AMT RESEARCH

# IEEE Volume 36 January

# Automatic Music Transcription

## Overview

The nature of music signals, which often contain several sound sources that are highly correlated over both time and frequency, AMT is still considered an open problem in the literature. Usually and AMT system takes an audio waveform as input, computes a time-frequency representation and outputs pitches over time or ideally a typeset music score. Percussion and unpitched sounds will be outside the scope of this research. [1], [2]

## Applications

A successful AMT system would aid in musical education, music creation, music production, music searching and musicology. It is considered a fundamental problem in the field and is closely related to audio source separation [3] and music information retrieval [4] because knowing the contents of the piece can greatly assist with these tasks. It provides the main link between music signal processing and symbolic music processing.

AMT has close relations with speech processing as both tasks involve converting acoustic signals to symbolic sequences. [5] Both disciplines benefit from language modelling components that are combined with acoustic components. The methodologies used are also very similar in both fields. [6] One of the key differences of the two disciplines are that musical sources are highly correlated in time and in frequency.

Furthermore, AMT is related to image processing and computer vision as musical objects can be recognized in two-dimensional time-frequency representations.

## Key Challenges

1. Polyphonic music contains a mixture of simultaneous sources. Inferring musical attributes from the mixture signal is an underdetermined problem
2. Overlapping sound events exhibit harmonic relations with each other. For any consonant musical interval, the fundamental frequencies form small integer ratios, so that their harmonics overlap in frequency.

For example the fundamental frequency ratio of its three notes C:E:G is 4:5:6.

1. Statistical independence of sources does not apply in music source separation, due to the synchronization of onsets and offsets between different voices.
2. The annotation of ground truth transcriptions for polyphonic music is time consuming and requires high expertise. The lack of such annotations has limited powerful supervised learning techniques to specific contexts. There are some approaches to circumvent this problem [7] but they require professional music performers and thorough pre- and postprocessing. It is also noted sheet music is often considered a weak label for a number of reasons: they are not time-aligned to the audio signal, there are different versions and interpretations of musical pieces.

## Commercial AMT software

* Melodyne - <http://www.celemony.com/en/melodyne>
* AudioScore- <http://www.sibelius.com/>
* ScoreCloud- <http://scorecloud.com/>
* AnthemScore - <https://www.lunaverus.com>
* Transcribe! - <https://www.seventhstring.com/xscribe/>

## Issues with current AMT systems

* Octave errors
* Semitone errors
* Missed notes
* Extra notes
* Merged/fragmented notes
* Incorrect offsets/onsets
* Misassigned streams

## Overview of AMT methods

Most approaches are designed to achieve an intermediate goal in AMT which does not actually resemble musical notation.

### MPE (Multiple Pitch Estimation/ Frame Estimation)

### 

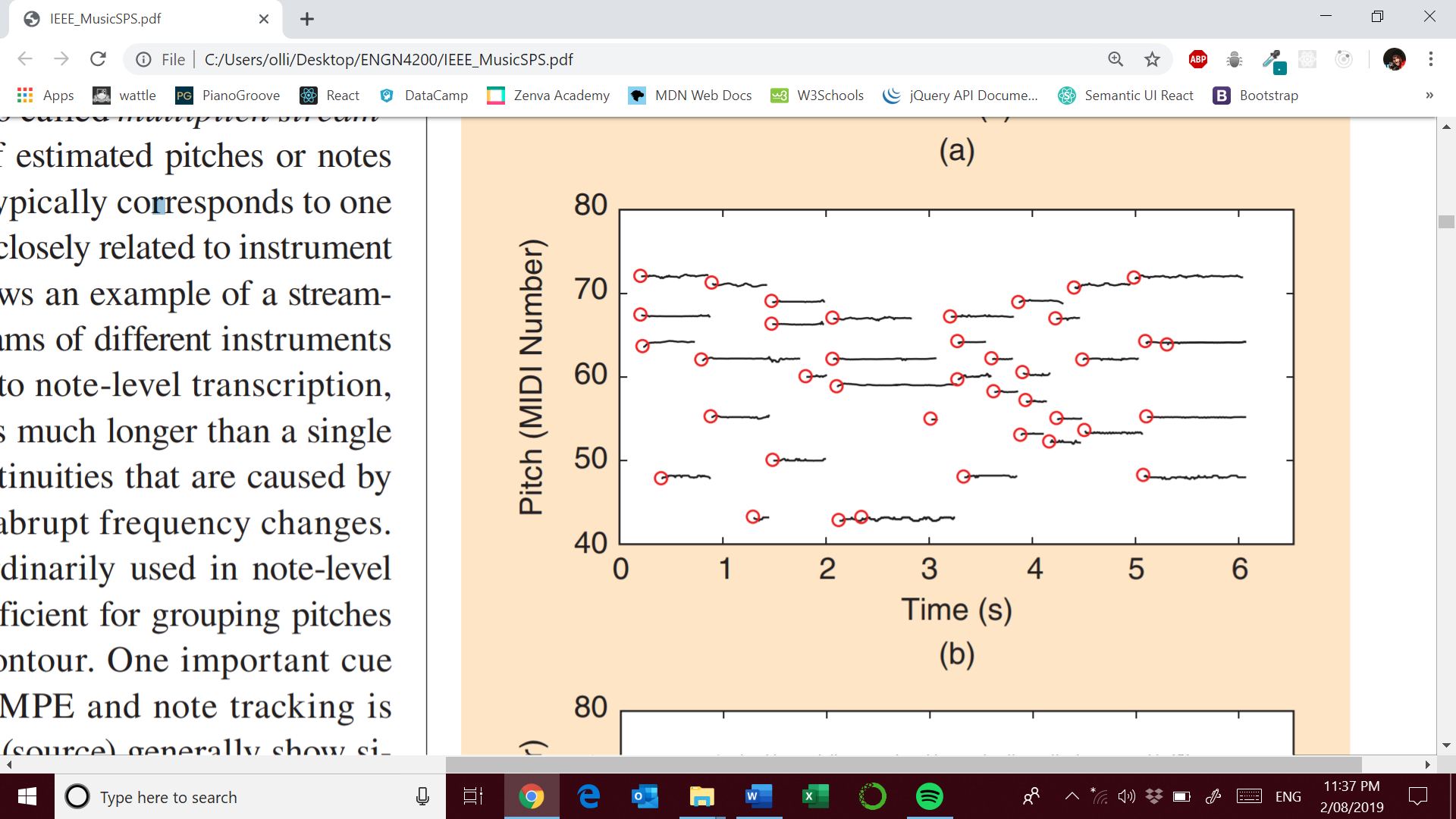
The estimation of the number and pitch of notes that are present in each time frame ( ~10ms). This is usually performed independently in each frame although contextual information is considered in filtering estimations in a post processing stage. A number of approaches operate at this level including:

* **Traditional Signal processing methods** [11], [12]
  + Simple and fast and generalize better to different issues
* **Probabilistic modeling** [8]
* **Bayesian approaches** [13]
  + Comprehensive modeling of the sound generation but can be very slow and complex
* **NMF** [14] – [17]
* **Neural Networks** [18],[19]
  + High accuracy on specific instruments

For **comparison of the models** one can refer to the **Multiple Fundamental Frequency Estimation and Tracking task** <https://www.music-ir.org/mirex/wiki/MIREX_HOME>

These types of methods however, do not form concepts of musical notes and rarely model any high-level musical structures.

### Note level transcription



This level of transcription is one level higher then MPE but also considers pitch estimates over time into notes. In the literature a musical note is characterized by three elements: pitch, onset time, and offset time [1]. As note offsets can often be ambiguous they are often neglected in the evaluation of note-tracking approaches. Many note tracking approaches form notes by post processing singular MPE outputs. Some of the techniques used in this context that **do not consider note interactions**:

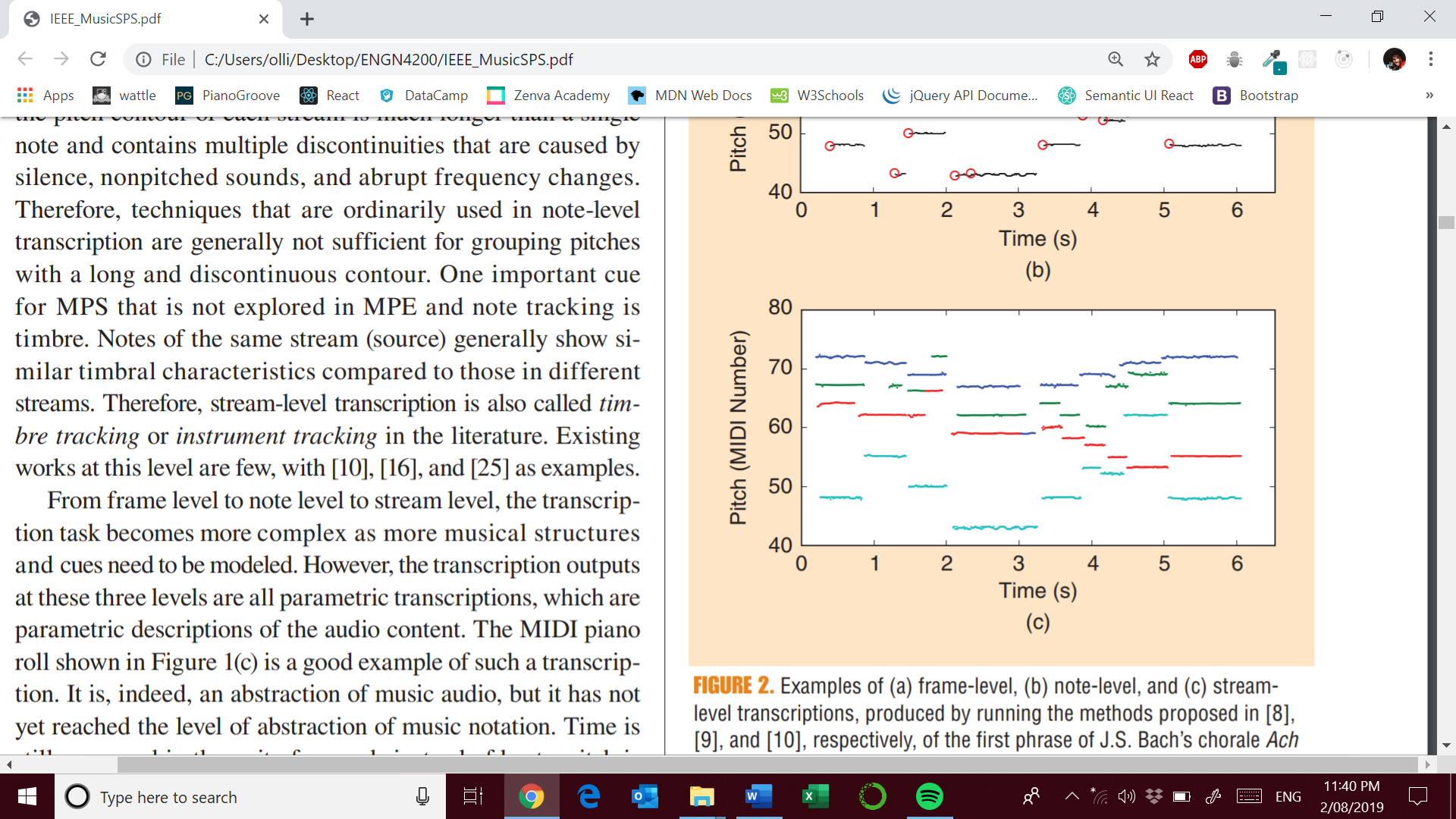
* **Median filtering** [12]
* **Hidden Markov models** [20]
* **NNs** [5]

As note interactions are not considered there are often spurious or missing notes that share harmonics with correctly estimated notes.

Some approaches have been considered that **do consider note interactions**:

* Spectral likelihood model [9] and music language models (MLM) [5], [18]
* Note detection onsets with post processing for pitch estimation within each interonset interval [21]
* Estimation of pitch, onset and offset in the same framework [22]-[24]

### Stream-level transcription/ Multipitch streaming



Works by grouping estimated pitches into streams/ instrument voices. The advantage of this level of transcription is that timbral characteristics are considered. Existing works in instrument tracking are limited: [10],[16],[25]

### Notation-level transcription

Concepts of beat, bar, meter, key and harmony are lacking in the approaches thus far considered. MIDI pitch outputs do not consider these concepts. The next level is to transcribe the audio into human readable musical scores. Transcription at this level requires deep understanding of musical structures. Some works in this field based upon analysis of MIDI outputs are listed below:

* Pitch spelling [26]
* Timing quantization [27]
* Voice separation [28]

Little work has been done on integrating these structures into a complete musical notation transcription especially for **polyphonic music**.

There are several packages that provide this functionality, but the results are unsatisfying:

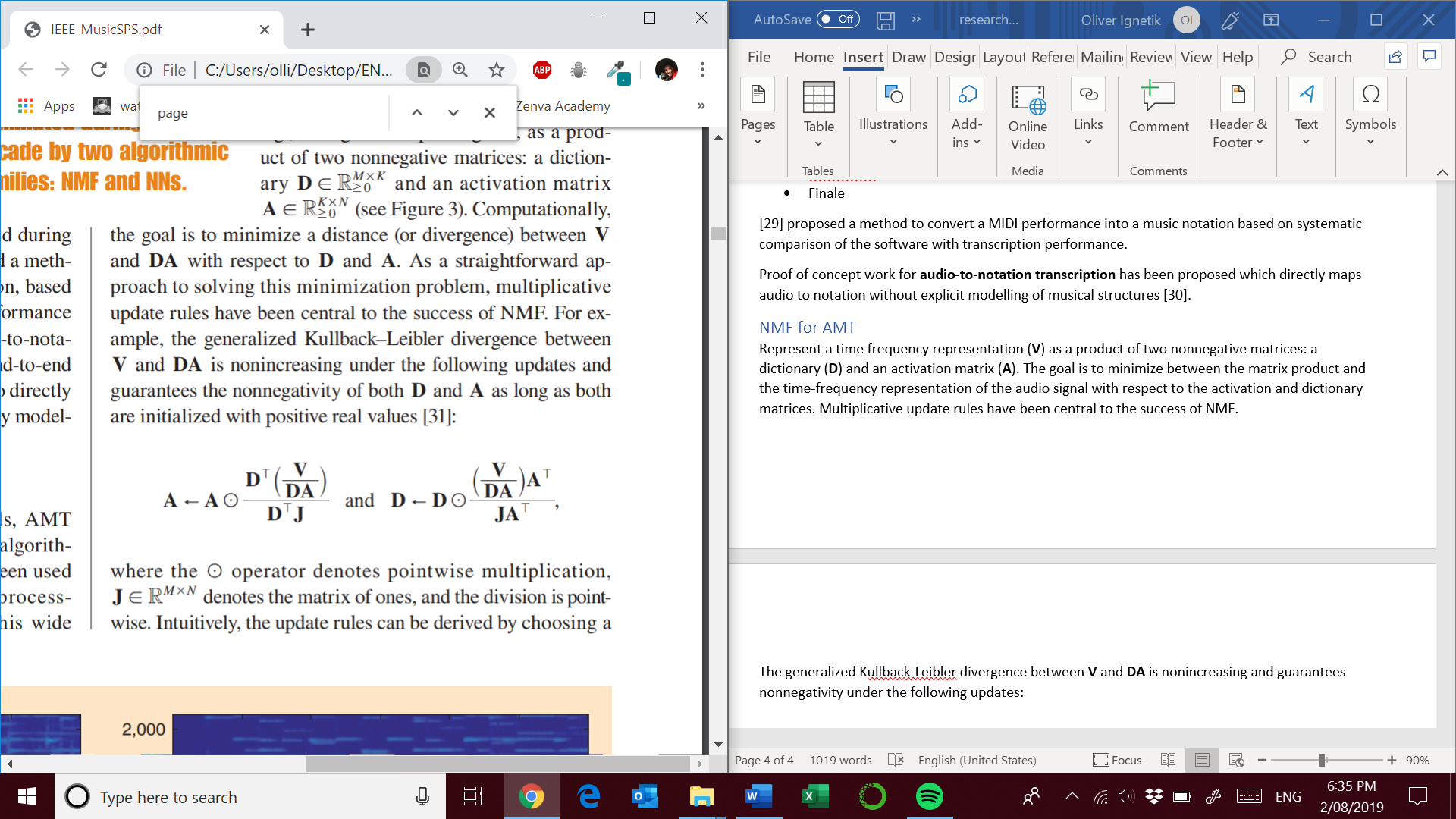
* GarageBand
* MuseScore
* Finale

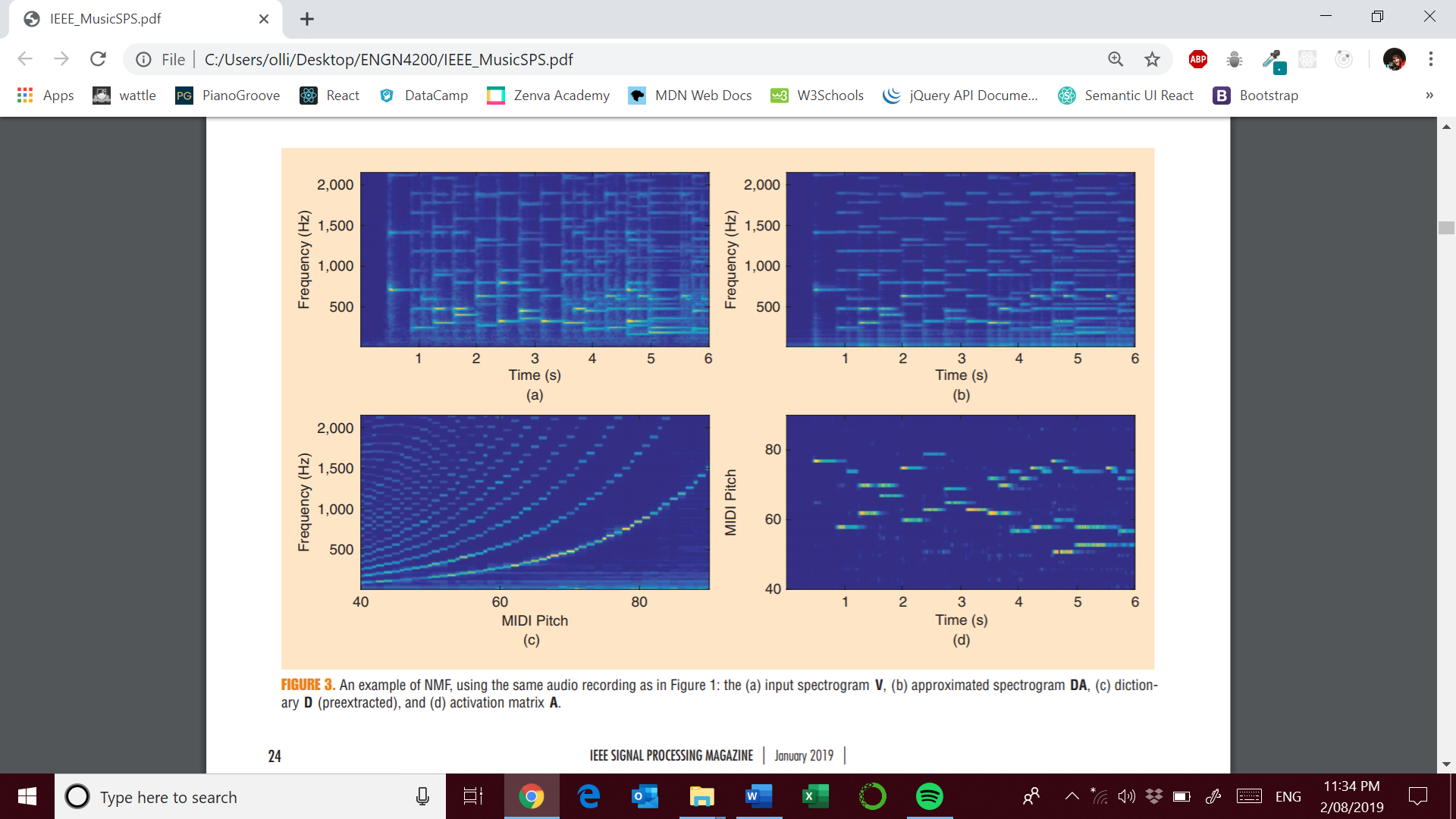
[29] proposed a method to convert a MIDI performance into a music notation based on systematic comparison of the software with transcription performance.

Proof of concept work for **audio-to-notation transcription** has been proposed which directly maps audio to notation without explicit modelling of musical structures [30].

# NMF for AMT

Represent a time frequency representation (**V**) as a product of two nonnegative matrices: a dictionary (**D**) and an activation matrix (**A**). The goal is to minimize between the matrix product and the time-frequency representation of the audio signal with respect to the activation and dictionary matrices. Multiplicative update rules have been central to the success of NMF.

The update rules can be derived by choosing a specific step size in a gradient descent-based minimization of the divergence [31]. One such solution is based on a Kullback-Leibler divergence between **V** and **DA** whichis **nonincreasing** and guarantees **nonnegativity** under the following updates provided both are initialized with positive real values: 

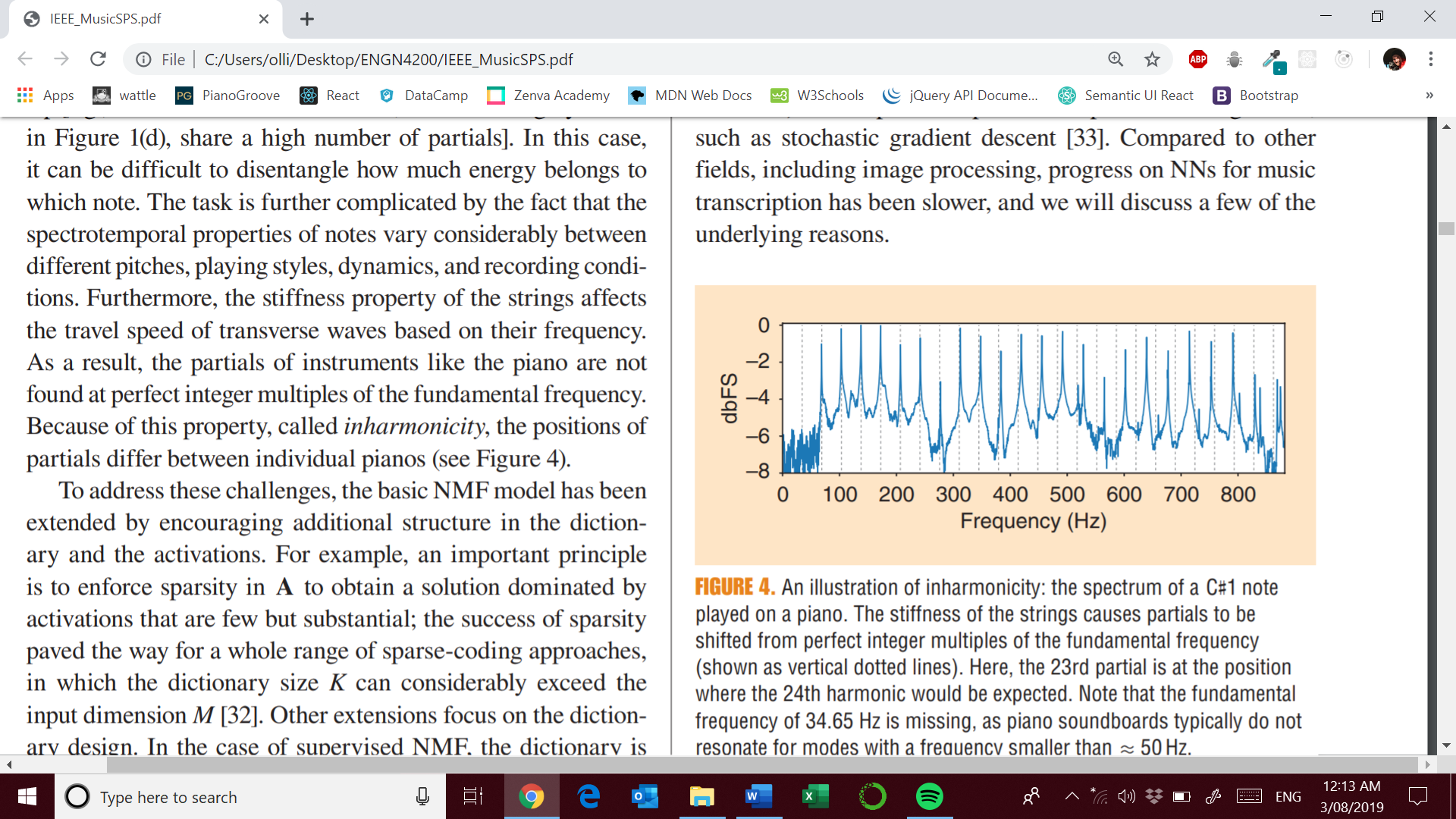


In the context of AMT:

* The nth column of **V** (the spectrum at the time point n) is modeled as a linear combination of the K columns of **D**, and the corresponding coefficients are given by the nth column of **A**.
* Each column of **D** is the spectral template/ expected spectral energy distribution with a specific note
* For each template, the corresponding row in **A** is referred to as the associated activation. It encodes when and how intensely the note is played over time.
* NMF yields a purely constructive representation where spectral energy modeled by one template cannot be canceled by another [31].

## Difficulties in AMF

* Sound objects are highly correlated. **Consonant intervals** are frequently found in most pieces of music. Consonant intervals share many **overlapping partials** which makes disentangling the mixture signal more difficult. How do you determine how much energy belongs to which note?
* Spectro temporal properties of notes vary considerably between different pitches, playing styles, dynamics and recording conditions. Due to the stiffness properties of strings the partials of different instruments are not found at perfect integer multiples of the fundamental frequency. This is known as **inharmonicity**.



## Possible solutions

* Enforcing sparsity in the activation matrix **A** to obtain a solution dominated by activations that are few but substantial. [32]
* Supervised NMF uses a precomputed dictionary by extracting templates from recordings of each note thus ensuring the absence of interference.
* Representing each NMF template as linear combinations of **fixed narrow-band subtemplates** [15] which **enforces harmonic structure** for all NMF templates
* **Shift-invariant dictionaries** use single templates to represent a **range of different fundamental frequencies**. Using a logarithmic frequency axis, the distances between individual partials are fixed. Shifting the template in frequency allows the modelling of sounds of varying pitches. Sharing parameters in this way has led to increased model capacity [16], [17].
* **Spectrotemporal dictionaries** allow the modelling of temporal evolution of notres using a **Markov process** which governs the use of different subsets of templates. [16], [23]

# NNs for AMT

Some of earliest approaches based on NNs was **Marolt’s Sonic system** [21]. A central component of this approach is to use **time-delay networks** which resemble convolutional networks in the time direction [33]. They are employed to analyze the output of **adaptive oscillators** to track and group partials in the output of a **gammatone filterbank**. This approach is very competitive and still appears in comparisons [23].

Recently Bock and Schedl [34] used two spectrograms as input to enable their network to exploit high time and frequency accuracy. The input is processed using multiple long short-term memory layers [33]. There are several benefits of using this approach:

* Spectral properties of a note evolve across input frames allowing LTSM networks to model sequences of note evolution.
* Long range dependencies between notes can be captured by predicting popular chord progressions

[18] focused on long-range dependencies by combining an acoustic front end with a symbolic language model. Using information obtained from MIDI files, a **recurrent NN** (RNN) was trained to **predict** the active **notes** in the **next time frame**. This approach needed to learn a very large probability distribution of all the possible notes. To help with modelling such an enormous probability space they employed a **NADE architecture** which represented the distribution as a long product of conditional probabilities. This approach has been popularized recently by the **WaveNet architecture**. Only modest improvements over HMM baselines where observed thus this approach remains an **open question.**

Kelz et al. [19] focuses on acoustic modelling, by reporting on the results of a larger scale hyperparameter search and describing the influence of individual system components. Trained using this procedure this models considerably outperformed existing models.

The current **state of the art method** for general purpose piano transcription was proposed by Google Brain [24]. In this approach **one network detects note onsets** which informs a second **network which focuses on note lengths**. This effectively splits the complex joint probability distribution into a probability over onsets and a probability over frame activities. The **temporal dynamics** of **onsets and frame activities** are quite **different**.