ECS7006 Music Informatics

Week 12 - Cross-Modal and Multimodal Music Informatics

School of Electronic Engineering and Computer Science
Queen Mary University of London

prepared by Emmanouil Benetos
adapted from material by Meinard Müller, Zhiyao Duan, Ichiro Fujinaga, and Jorge Calvo-Zaragoza

emmanouil.benetos@qmul.ac.uk

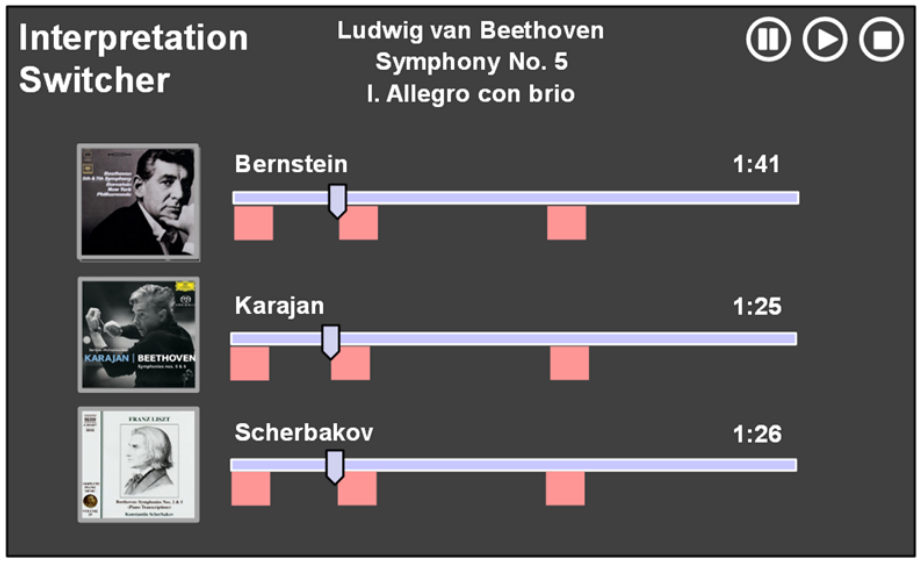
2023

Queen Mary University of London

Week 10 recap

Audio Matching & Cover Song Detection

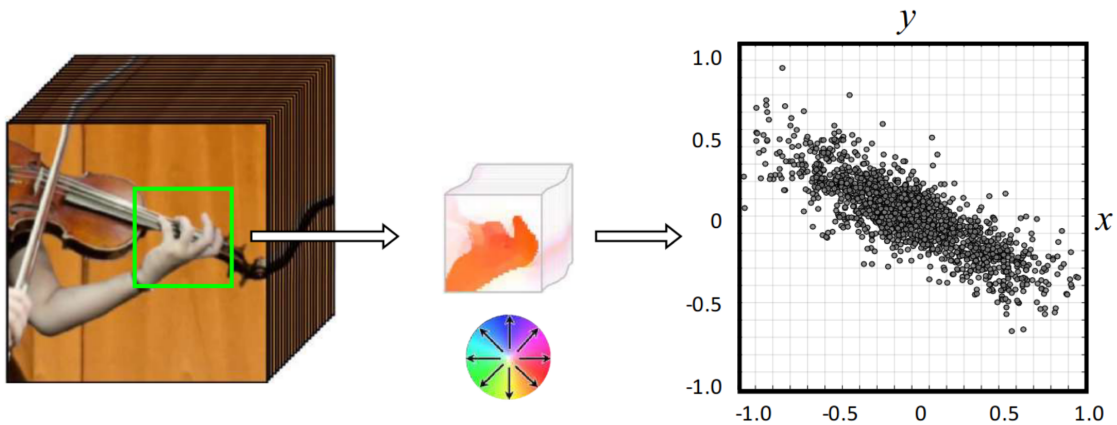
- Audio matching: requirements and feature design
- Diagonal matching
- DTW-based matching
- Cover song detection requirements
- Cover song detection procedure



This week's content

Cross-Modal and Multimodal Music Informatics

- Introduction
- Audio-symbolic music retrieval
- Audiovisual analysis of music
- Optical music recognition
- Lyrics alignment and transcription



Reading

M. Mueller, A. Arzt, S. Balke, M. Dorfer and G. Widmer, “Cross-Modal Music Retrieval and Applications: An Overview of Key Methodologies,” IEEE Signal Processing Magazine, vol. 36, no. 1, pp. 52-62, 2019.

Z. Duan, S. Essid, C. C. S. Liem, G. Richard and G. Sharma, “Audiovisual Analysis of Music Performances: Overview of an Emerging Field,” in IEEE Signal Processing Magazine, vol. 36, no. 1, pp. 63-73, 2019.

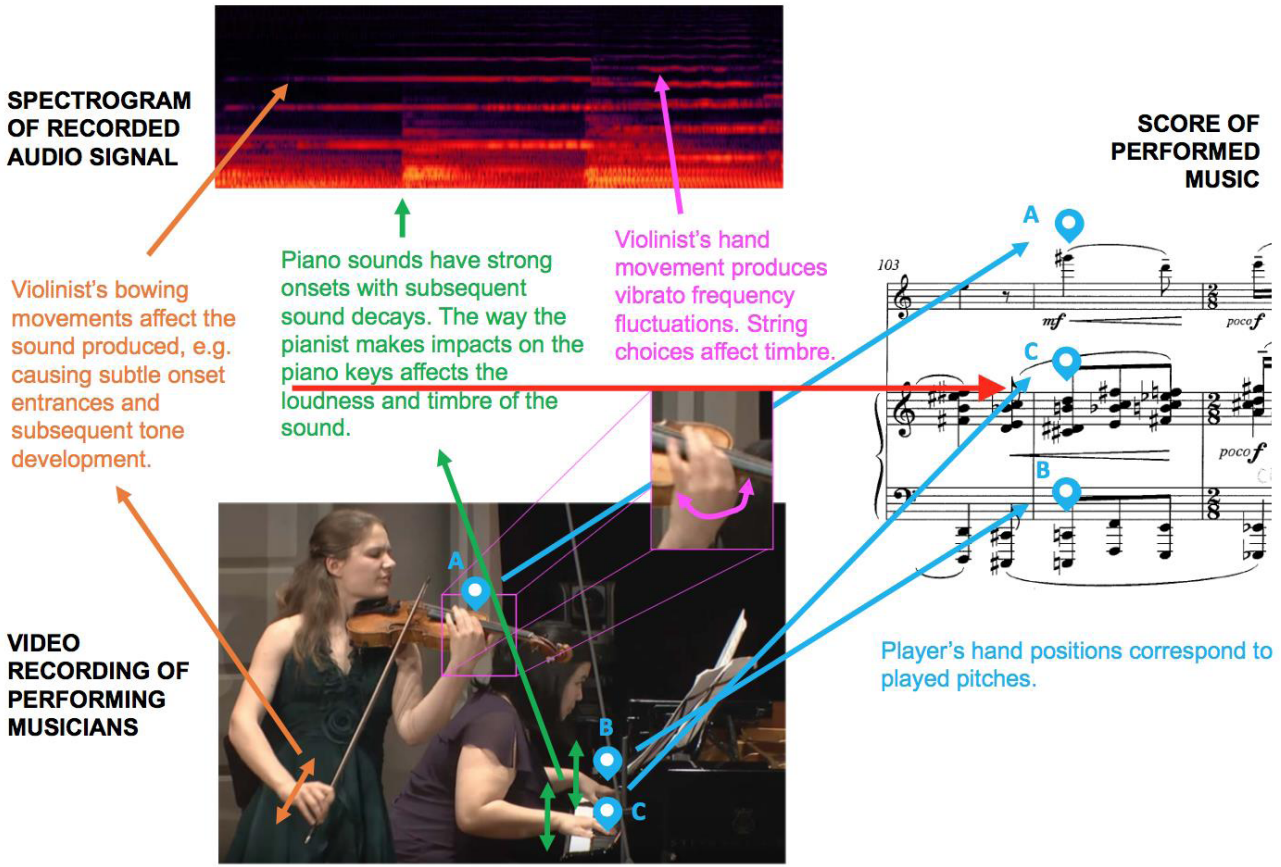
A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. S. Marcal, C. Guedes, and J. S. Cardoso, “Optical music recognition: state-of-the-art and open issues”, International Journal of Multimedia Information Retrieval, vol. 1, pp 173-190, 2012.

A. Mesaros and T. Virtanen, “Automatic Recognition of Lyrics in Singing”, EURASIP Journal on Audio, Speech, and Music Processing, Article ID 546047, 2010.

Introduction

Introduction

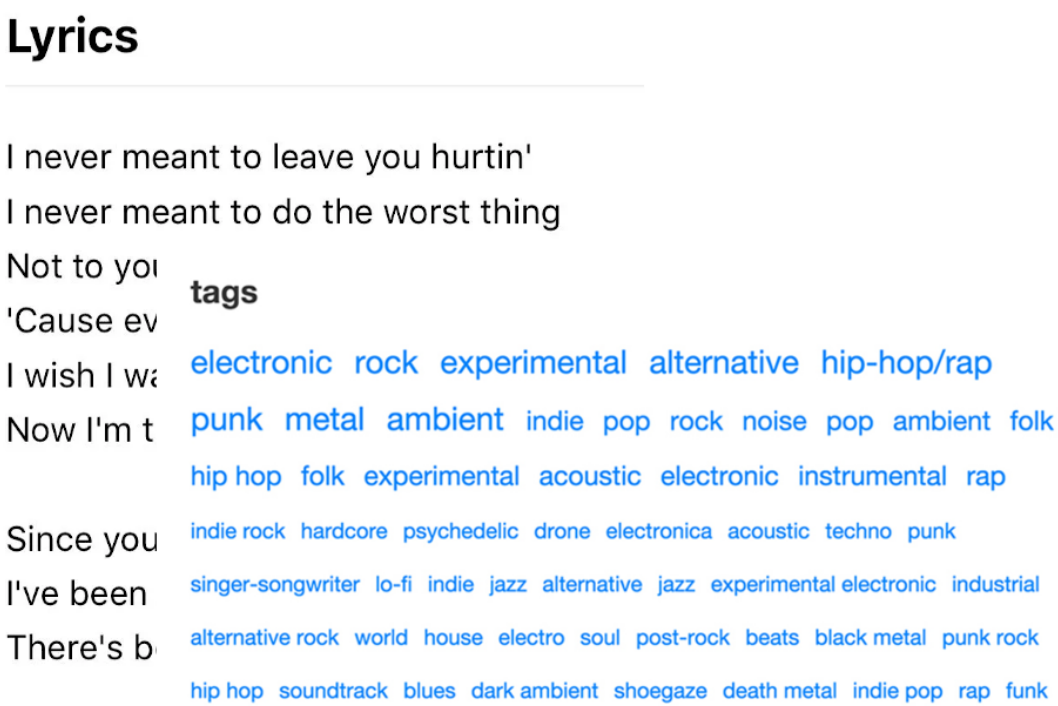
- So far we have been focusing on analysing music audio (and converting music audio into symbolic representations)
- However, music is a multimodal art form



(Duan et al., 2019)

Introduction

Modalities include but are not limited to: audio, video, images, symbolic representations, lyrics, metadata...



Introduction

One key issue in music informatics concerns the development of methods for identifying and establishing semantic relationships across various music representations and formats.

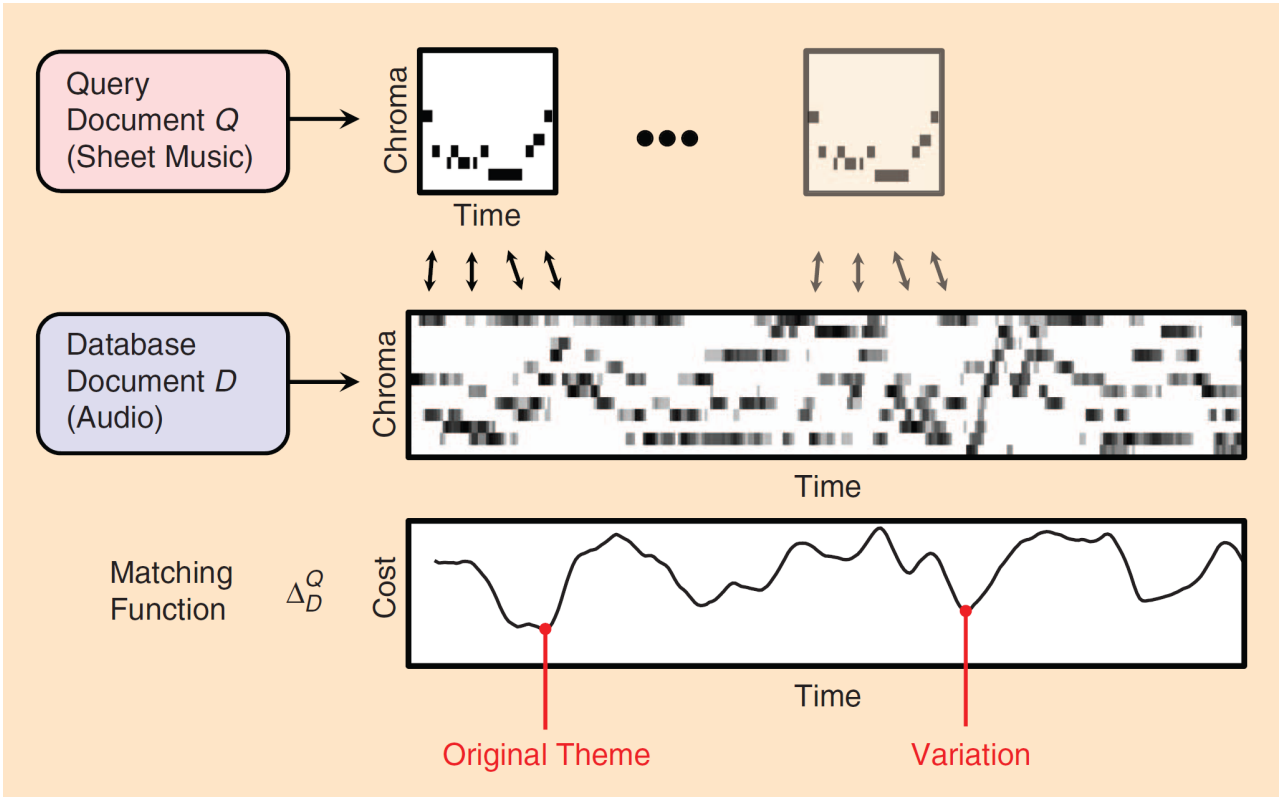
There is a need for cross-modal retrieval algorithms that, given a query in one modality (e.g. short audio excerpt), find corresponding information in another modality (e.g. the name of the piece and sheet music).

Audio-symbolic music retrieval

Audio-symbolic music retrieval

Task: Audio retrieval using symbolic musical themes

Let $\mathcal{Q}$ be a collection of musical themes in symbolic format and let $\mathcal{D}$ be database of audio recordings. Given a query Q, the retrieval task is to identify the semantically corresponding database items $D \in \mathcal{D}$.



Audio-symbolic music retrieval

Task: Audio retrieval using symbolic musical themes

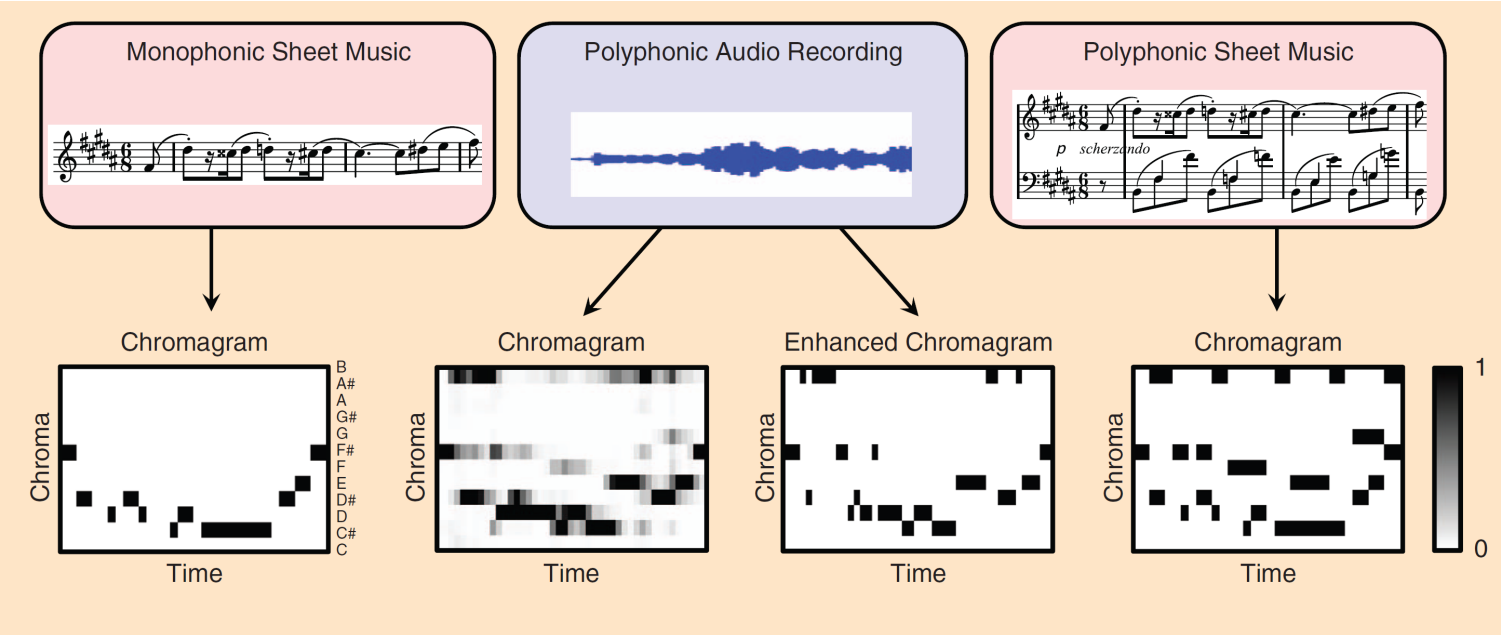
Challenges:

- Tempo deviations
- Tuning deviation
- Key transposition
- Difference in polyphony level between query and database documents

Solutions to the above include sequence alignment techniques, transposition-invariant features, melody extraction...

Symbolic fingerprinting

- Chroma features are very convenient representations for comparing music data of different modalities, and are highly robust to musical and acoustical variations.
- However, reduction into chroma also leads to a loss of information (e.g. accurate timing, pitch parameters), and renders the comparison of monophonic and polyphonic versions difficult.
- Solution: transcribe acoustic data into the symbolic domain.



Symbolic fingerprinting

Arzt et al. (2012, 2014) introduced a symbolic fingerprinting approach that allows not only for audio identification but for audio-score matching.

Starting with a symbolic representation, assume that each note event $e = (t, p)$ is specified by an onset time t and pitch p. To obtain fingerprints, triples are considered:

$$e_1 = (t_1, p_1) \quad e_2 = (t_2, p_2) \quad e_3 = (t_3, p_3) \quad \text{with } t_1 < t_2 < t_3$$

For each such triple, their time and pitch differences are defined:

$$\Delta_t^{1,2} = t_2 - t_1 \quad \Delta_t^{2,3} = t_3 - t_2$$
$$\Delta_p^{1,2} = p_2 - p_1 \quad \Delta_p^{2,3} = p_3 - p_2$$

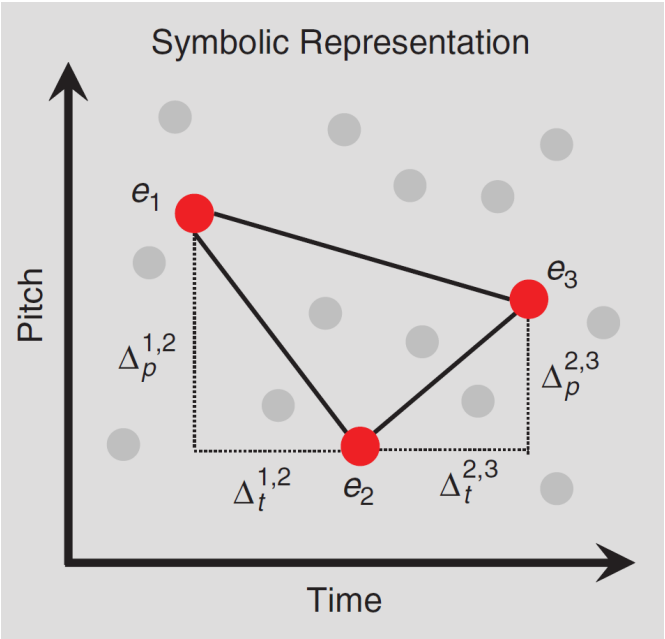
Furthermore, the ratio of time differences is defined as $\tau = \Delta_t^{2,3} / \Delta_t^{1,2}$.

Symbolic fingerprinting

A symbolic fingerprint is defined to be a list of the following numbers:

$$[\Delta_p^{1,2}, \Delta_p^{2,3}, \tau]$$

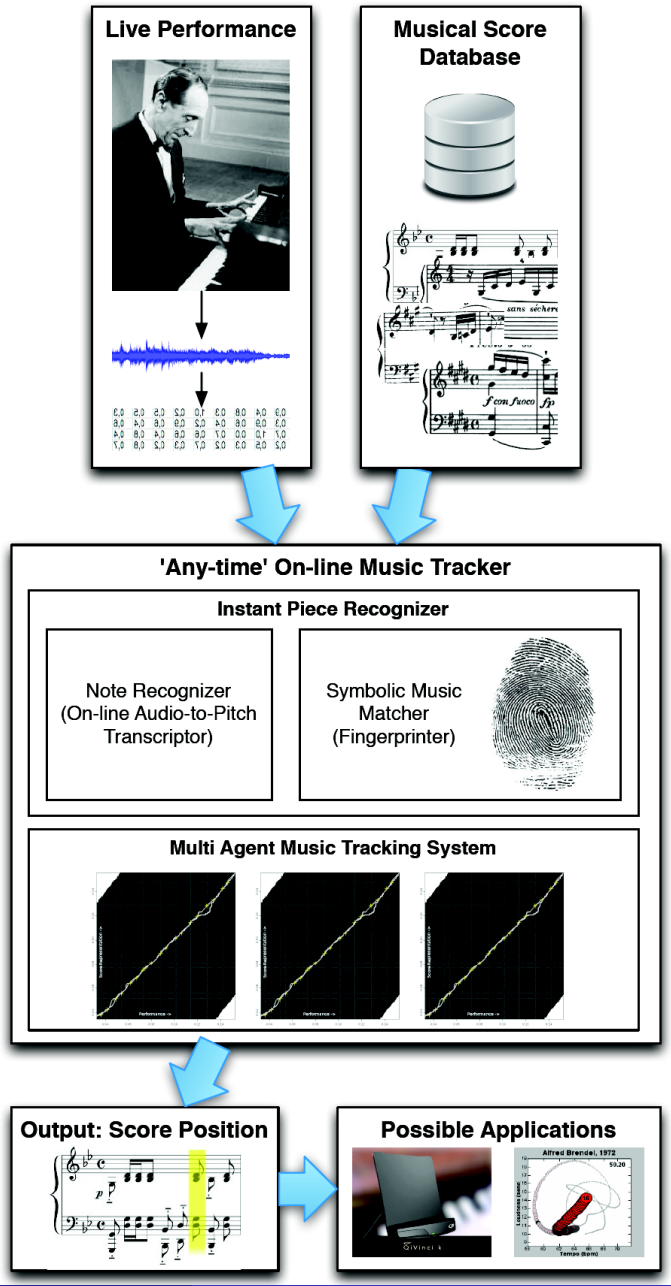
The above fingerprint is invariant to transpositions and tempo changes.



Symbolic fingerprinting

Task: given a short audio excerpt, identify the exact position in a machine-readable score.

Application: score retrieval and score following, e.g. “Piano Music Companion” (Artz et al., 2014).



Symbolic fingerprinting

Process:

1. Compute symbolic fingerprints for database items $\mathcal{D}$.
2. For audio query Q, perform automatic music transcription (i.e. converting audio into a machine-readable score).
3. Compute symbolic fingerprints for query.
4. Use identification and indexing techniques (similar to those used in audio identification) to match the automatically transcribed segment with a database item, and the exact score position.

Requirements:

- Good automatic music transcription performance
- Real-time transcription and fingerprint search

Symbolic fingerprinting - future directions

- Current technologies for audio-symbolic retrieval focus on specific use cases (e.g. classical piano performance).
- Automatic transcription typically carried out using deep learning approaches, which are often non-causal (e.g. using bi-directional recurrent neural networks).
- Audio-symbolic retrieval is still an open problem for multi-instrument music and noisy recordings.
- Next step: linking audio segments with digitised sheet music.

Audiovisual Analysis of Music

Motivation

Musicians use audiovisual cues to coordinate with each other

Motivation

Audiences enjoy audiovisual expressions:

- Music video streaming services

- Visual aspect is an important factor in the communication of meanings (Platz & Kopiez, 2012)
- Sight over sound in the judgement of music performance (Tsay, 2013)

Applications

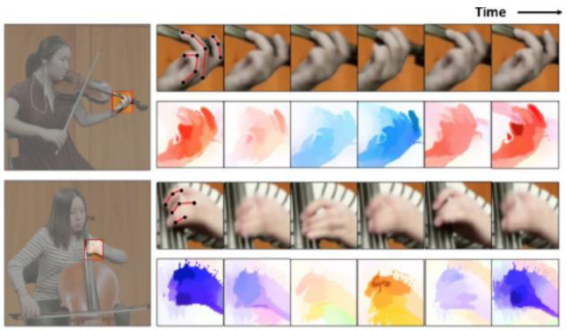
- Concerts: Automatic camera/light/sound control, augmented concerts with visual displays, virtual concerts

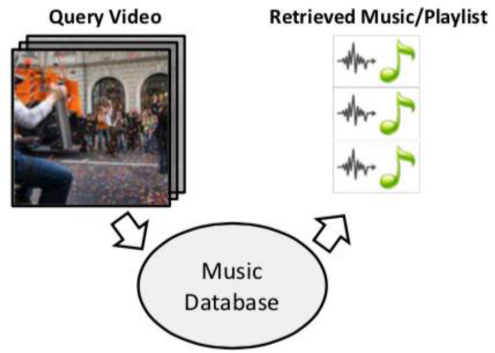
- Music Interaction: human-computer collaborative music making, visually informed automatic accompaniment
- Music Education: pose analysis and feedback, automatic visual demonstration, automatic fingering annotation

Problem Categorisation

Analysis:

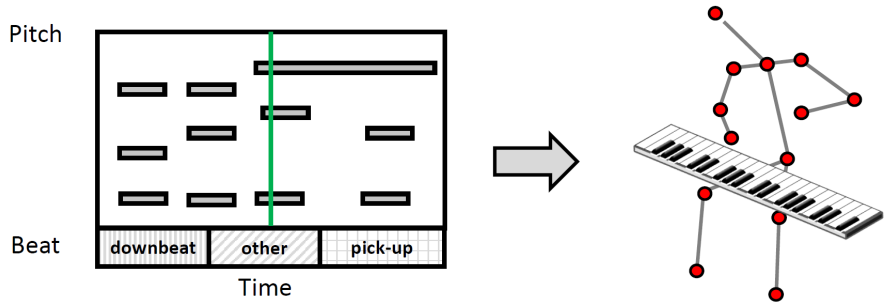
- Association
- Transcription
- Separation





Classification and retrieval:

- Genre classification
- Mood recognition
- Audiovisual matching
- Recommendation



Problem Categorisation

According to instrument/source type:

- Percussion: large-scale motions related to sound articulation.
- Piano: subtle hand/finger motions, related to sound articulation, indicative of notes being played.
- Strings: one hand indicates notes being played, other hand indicates note articulation.
- Winds: Note articulations are difficult to see; visible motions are about notes.
- Vocals: mouth shapes reveal phonemes/diphones, but not the pitch. Body movement can be correlated with musical content.

Levels of audiovisual correspondence

Static

- Fixed image $\leftrightarrow$ audio frame
- e.g. posture of a flutist $\leftrightarrow$ play/nonplay activity
- e.g. piano fingering $\leftrightarrow$ music transcription

Dynamic

- Dynamic movement $\leftrightarrow$ audio feature fluctuation
- e.g. guitarist's strumming hand $\leftrightarrow$ rhythmic pattern
- e.g. violinist rolling left hand $\leftrightarrow$ vibrato

Static audiovisual correspondence

Typical static audiovisual correspondences:

- Musicians' positions
- Parts of the instrument that lead to sound production (e.g. fingering analysis)

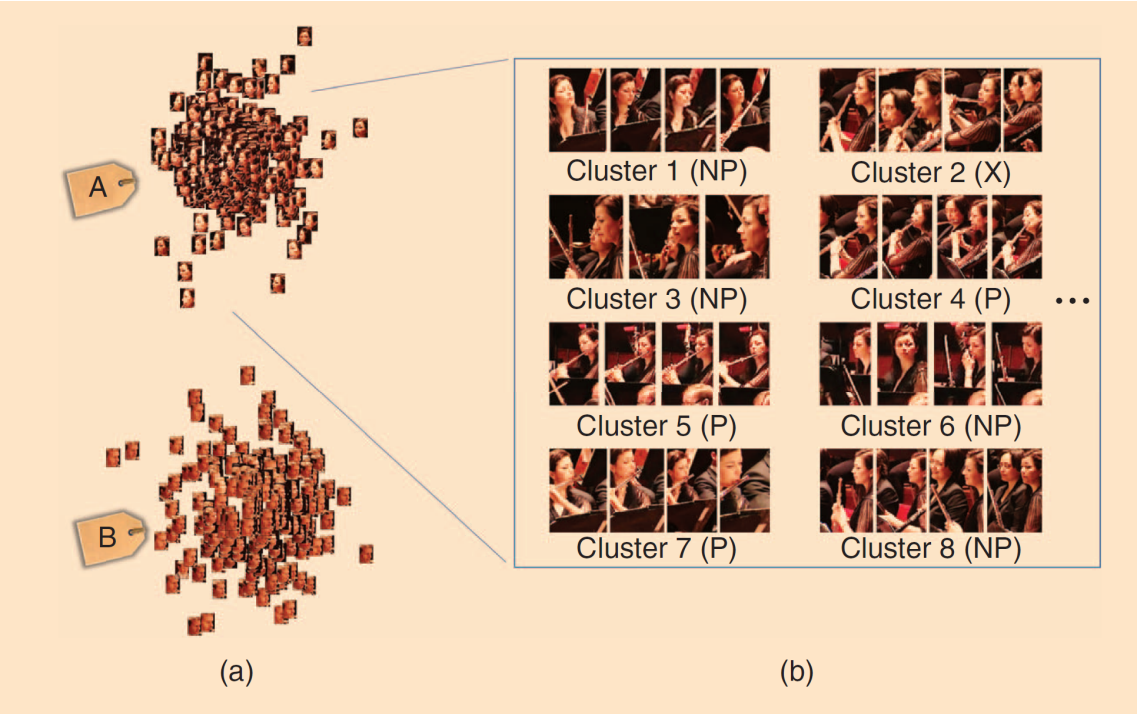
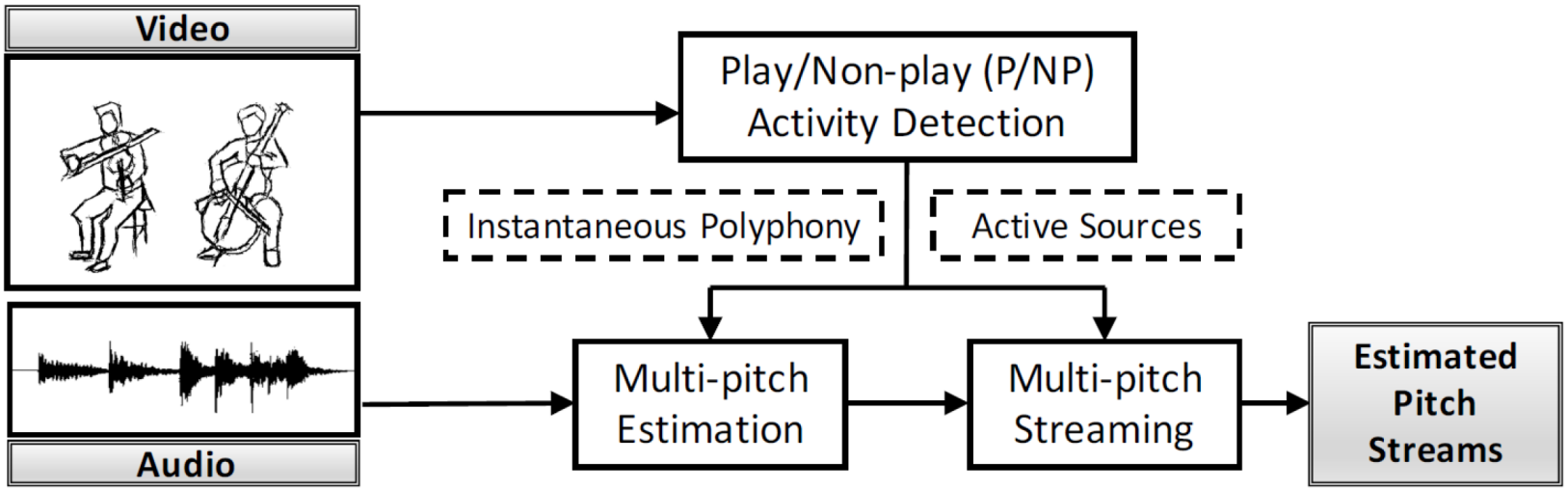


Figure: (a) Clustering musicians; (b) Clustering play/nonplay activity.

Static audiovisual correspondence

Play/nonplay detection can also improve multi-pitch detection and multi-pitch streaming performance, by respectively estimating instantaneous polyphony and active sources of an ensemble performance.



(Dinesh et al., 2017)

Dynamic audiovisual correspondence

Existing methods rely on knowledge of instrument type and playing techniques:

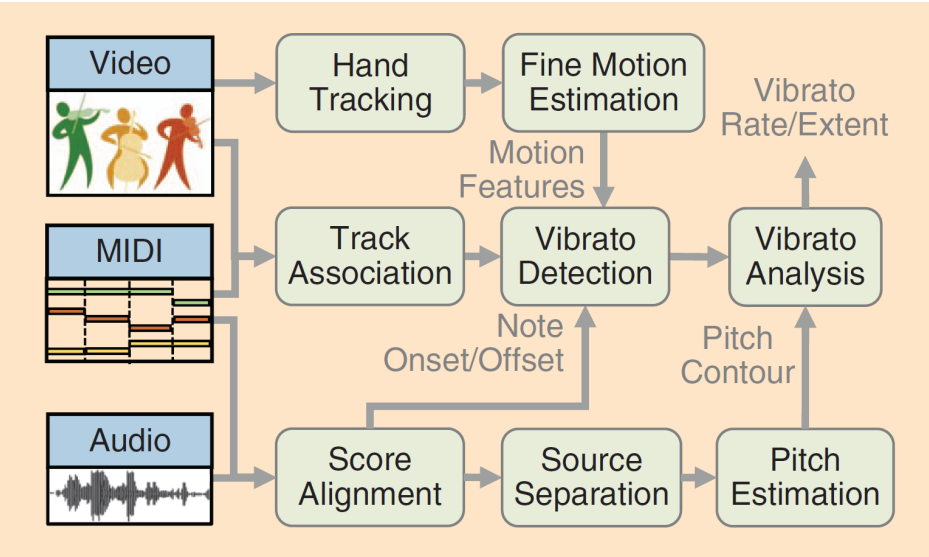
- Correspondence between bowing motions and note onsets of string instruments.
- Hitting actions and drum sounds in percussive instruments.
- Left-hand rolling motions of string vibrato notes.

Challenge: many irrelevant motions in the visual scene; interference from multiple sound sources in the audio signal.

Dynamic audiovisual correspondence

Vibrato analysis of string instruments

- Vibrato is characterised by a periodic fluctuation of pitch with a rate of 5-10 Hz.
- Vibrato is difficult to detect from the audio signal in an ensemble, due to interference from other sound sources.
- For string instruments, vibrato is produced through the rolling motion of the left hand.

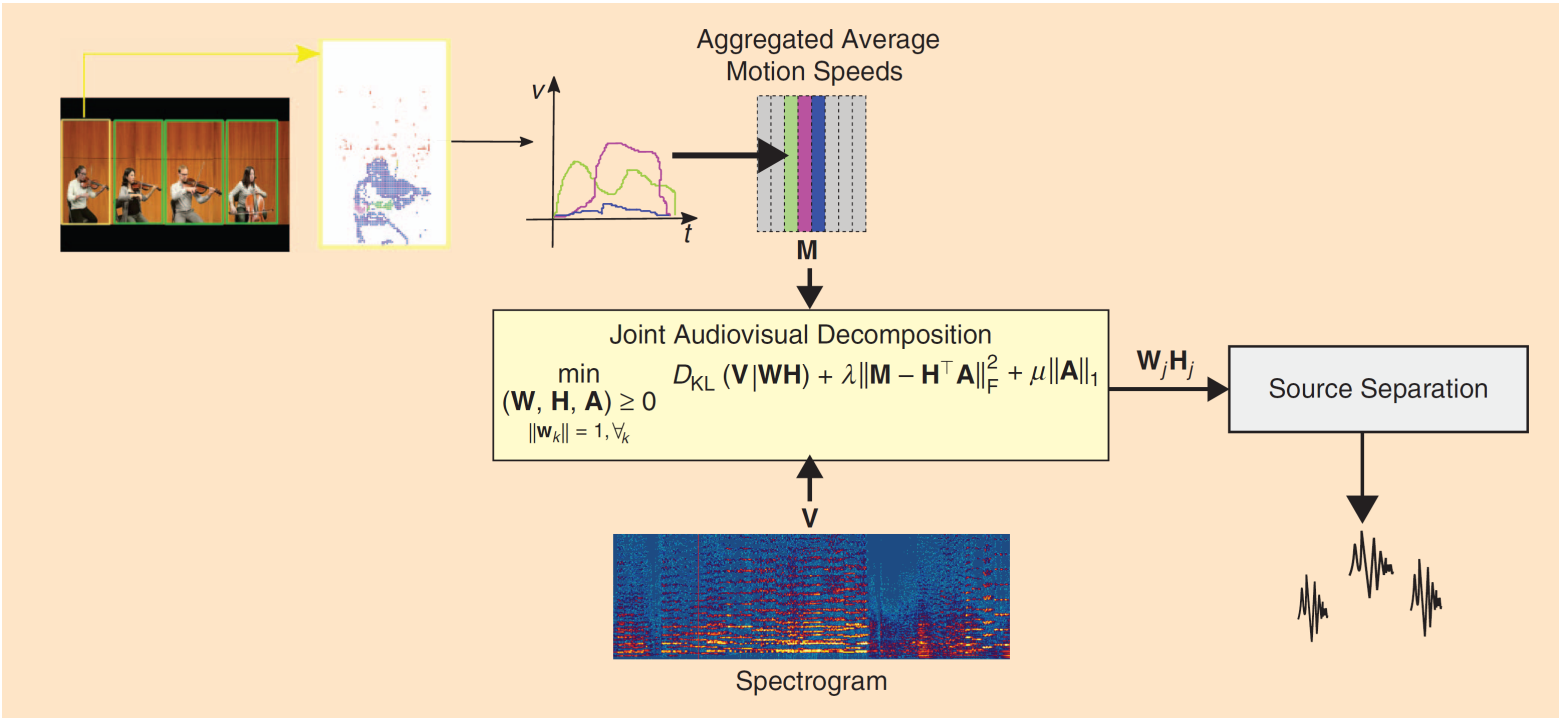


Audiovisual vibrato detection system from (Li et al., 2017).

Dynamic audiovisual correspondence

Motion-driven audio source separation

- Players' motions are highly correlated to the sound characteristics of audio sources
- Approach: gather motion speeds in predefined regions; jointly analyse audio signal with motion speeds.



Audiovisual source separation system from (Parekh et al., 2017).

Audiovisual Analysis of Music - perspectives

Challenge: inherent mismatch between audio content and corresponding image frames - correspondences can be partial and intermittent.

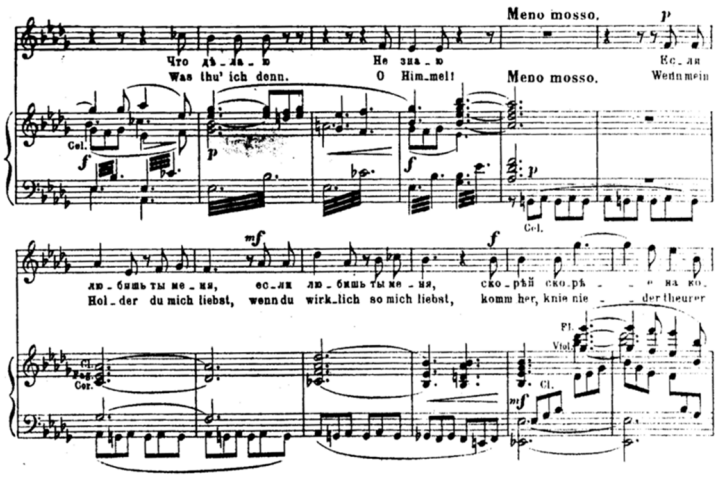
Future directions:

- Audiovisual emotion analysis of music videos
- Video-based tutoring for music lessons
- Improving other tasks (e.g. score following, automatic accompaniment)
- Sound-to-image and audiovisual sequence generation

Optical Music Recognition

Optical Music Recognition (OMR)

The process of converting images of music scores into a machine-readable music score, such as MIDI, MusicXML, or MEI.



OMR

OMR Applications

- Creating auditory versions of handwritten drafts
- Musicological analysis at scale
- Music education
- Braille output
- Preservation/accessibility of historical manuscripts
- Re-editing and creation of derived works
- Converting scores to different formats

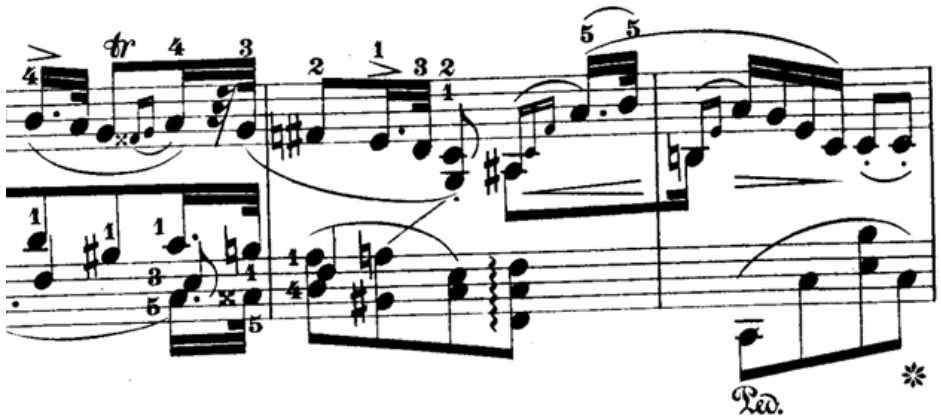
OMR Challenges

Graphical complexity (e.g, dense scores, artifacts from image capturing procedure, overlapping symbols, ambiguities)

Structural complexity

- Inherently complicated content
- Syntactic and semantic rules
- Violations of these rules

Natural limit of OMR: where musicians disagree



OMR Challenges

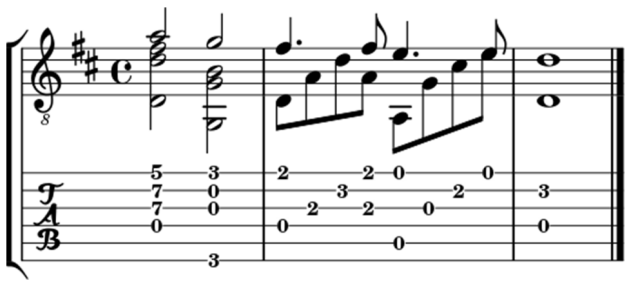
- Many historical and modern music notation systems
- Mensural notation: describes measured rhythmic durations in terms of numerical proportions between note values

Note values

Name		Century			
		13th	14th	15th	17th
Maxima	Mx				
Longa	L				
Breve	B				
Semibreve	Sb				
Minim	Mn				
Semiminim	Sm				
Fusa	F				
Semifusa	Sf				

Mensural Notation

Guqin Notation



Tablature Notation

OMR Challenges

Engraving mechanism

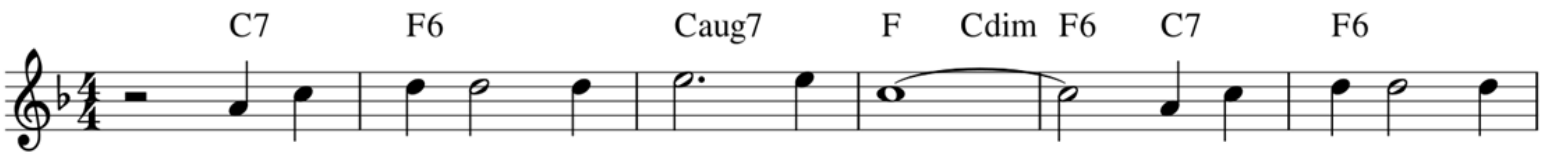


Figure: Notation typeset by a machine (“printed”).

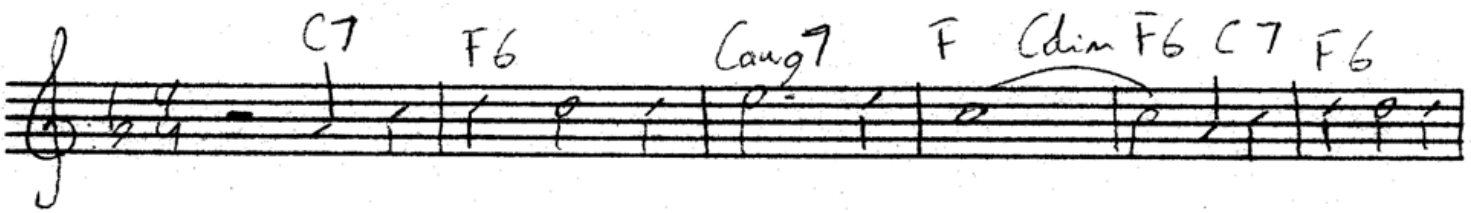
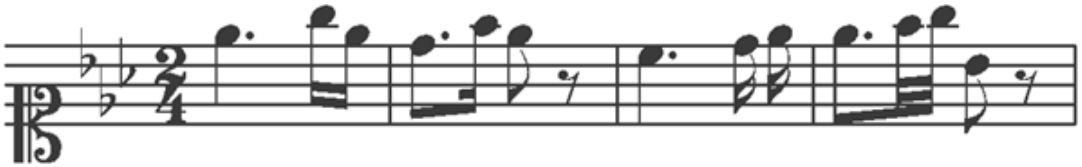


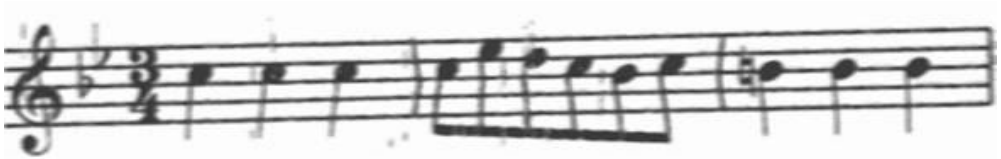
Figure: Handwritten notation (typically over printed staff lines).

OMR Challenges

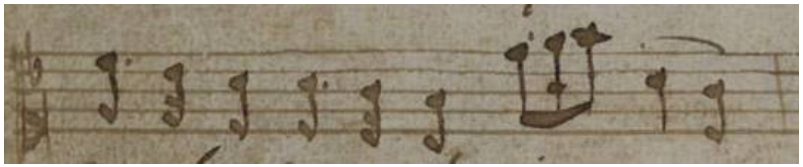
Graphical complexity



Ideal conditions



Camera-based scenario (blurring, skewing, 3D distortions, low resolution...)



Degraded documents (heterogeneous background, inkblots, mold...)

OMR state-of-the-art

OMR research dates back to the 1960s (see Rebelo et al., 2012 for a survey)

A universal solution for OMR is out of reach

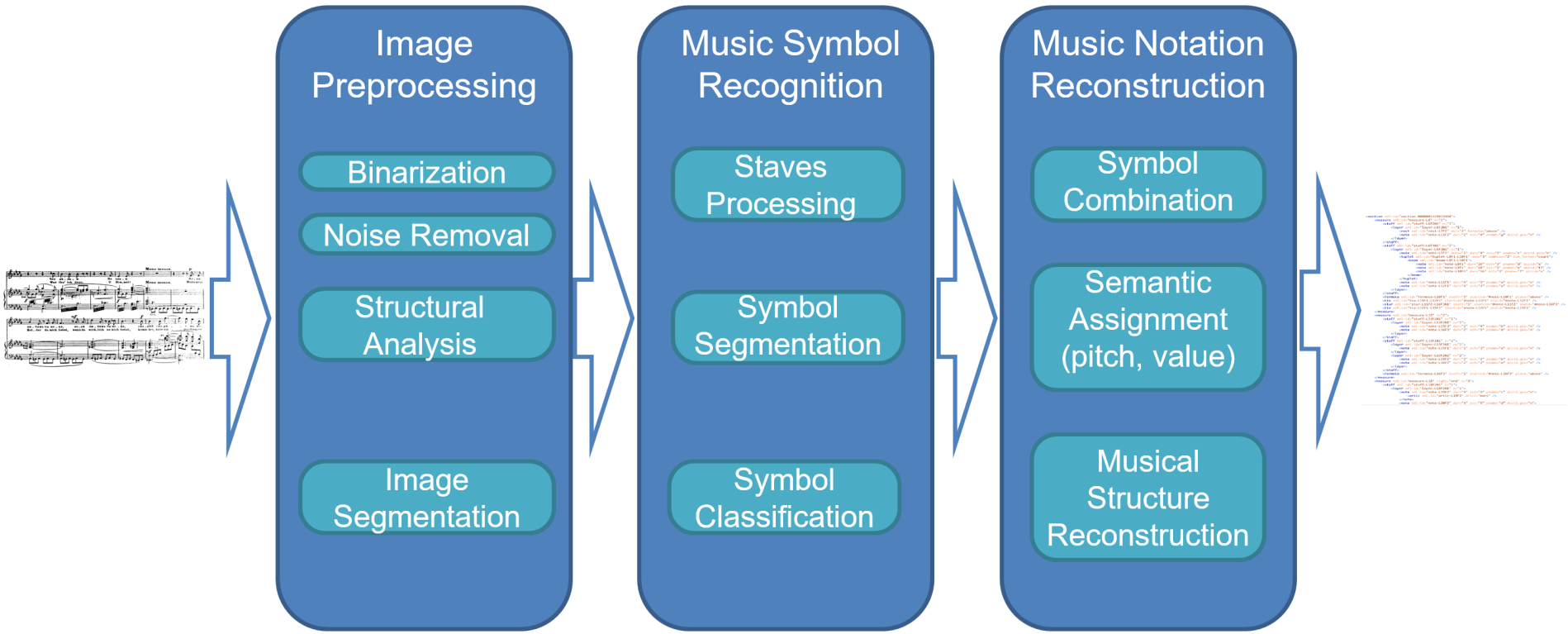
New OMR systems are built on top of previous research:

- Develop general OMR workflows
- Enforce interoperability and common standards

The current state of the art can be organized into:

- Static scenario: full-pipeline or end-to-end approaches
- Interactive scenario

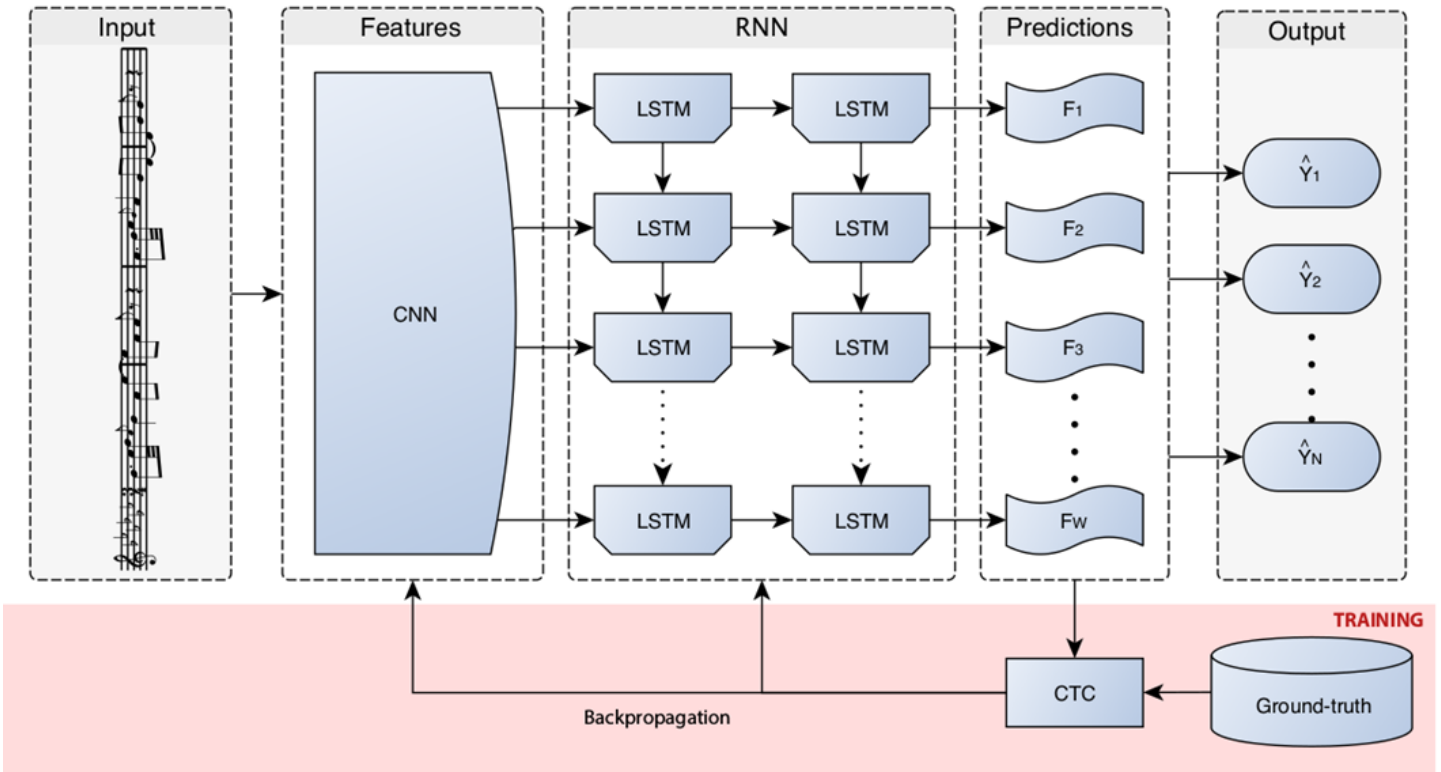
OMR full-pipeline approaches



OMR end-to-end approaches

Formulate the task without any further subdivision

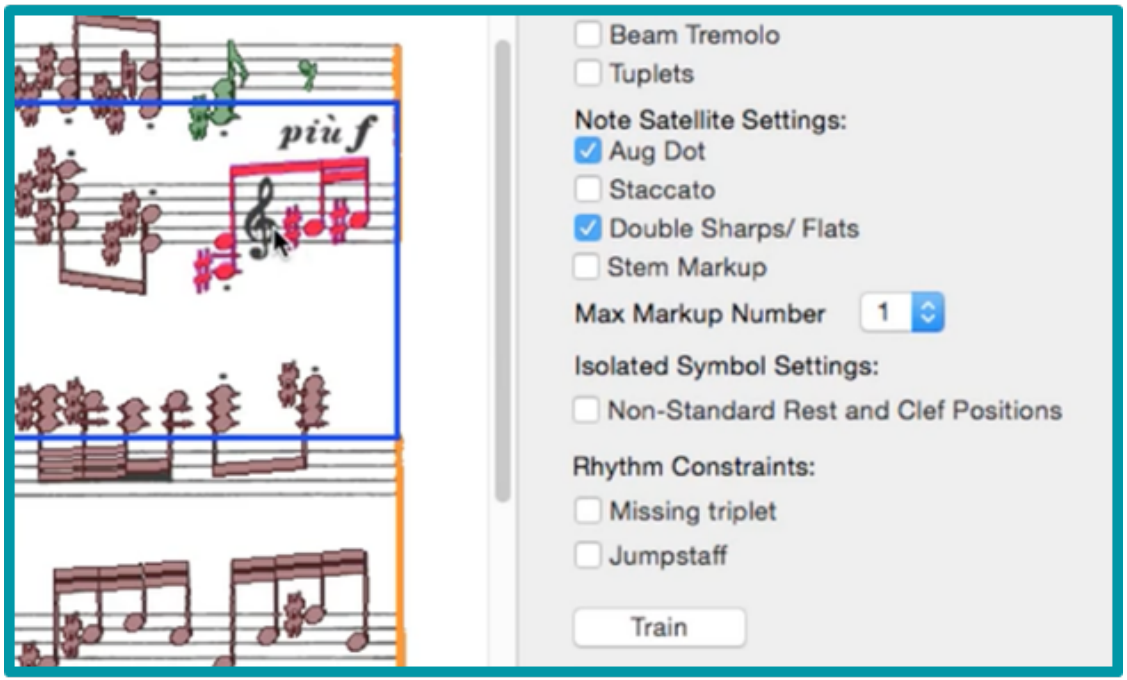
However, current models can only support tasks whose output can be expressed as a unidirectional sequence (e.g. single staff)



(Calvo-Zaragoza et al., 2017)

OMR interactive approaches

- OMR systems are far from being error-free.
- If perfect recognition is pursued, users must correct remaining errors.
- Interactive approaches integrate the user in the recognition loop → corrections are used to dynamically improve the recognition model.



(Chen et al., 2017)

OMR - conclusions

- OMR is becoming useful for more and more practical purposes
- The OMR community is growing (datasets, code, dedicated workshop – WoRMS)
- Machine learning methods make OMR methods transfer easier to new use-cases
- Open problems: domain / manuscript adaptation, integration into musicological workflows

`https://omr-research.net/`

Lyrics alignment and transcription

Lyrics in music informatics

Lyrics are one of the most important aspects of singing:

- The lyrics of a song represent its theme and story and are essential for creating an impression of the song.
- Knowing part of the lyrics of a song can help identify the song and its composer/songwriter.

Lyrics in music informatics:

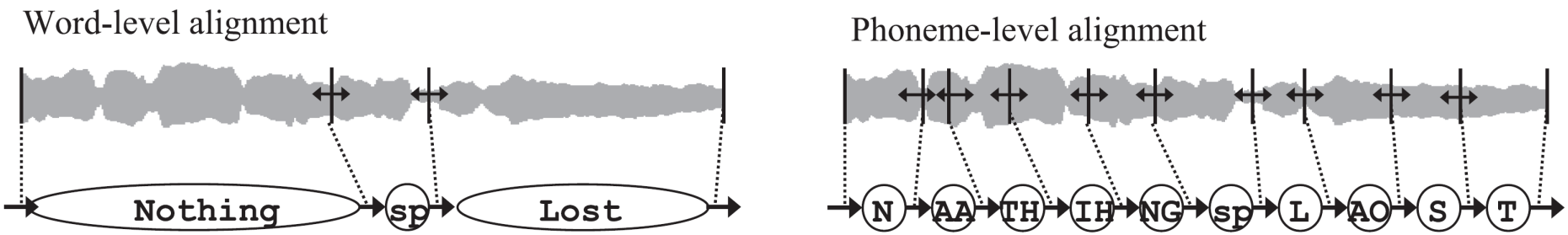
- Automatic search/indexing of music using lyrics
- Automatic recognition of lyrics from audio
- Automatic synchronisation of lyrics with audio ("lyrics-to-audio alignment")
- Music retrieval/navigation using query-by-singing

Lyrics-to-Audio Alignment

Problem Definition

Given audio signals of singing voices and corresponding textual lyrics as input data, lyrics-to-audio alignment can be defined as the problem of estimating the temporal relationship between them.

The start and end times of every block of certain length in lyrics are estimated; a block could be a phoneme, syllable, word, phrase, line, or paragraph.



Lyrics-to-Audio Alignment

The problem of lyrics-to-audio alignment bears a relationship to text-to-speech alignment used in spoken language technologies, which is generally conducted by using forced alignment techniques (e.g. DTW, HMMs).

Challenges:

- Fluctuation of acoustic characteristics: singing voice has a wider range of pitch and loudness characteristics compared to speech (e.g. phoneme lengths are less predictable in singing).
- Influence of accompaniment sounds: singing voice overlaps over frequency and time with accompaniment.
- Incomplete lyrics: available textual lyrics do not always correspond exactly to what is sung in a song (e.g. repetitions, utterances).

Lyrics-to-Audio Alignment

Lyrics-to-audio alignment systems are language-specific; most research has been carried out for lyrics in English, Cantonese, and Japanese.

Commonly used features:

- Mel-frequency cepstral coefficients (MFCCs)
- F0 contours
- Phoneme durations

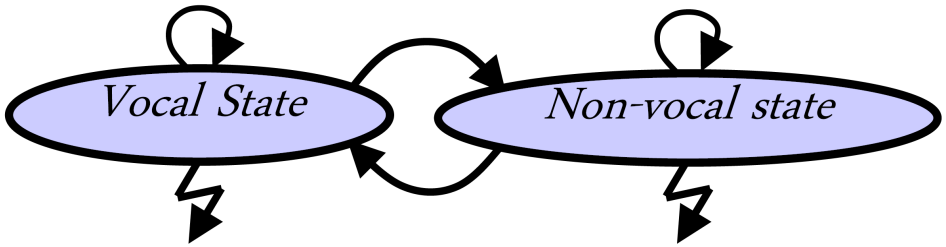
Additional information used to improve alignment:

- Singing voice detection
- Song structure
- Onset/beat detection

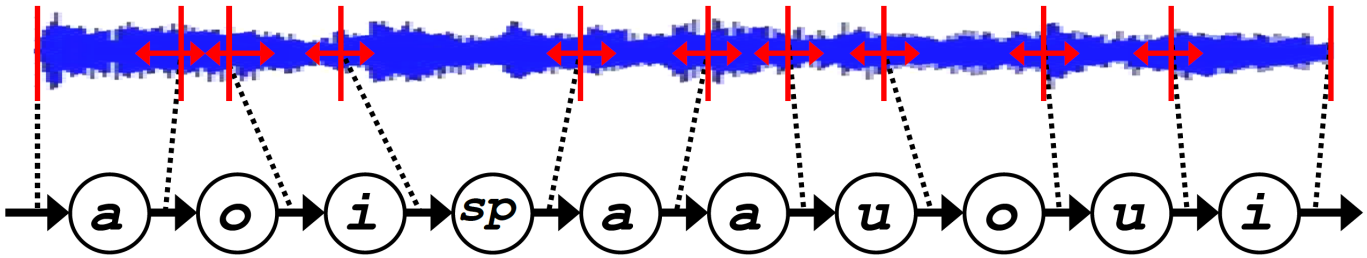
Lyrics-to-Audio Alignment

Case study: LyricSynchronizer (Fujihara and Goto, 2011)

1. Vocal extraction (using melody extraction algorithm)
2. Singing voice detection using HMMs



3. Phoneme-to-audio alignment using Viterbi encoding, combining information from lyrics and resynthesized vocals



Lyrics-to-Audio Alignment - Conclusions

It is possible to align lyrics and audio with satisfactory accuracy for songs in which words are pronounced clearly and the sounds of vocals are mixed louder.

Areas of active research:

- Integrating content-based music analysis methods to improve alignment performance (e.g. singer identification, style/genre classification).
- Joint inference of music/singing cues as opposed to linear pipeline (e.g. Stoller et al, 2019)
- Utilising additional sources of external information (e.g. score-informed lyrics-to-audio alignment)

Lyrics Transcription

- Most computational research on lyrics has focused on audio-to-lyrics alignment.
- There are however cases where lyrics transcription is needed, e.g. when lyrics are not available or to aid an alignment-based system when untimed lyrics are inaccurate.
- The problem of lyrics transcription is analogous to that of automatic speech recognition (ASR) in spoken language; most systems for lyrics transcription are adapted from ASR systems.

Lyrics Transcription

Differences between speech and singing:

- In singing, intelligibility is often secondary to the intonation and musical qualities of the voice.
- Vowels are sustained much longer in singing than in speech.
- In speech, pitch/loudness changes express emotions; in singing, the singer is required to control the pitch, loudness, and timbre.
- The pitch range in singing is usually higher than in speech; however, in speech the pitch varies all the time, whereas in singing it stays approximately constant during a note.

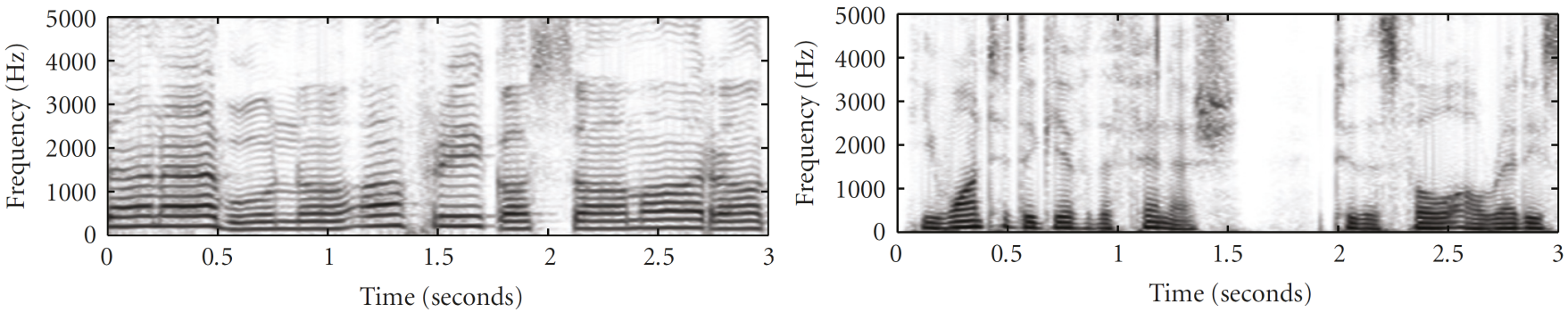


Figure: Example spectrograms of singing (left) and speech (right).

Lyrics Transcription

Benchmark lyrics transcription system (Mesaros and Virtanen, 2011)

1. Feature extraction: Mel-frequency cepstral coefficients (MFCCs) & delta-MFCCs

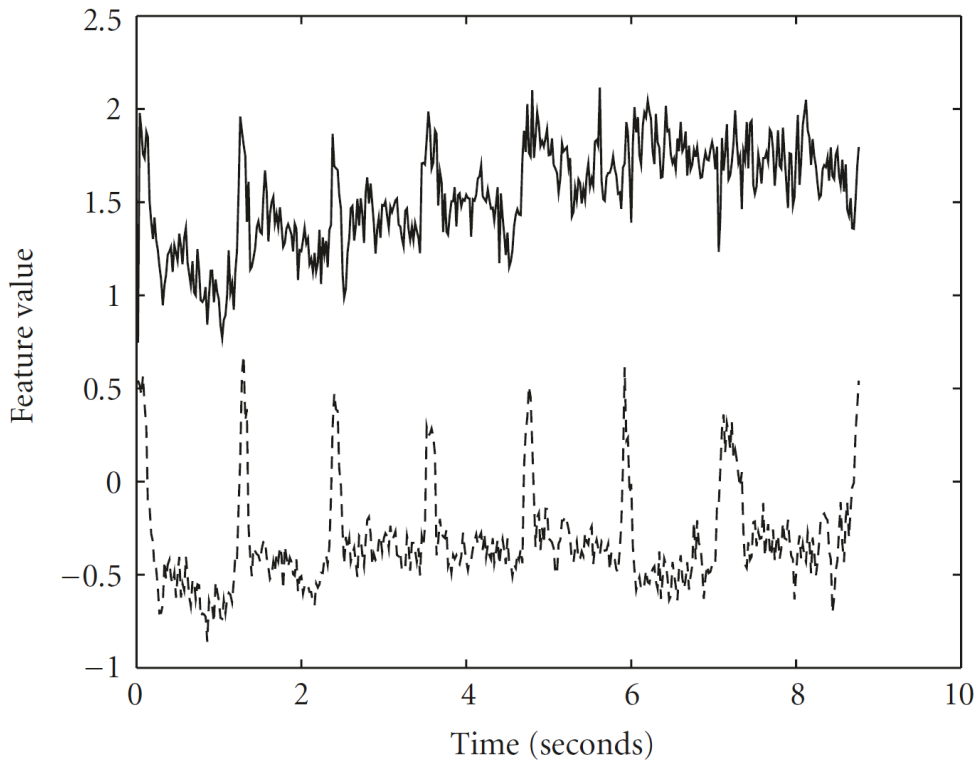
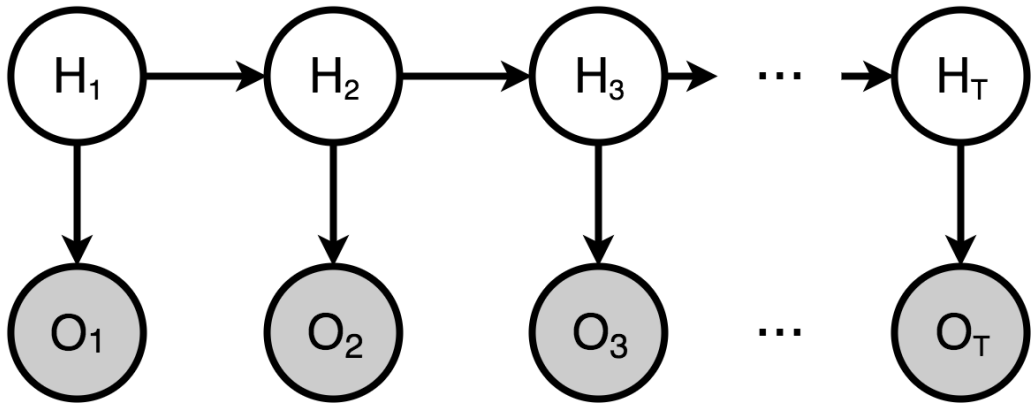


Figure: MFCC features calculated from a descending scale of notes from G#4 to F#3 with the phoneme /m/. Solid line: 3rd MFCC; dashed line: 7th MFCC.

Lyrics Transcription

2. Phoneme recognition: GMM-HMM (Gaussian Mixture Model - Hidden Markov Model) - one model per phoneme, plus models for silence and short pauses.



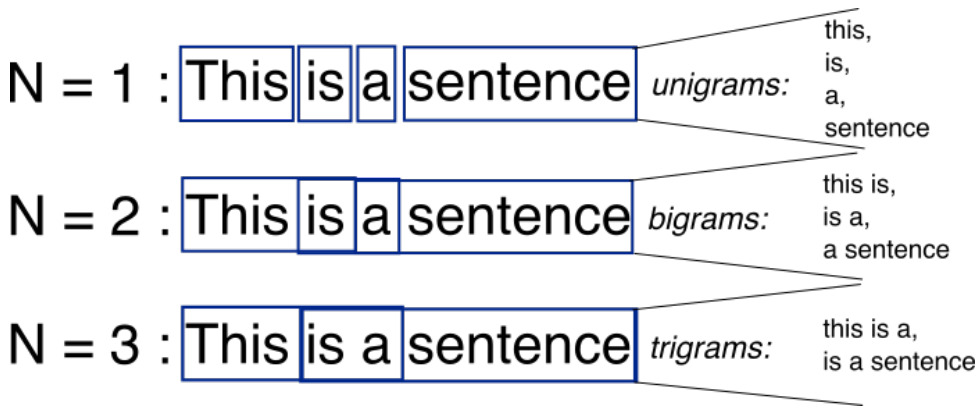
3. Adaptation to singing: training models for speech data and then adapting model parameters using a small set of singing data.

Lyrics Transcription

4. Language modelling: provides a probability over sequences of units (phonemes, letters, syllables, words). Common language models include n-grams and HMMs.

   n-gram: models the probability of observing the sequence of units $\{w_1, \ldots, w_M\}$ as:

$$P(w_1, \ldots, w_M) = \prod_{i=1}^{M} P(w_i | w_{i-(N-1)}, \ldots, w_{i-1})$$



Lyrics Transcription - Conclusions

- Open and active area of research
- Increasing number of works utilising musical knowledge, e.g. McVicar et al, 2014 - using repetitions to improve lyrics recognition performance.
- Cross-cultural studies for lyrics transcription (e.g. Gupta et al., 2018)
- Deep learning-based methods for automatic speech recognition are being adopted for lyrics transcription (e.g. Demirel et al., 2020)