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 an 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)

### 

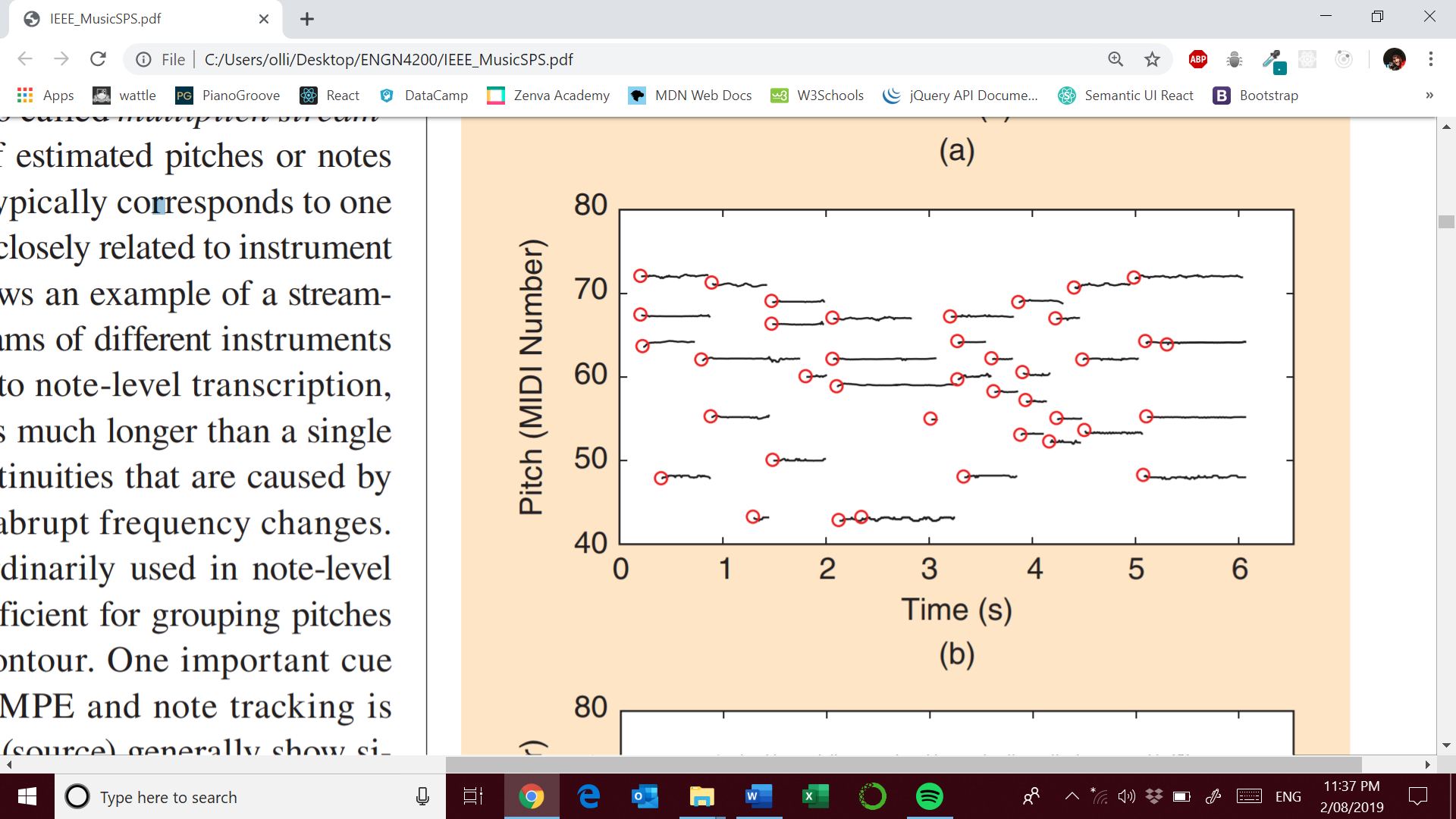
The estimation of the number and pitch of notes that are present in each time frame ( ~10ms). This is usually performed independently in each frame although contextual information is considered in filtering estimations in a post processing stage. A number of approaches operate at this level including:

* **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**:

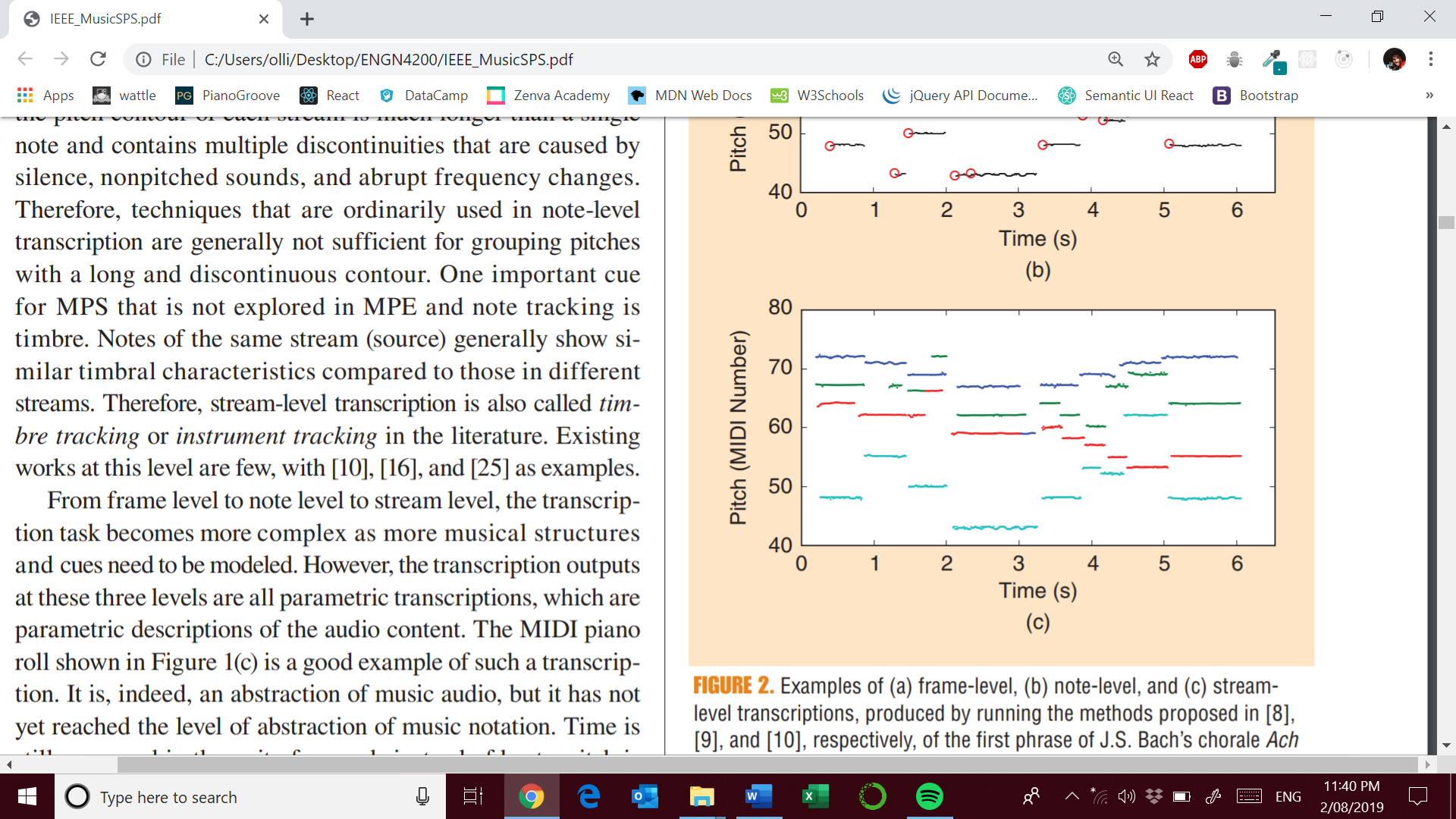
* **Median filtering** [12]
* **Hidden Markov models** [20]
* **NNs** [5]

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:

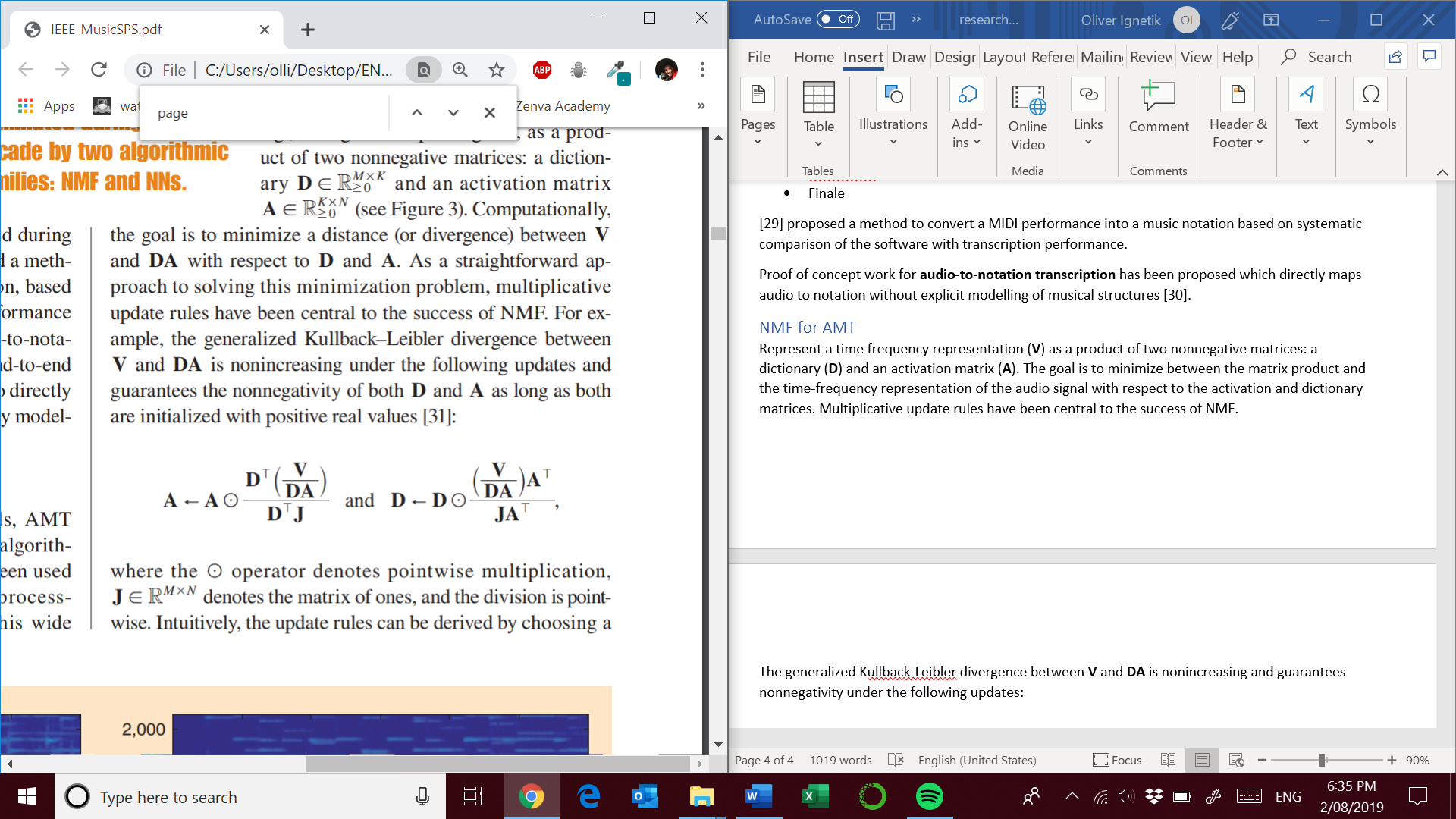
* GarageBand
* MuseScore
* Finale

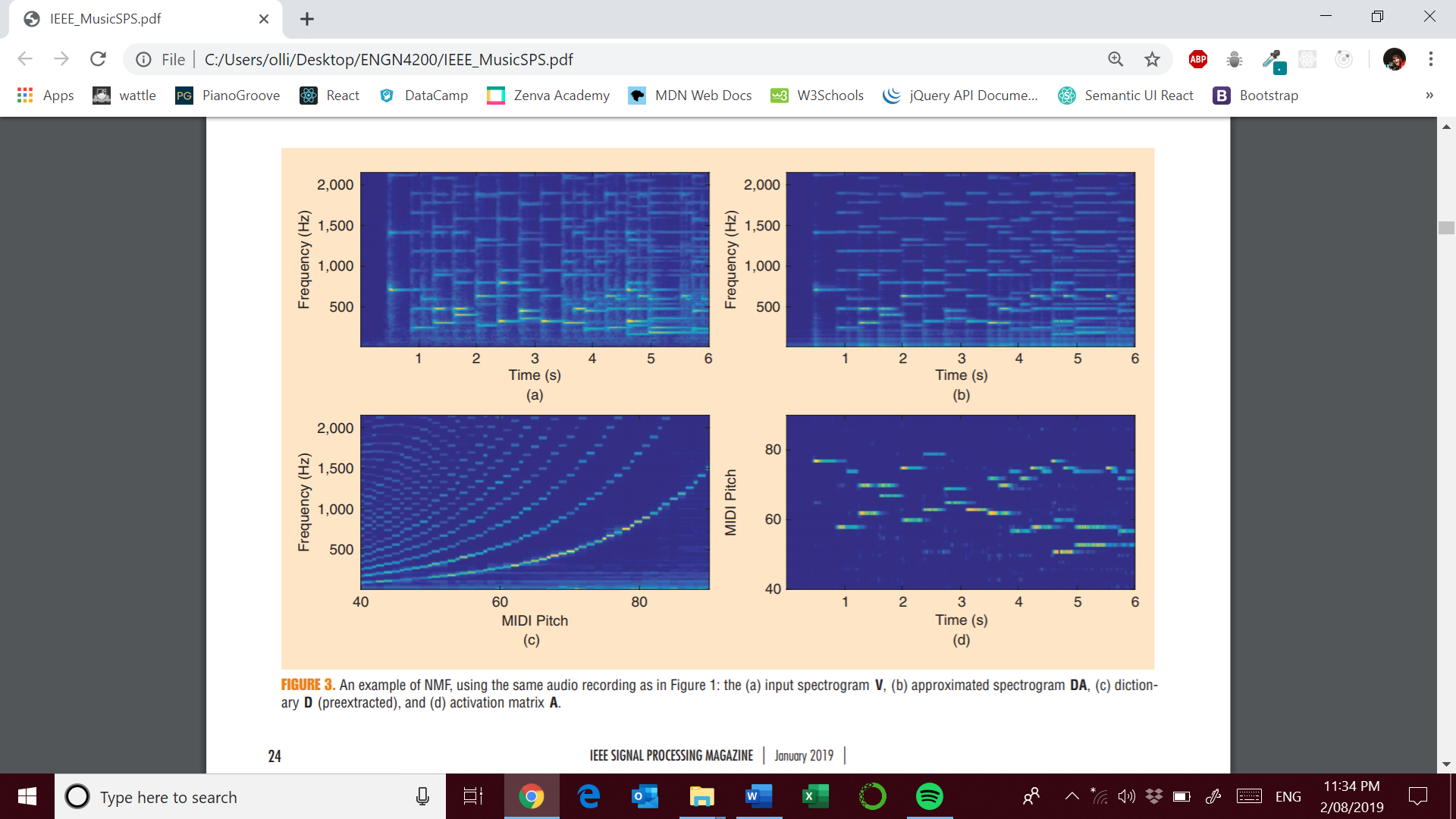
[29] proposed a method to convert a MIDI performance into a music notation based on systematic comparison of the software with transcription performance.

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: 

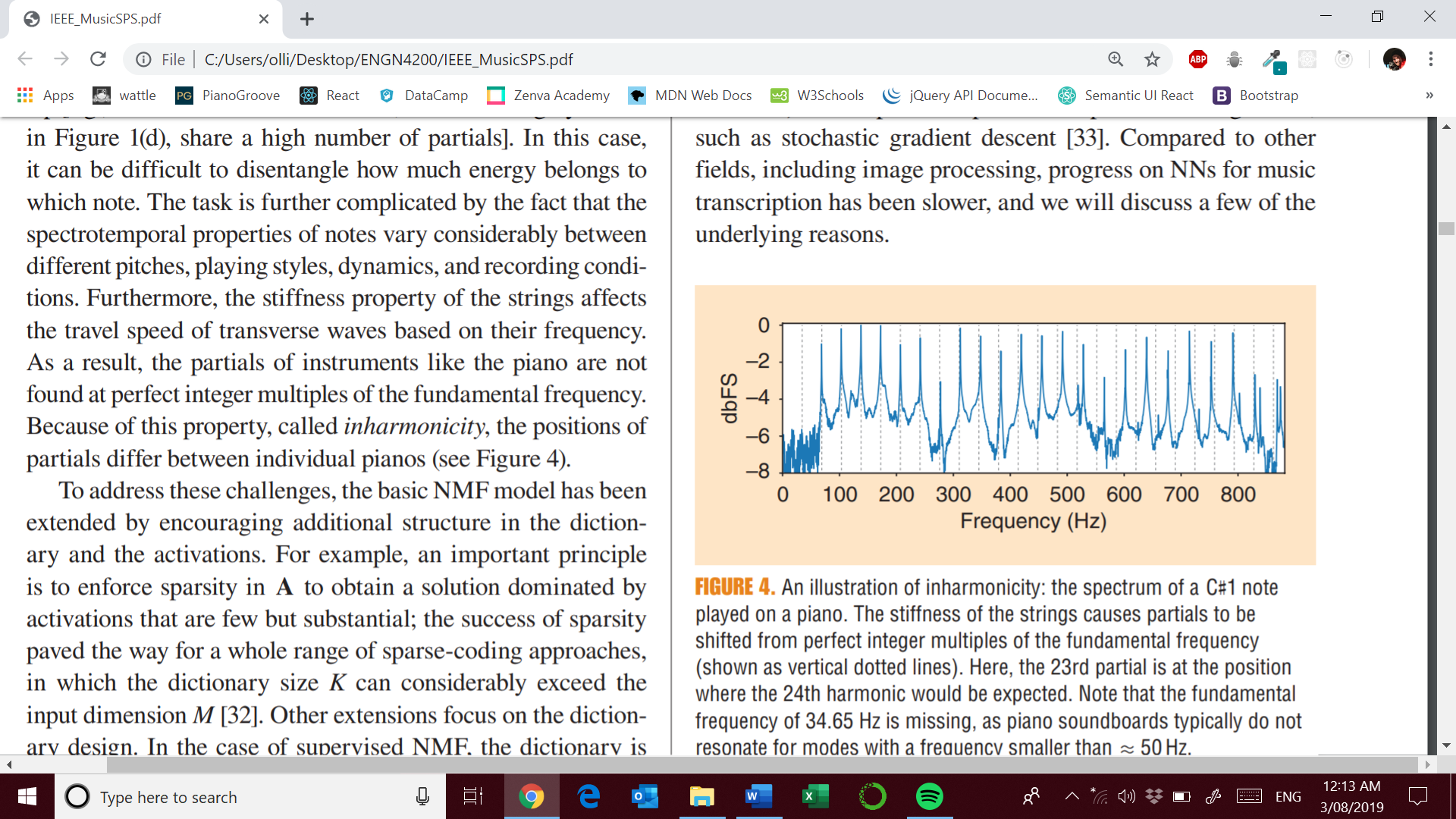


In the context of AMT:

* 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**.

# Comparison of NMF and NN models

NMF models are being increasingly replaced with NN based methods. So how is linearity a constraint?

## NMF

* Linear generative model
* To represent a single pitch, we need to linearly interpolate the given spectral templates in a dictionary which may not represent the real-world recording
* The number of invalid spectra increases much faster than valid spectra as we increase the number of templates.
* They remain competitive or even exceed NN results.
* They do not require large data sets for training.
* Highly adaptable to new acoustic conditions
* Models more spurious notes due to misrepresentation of pitches via interpolation of spectral templates

## DNN

* In recent years deep networks have shown significant potential to represent these template manifolds in a robust and efficient way [33].
* NNs can be trained in an end-to-end fashion, which means that model parameters do not require post processing. **Why do NMF activations require post processing of model parameters ?**
* Require large data sets for training which are subject to bias [7].
* Largest publicly available data set contains over several hours of piano music [11]. However, only a small number of different instruments are used thus overfitting becomes a problem
* A lot of instruments don’t even have any available datasets.
* Other biases include harmonic and rhythmic tendencies
* Poor adaptability to new acoustic conditions. Providing just a few examples of isolated notes of the instrument to be transcribed NMF models perform considerably better. This means the error rate for nonadopted networks can be an order of magnitude higher than that of an adapted NMF system. **Is this fair to compare the methods in this fashion?**
* LSTM layers can learn how notes evolve over time thus reducing the number of spurious notes
* Networks can learn the likelihood of occurring notes suggesting the development of a MLM

# Future Work

## MLMs

* Possibility of polyphonic transcription with the use of DL as postprocessing of AMT output Combination of a restricted Bolzmann machine (RBM) with an RNN for language predictions. [5].
* Combination of acoustic and language models using a probabilistic graphical model, showed promising results [18]
* Modelling other musical cues such as chords, key and meter may help improve the language predictions.

## Score informed transcription

* Scores provides a strong prior for the transcription [35]
* Score-to-audio alignment as a preprocessing step to align the music score with the input audio prior to the transcription. **This approach needs generalization to all instruments**
* **Lead sheet** informed transcription is **unexplored**

## Context-specific transcription

* Prior knowledge of the instruments and sound recording environment lead to considerable performance improvements. For context specific piano transcription, multipitch detection accuracy can exceed 90% [22],[23].
* Open work in this field is focused on the creation of context specific AMT systems for multiple instruments

## Non-Western Music

* The vast majority of approaches target Western music with a number of assumptions based on modes and standard tuning frequency.
* This topic remains an open problem with several challenges such as : lack of annotated datasets and the design of music notation and signal representations that avoid Western bias [36]

## Expressive pitch and timing

* Research on AMT has followed the view describing notes in terms of discrete pitches
* **No suitable representation exists for singing** where singers use microtonality and micro timing as a form of expression.
* **Better representations of microtonality** are required which can be reduced to Western score notation

## Percussive sounds

* An **active problem** [1, Ch.5] most often referred to as drum transcription
* Drum sets can consist of a bass drum, hi-hat, cymbals and toms.
* How does one classify the percussive sounds into these classes?
* There are a number of key challenges such as : the presence of several harmonic, inharmonic and nonharmonic sounds in the music signal, need for increased temporal resolution.

## Evaluation metrics

* Most AMT approaches are evaluated using the set of metrics proposed for the MIREX-F0 Estimation and Note Tracking public evaluation tasks (<http://www.music-ir.org/>)
* Three metrics are included: frame based, note based and stream based
* These metrics do not translate to an accurate representation of human perception of transcription accuracy
* **Creation of evaluation metrics** that corresponds with **human perception** is an **open problem**

# Acknowledgements

Emmanouil Benetos ([emmanouil.benetos@qmul.ac.uk](mailto:emmanouil.benetos@qmul.ac.uk))

Simon Dixon ([s.e.dixon@qmul.ac.uk](mailto:s.e.dixon@qmul.ac.uk))

Zhiyao Duan ([zhiyao.duan@rochester.edu](mailto:zhiyao.duan@rochester.edu))

He co-presented a tutorial on automatic music transcription at the International Society for Music Information Retrieval (ISMIR) Conference in 2015. (<http://c4dm.eecs.qmul.ac.uk/ismir15-amt-tutorial/>)

Sebastian Ewert ([sewert@spotify.com](mailto:sewert@spotify.com))

# Useful groups

ISMIR <http://ismir2015.uma.es/index.html>

# References

[1] A. Klapuri and M. Davy, Eds., Signal Processing Methods for Music Transcription. New York: Springer-Verlag, 2006.

[2] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, “Automatic music transcription: Challenges and future directions,” J. Intelligent Inform. Syst., vol. 41, no. 3, pp. 407–434, 2013.

[3] M. Müller, D. P. Ellis, A. Klapuri, and G. Richard, “Signal processing for music analysis,” IEEE J. Sel. Topics Signal Process., vol. 5, no. 6, pp. 1088–1110, 2011.

[4] M. Schedl, E. Gómez, and J. Urbano, “Music information retrieval: Recent developments and applications,” Foundations Trends Inform. Retrieval, vol. 8, pp. 127–261, 2014. doi: 10.1561/1500000042.

[5] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, “Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription,” in Proc. Int. Conf. Machine Learning, 2012, pp. 1159– 1166.

[6] T. Virtanen, M. D. Plumbley, and D. P. W. Ellis, Eds., Computational Analysis of Sound Scenes and Events. New York: Springer-Verlag, 2018.

[7] L. Su and Y.-H. Yang, “Escaping from the abyss of manual annotation: New methodology of building polyphonic datasets for automatic music transcription,” in Proc. Int. Symp. Computer Music Multidisciplinary Research, 2015, pp. 309–321.

[8] Z. Duan, B. Pardo, and C. Zhang, “Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions,” IEEE Trans. Audio, Speech, Language Process. (2006–2013), vol. 18, no. 8, pp. 2121–2133, 2010.

[9] Z. Duan and D. Temperley, “Note-level music transcription by maximum likelihood sampling,” in Proc. 15th Int. Society Music Information Retrieval Conf., 2014, pp. 181–186.

[10] Z. Duan, J. Han, and B. Pardo, “Multi-pitch streaming of harmonic sound mixtures,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 22, no. 1, pp. 138–150, 2014.

[11] V. Emiya, R. Badeau, and B. David, “Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle,” IEEE Trans. Audio, Speech, Language Process. (2006–2013), vol. 18, no. 6, pp. 1643–1654, 2010.

[12] L. Su and Y.-H. Yang, “Combining spectral and temporal representations for multipitch estimation of polyphonic music,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 23, no. 10, pp. 1600–1612, 2015.

[13] P. H. Peeling, A. T. Cemgil, and S. J. Godsill, “Generative spectrogram factorization models for polyphonic piano transcription,” IEEE Trans. Audio, Speech, Language Process. (2006–2013), vol. 18, no. 3, pp. 519–527, 2010.

[14] P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in Proc. IEEE Workshop Applications Signal Processing Audio and Acoustics, 2003, pp. 177–180.

[15] E. Vincent, N. Bertin, and R. Badeau, “Adaptive harmonic spectral decomposition for multiple pitch estimation,” IEEE Trans. Audio, Speech, Language Process. (2006–2013), vol. 18, no. 3, pp. 528–537, 2010.

[16] E. Benetos and S. Dixon, “Multiple-instrument polyphonic music transcription using a temporally-constrained shift-invariant model,” J. Acoust. Soc. Amer., vol. 133, no. 3, pp. 1727–1741, 2013.

[17] B. Fuentes, R. Badeau, and G. Richard, “Harmonic adaptive latent component analysis of audio and application to music transcription,” IEEE Trans. Audio, Speech, Language Process. (2006–2013), vol. 21, no. 9, pp. 1854–1866, 2013.

[18] S. Sigtia, E. Benetos, and S. Dixon, “An end-to-end neural network for polyphonic piano music transcription,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 24, no. 5, pp. 927–939, 2016.

[19] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, “On the potential of simple framewise approaches to piano transcription,” in Proc. Int. Society Music Information Retrieval Conf., 2016, pp. 475–481.

[20] J. Nam, J. Ngiam, H. Lee, and M. Slaney, “A classification-based polyphonic piano transcription approach using learned feature representations,” in Proc. Int. Society Music Information Retrieval Conf., 2011, pp. 175–180.

[21] M. Marolt, “A connectionist approach to automatic transcription of polyphonic piano music,” IEEE Trans. Multimedia, vol. 6, no. 3, pp. 439–449, 2004.

[22] A. Cogliati, Z. Duan, and B. Wohlberg, “Context-dependent piano music transcription with convolutional sparse coding,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 24, no. 12, pp. 2218–2230, 2016.

[23] S. Ewert and M. B. Sandler, “Piano transcription in the studio using an extensible alternating directions framework,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 24, no. 11, pp. 1983–1997, 2016.

[24] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. S. C. Raffel, J. Engel, S. Oore, and D. Eck, “Onsets and frames: Dual-objective piano transcription,” in Proc. Int. Society Music Information Retrieval Conf., 2018, pp. 50–57.

[25] V. Arora and L. Behera, “Multiple F0 estimation and source clustering of polyphonic music audio using PLCA and HMRFs,” IEEE/ACM Trans. Audio, Speech, Language Processing, vol. 23, no. 2, pp. 278–287, 2015.

[26] E. Cambouropoulos, “Pitch spelling: A computational model,” Music Perception, vol. 20, no. 4, pp. 411–429, 2003.

[27] H. Grohganz, M. Clausen, and M. Mueller, “Estimating musical time information from performed MIDI files,” in Proc. Int. Society Music Information Retrieval Conf., 2014, pp. 35–40.

[28] I. Karydis, A. Nanopoulos, A. Papadopoulos, E. Cambouropoulos, and Y. Manolopoulos, “Horizontal and vertical integration/segregation in auditory streaming: A voice separation algorithm for symbolic musical data,” in Proc. Sound and Music Computing Conf., 2007, pp. 299–306.

[29] A. Cogliati, D. Temperley, and Z. Duan, “Transcribing human piano performances into music notation,” in Proc. Int. Society Music Information Retrieval Conf., 2016, pp. 758–764.

[30] R. G. C. Carvalho and P. Smaragdis, “Towards end-to-end polyphonic music transcription: Transforming music audio directly to a score,” in 2017 IEEE Workshop Applications Signal Processing Audio and Acoustics, 2017, pp. 151–155.

[31] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in Proc. Advances Neural Information Processing Systems, 2001, pp. 556–562.

[32] S. A. Abdallah and M. D. Plumbley, “Unsupervised analysis of polyphonic music by sparse coding,” IEEE Trans. Neural Netw. (1990–2011), vol. 17, no. 1, pp. 179–196, 2006.

[33] I. Goodfellow, A. Courville, and Y. Bengio, Deep Learning. Cambridge, MA: MIT Press, 2016.

[34] S. Böck and M. Schedl, “Polyphonic piano note transcription with recurrent neural networks,” in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2012, pp. 121–124.

[35] S. Wang, S. Ewert, and S. Dixon, “Identifying missing and extra notes in piano recordings using score-informed dictionary learning,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 25, no. 10, pp. 1877–1889, 2017.

[36] X. Serra, “A multicultural approach in music information research,” in Proc. 12th Int. Society Music Information Retrieval Conf., 2011, pp. 151–156.