Phoneme Detection for Lyrics Synchronization

Matthias Gruhne, Christian Dittmar

Fraunhofer Institut für Digitale Medientechnologie, 98693 Ilmenau, Deutschland, Email: {ghe,dmr}@idmt.fraunhofer.de

Introduction

Automated detection of phonemes in polyphonic music is an important prerequisite for synchronizing music and corresponding lyrics. This paper describes a novel approach of automatic phoneme estimation within digitized music pieces. Since there are already a number of publications (e.g. [1], [5]), aiming at distinguishing singing passages and non-singing passages in music, this paper only concentrates on detecting voiced phonemes in singing passages of music. In a first step, the leading melody of a segment is recognized based on an efficient implementation of a Multiresolution Fast Fourier Transform (MRFFT) [2] and subsequently the harmonics are extracted and a subset is selected for further processing. Thereafter, the harmonics are resynthesized to an audible signal. Finally, different common feature extraction methods are applied and the performance of the classification algorithms Gaussian Mixture Models (GMM), Support Vector Machines (SVM) and Multilayer Perceptron (MLP) is compared in order to detect the phonemes based on a manually established music phoneme database. The evaluation section shows the results by using the previously described classifiers and by enabling/disabling a harmonics extraction as a preprocessing step. Furthermore the phoneme classification rate dependent on the fundamental frequency is given.

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 and the feature space is decorrelated by using a linear discriminant analysis. The resulting feature matrix is classified with common classifier techniques, GMM, SVM and MLP.

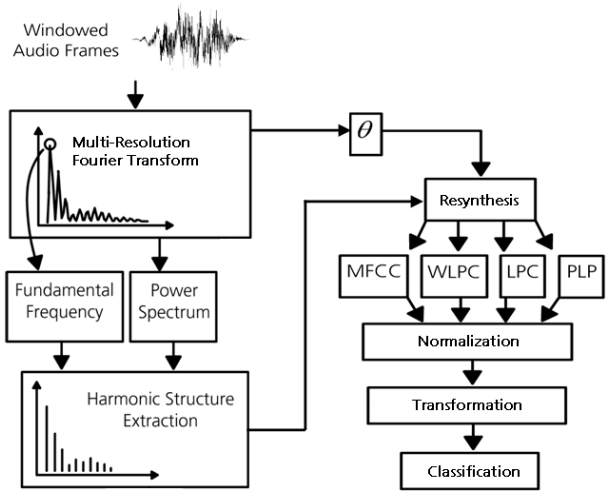

Figure 1: This figure depicts the overall setup of the phoneme recognition system.

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 test set 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).

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.

The occurrence ratio of the phonemes between test and training set is equal. Concerning the parameters of the feature and harmonic extraction algorithm, besides the in Dressler[**?**] 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.

Test Results

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, reaching 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%. Figure 2 depicts the results based on the fundamental frequency. As one can see, the recognition rate of the phonemes decreases at fundamental frequencies above 300 Hz significantly. This is, because the LPC filter coefficients are not fully independent from the fundamental frequency. A higher fundamental frequency results in convergence of the LPC transfer function to the spectral lines of the harmonics and the recognition rate decreases.

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.

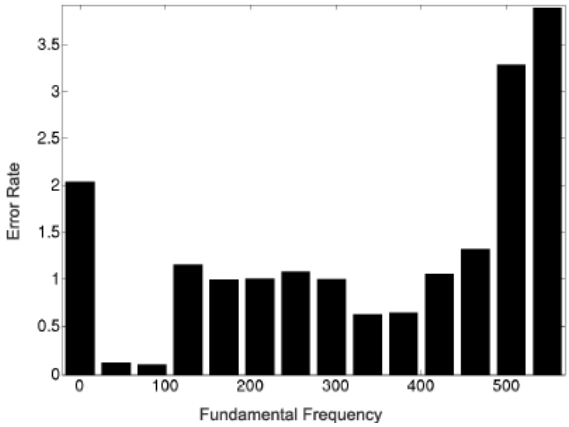

Figure 2: This figure depicts dependency of the phoneme recognition based on the fundamental frequency.

Acknowledgements

This work has been partly supported by the PHAROS and the DIVAS project, funded under the EC IST 6th Framework Program as well as by grant No. 01MQ07017 of the German THESEUS program.

References

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

[3] Fujihara H. et al.: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] Harma A. et al.:A comparison of warped and conventional linear predictive coding. In IEEE Transaction on Acoustics, Speech and Signal Processing, 9 (2001), 579 – 588, 2001.

[5] Wang Y. et al.:Automatic synchronization of acoustic musical signals and textual lyrics. In Proceedings of the 12th annual ACM international conference on Multimedia, 2004.