PHONEME RECOGNITION IN POPULAR MUSIC

Matthias Gruhne, Konstantin Schmidt and Christian Dittmar
Fraunhofer IDMT
Langewiesener Str. 22
98693 Ilmenau
{ghe,schmkn,dmr}@idmt.fraunhofer.de

ABSTRACT

Automatic lyrics synchronization for karaoke applications is a major challenge in the field of music information retrieval. An important pre-requisite in order to precisely synchronize the music and corresponding text is the detection of single phonemes in the vocal part of polyphonic music. This paper describes a system, which detects the phonemes based on a state-of-the-art audio information retrieval system with harmonics extraction and synthesizing as pre-processing method. The extraction algorithm is based on common speech recognition low-level features, such as MFCC and LPC. In order to distinguish phonemes, three different classification techniques (SVM, GMM and MLP) have been used and their results are depicted in the paper.

1 INTRODUCTION

During the last years, users of personal computers have aquired huge amounts of digital music due to efficient audio compression techniques. One of the most fascinating leisure activities, especially in asian countries is the karaoke application. It is however, difficult and time-consuming to create karaoke files at the moment. Since a manual tagging of the song is required, in order to convey the information, when a certain word is played to display and mark it on the screen. It might be commercially interesting to create karaoke applications for the home user, that is able to create karaoke files automatically from the personal music collection and corresponding texts from the internet. A basis of such system is the automatic synchronisation between music and corresponding lyrics. There have been scientific papers in the past, which described such methods [5] [1], but there is still room to improve the results significantly. The authors think, that phoneme recognition is a basis for synchronizing lyrics and corresponting text, since the most prominent phonemes give a good indication of the time label within a currently sung word. Thus, the system presented in this paper describes a novel approach for automatically detecting pho-nemes in music. Section 2 describes the setup of the overall system, illustrates the state of the art and shows own extensions. Subsequently, the test setup is described and results are depicted. The paper finishes with conclusions and an explanation of the future work.

2 PROPOSED SYSTEM

The proposed system uses techniques of a common state of the art information retrieval system, but makes additionally use by a harmonics extraction algorithm at the beginning. The overall design has been inspired by the method used by Fujihara [3], who described a singer identification method based on a music information retrieval system with a previous harmonics extraction. Since singer identification addresses a similar task, the results of detecting phonemes from polyphonic music have been expected to increase as well. The proposed method starts with a fundamental frequency estimation as described in [2] in order to improve the harmonic extraction results. Dressler uses a Multi-Resolution Fast Fourier Transform (MRFFT) to compute the spectra in different time-frequency resolutions efficiently. In order to discriminate frequencies, an instantaneous frequency (IF) is estimated from successive phase spectra. Due to the fact, that sinusoidal components of the audio signal contain the most relevant melody information, the harmonics are identified using a psychoacoustic model under distinction of spectral features. After estimating the fundamental frequency, the partials are retrieved from a spectrogram. The final sinusoidal resynthesizing of the audio signal is determined by transforming the spectrum into the time domain by using an Inverse Discrete Fourier Transform (IDFT). Only the previously calculated harmonic components of the spectrum are considered for a resynthesis. After constructing the signal, common speech recognition features had been extracted and assembled. The applied features are Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), Perceptual Linear Prediction (PLP) and Warped Linear Prediction Coefficients (WarpedLPC) [4]. Before the actual classification the dimensions are reduced by using a linear discriminant analysis. The resulting feature matrix is classified with common classifier techniques, namely Gaussian Mixture Models (GMM), Support Vector Machines (SVM) and Multilayer Perceptron (MLP).

Classifier	Pr.	Rc.	CCI
Results **with** harmonics analysis			
MLP	0.335	0.338	54.42 %
SVM	0.333	0.340	57.68 %
GMM	0.309	0.300	49.13 %
Results **without** harmonics analysis			
MLP	0.186	0.187	34.16 %
SVM	0.167	0.184	28.34 %
GMM	0.178	0.191	31.45 %

Table 1. Accuracy of GMM, SVM and MLP with and without harmonics analysis.

3 EVALUATION

In order to estimate the phonemes in the vocal parts of the popular music, an extensive database has been established. Overall, 2244 phonemes have been manually labeled from vocal parts of popular music. Since the genre of the most songs in karaoke applications concentrate on popular music, the testset in the proposed system uses only audio items in the genre Pop from the last fifty years. Due to the fact, that the vocal part of music contains a large amount of residual "distortion" besides the plain voice, this paper concentrates only on extracting the 15 most discriminative voiced phonemes. These phonemes have been labeled from 37 popular music songs, 21 songs performed by male singers and 16 songs performed by female singers. The items have been split into training (51 percent) and test set (49 percent). The occurrence ratio of the phonemes between test und training set is equal.

The considered phonemes can be divided into three different classes (approximants, nasals and vowels). Some of the usually distinguished vowels have been combined, because they refer to the same character and are sometimes so similar, that they even confound by non-native speakers. Concerning the parameters of the feature and harmonic extraction algorithm, besides the in Dressler[2] described fundamental frequency analysis and the in Fujihara[3] described synthesis, eight LPC features have been used, because they turned out to deliver most reliable results. Furthermore eight WLPC coefficients and nine PLP coefficients have been used. The frequency of MFCC features ranged between 50 Hz and 5 kHz, 13 coefficients were utilized.

4 TEST RESULTS

The test results, that have been received during the tests and which are described in this section more in detail. Table 1 shows precision (Pr), recall(Rc) and percentage of correct classified instances (CCI) of the tested classifiers. Table1 is divided in two main blocks. The upper block shows the performance of the system by performing a previous harmonics analysis and the lower block depicts the results without harmonics analysis, showing the results of the GMM, SVM and MLP classifiers.

The best result could be obtained by performing a harmonics analysis and using an SVM classifier. With this configuration, the proposed system reached an average precision of 0.33 and an average recall of 0.34 (58% CCI). By not executing a previous harmonics analysis, the MLP classifier performed better than SVM and GMM (34% CCI). The difference between the results with resynthesis and without are significant, especially by considering the fact, that the fundamental frequency analysis reaches at the moment only an accuracy of about 70%.

5 CONCLUSIONS AND FUTURE WORK

This paper described a novel approach for detecting phonemes in vocal parts of polyphonic music. The described method incorporates state of the art techniques in feature extraction and classification used in music and speech recognition and performs melody detection algorithms for reducing influences from accompanying sounds. The results show, that the best classifier with an overall performance (with harmony extraction) of 58% (CCI). The results with resynthesis by fundamental frequency and without are not significant, but due to the accuracy of 70 percent of the fundamental frequency estimation, the results could improve enormously with an improvement of preprocessing algorithms. In order to improve the performance of the system, it is planned to extend the test set by manually labelling more songs as well as using additional features and to test different classifier.

6 REFERENCES

[1] K. Chen, S. Gao, Y. Zhu, and Q. Sun. Popular song and lyrics synchronisation and it's application to music. In *Proceedings of the 13th Annual Conference on Multimedia Computing and Networking (MMCN)*, 2006.

[2] K. Dressler. Sinusoidal extraction using an efficient implementation of a multi-resolution fft. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2006.

[3] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Singer identification based on accompaniment sound reduction and reliable frame selection. In *Proceedings of the 6th International Symposium on Music Information Retrieval (ISMIR)*, pages 329–336, 2005.

[4] A. Harma and U. Laine. A comparison of warped and conventional linear predictive coding. In *IEEE Transaction on Acoustics, Speech and Signal Processing Vol. 9 No. 5*, pages 579 – 588, 2001.

[5] Y. Wang, M. Y. Kan, T. L. Nwe, A. Shenoy, and Y. Yin. Lyrically: Automatic synchronization of acoustic musical signals and textual lyrics. In *Proceedings of the 12th annual ACM international conference on Multimedia*, 2004.