. and McGonegal C.A., "A comparative performance study of several pitch detection algorithms", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, no. 5, pp. 399-418, July 1976.

[9] Sondhi, M.M., "New methods of pitch extraction", IEEE Trans. Audio Electroacoust. Special issue on Speech Communication and Processing – Part II, AU-16, pp.262-266, June 1968.

[10] Lemström, K. and Perttu, S., "SEMEX - An efficient music retrieval prototype", Proceedings of Music IR 2000, Plymouth, MA, USA, Oct. 2000.

[11] Myers, G., "A fast bit-vector algorithm for approximate string matching based on dynamic programming", Proceedings of the 9th Annual Symposium on Combinatorial Pattern Matching, Piscataway, USA, 1998.