AUTOMATIC UNDERSTANDING OF LYRICS FROM SINGING

Annamaria Mesaros, Tuomas Virtanen

Tampere University of Technology
Korkeakoulunkatu 1, 33720, TAMPERE
annamaria.mesaros@tut.fi, tuomas.virtanen@tut.fi

1 INTRODUCTION

When listening to a song, we automatically recognize the words and interpret the linguistic information. We can identify an artist by its voice and a song by its lyrics. Performing some of these tasks in an automated manner has a significant potential in music classification, because the increased amount of available storage and music players have led to large music collections that need to be organized.

Automatic lyrics transcription could provide useful tools for organizing music collections and retrieval of songs based on sung queries or retrieval of songs containing certain key terms. Fragments of lyrics of a song can be used in text searches for identifying a song and its author. Retrieving information about music pieces based on audio only include musical genre [1], artist identity [2, 3], query-by-humming [4] and lyrics [5].

The basis for the techniques in this paper is in automatic speech recognition. Because of the difficulty of lyrics recognition, many studies have focused on a simpler task of audio and lyrics alignment [6, 7, 8], where the textual lyrics to be synchronized with the singing are known. This paper presents construction of a lyrics recognition system and its uses for automatic alignment of music and lyrics and lyrics transcription using commercial pop music.

2 PHONETIC RECOGNITION SYSTEM FOR SINGING VOICE

The singing voice has much higher dynamics than speech and higher frequency range. People can have specific speaking style, but in singing, the melody and rhythm are imposed by musical notes. The spectral properties of a sung vowel depend on formants, which are controlled by the length and shape of the vocal tract and the articulators. Skilled singers can control the pitch and the formant frequencies very accurately and sustain the vowels much longer than in speech, keeping the pitch, loudness and timbre under conscious control.

In singing, the pitch stays mostly constant during a note, with vibrato being used for artistic singing, while in speech the pitch has rapid variations, as can be seen in Figure 1. As a consequence, the variance of the spectrum of a sung vowel is smaller compared to speech, but the difference between the same vowel at different pitches can be significant. Intonation and musical quality of the produced sounds are considered more important than intelligibility, making singing a challenging signal for speech analysis methods.

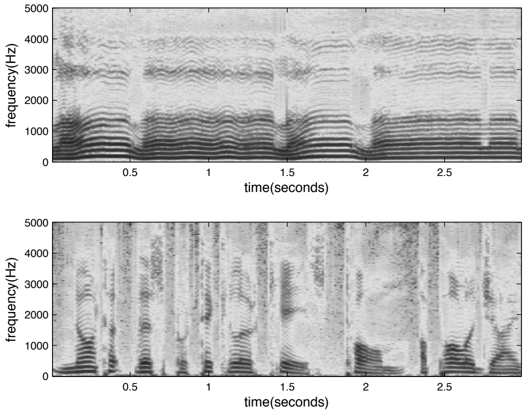

Figure 1: Example spectrograms of male singing (upper panel) and speech (lower panel)

The common properties of speech and singing make it possible to use hidden Markov models (HMM) for singing recognition as in speech recognition. The observed sequence of speech feature vectors is considered to be generated by a hidden Markov model consisting of a number of states with associated observation probability distributions and a transition matrix defining transition probabilities between the states. The emission probability density function of each state is modeled by a Gaussian mixture model (GMM). In the training process, the transition matrix and the means and variances of the Gaussian components in each state are estimated to maximize the likelihood of the observation vectors in the training data.

The HMMs need to be trained on large databases. Due to lack of a large enough database with singing, more steps are needed in order to obtain acceptable acoustic models of singing phonemes. We trained coarse models using speech and adapted them to singing using maximum likelihood linear regression (MLLR) adaptation [9]. The constructed recognition system consists of 39 monophone HMMs plus silence and short pause models. Each phoneme is represented by a left-to-right HMM with three states. As features we use 13 mel-frequency cepstral coefficients (MFCCs) plus delta and acceleration coefficients, calculated in 25 ms frames with a 10 ms hop between adjacent frames.

3 LYRICS TRANSCRIPTION FOR QUERY-BY-SINGING

Additional robustness to the acoustic models of a speech recognizer is provided by using language models. A language model restricts and models probabilities of possible word sequences. The language model consists of a *vocabulary* and a set of rules describing how the units in the vocabulary can be connected into sequences. Usually the vocabulary is chosen as the most frequent words from a training text corpus. It is important to choose training text with similar topic as the desired test data, to have a good coverage of vocabulary and words combinations. For our work we chose to use song lyrics text.

For constructing a word language model we used the lyrics text of 4470 songs, containing over 1.2 million word instances, retrieved from www.azlyrics.com [10]. From a total of approximately 26000 unique words, a vocabulary of 5167 words was chosen by

correct transcription	recognized
I heard you crying loud	I heard you crying alone
all the way across town	all away across the sign
you've been searching for that someone	you been searching for someone
and it's me out on the prowl	I miss me I don't apologize
as you sit around	you see the rhyme
feeling sorry for yourself	feelin' so free yourself

Table 1: Examples of errors in recognition

keeping the words that appeared at least 5 times. This is a surprisingly small vocabulary content of the mainstream songs lyrics.

3.1 Word recognition results from clean singing

For testing the recognition system we used a database consisting of monophonic singing recordings – 49 fragments (19 male and 30 female) of popular songs. The lengths of the sung phrases are between 20 and 30 seconds and usually consist in a full verse of a song. The total amount of singing material is 30 minutes. We used 5-fold cross-validation for the model adaptation and word recognition tests.

Recognition performance of phonemes and words in singing are very modest. Using the adapted models and the bigram language model we obtained an average of 24% word recognition rate and high rate of substitution errors [10]. The output of the recognition system offers sometimes words that are acoustically very similar with the correct ones, as the examples shows in Table 1. Such incompletely transcribed lyrics can be used in retrieval of songs based on sung queries or retrieval of songs containing certain key terms, as it will be presented in the next section.

3.2 Query-by-singing retrieval

Query-by-humming/singing aims to identify a piece of music from its melody/lyrics. In a query-by-humming application, the search algorithm will transcribe the melody sung by the user and will try to find a match of the sung query with a melody from the database. Assuming that we also have the lyrics of the songs we are searching through, the words output from a phonetic recognizer can be searched for in the lyrics text files. This will provide additional information and narrow down the melody search space. Furthermore, lyrics will be more reliable than the melody in the case of less skilled singers.

We built a proof of concept retrieval system based on sung queries [10]. We used as test queries the 49 singing fragments of the clean singing database (with an average recognition rate 24%). A database with lyrics was constructed, consisting of the correct text lyrics of the sung queries and additional 100 files. For retrieval we use a bag-of-words approach, simply searching for each recognized word in all the text files and ranking the songs according to the number of matched words. We consider a song being correctly identified when the queried fragment appears among the first N ranked lyrics

	Top 1	Top 5	Top 10
recognized [%]	57%	67%	71%

Table 2: Query-by-singing retrieval results

files. Table 2 presents the retrieval accuracy for N being 1, 5 and 10. The application shows promising results, the first retrieved song being correct in 57% of the cases.

4 AUTOMATIC SINGING-TO-LYRICS ALIGNMENT

Alignment of singing to lyrics refers to finding the temporal relationship between a possibly polyphonic music audio and the corresponding textual lyrics. A straightforward way to do alignment is by creating a phonetic transcription of the word sequence comprising the text in the lyrics and aligning the corresponding phoneme sequence with the audio using the HMM recognizer. For alignment, the possible paths in the Viterbi search algorithm are restricted to just one string of phonemes, representing the input text. We presented an alignment system in [7]. Such systems have applications in automatic production of material for entertainment purposes, such as karaoke.

In commercial music, accompanying instruments often follow the same melody and notes as the singing and their overlapping represents a challenging sound separation problem. In order to apply the HMMs on singing in polyphonic music, we employ a vocal separation algorithm [11]. The algorithm separates the vocals using the time-varying pitch and enhances them by subtracting a background model.

4.1 Alignment performance evaluation and results

For testing the alignment, we chose 17 songs from commercial music collections. The songs were manually segmented into structurally meaningful units (verse, chorus). We obtained 100 fragments of polyphonic music containing singing and instrumental accompaniment. Each fragment has a corresponding lyrics text input containing a number of lines of text, each line corresponding roughly to one singing phrase. The lyrics were processed to obtain a sequence of words with optional silence (sil), pause (sp) and noise (noise) between them, that will be aligned with the audio using the acoustic models. An example of resulting sequence for one of the test songs is:

[sil | noise] I [sp] BELIEVE [sp] I [sp] CAN [sp] FLY [sil | noise] I [sp] BELIEVE [sp] I [sp] CAN [sp] TOUCH [sp] THE [sp] SKY [sil | noise] I [sp] THINK [sp] ABOUT [sp] IT [sp] EVERY [sp] NIGHT [sp] AND [sp] DAY [sil | noise] SPREAD [sp] MY [sp] WINGS [sp] AND [sp] FLY [sp] AWAY [sil | noise]

The alignment performance is evaluated by the average of the absolute alignment errors in seconds at the beginning and at the end of each lyric line. The timestamps for beginning and end of each line in the lyrics were manually annotated. The absolute alignment errors range from 0 to 9 seconds, with 0.94 seconds average absolute alignment error. Figure 2 illustrates examples of line-level alignment for song fragments.

music piece 1
music piece 2

Figure 2: Automatic alignment examples. The black lines represent the manual annotations, the gray lines the automatic alignment output. The errors are calculated at each end of the black segments.

4.2 Discussion

Misalignment of music and lyrics can result from a faulty output of the vocal separation stage. In some cases, the output contains a mixture of the vocals with some instrumental sounds, but the voice is usually too distorted to be recognizable. In other cases, the errors appear when the text lines in the lyrics do not correspond to singing phrases, so there are breathing pauses in the middle of a text line. In these cases even the manual annotation of the lyrics can have ambiguity.

A demo of the lyrics alignment system using full length songs can be found online [1].

5 CONCLUSIONS

This paper presented methods from speech recognition applied to recognize lyrics from singing voice. We attempt to recognize words in singing voice for song retrieval and automatic alignment applications. Due to the lack of large enough singing databases to train a phonetic recognizer on singing data, we used speech data to train a set of models and adapted them to singing using monophonic singing examples.

Word recognition using the adapted models and a bigram language model built from lyrics text allows recognition of approximately one fifth of the sung words in the monophonic singing. Despite the low performance, the method has potential in music information retrieval. Our query-by-singing experiment indicates that a song might be retrieved based on words that are correctly recognized from a user query. We also demonstrated the capability of the recognition methods in automatic alignment of singing from polyphonic audio and text. In order to suppress the effect of the instrumental accompaniment, a vocal separation algorithm was applied. The alignment performance is very good, with an average alignment error of 0.94 seconds for our test data.

5.1 Acknowledgments

This research has been funded by the Graduate School in Electronics, Telecommunications and Automation (GETA) and the Academy of Finland.

[1] A demo of the lyrics alignment system can be found at http://www.cs.tut.fi/~mesaros/demos.html

REFERENCES

[1] TZANETAKIS G & COOK P, Musical genre classification of audio signals, *IEEE Transactions on Speech and Audio Processing*, **10**(2002) 5.

[2] TSAI W H & WANG H M, Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals, *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(2006) 1.

[3] MESAROS A, VIRTANEN T, & KLAPURI A, Singer identification in polyphonic music using vocal separation and pattern recognition methods, in *Proceedings of International Conference on Music Information Retrieval*, 2007.

[4] TYPKE R, WIERING F, & VELTKAMP R C, MIREX symbolic melodic similarity and query by singing/humming, in *International Music Information Retrieval Systems Evaluation Laboratory(IMIRSEL)*, URL http://www.music-ir.org/mirex2006/.

[5] SUZUKI M, HOSOYA T, ITO A, & MAKINO S, Music information retrieval from a singing voice using lyrics and melody information, *EURASIP Journal on Advances in Signal Processing*, (2007).

[6] FUJIHARA H, GOTO M, OGATA J, KOMATANI K, OGATA T, & OKUNO H G, Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals, in *ISM '06: Proceedings of the 8th IEEE International Symposium on Multimedia*, 2006.

[7] MESAROS A & VIRTANEN T, Automatic alignment of music audio and lyrics, in *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, 2008.

[8] WONG C H, SZETO W M, & WONG K H, Automatic lyrics alignment for Cantonese popular music, *Multimedia Systems*, **12**(2007) 4-5.

[9] MESAROS A & VIRTANEN T, Adaptation of a speech recognizer to singing voice, in *Proceedings of 17th European Signal Processing Conference*, 2009.

[10] MESAROS A & VIRTANEN T, Automatic recognition of lyrics in singing, *EURASIP Journal on Audio, Speech, and Music Processing*, (2010).

[11] VIRTANEN T, MESAROS A, & RYYNÄNEN M, Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music, in *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition SAPA*, 2008.