Towards Statistical Unsupervised Online Learning for Music Listening with Hearing Devices

Purwins, Hendrik; Marchini, Marco; Marxer, Richard

Publication date:
2017

Document Version
Other version

Citation for published version (APA):
Purwins, H., Marchini, M., & Marxer, R. (2017). *Towards Statistical Unsupervised Online Learning for Music Listening with Hearing Devices*. 41. Paper presented at The Music & Cochlear Implants Symposium, Snekkersten, Denmark.

Towards Statistical Unsupervised Online Learning for Music Listening with Hearing Devices

Hendrik Purwins , Marco Marchini , and Ricard Marxer

The same piece of music maybe be experienced differently by different subjects, depending on various factors: The *perceptual processing* of the piece (e.g. 3500 inner hair cells vs. a few electrodes in a cochlear implant) may allow for a representation with various degrees of frequency and temporal resolution. The listener's *cognitive capabilities*, e.g. memory, voluntary attention direction, abstraction and inference (normal listener vs. a person with dementia) will cause different cognitive processing and musical experience. The *familiarity* with a particular musical style, instrument, piece may affect the emotional impact and intellectual understanding of a piece.

It could be desirable for a hearing device (hearing aid or cochlear implant) to perform musical signal processing to transfer the musical experience of a piece for a normal listener to a similar musical experience of the same piece in a listener with severe hearing limitations. Such a transformation would require a formal representation of the music piece that is scalable in complexity. Such a scalable representation is provided in [1], where the number of different sounds (e.g. considering open and closed hi-hats as one or two distinct sound categories) as well as the temporal context horizon (e.g. storing up to 2-note sequences or up to 10-note sequences) is adaptable. The framework in [1] is based on two cognitively plausible principles: unsupervised learning and statistical learning. Opposed to supervised learning in primary school children, where the school teacher points at a written letter and articulates its phonetic pronunciation, infants perceptually organise the phonemes they are exposed to into groups, based on the phoneme's similarity and context (*unsupervised learning*). 8-month-old infants are able to learn statistical relationships between neighboring speech sounds [3] (*statistical learning*). In [1], Figure 1, unsupervised learning is implemented as agglomerative clustering, informed by the Gestalt principle of regularity. The model [1] performs statistical learning, applying variable length Markov chains. In [2], grouping of sounds into phonetic/instrument categories and learning of instrument event sequences is performed jointly using a Hierarchical Dirichlet Process Hidden Markov Model.

Whereas machines often learn by processing a large data base and subsequently updating parameters of the algorithm, humans learn instantaneously, i.e. the mental representation is continously changed after every exposure to small batches of sound events (*online learning*). In [2] online learning is implemented via the interplay of Cobweb clustering and a hierarchical n-gram instantaneously updating the number of timbre groups and their respective transition counts. We propose to use online learning for the co-evolution of both CI user and machine in (re-)learning musical language.

H. Purwins is with Audio Analysis Lab and Sound and Music Computing Group, Department of Architecture, Design & Media Technology, Aalborg Universitet København. hpu@create.aau.dk

M. Marchini is with Sony CSL, 6, Rue Amyot, 75005 Paris, France. marco.marchini3@gmail.com

R. Marxer is with the Speech and Hearing Group, Department of Computer Science, University of Sheffield, Sheffield S1 4DP, U.K. (e-mail: r.marxer@sheffield.ac.uk.

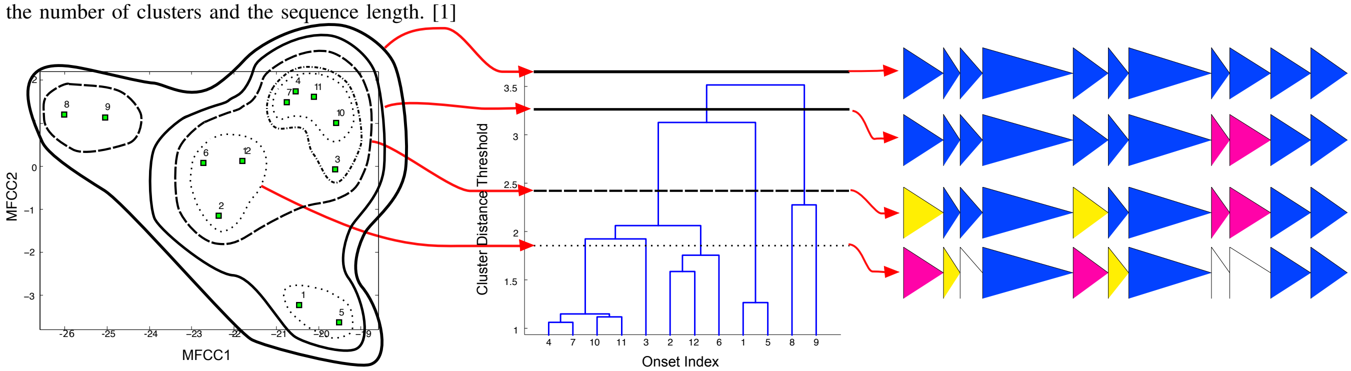

Figure 1. Percussive sound events are grouped into timbre categories (left), using agglomerative clustering (unsupervised learning), yielding a clustering tree (middle) that can be cut at different resolutions, e.g. following a Gestalt principle of rhythmical regularity (dashed line), resulting in corresponding discrete sequences (right) that are further analyzed using variable length Markov chains. Such a model is scalable in complexity parameterised by the number of clusters and the sequence length. [1]

REFERENCES

[1] Marco Marchini and Hendrik Purwins. Unsupervised analysis and generation of audio percussion sequences. In *International Symposium on Computer Music Modeling and Retrieval*, pages 205–218. Springer, 2011.

[2] R. Marxer and H. Purwins. Unsupervised incremental online learning and prediction of musical audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 24(5):863–74, 2016.

[3] J. Saffran, R. Aslin, and E. Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996.