Phoneme Detection in Popular Music

Fraunhofer Institut Digitale Medientechnologie

Matthias Gruhne, Konstantin Schmidt and Christian Dittmar

Fraunhofer Institute for Digital Media Technology IDMT, Germany

Introduction

Motivation

- Phoneme detection in polyphonic music is an important prerequisite for lyrics synchronization for karaoke applications or browsing in music catalogues.
- Phoneme detection might be furthermore used for the singing recognition in the polyphonic music.

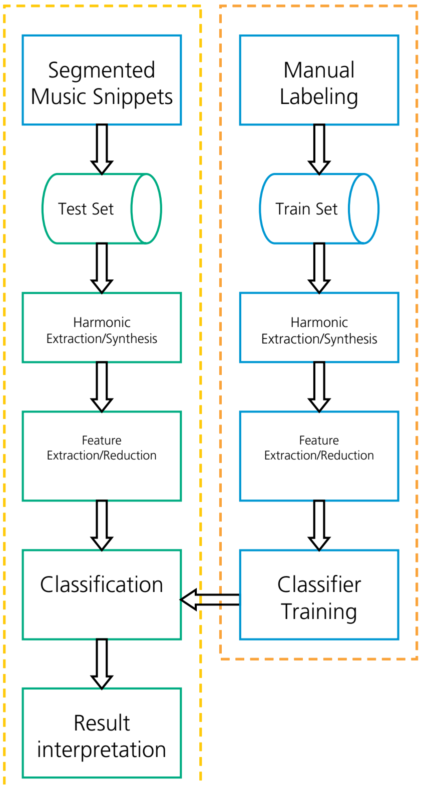
Challenges

- Finding voiced phonemes in popular music is an ambitious task due to interference with other instruments playing simultaneously.
- Vocal/Nonvocal detection is disregarded in this paper and borrowed from available technology.

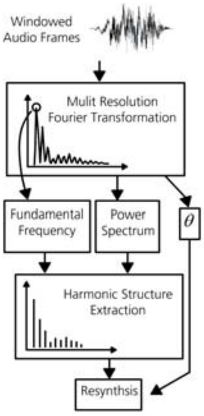
Statement of the problem

- The goal is to automatically identify the sung phonemes of the previously segmented music snippets.

Outline


Segmented Music Snippets
Test Set
Harmonic Extraction/Synthesis
Feature Extraction/Reduction
Classification
Result interpretation
Manual Labeling
Train Set
Harmonic Extraction/Synthesis
Feature Extraction/Reduction
Classifier Training

Harmonics Analysis


Windowed Audio Frames
Mulit Resolution Fourier Transformation
Fundamental Frequency
Power Spectrum
θ
Harmonic Structure Extraction
Resynthsis

- Multiresolution Short-Time Fourier Transformation has been performed in order to receive a high resolution spectral representation in short processing time.
- The fundamental frequencies are detected based on [2] and the harmonic structure is extracted thereafter.
- [3] Synthesized 20 partials. Performed listening tests showed unsuitable performance for male singers. Therefore the number of partials is based on the fundamental frequency (if the frequency is lower, more partials extracted).

Feature Extraction and Classification

Feature Extraction

- A number of different common Low-Level features have been extracted.
- Used features are MFCC, LPC, PLP and Warped LPC.
- All feature vectors have been appended in order to receive one large vector.
- In order to reduce the dimensions and consecutively decorelating the features, the Linear Discriminant Analysis (LDA) has been performed.

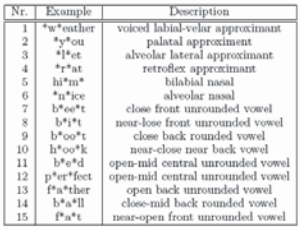
Feature Classification

- All features have been performed by using different classifiers.
- As classifiers, Gaussian Mixture Models (GMM), Support Vector Machines (SVM) and Multilayer Perceptron (MLP) have been tested.

Database

Labeled Data

- 30 second-snippets from 37 songs have been automatically extracted.
- All phonemes from this pieces have been manually labeled.
- The considered genre was popular music.
- 21 songs have been performed by male singers.
- 16 songs have been performed by female singers.
- Altogether 2244 phonemes have been manually labeled and can be used by the system.
- Only 15 voiced phonemes have been distinguished, because they are significantly separable for lyrics synchronization.



Nr.	Example	Description
1	*w*eather	voiced labial-velar approximant
2	*y*ou	palatal approximent
3	*l*et	alveolar lateral approximant
4	*r*at	retroflex approximant
5	hi*m*	bilabial nasal
6	*n*ice	alveolar nasal
7	b*ee*t	close front unrounded vowel
8	b*i*t	near-lose front unrounded vowel
9	b*oo*t	close back rounded vowel
10	h*oo*k	near-close near back vowel
11	b*e*d	open-mid central unrounded vowel
12	p*er*fect	open-mid central unrounded vowel
13	f*a*ther	open back unrounded vowel
14	b*a*ll	close-mid back rounded vowel
15	f*a*t	near-open front unrounded vowel

Table 1. Phonemes recognized in this system

Evaluation

Settings and Evaluation

- Feature Extraction: 8 LPC, 8 WLPC, 8 PL and 13 MFCC features.
- Windowing: Hamming, 46ms, 23ms overlap.
- Feature Classification: GMM (4 mixtures), SVM (radial basis function), MLP (1 layer,30 neurons) compared.
- For decorrelation and dimension reduction, the LDA has been used (20 dimensions).
- Evaluation: different songs in training and test set.
- The harmonics have been extracted and synthesized from all phonemes and used for feature extraction.
- Half of the set have been used for training and the other half have been used for comparing the classification performance.

Artist	Title
Michael Jackson	Billy Jean
Zoot Woman	It's automatic
The Beatles	She loves you
Ozzi Osbourne	Dreamer
Mando Diao	Mr. Moon

Table 2. Songs in the database

Results

- The result measures are precision (relevant data to all data), recall (ratio of found data to relevant data) and the value of correct classified instances (ratio between correct classified entities and all entities)
- The value of the correct classified instances (CCI) showed best performance by using Support Vector Machines (with harmonic extraction).
- The results between MLP and SVM are similar and GMM's perform slightly worse.
- There are significant differences between phoneme recognition results if harmonics analysis is used as preprocessing and not.
- The average precision and recall are twice as high, if a previous harmonics analysis is performed.

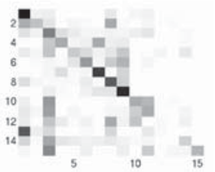
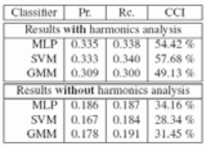


Figure 3. Confusion matrix of each phoneme using SVM classifier



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 3. Results of the classification with and without previous harmonics extraction

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 Net working, 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, 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, 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, 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.