# Musical Source Separation

Reference: E. Cano, D. Fitzgerald, A. Liutkus, M. Plumbley, F. Stoter, “Musical Source Seperation” in IEEE SPS Journal Vol. 36 January

**Task**

If we only have access to the final recording mix, we must perform musical source separation (MSS) to estimate the original sources. The source of interest is called the **target source.**



Above is an example of a mixture of different sources in a two-channel mixture.

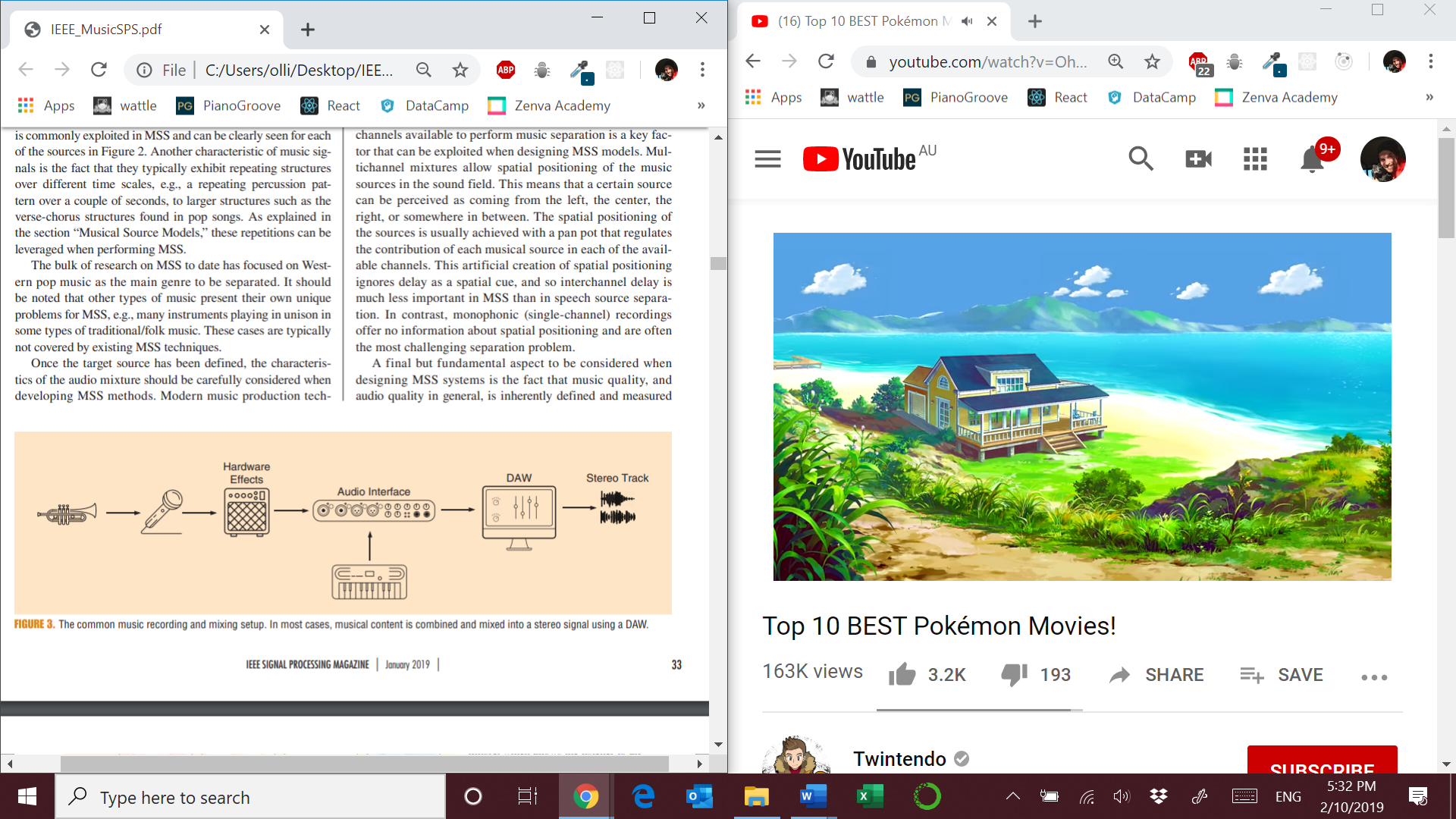
Music sources can be :

* Harmonic – f0 and harmonic series, continuous in time as horizontal components. *Common* *Fate* of harmonic overtones. The way the harmonics evolve, and the relative strengths contribute to the timbre of the instrument.
* Percussive – continuous in frequency, vertical components with transient like characteristics
* Singing voice – combination of harmonic and percussive like components. The components are then filtered by the vocal fold.

**Properties of musical sources:**

* Sparsity – not many points over time-frequency
* Repetition – in rhythm and piece structure

**Typical musical source mixture flow**

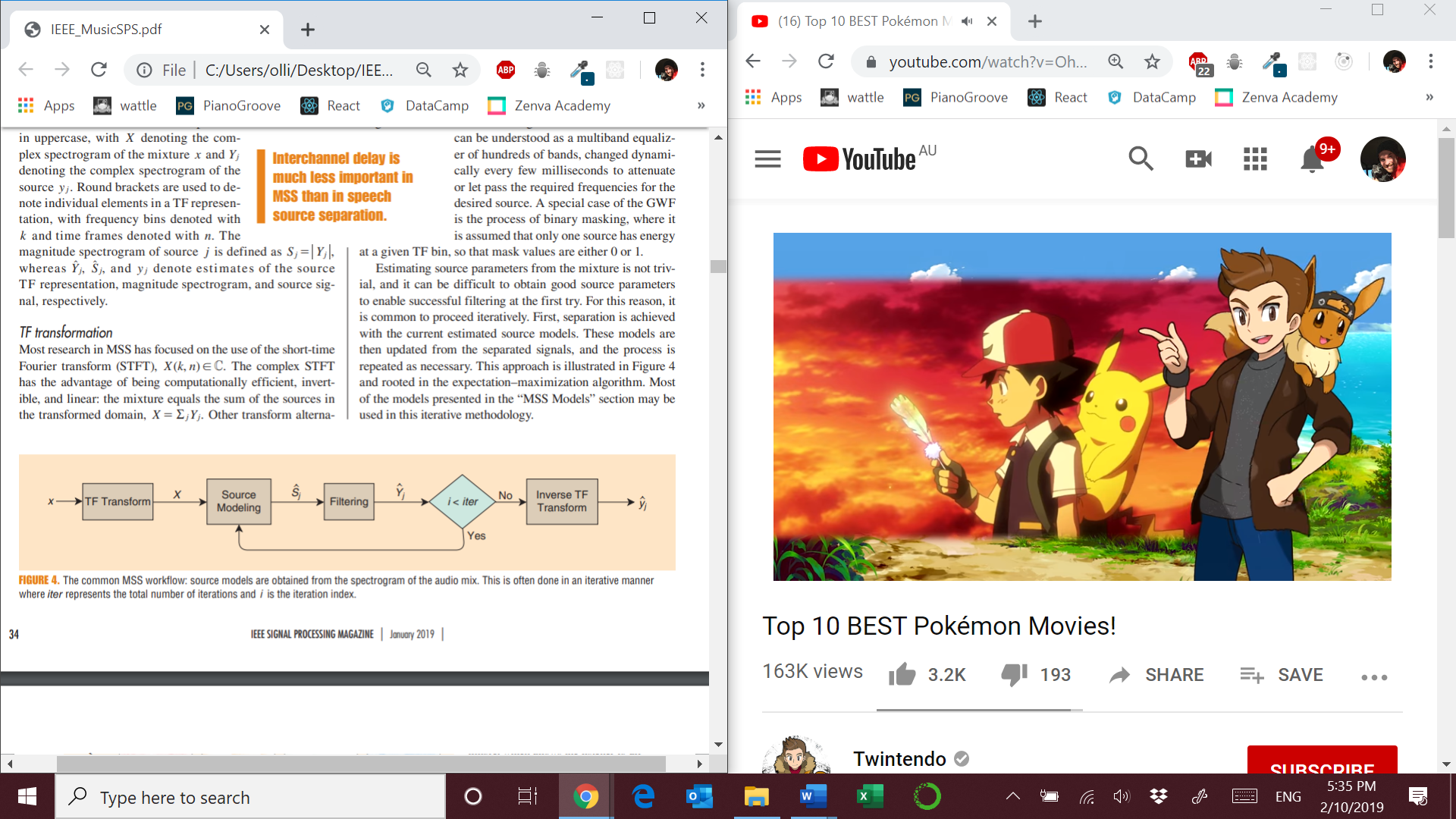


**Musical formats**

* Monophonic – most difficult with no spatial positioning information
* Stereo – 2 channels
* 5.1 – 5 channels + low frequency channel

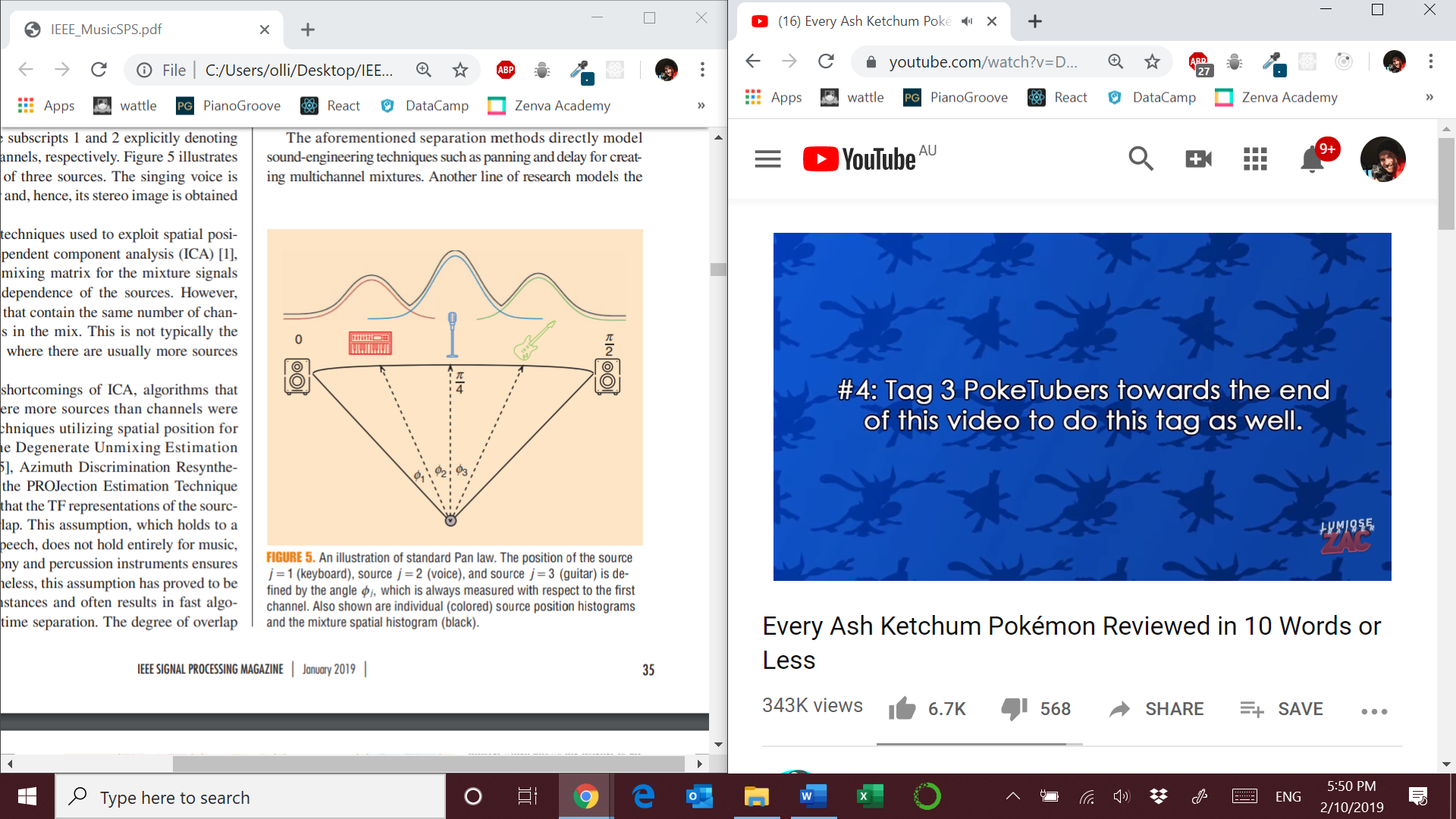
There are usually more sources then channels : ie. common to have 5 instruments and 2 channels

**MSS workflow**



* Use of GWF
* Essentially, each TF point in the original mixture is weighted with the ratio of the source magnitude to the sum of the magnitudes of all sources

**Musical source position models**



* Stereo image – is the sum of the projection of each source into the vector space defined by the panning angle
* **Algorithms** : DUET [5], ADRess [6], PROJET [7] assume little overlap
* **Goal** : Mixture spatial histogram with the assumption that there is little overlap.
* **Local Gaussian modeling [3], [8]** : model spatial configuration of a source through interchannel correlations. The STFT coefficients of each source are calculated and encoded in the **spatial covariance matrix**

**Musical Source models – what if the sources are all in the same location ?**

**Kernel models**

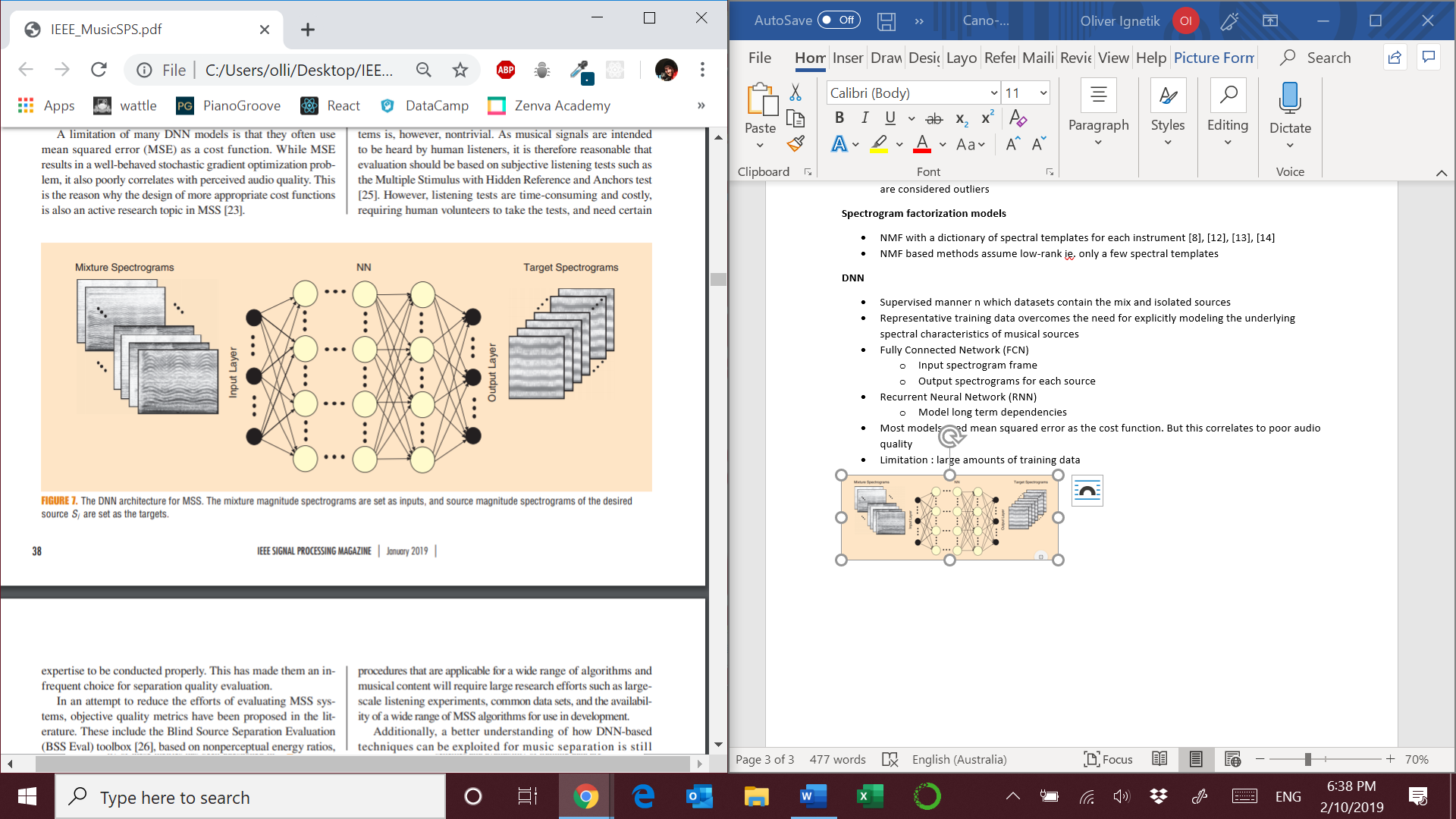
* Based on modeling the spectral characteristics of each source
* Kernel additive models (KAM) exploit local features in spectrograms [9]
* For example, harmonic sounds have similar values over time. Suitable proximity kernel – adjacent time frames over the same frequency
* KAM models assume interference due to other sources is spars, so TF bins with interference are considered outliers

**Spectrogram factorization models**

* NMF with a dictionary of spectral templates for each instrument [8], [12], [13], [14]
* NMF based methods assume low-rank ie. only a few spectral templates

**DNN**

* Supervised manner n which datasets contain the mix and isolated sources
* Representative training data overcomes the need for explicitly modeling the underlying spectral characteristics of musical sources
* Fully Connected Network (FCN)
  + Input spectrogram frame
  + Output spectrograms for each source
* Recurrent Neural Network (RNN)
  + Model long term dependencies
* Most models used mean squared error as the cost function. But this correlates to poor audio quality
* Limitation : large amounts of training data



**Evaluation**

* Human listener evaluation : Multiple Stimulus with Hidden Reference and Anchors test [25]
* SiSEC [29] – community for source separation
* Alternatives for evaluation [30]

**Future research directions**

* Phase retrieval techniques to estimate the phase of the target source
* Feature representations that better match human perception, this allows models to focus on aspects that are most notable to humans
* Separation of voices from the same instrument family
* Perceptually based optimization schemes in DNNs