Time-series segmentation and latent representation of musical instruments

Gregory Szep

King's College London

September 4, 2018

Abstract

Music information retrieval tasks serve as faithful benchmarks for time-series analysis pipelines due to the availability of strongly labelled training data such as MusicNet. Clustering algorithms in spectral sub-spaces, hidden Markov models and causal convolutional neural networks are compared in their ability to transform time-series to a continuous latent space that clusters eleven orchestral instruments. The latent space is evaluated quantitatively with precision-recall metrics obtained by comparing the instrument prediction from a segment of audio to the ground truth obtained from musical scores, and qualitatively by generating samples of audio for given regions in the latent space.

1 Methodology Outline

1.1 Mapping time-series to latent space

The input data are single channel time-series points $\mathcal{D} = \{x(t_1)\ldots x(t_N)\}$ sampled at frequency f from an underlying continuous state-time process $x(t)$, that is the oscillating sound waves emitted by a live orchestra.

For each time point there is a polyphonic score matrix $\mathbf{L}(t) \in \{0,1\}^{N\times L}$ which representing the N midi notes played by L instruments. Instrument activations $\bar{\mathrm{y}}(t) \in \{0,1\}^{L}$ and N note activations $\bar{\mathrm{n}}(t) \in \{0,1\}^{N}$ can be individually considered as well as only change-points between states. The most difficult task is to recover the polyphonic score matrix

$\mathbf{L}(t) \in \{0,1\}^{N\times L}$ from the input signal $x(t)$. The score matrix is generally sparse as playing too many notes with too many instruments at the same time sounds terrible. Sparse labels lead to sparse learning signals which hamper the convergence of machine learning algorithms. Thus we first look to recover changepoints and then the more densly labelled $\bar{\mathrm{n}}(t) \in \{0,1\}^N$ and $\bar{\mathrm{y}}(t) \in \{0,1\}^L$. From this we attempt to recover the full matrix in post-processing.

Within unsupervised methods this task is known as under-determined blind source separation. When the number of input channels equals to the number of sources this problem is fully determined and can be solved using independent component analysis [1] and other algorithms that search for sparse representations. Through the lens of supervised approaches this problem can be seen as an audio segmentation task. In recent years convolutional networks have demonstrated success in image segmentation [], whos architectures are adaptable to audio data.

The principal assumption in this task is that $x(t)$ is a linear superposition of sources and that in some latent space $\bar{\mathrm{z}}(t)$ these sources are linearly separable. The optimal latent space mapping f should minimise the cost function $J[f]$, which in general is some norm $\| \; \|$ of the error subject to some complexity measure $S[f]$. The data in the latent space should be separable by a suitable choice of unmixing matrix $\mathbf{W}$.

$$J[f] = \|\bar{\mathrm{y}}(t) - \mathbf{W}\,\bar{\mathrm{z}}(t)\| + S[f] \quad \text{where} \quad \bar{\mathrm{z}}(t) = f(x,t) \tag{1.1}$$

$$\exists\,\mathbf{W}, f \;:\; f = \underset{f'}{\operatorname{argmin}}\, J[f'] \tag{1.2}$$

Note that the map f is most likely not one-to-one respect to t. It is not a single time point but a time window that encodes which instrument is being played. While linear separability via $\mathbf{W}$ by is sufficient it is not necessary; one could instead simply look for clusters.

1.1.1 Wavelet transforms and spectral speed

The naive approach would be to attempt to guess the mapping f via suitable phenomenological agruments. It is reasonable to suppose that one could separate instruments based on their spectral signature at any moment in time. The decibel spectrogram $S(\omega,t)$ is obtained by taking the $\log_{10}$ absolute value of a short-time fourier transform.

$$S(\omega,t) := \log_{10}\left[\left|\int_{-\infty}^{\infty} x(\tau)\mathrm{e}^{-\mathrm{i}\omega\tau} h(\tau - t)\,\mathrm{d}\tau\right|\right] \tag{1.3}$$

A suitable window function $h(t)$ must be chosen; different choices lead to different amounts of spectral leakage and time-frequency domain resolution []. For demonstration purposes a 46ms wide Hann window is chosen with a stride of 3ms, which is on the order of the shortest duration of a musical sound.

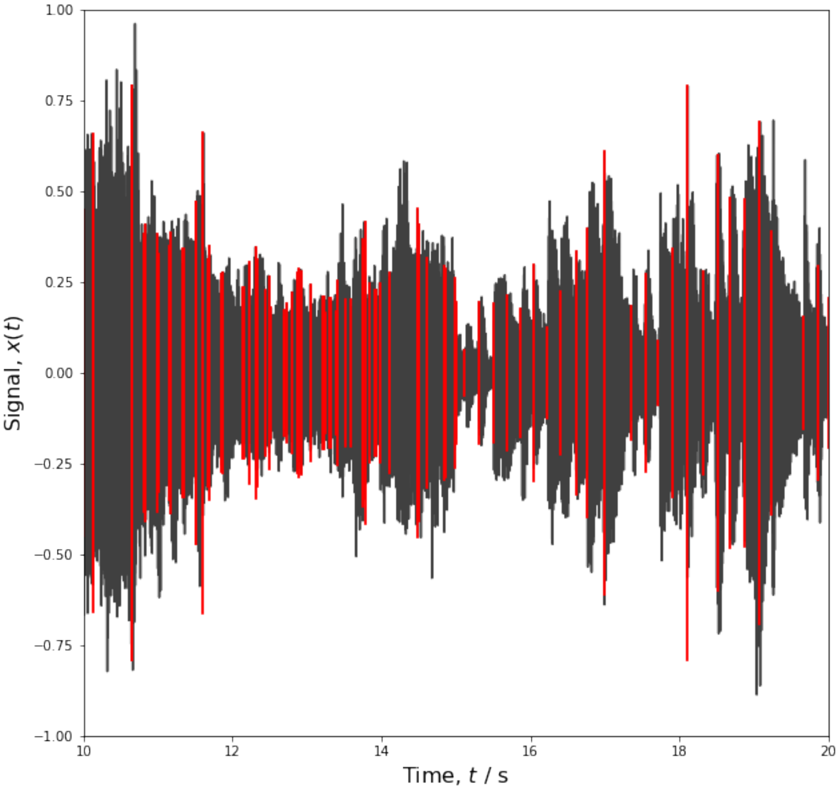

Figure 1: Input signal labelled with changepoints in polyphonic score $\mathbf{L}(t)$

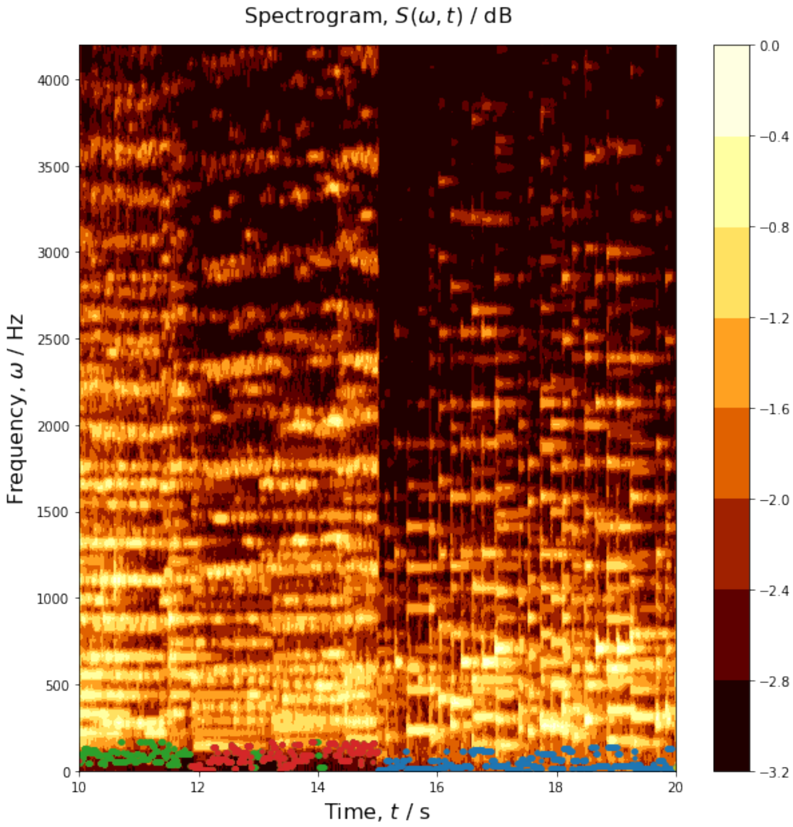

Figure 2: Spectrogram $S(\omega, t)$ with dominant frequency labelled by instrument $\bar{\mathrm{y}}(t)$

The spectral signature of the input signal at a given time t can be represented as an F dimensional point, where F is the number of frequency bins up to a given cutoff frequency calculated in the spectrogram. The spectrogram in Figure 2 shows a promising variety of signatures within domain Ω that may differentiate between instruments. This observation motivates the calculation of spectral speed $v(t)$ defined as

$$v(t) := \frac{1}{|\Omega|} \int_{\Omega} \left| \partial_t S(\omega, t) \right| \mathrm{d}\omega \tag{1.4}$$

Figure 3 reveals that peaks in spectral speed $v(t)$ typically coincide with changepoints in notes and stepwise changes may indicate changepoints in instruments. Peaks can be detected using wavelet transform approaches [2] and true positives are counted when a detected peak lies within threshold time τ of a ground truth changepoint. Evaluations of 1000 clips of 5s from MusicNet in Figure 4 show that within 60ms a window approximately 60% of the changepoints can be recalled with 60% precision. The duration of a quaver amongst Beethoven symphonies range 40-125ms. This suggests that note changepoints more readily distinguised in slower symphonies. Such a claim is supported by the example in 3, which shows a higher prevelance of false positives and negatives in the fast paced signal on the left-hand side. This approach is not precise enough to proceed with classification of the segments between changepoints.

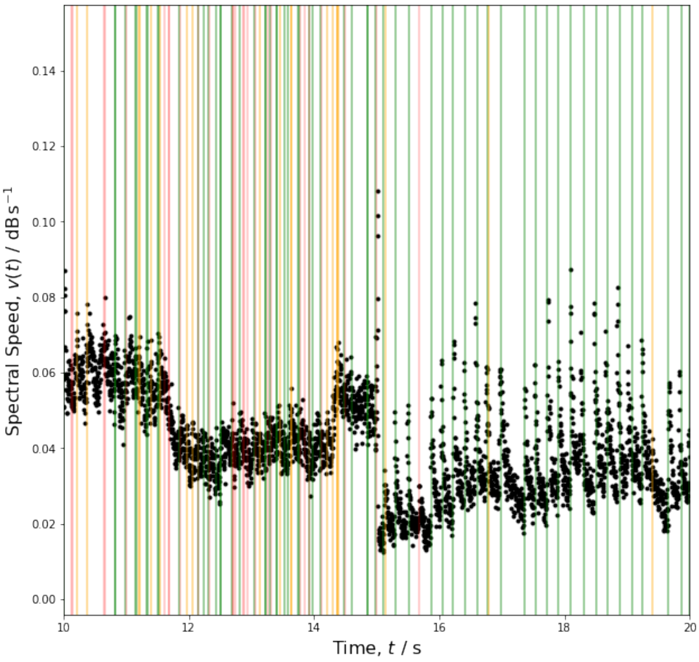

Figure 3: Spectral speed $v(t)$ labelled by true positive false positive and false negative changepoint detections

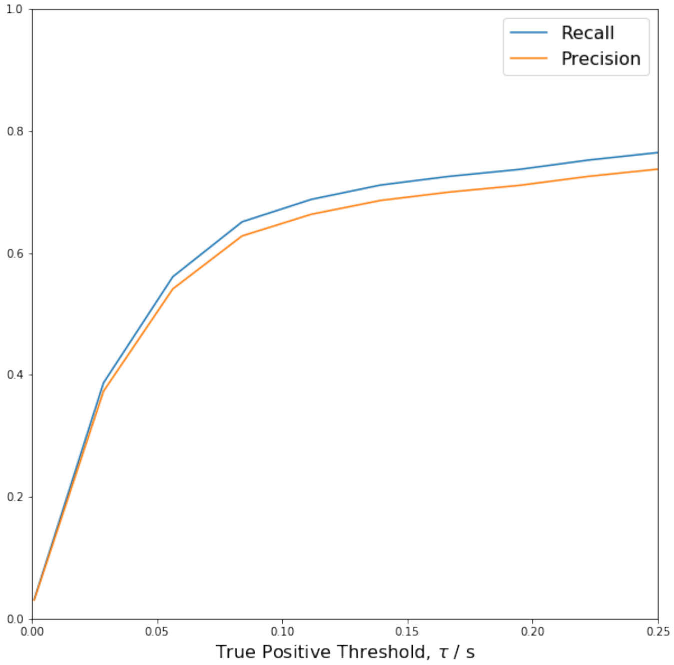

Figure 4: Changepoint precision-recall curves given by peaks in spectral speed $v(t)$

1.1.2 Markov models and expectation-maximisation

1. discuss implementations of random Markov fields in image segmentation and how they can be adapted to audio segmentation

2. random Markov field vs hidden Markov model?

3. expectation-maximisation vs error back-propagation

1.1.3 Feature extraction with causal convolutions

Convolutional architectures have become popular due to their ability to compress spatio-temporal information for discrimination and generation tasks [3, 4]. A causal convolutional network [5] — which encodes the arrow of time in its architecture — is trained for the audio segmentation task.

1. dilated causal convolutions as a merge between a feature extractor and a dimensionality reduction technique. This is a supervised method

2. Compare clusters to those obtained in Section 1.1.1

1.2 Clustering in latent space

1. We have a hierarchical clustering problem: there are 11 instruments each of which can play 28-83 notes. The easier problem is to only cluster instruments, the harder problem is to cluster both instruments and notes.

1.2.1 K-means

1. Since we know how many instruments there are, we can apply a naive K-means and see what happens. Here the disadvantage is that time-ordering may be ignored, which can lead to noisy/discontinuous audio segment classifications
2. Discuss de-noising strategies in post-classification: possibly Markov random fields from section 1.1.2?

1.2.2 Fully convolutional networks

1. I shall attempt to adapt the the fully convolutional architecture [6] for the audio segmentation task. Discuss advantages of end-to-end trained solution.

1.3 Evaluation methods

1.3.1 Audio segment retrieval

1. outline of object detection / segmentation in image analysis
2. precision-recall metric applied to audio

1.3.2 Instrument generation

1. introduction to encoder / decoder pipelines as generative models which can produce data given activation of input in latent space.
2. attempt to produce sounds that are interpolations between existing instruments, assess qualitatively how realistic they sound

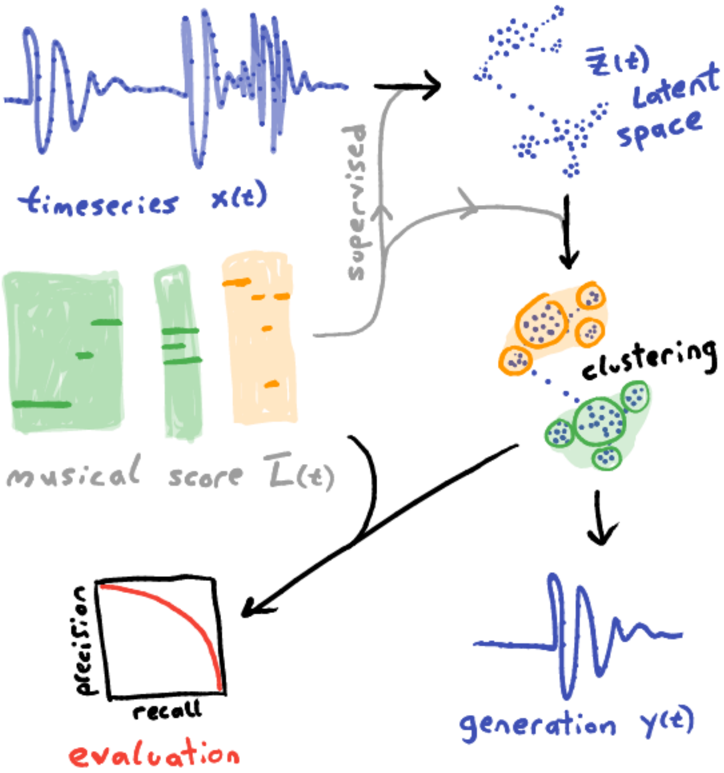

Figure 5: Summary of methodology showing all stages of the audio segmentation task. Each transition between sub-figures can be achieved with appropriate algorithms

2 Dataset Description

2.1 Input and Labels

1. small summary table of the MusicNet dataset [7], advantages over EEG and other bio-sensory data for bench-marking signal processing algorithms
2. raw labels are time aligned transcripts of the sheet music. How to we parse that into instrument activations and note activations.
3. cross-validation — if any

2.2 Data Partitioning

1. test, validation and train sets
2. cross-validation — if any

3 Results, Protocols and Conclusions

1. all relevant figures will be presented here

3.1 Learned feature maps

3.2 Clustering performance

3.3 Generating instruments

References

[1] J. Platt and S. Haykin, “Information-Maximization Approach to Blind Separation and Blind Deconvolution,” *Technology*, vol. 1159, no. 6, pp. 1129–1159, 1995.

[2] P. Du, W. A. Kibbe, and S. M. Lin, “Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching,” *Bioinformatics*, vol. 22, pp. 2059–2065, sep 2006.

[3] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel Recurrent Neural Networks,” in *ICML, International Conference on Machine Learning*, vol. 48, pp. 1747–1756, 1 2016.

[4] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, “Generative Adversarial Networks,” *arXiv preprint*, pp. 1–9, 2014.

[5] H. Kameoka, Nobutaka Ono, Kunio Kashino, and Shigeki Sagayama, “Complex NMF: A new sparse representation for acoustic signals,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3437–3440, IEEE, 4 2009.

[6] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, IEEE, 6 2015.

[7] J. Thickstun, Z. Harchaoui, and S. Kakade, “Learning Features of Music from Scratch,” *arXiv*, 11 2016.