Reference: 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.

Accuracies reported in recent years have reached a limit as can be seen in MIREX. Current methods use general purpose models which are unable to capture the rich diversity found in music signals.

**Ways forward**

* Tailor algorithms to specific use cases
* Semi-automatic approaches like score informed audio aligned

**AMT tasks**

1. MPE – finding multiple f0 in a time frame
2. Note onset/offset
3. Loudness estimation/quantisation
4. Instrument recognition
5. Extraction of rhythmic information
6. Time quantisation

**Methods**

**MPE**

They can either be joint/iterative approaches. Iterative approaches work by extracting the most prominent f0 in each iteration until no additional pitches can be found. Accumulates errors at each iteration step. Joint approaches evaluate f0 combinations at the expense of computational cost. Most approaches are joint estimations.

**Best approaches**

**Reference for evaluation of MiReX**

Bay, M., Ehmann, A.F., Downie, J.S. (2009). Evaluation of multiple-F0 estimation and tracking systems. In 10th int. society for music information retrieval conf. (pp. 315–320).

**FEATURE BASED**

A pitch candidate set score function or pitch salience function is used to estimate f0 from time-frequency representations.

**Best approach (2012)**

Dressler, K. (2012). Multiple fundamental frequency extraction for MIREX 2012. In Music information retrieval evaluation eXchange. http:www.music-ir.org/mirex/abstracts/2012/KD1.pdf.

Based on FFT multi resolution analysis. Each spectral bin is multiplied by the bins instantaneous frequency. Pitch estimation is made by identifying spectral peaks and performing pair-wise analysis. Rank peaks by harmonicity, smoothness, appearance of intermediate peaks and harmonic number. System also tracks tones over time using an adaptive magnitude and a harmonic magnitude threshold.

**STATISTICS BASED**

Viewed as a MAP (maximum a posteriori) estimation problem. Maximum likelihood models based on gaussians that represent the partials.

Synchronous evolution of Gaussian partials is modelled by Gaussian mixtures.

**FACTORISATION BASED MPE**

Incorporated modifications to account for inharmonicity

**NOTE TRACKING**

Onset/offset times considered. Most factorisation method are based around the activation matrix. Some ideas for note pruning:

* If a note is smaller then some threshold then discards
* Rules to address spurious notes such as small gaps

HMM approaches:

* Note inactivity and activity as states
* Need ground truth training sets

Dynamic Bayesian networks:

* Note layer and note combination layers, were the model parameters, were learned using Chopin pieces

**SUBTASKS**

Instrument identification is rendered very difficult by harmonic overlaps and is closely related to sound source separation. Turns into a classification process. A number of studies have been based around this. If classification is performed on individual items they are likely to have better results. Some ideas include :

* Focusing on areas with isolated partials in the spectrogram
* Represent time frequency masks that indicate areas of the spectrogram which belong to single instruments
* Jointly separate and recognize instruments by either utilizing mid-level representation of the signal and modelling it as a sum of instruments and pitch specific active atoms.

Onset detection is the first set to understanding underlying periodicity. Usually performed by checking features like energy levels and steady-state regions.

Offset detection is comparatively less explored. Partials that decay exponentially make the problem more difficult.

Tempo is measured using auto correlation, comb filterbanks, inter-onset interval histograms, Fourier transforms and periodicity transforms. Beat tracking is the task of finding the tempo. Some ideas include :

* Rule based methods
* Adaptive oscillators
* Agent-based or multiple hypothesis trackers

Key harmony and definition tasks are done with HMM and dynamic Bayesian network.

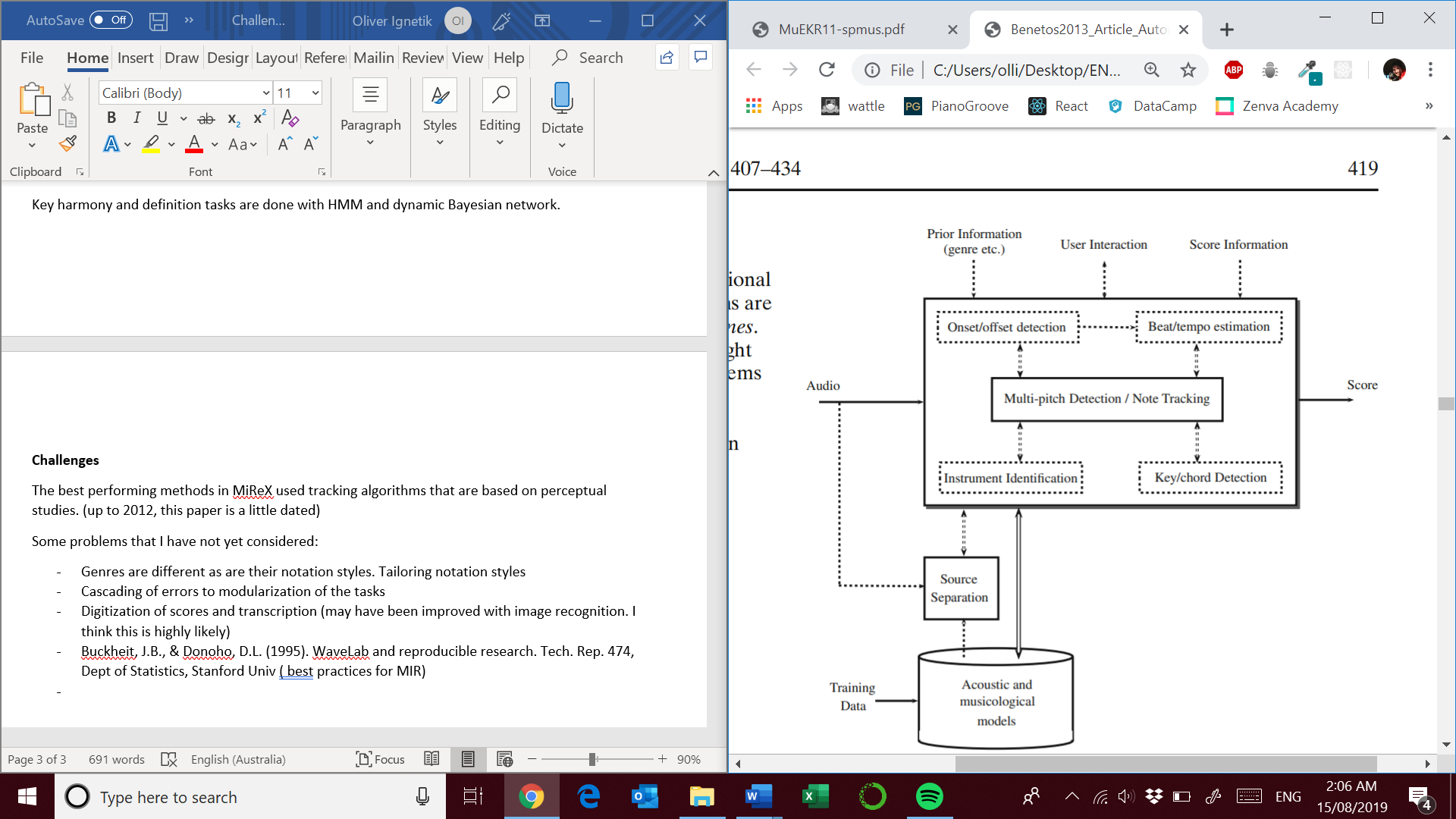
**Challenges**

The best performing methods in MIREX used tracking algorithms that are based on perceptual studies. (up to 2012, this paper is a little dated)

Some problems that I have not yet considered:

* Genres are different as are their notation styles. Tailoring notation styles
* Cascading of errors to modularization of the tasks
* Digitization of scores and transcription (may have been improved with image recognition. I think this is highly likely)
* Buckheit, J.B., & Donoho, D.L. (1995). WaveLab and reproducible research. Tech. Rep. 474, Dept of Statistics, Stanford Univ ( best practices for MIR)

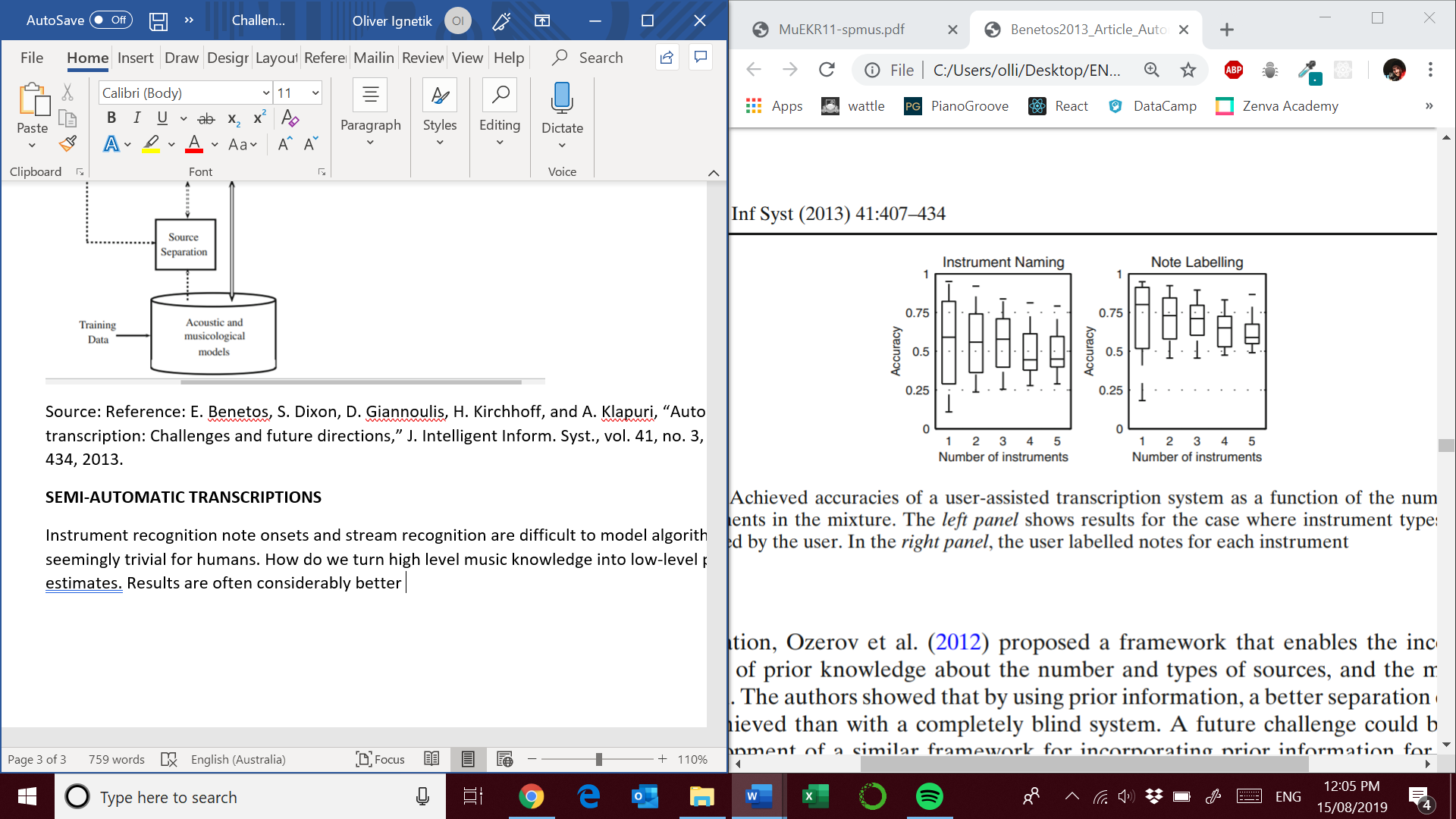
**ARCHITECTURE**



Source: Reference: 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.

**SEMI-AUTOMATIC TRANSCRIPTIONS**

Instrument recognition note onsets and stream recognition are difficult to model algorithmically but seemingly trivial for humans. How do we turn high level music knowledge into low-level parameter estimates. Results are often considerably better



Source: Reference: 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.

Using the labels identified by a user the results were considerably better.

**Score informed**

Automatic music tutors like yousician, fenderplay etc. Automatically aligned scores. Combination of multiple genre/instrument specific systems and filtered based on preliminary classification process.

**Information integration**

Extensive research has been performed into beat tracking and rhythm parsing but it is rarely taken advantage of. Extraction of musically meaningful temporal constraints can increase accuracy. Some approaches :

* 2 dimensional hierarchical tree-structured Bayesian model
* Temporal constraints applied to NMF approach

Key and local harmony can be thought of imposing a probability distribution over notes and chords. Determining the chord given the set of detected notes involves selecting the subset of notes which are deemed to have harmonic function.

**Joint transcription and source separation**

Allow instrument identification to be performed on different streams which generally have some characteristic timbre. Some ideas :

* Include source separation spatial algorithms
* Reference transcriptions of solo instruments could help with source separation after the final mix

**Creating training data**

A large subset of AMT approaches experiments only on piano data because its easier to create recordings with aligned ground-truth using a Diskalvier. Some ideas:

* Crowd sourced transcriptions
* There are not enough ground truth transcriptions for multiple instrument situations
* More ground truth transcriptions would reduce the task of transcription to alignment with existing scores

**Towards a complete transcription**

Most of the transcription methods mentioned to not produce musical notation. Additional issues need to be addressed if this goal is to be achieved:

* Typesetting: are the results satisfying ?
* Estimation of dynamics: score informed scenarios have been proposed, how do we evaluate their performance? This is much more difficult in MIDIs with multiple instruments due to the lack of attack phase information
* Fingering: programming based methods for assigning fingers to notes, work has been done on polyphonic piano and guitar as well as violin to determine most likely finger locations given the input signal
* Expressive notation and articulation: score informed transcription has been done. Used an AMT system before extracting such features such as loudness and timing (rubato, legato)
* Genre specific transcriptions
* Expressing the meter of music signals

Work needs to be done on providing mapping between mid-level features and the expressive marking which are used in a music score.