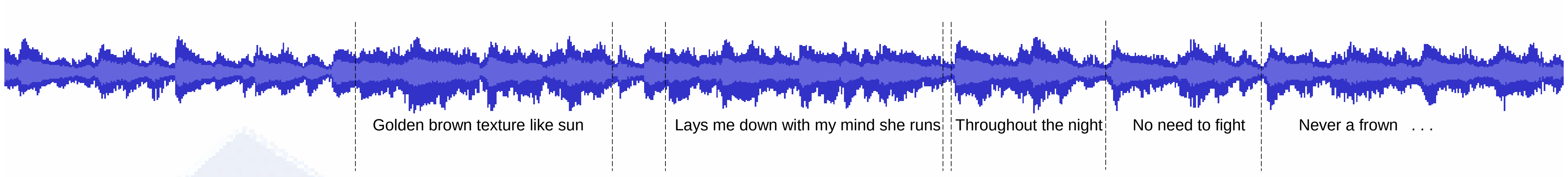
Audio Research Group

Automatic Alignment of Music Audio and Lyrics

Annamaria MESAROS, Tuomas VIRTANEN

Tampere University of Technology

annamaria.mesaros@tut.fi tuomas.virtanen@tut.fi

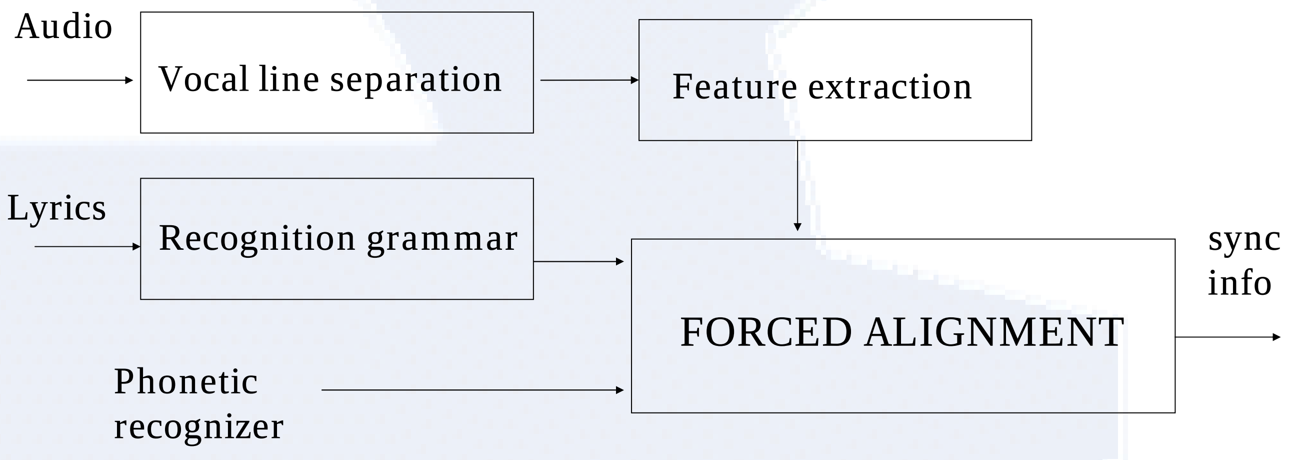


General idea: use a phonetic recognizer on the singing voice

Difficulties of the approach:

- polyphonic audio - unlikely to have a reliable phonetic speech recognizer on such a complex signal
- singing voice is quite different from speech

System overview



Phonetic recognizer

-HMM based speech recognizer: 39 English phonemes, silence, short pause and instrumental (noise) models

- phonemes, silence and short pause HMMs trained on speech data

- noise HMM trained separately on instrumental sections

- adaptation of the phoneme HMMs to clean singing voice: standard supervised MLLR speaker adaptation

Vocals separation

- melody transcription (Ryynänen&Klapuri) - estimation of the notes of the main melodic line

-sinusoidal modeling - representation and separation of the acoustic signal that corresponds to the main melodic line

- it assumes that the main melodic line is the voice

Lyrics processing

-the lyrics determine a sequence of words

- breathing pauses between words and lines in the text

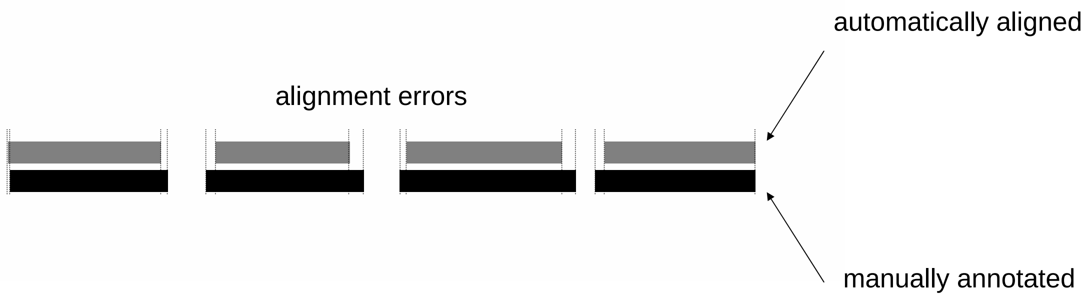
- create a recognition grammar consisting in a sequence of phonemes, pauses and instrumental parts

. . . B R AW N sp T EH K S CH ER sp L AY K sp S AH N [sil | noise] L EY Z sp M IY sp D AW N . . .

Use **forced alignment** procedure - the sequence of phonemes to be aligned with the sequence of features extracted from the acoustic signal

Evaluation procedure:

- sections (verse, chorus, containing vocals and instrumental accompaniment) from 17 songs
- ground truth: manually annotated start and end of each line
- total 100 sections comprising over 1000 lines of text



absolute error between annotated and automatically aligned text:

mean **1.4 s**, median 0.64 s

Errors :

- alignment related: long vowels, distorted voice mixed with instrumental sounds
- missing sounds: failure to reconstruct consonants (incomplete words in the vocal line)
- no perfect alignment - there is certain ambiguity in manual annotation

TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Signal Processing