Language Identification in Vocal Music

Jochen Schwenninger
Department of Electrical Engineering
University of Ulm
Ulm, Germany
jochen.schwenninger@uni-ulm.de

Raymond Brueckner, Daniel Willett, Marcus Hennecke
Harman/Becker Automotive Systems
Speech Dialog Systems
Ulm, Germany
{rbrueckner,dwillett,mhennecke}@harmanbecker.com

Abstract

Language identification is an important field in spoken language processing. The identification of the language sung or spoken in music, however, has attracted only minor attention so far. This, however, is an important task when it comes to categorizing, classifying and labelling of music data.

In this paper, we review our efforts of transferring well-established techniques from spoken language identification to the area of language identification in music. We present results of distinguishing German and English sung modern music and propose and evaluate techniques designed for improving the classification performance. These techniques involve limiting the classification on song segments that appear to have vocals and on frames that are not distorted by heavy beat onsets.

1. Introduction

A lot of effort has been put into the categorization of audio data. However, only in recent times the lyrics have attracted some interest [1].

Although the text caries the major part of the artist's message, automatically extracting it is difficult, as most speech recognizers are established for spoken speech in quiet or with only little random noise, not for sung language convoluted with instrumental music. Besides, an automatic lyric transcription is difficult because of the fact that different speech recognizers have to be used for different languages.

This difficulty is partly caused by the fact that different speech recognizers have to be used for different languages. To avoid using them in parallel, one possibility is to use a simple scheme to determine the most probable language and use only the corresponding recognizer.

The rest of the paper describes such a scheme for language identification (LID) that is based on ideas from the area of LID in spoken language. Section 2 deals with the techniques devised to improve the performance for music files, compared to spoken utterances. In Section 3 the experimental setup and in Section 4 the results are presented which lead to the conclusions in Section 5.

2. Language identification

Traditional language identification for spoken language is a two-stage process: First different features are extracted from known utterances and their distribution is approximated using statistical models. After this training phase, unknown utterances can be classified by the system by extracting the features, modeling their distribution and then comparing this model to the already learned ones. Most probably the utterance in question belongs to the language with the most similar distribution.

But since the identification accuracy is only at 64 % when used on modern polyphonic music, different adaptations are tested. These approaches aim at separating the background music from the singing voice.

2.1. Segmentation

A typical music track consists of passages where singing is present (chorus, verse) as well as purely instrumental parts (intro, soli etc.). The latter degrade the accuracy of the modeling, since these additional sounds have to be included in the model as well. Since we are only interested in a fraction of the frames, those with the highest probability of singing, no perfect discrimination is needed. According to Nwe and Wang [2], a good indicator for the presence of singing voice is the energy in high frequency regions.

So a simple discrimination system can be constructed by thresholding the log energy in high freqency bands. Tests proved that for a 512 point FFT and a sample rate of 22 kHz the last 50 frequency bins yield good results.

In order to maintain contextual information between successive frames, the stream of frames was divided into 300 windows of about 1s duration. This length ensures that some context for the statistical modelling is available. Then all frames were sorted by their energy at high frequencies. Eventually a window was considered to contain singing voice if more than half of its frames showed a feature value among the top 30 % of all frames. This resulted in approximately 20 - 30 seconds of audio data when applied to whole songs. In the case of pure instrumental pieces, the frames with top values are taken and considered as singing voice.

2.2. Distortion reduction

In modern music most of the rhythm is generated by either the drums or the bass guitar. Therefore, most of the time they are very prominent and furthermore they are often

found in the center of the stereo mix. In the case of a typical onset, they dominate the whole spectrum, effectively masking most of the voice. Hence, the assumption is that the classification rate can be improved by ignoring the frames during such an onset. To find the onsets, a Mel-based approach first described by West and Cox [4] was used. The detected frames are ignored in the following processing steps.

2.3. Azimuth discrimination

Even if the segments which contain singing are identified perfectly, still the background music detoriates the models. One possible solution for stereophonic signals is a kind of beamforming. Since the vocalist is often panned to the center of the virtual stage, a concentration in this direction should increase the overall performance.

Barry proposes an algorithm for this purpose: Azimuth discrimination and resynthesis. He first determines the frequency bands that originate in a specific direction and then reconstructs them. For details see [3]. In the experiments a FFT-length of 512 was used to match the preprocessing for the MFCCs. The auditory scene was divided into 21 distinct directions and a reconstruction window of size 4 was used.

3. Experimental Setup

During the experiments three different databases were used. The first, denoted by "SPE", consisted of digit and short free speech utterances, totaling approximately 1:45 hours of speech in German (752 files) and English (776 files) each. The average utterance's length is 8s. This approach was used to define an upper limit for language identification in the case of perfect separation of singing and music.

The second database "MUS" was build from 205 modern pop and rock songs, 103 German and 102 English ones. They were manually selected to cover the same musical genres, in order not to differentiate between languages by means of different instrumentation. The last dataset "ACAP" consisted entirely of A Capella pieces, 40 German und 39 English. They were used to test the influence of the background instruments. Unfortunately, not enough solo pieces were available, so that in most songs background singing is present, with probably the same negative effect as background instrumentation.

The features used for modeling are MFCC extracted from stereophonic audio files sampled at 22050 Hz. To keep computation time low, only 30 seconds after the first minute were selected. A 512 point FFT was performed on hamming-windowed frames with 50 % overlap. A DCT for decorrelation reduced the 32 Mel-filterbank coefficients to 10 values after discarding the energy term. Finally a cepstral mean normalization was performed for each track.

4. Results

The results for the baseline system for the different datasets are listed in Table 1. But in spite of the expectations, none of the signal processing techniques described in Section 2 improved the poor accuracy on the "MUS" dataset.

The concentration on passages with high probability of singing yielded comparable accuracy. Ignoring the major onsets to avoid distortion of speech caused a slow decrease of performance. The azimuth discrimination appeared highly unstable and resulted in almost all tracks being either classified as German or English, depending on the parameter settings. Even filtering the spectrum to the human speech range (200-2500 Hz) did not improve accuracy.

Table 1. LID accuracy

Dataset	Acc [%]
SPE	84
ACAP	68
MUS Baseline	64
MUS with Segmentation	63
MUS with Distortion Reduction	61
MUS with Azimuth Discrimination	51

5. Conclusion and Outlook

In comparison to Tsai and Wang [6] who devised a system discriminating between Mandarin and English, the setup presented in this paper is much simpler. The results achieved in identifying spoken language looked very promising, but the application to music was not as successful with recognition accuracy at only 64%. Unfortunately, none of the implemented preprocessing steps helped to improve this performance.

To increase the performance on the examined task, different aspects have to be improved: The discrimination between singing and pure instrumental passages, the extraction of the voice from the background music and finally the modeling of language together with the extraction of features. Such a system, although computationally much more demanding, might yield satisfactory results in the task of language identification in vocal music.

References

[1] J. P. G. Mahedro, A. Martinez, P. Cano, M. Koppenberger and F. Gouyon, "Natural language processing of lyrics," in *ACM Int. Conf. on Multimedia Proc.*, pp. 475-478, 2005.

[2] T. L. Nwe and Y. Wang, "Automatic Detection of Vocal Segments in Popular Songs," in *Int. Conf. on Music Information Retrieval Proc.*, 2004, pp. 138-145.

[3] D. Barry, B. Lawlor and E. Coyle, "Sound Source Separation: Azimuth Discrimination and Resynthesis," in *Int. Conf. on Digital Audio Effects*, October 2004.

[4] K. West and S. Cox, "Finding an Optimal Segmentation for Audio Genre Classification," in *Int. Conf. on Music Information Retrieval Proc.*, 2005, pp. 680-685.

[5] http://music-ir.org/evaluation/m2k

[6] W. H. Tsai and H. M. Wang, "Towards Automatic Identification of Singing Language in Popular Music Recordings," in *Int. Conf. on Music Information Retrieval Proc.*, 2004, pp. 568-576.