EXPERIMENT DESIGN FOR PSYCHOACOUSTIC ANALYSES OF AUDIO-SCORE COMPOSITIONS

Katharina Pollack
Acoustics Research Institute
Academy of Sciences, Vienna
katharina.pollack@oeaw.ac.at

Piotr Majdak
Acoustics Research Institute
Academy of Sciences, Vienna
piotr.majdak@oeaw.ac.at

Barbara Maria Neu
University of Music and
Performing Arts, Vienna
neu@mdw.ac.at

ABSTRACT

Psychoacoustics and the perception of sound play a crucial role in the process of composing a musical piece. The interpretation by the musician, despite a-priori knowledge about the musical piece, and, if applicable, the interaction between the musician and the composer in a live scenario, both heavily depend on their auditory systems and auditory environment. To better understand the underlying psychoacoustic processes, we conducted an experiment using isolated parts of audio-score recordings of a clarinet. In the experiment, we focused on perceptive parameters involved in the musical piece. We relate our findings to existing psychoacoustic models, considering the limits of the models' auditory stages and the possibility of incorporating the models in the process of audio-score compositions. To this end, we give an outlook on what a psychoacoustic analysis of audio-score compositions could look like.

1. INTRODUCTION

A score contains information the composer wants to convey to the musician. It can comprise instructions regarding pitch, dynamics, and tempo, even instructions open to interpretation and possibly relying on mimesis, e.g., instructions regaring timbre. There are various interpretations of how an audio score is defined. For example, Bell distinguishes between reactive and rehearsed audio scores [1]. Bhagwati distinguishes between situative and fixed, and five conveyance subcategories [2]. He also states that "Musical precision is conveyed most effectively through music itself." in the context of musical pedagogy. Similar to situative and fixed audio scores, Sdarulig et al. distinguish between online and offline audio scores [3]. And Schimana explores the concept of audio scores based on acoustic memory [4].

An audio score is not just a different visual representaton of sound, e.g., a spectrogramme or a sketch of the score beyond classical notation. From a basic research perspective, a visual score represents the score in space, and an audio score represents the score in time. The score itself is transferred from the composer(s) to the musician(s) via at least one acoustic transfer path. There can be multiple acoustic transfer paths, sometimes on purpose and sometimes unpredictable or even unwanted ones, because, e.g., the score can be conveyed by a loudspeaker to a musician, but the acoustic properties of a room and the number of instruments and louspeaker playing simultaneously add another acoustic transfer path.

Using a computational psychoacoustic model to simulate a virtual listener can be advantageous for a number of reasons, one of which is to estimate the behaviour of a person prior to conducting an experiment. However, all psychoacoustic models are an approximation of reality and can focus on different stages of the auditory processing chain. Additionally, the virtual listener is not represented in the same way in all psychoacoustic models. One way would be to model the listener as an agent in a perception-action cycle [5]. In this perception-action cycle, the agent, in this case a musician, is interpreted as agent whose sensors receive signals from the environment, e.g., auditory, visual, or somatic signals. These sensations update the internal beliefs of the agent and influence movement, e.g., eye, head, or limbs, which in turn influence the environment and, thus, close the perception-action cycle.

In order to apply psychoacoustic models as composition tools, they have to be evaluated on synthetic sounds. In this study, we present a pilot study to help design an experiment setup for psychoacoustic analysis of audio-score compositions.

2. METHODS

In order to analyse why the musician played what they heard, we set up an experiment, in which we examined simple audio scores played by a single musician repeating isolated musical sequences. A video recording of the musician commenting on each take was made immediately after the recorded sequence in order to post-hoc identify conscious decisions.

2.1 Signals

Sequences were implemented in MaxMSP (8.6.3) on a laptop (MacOSX Sequoia 15.5). Using a sound pressure level meter, it was ensured that sound pressure levels at the ear of the musician would not exceed 85 dB(A). Three different tone combinations were created: [d',d#',f'], [a', h', c''], and [g#''', a''', b''']. Each combination was chosen to be within a critical band [6, 7], hence, the combinations

were called CB1, CB2, and CB3, respectively. For each of the combinations, four different sequences were created: *Long tones*, in wich every tone lasted a few seconds, *long to short tones*, in which the duration of tones was shortened in an arhythmical way, *short tones*, in which every tone lasted less than one second, and *long tones with noise*, in which every tone lasted a few seconds but was accompanied with noise. For each of the sequences, the tones were played in a partially overlapping, partially alternating fashion with varying attack, decay, sustain, and release times. For the experiment, the following thirteen sequences were recorded:

1. CB1 long tones
2. CB2 long tones
3. CB3 long tones
4. CB3 long tones
5. CB1 long tones
6. CB1+CB2+CB3 long tones
7. CB1 long to short tones
8. CB2 long to short tones
9. CB3 long to short tones
10. CB1+CB2+CB3 short tones
11. CB1 long with noise
12. CB2 long with noise
13. CB3 long with noise

Basically, the sound was moving from one pitch to the next with varying speed, duration, and overlap time.

2.2 Participant and apparatus

The signal flow in the experiment was as follows: The signal generated by the computer of the composer was routed to an interface (RME Fireface UCXII 24110382) and played to the musician – the participant – over a loudspeaker (Genelec 8020B active monitor) placed 25 cm away from the musician's right ear. A microphone was placed just above the musician's ear and held in place by clipping it to an elastic headband. A second microphone was placed in the room in order to record the musician's thoughts after the recordings, alongside the video recording of the camera. On the clarinet, one microphone was inserted by 5 cm into the cone of the instrument, and a gooseneck microphone 5-cm away from the cone. These two microphones were placed such that different timbre features may be extracted for further processing.

The room (3.5 x 5.4 x 3.2 m, width x length x height) was acoustically treated with Basotect plates on the walls. Using a sound-pressure level metre, a noise floor of 27 dB(A) was measured. In total, four people were in the room: audio recording engineer, camera operator, composer, musician. Towards one side of the room, the recording engineer was placed behind a curtain. In the remaining part of the room, the composer, the musician, and the camera operator were placed in a triangular formation. The space between the camera operator and the composer was 1.5m, as was the space between the camera operator and the musician. The space between the musician and composer was 2.5 m, and the space between the musician to the wall was 80 cm. A loudspeaker (Genelec 8020B) was placed at the height and the distance of 25 cm from the musician's ear with the musician sitting in a rigid position. During the recording, the distance between the musician's ear and the loudspeaker changed marginally. In this scenario, no eye-contact during the recording was made between the composer and the musician in order to replicate a live-performance scenario as close as possible.

2.3 Procedure

Before the recordings, the instruction "play what you hear" was given by the composer to the musician. After each stimulus was recorded, the musician reported what they heard and what sound(s) they were trying to create. Sometimes, but not often, the composer would clarify on what part of the sound the musician should focus on and try to reproduce using their instrument.

3. RESULTS AND DISCUSSION

The total duration of the recording was two hours and 23 minutes and an analysis of the complete recording is out of the scope for this article. Here, we analysed three short clips from three different sequences, lasting only a few seconds. The first clip was from sequence 1 (CB1 long tones) in which the three tones were played ordered [f', d', d#']. Figure 1 shows the waveforms with normalised amplitude of the sequence in various stages of the processing flow.

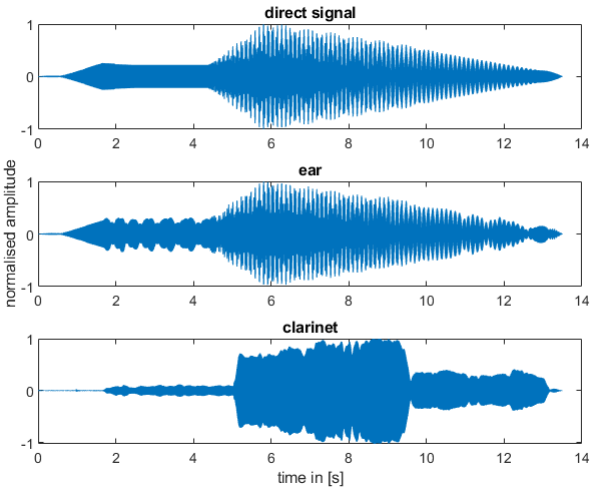

Figure 1. CB1, long tones. Top waveform is the direct signal, middle waveform is the signal arriving at the musician's ipsilateral ear, bottom waveform is the signal measured 5 cm outside of the clarinet's cone.

Direct signal was the output of the interface, the *ear* was

the microphone signal placed close to the musician's ipsilateral ear, and *clarinet* was the microphone signal placed outside the cone of the instrument. It can be seen that the direct signal from the interface is a synthetically generated one, containing linear amplitude changes. The amplitude of the ear signal is different than in the direct signal, because of reflections of the musician's head, shoulders, and ear. And in the clarinet signal, the musician started playing approximately 400 ms after the ear signal, a delay which consists of the musician's reaction time, their time to make a decision on what to play, and also the time to breathe in and start playing.

We also analysed the signals in the time-frequency domain. Figures 2, 3, and 4 show the spectrogrammes of the clip. These figures were created with mulaclab from the large time-frequency analysis toolbox (LTFAT) v2.6.0 [8, 9]. The window length was set to 4096 samples in order to increase the frequency resolution, other frame parameters remained at their default values. The beat can be seen clearly as well as the changing frequency (slow between f' and d' and fast between d' and d#').

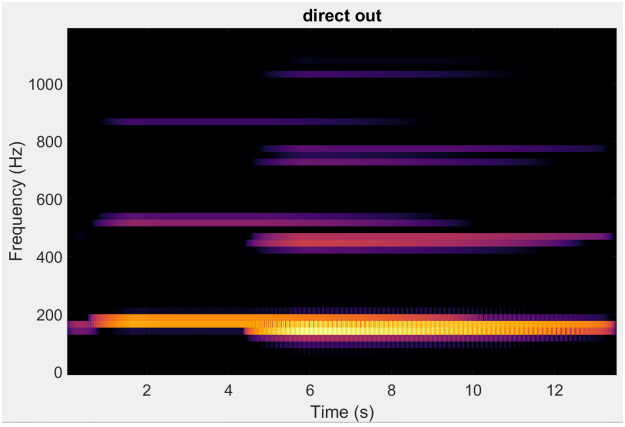

Figure 2. The signal *direct out* from clip 1 in the frequency domain.

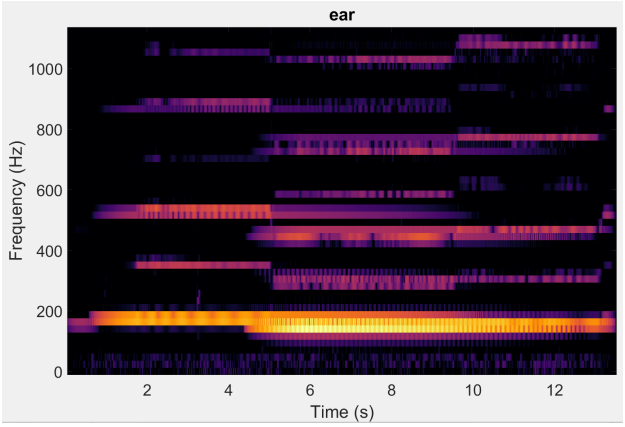

Figure 3. The signal *ear* from clip 1 in the frequency domain.

Changes in timbre can be seen between the direct-out and the clarinet signals (independent of the microphone position). In the *ear* signal, the beat started after two seconds

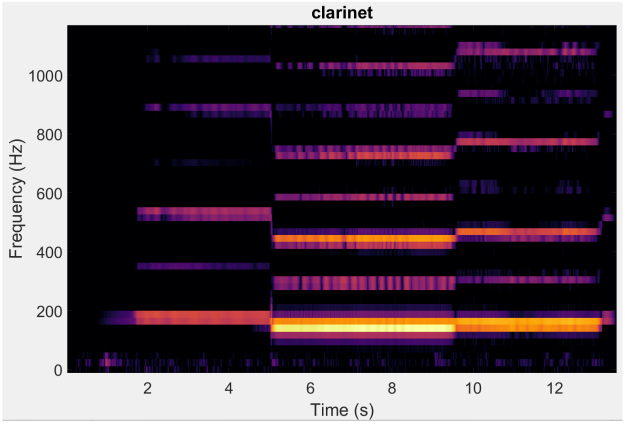

Figure 4. The signal *clarinet* from clip 1 in the frequency domain.

between f' and d', and increased its frequency as the interval progressed to a minor second (d' and d#'). The beat was not played by the musician, the musician only decided to switch notes once the next tone was louder.

Then we analysed two additional clips, one taken from sequence 7 (CB1 long to short) and the other from sequence 13 (CB3 long with noise). Figure 5 shows the clip from sequence 7. In the beginning of this clip, the musician played significantly shorter tones than what they heard, and it seems that later in the sequence, the amount of tones subsequently played were too much to play back the same way. Instead, the musician played only the tone (sometimes multiple tones) they could hear once they stopped playing.

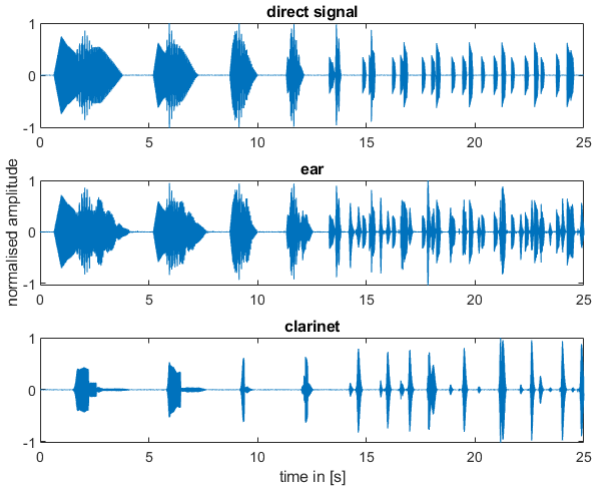

Figure 5. Clip from sequence 7 (CB1 long to short). Order of waveforms same as in Fig. 1.

Figure 6 shows the clip from sequence 13. In this clip, the direct-out signal contained noise underlying long tones. The clarinet signal did not contain noise, because apparently the musician focused on playing high-pitch tones quietly, a design choice by the composer which itself was already a challenge on this instrument. Because of the low volume of the speaker signal, the sound from the clarinet

was captured not only in the clarinet microphone but also in the ear microphone, which complicated the signal separation and, thus, the analysis.

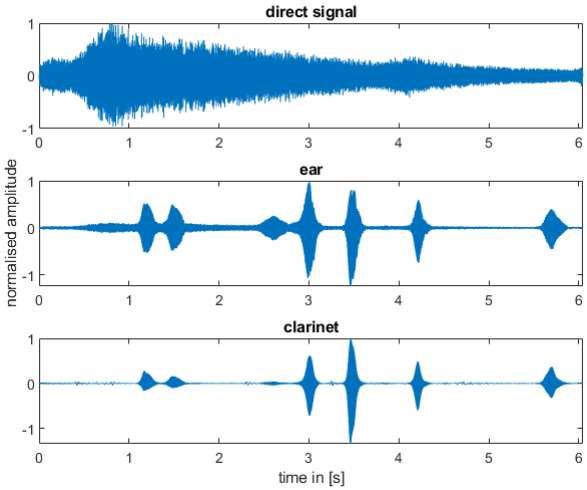

Figure 6. Clip from sequence 13 (CB3 long with noise). Order of waveforms same as in Fig. 1.

There are a number of psychoacoustic models that would be useful for compositions, for example, models covering temporal [12] or spectral masking [13, 14, 15, 11]. They could tell at what time or level differences two distinct sounds can be perceived.

With respect to spectral changes, there are two models that might be able to cover the prediction of whether or not a virtual listener could detect a change in timbre. One is a monaural model, i.e., not considering any binaural effects (e.g., location of a sound source) [15, 16, 17]. The other one is a binaural model [18].

One current limitation of these models is that they consider only static sounds. For example, a change can be detected in a sound that contains two clicks which are 40-ms delayed, but the model would not be able to handle a sound that contains more clicks with varying delays. Another known limitation of the perceptual similarity models is that they have a binary output whether or not a listener be able to detect a difference between the two sounds presented. However, there is no absolute threshold for change detection in timbre that is the same for each individual, and a probabilistic output, i.e., how likely is a listener able to detect a difference between two sounds, would make more sense in a realistic scenario.

4. CONCLUSIONS

In general, a new pitch introduced in the direct signal resulted in acoustic phenomena in the ear signal, such as, e.g., amplitude modulation. Exploiting such (psycho-)acoustic phenomena explaining whether a musician is able to hear a difference in audio scores can help composers to experiment with the auditory system of the musicians. In this contribution, we proposed an experiment setup that was piloted on one musician and instrument, and we covered a first insight into the potentials of psychoacoustic analysis of audio scores. An application of said models in the context of audio scores seems to be required to better understand the interpretation of audio scores from the perspective of psychoacoustics.

Acknowledgments

This research was funded by the Austrian Science Fund (FWF) [grant DOI 10.55776/AR824]. The authors thank Elisabeth Schimana for providing the audio-score compositions.

5. REFERENCES

[1] J. Bell, "Audio-scores, a resource for composition and computer-aided performance," Ph.D. dissertation, Guildhall School of Music and Drama, 2017.

[2] S. Bhagwati, "Elaborate audio scores: Concepts, affordances and tools," in *Proceedings of the 4^{th} International Conference on Technologies for Music Notation and Representation–TENOR*, vol. 18, 2018, pp. 24–32.

[3] C. Sdraulig and C. Lortie, "Recent audio scores: Affordances and limitations," in *Proceedings of the 5^{th} International Conference on Technologies for Music Notation and Representation–TENOR*, vol. 19, 2019, pp. 38–45.

[4] E. Schimana, "Sound as score: Live generated audio score and audio score based on acoustic memory," in *Proceedings of the 5^{th} International Conference on Technologies for Music Notation and Representation–TENOR*, 2019.

[5] P. Majdak, R. Baumgartner, and C. Jenny, "Formation of three-dimensional auditory space," in *The technology of binaural understanding*. Springer, 2020, pp. 115–149.

[6] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *Bell System Technical Journal*, vol. 12, no. 4, pp. 377–430, 1933.

[7] H. Fletcher, "Auditory patterns," *Reviews of modern physics*, vol. 12, no. 1, p. 47, 1940.

[8] P. L. Søndergaard, B. Torrésani, and P. Balazs, "The linear time frequency analysis toolbox," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 10, no. 04, p. 1250032, 2012.

[9] Z. Průša, P. L. Søndergaard, N. Holighaus, C. Wiesmeyr, and P. Balazs, "The large time-frequency analysis toolbox 2.0," in *International Symposium on Computer Music Multidisciplinary Research*. Springer, 2013, pp. 419–442.

[10] A. Osses Vecchi, R. García León, and A. Kohlrausch, "Modelling the sensation of fluctuation strength," in *Proceedings of Meetings on Acoustics*, vol. 28, no. 1. Acoustical Society of America, 2016, p. 050005.

[11] P. Majdak, C. Hollomey, and R. Baumgartner, "Amt 1. x: A toolbox for reproducible research in auditory modeling," *Acta Acustica*, vol. 6, p. 19, 2022.

[12] J. F. Culling, M. L. Hawley, and R. Y. Litovsky, "The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources," *The Journal of the Acoustical Society of America*, vol. 116, no. 2, pp. 1057–1065, 2004.

[13] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective"signal processing in the auditory system. i. model structure," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[14] ——, "A quantitative model of the "effective"signal processing in the auditory system. ii. simulations and measurements," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3623–3631, 1996.

[15] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2892–2905, 1997.

[16] ——, "Modeling auditory processing of amplitude modulation. ii. spectral and temporal integration," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2906–2919, 1997.

[17] A. Osses Vecchi and A. Kohlrausch, "Perceptual similarity between piano notes: Simulations with a template-based perception model," *The Journal of the Acoustical Society of America*, vol. 149, no. 5, pp. 3534–3552, 2021.

[18] T. McKenzie, C. Armstrong, L. Ward, D. T. Murphy, and G. Kearney, "Predicting the colouration between binaural signals," *Applied Sciences*, vol. 12, no. 5, p. 2441, 2022.