BLIND UPMIX FOR APPLAUSE-LIKE SIGNALS BASED ON PERCEPTUAL PLAUSIBILITY CRITERIA

Alexander Adami

International Audio Laboratories*,
Friedrich-Alexander Universität Erlangen
Erlangen, Germany
alexander.adami@audiolabs-erlangen.de

Lukas Brand

International Audio Laboratories*,
Friedrich-Alexander Universität Erlangen
Erlangen, Germany
lukas.brand@fau.de

Sascha Disch

Fraunhofer IIS,
Erlangen, Germany
sascha.disch@iis.fraunhofer.de

Jürgen Herre

International Audio Laboratories*,
Friedrich-Alexander Universität Erlangen
Erlangen, Germany
juergen.herre@audiolabs-erlangen.de

ABSTRACT

Applause is the result of many individuals rhythmically clapping their hands. Applause recordings exhibit a certain temporal, timbral and spatial structure: claps originating from a distinct direction (i.e, from a particular person) usually have a similar timbre and occur in a quasi-periodic repetition. Traditional upmix approaches for blind mono-to-stereo upmix do not consider these properties and may therefore produce an output with suboptimal perceptual quality to be attributed to a lack of plausibility. In this paper, we propose a blind upmixing approach of applause-like signals which aims at preserving the natural structure of applause signals by incorporating periodicity and timbral similarity of claps into the upmix process and therefore supporting plausibility of the artificially generated spatial scene. The proposed upmix approach is evaluated by means of a subjective preference listening test.

1. INTRODUCTION

Applause is a sound texture composed of many individual hand claps produced by a crowd of people [1]. It can be imagined as a superposition of discrete and individually perceivable transient foreground clap events and additional noise-like background originating from dense far-off claps as well as reverberation [2]. Due to the high number of transient events, applause-like signals form a special signal class which often needs a dedicated processing [3–5]. This is possible since applause sounds are well detectible among mixtures of other signal classes [6].

At first sight, the nature of applause sound textures may seem totally random [7]. However, previous publications and listening experiments indicate that a fully randomized applause texture, consisting of temporally and spatially randomized events with random timbre, is perceived as unnatural and non-plausible by listeners [8].

It was shown in [9] for sparse to medium dense applauses, or applause containing at least distinct dominant foreground claps, that listeners do expect a quasi-periodic occurrence of clapping events of a certain stable timbre [1, 10] originating from selected spatial locations in order to perceive a plausible spatial impression of being exposed to real-world applause. Phrased differently, listeners seem to be able to distinguish the clapping sounds from single individual persons, each having a distinct clap timbre, in the foreground signal from the much more dense background signal and perceive these as repeated events of similar timbre originating from the same source.

If these special properties are disturbed, listeners report perceptual quality degradations in the stability of the perceived spatial image and also a lack of plausibility of the spatial scene.

Building upon the findings in [9], we propose a blind (or non-guided) spatial upmix from mono to stereo of applause signals that preserves the important properties of quasi-periodicity and timbral similarity of foreground claps to be attributed to a distinct source. Thereby, the plausibility of the artificially generated spatial scene is strengthened.

The blind upmix proposed in this paper relies on a separation of transient and individually perceivable foreground claps and the noise-like background, being reminiscent of the guided applause upmix published in [3]. While the noise-like background is subjected to decorrelation, the separated foreground claps are distributed by panning to arbitrarily chosen positions in the stereo sound stage. Though operated in an unguided manner, as a novel contribution, our algorithm most importantly ensures that each position will be populated by a clap event of suitable timbre in a quasi-periodic fashion, thus supporting the notion of plausibility of the artificially generated spatial scene.

2. APPLAUSE SEPARATION

Before the actual upmix process can take place, the monophonic input applause signal $A(k, m)$ has to be decomposed into a signal part corresponding to distinctly and individually perceivable *foreground* claps $C(k, m)$, and a signal part corresponding to the noise-like *background* signal $N(k, m)$ [9], where k and m denote the discrete frequency and block index in short-time Fourier transform domain. The frequency transformation is done with high temporal resolution, i.e., a block size of 128 samples with 64 samples overlap is used. The corresponding signal model is given by Eq. 1:

$$A(k, m) = C(k, m) + N(k, m). \quad (1)$$

* A joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany

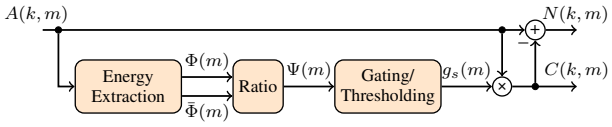

Figure 1: *Block diagram of applause separation into distinctive individually perceivable foreground claps $C(k,m)$ and noise-like background $N(k,m)$.*

The applause separation proposed in this paper is a modified version based on the approaches used in [9, 11]. Figure 1 depicts a block diagram describing the basic structure of the applause separation processing. Within the energy extraction stage, an instantaneous energy estimate $\Phi(m)$ as well as an average energy estimate $\bar{\Phi}(m)$ is derived from the input applause signal. The instantaneous energy is given by $\Phi(m) = \|A(k,m)\|_2$, where $\|\cdot\|_2$ denotes the L2-norm. The average energy is determined by a weighted sum of the instantaneous energies around the current block and given by

$$\bar{\Phi}_A(m) = \frac{\sum_{\mu=-M}^{M} \Phi_A(m-\mu)\cdot w(\mu+M)}{\sum_{\mu=-M}^{M} w(\mu+M)}, \tag{2}$$

where $w(\mu)$ denotes a weighting window (squared sine-window) with window index μ and length $L_w = 2M+1$. In the next stage, the ratio of instantaneous and average energy $\Psi(m)$ is computed. It serves as an indicator whether a discrete clap event is present and is given by

$$\Psi(m) = \frac{\Phi_A(m)}{\bar{\Phi}_A(m)}. \tag{3}$$

If the current block contains a transient, the instantaneous energy is large compared to the average energy and the ratio is significantly greater than one. On the other hand, if there is only noise-like background at the current block, instantaneous and average energy are almost similar and consequently, the ratio is approximately one. The basic separation gain $g_s(m)$ can be computed according to

$$\widehat{g}_s(m) = \sqrt{\max\left(1 - \frac{g_N}{\Psi(m)}, 0\right)}, \tag{4}$$

which is a signal adaptive gain $0 \leq \widehat{g}_s(m) \leq 1$, solely dependent on the energy ratio $\Psi(m)$. To prevent dropouts in the noise-like background signal, the constant g_N is introduced. It determines the amount of the input signal's energy remaining within the noise-like background signal during a clap. The constant was set to $g_N = 1$, which corresponds to the average energy remaining within the noise-like background signal. Since for the upmixing we are mainly interested in the directional sound component of a clap (i.e., the attack phase), thresholding or gating is applied to $\Psi(m)$ according to

$$g_s(m) = \begin{cases} \begin{cases} 0 & \text{if } \Psi(m) < \tau_{\text{attack}} \\ \widehat{g}_s(m) & \text{if } \Psi(m) \geq \tau_{\text{attack}} \end{cases} & \text{if } g_s(m-1) = 0 \\ \begin{cases} 0 & \text{if } \Psi(m) < \tau_{\text{release}} \\ \widehat{g}_s(m) & \text{if } \Psi(m) \geq \tau_{\text{release}} \end{cases} & \text{if } g_s(m-1) \neq 0. \end{cases} \tag{5}$$

This means, the separation gain $g_s(m)$ is only different from zero after the energy ratio surpassed an attack threshold τ_{attack} and only as long as it is above a release threshold τ_{release}. When $\Psi(m)$ falls below τ_{release}, the separation gain is set back to zero. For the separation attack and release threshold $\tau_{\text{attack}} = 2.5$ and $\tau_{\text{release}} = 1$ were used. The final separated signals are obtained according to

$$C(k,m) = g_s(m)\cdot A(k,m) \tag{6}$$

$$N(k,m) = A(k,m) - C(k,m). \tag{7}$$

Figure 2 depicts waveforms and spectrograms of an exemplary applause signal on the left and the corresponding separated clap signal in the middle, as well as the noise-like background signal on the right. Sound examples are available at https://www.audiolabs-erlangen.de/resources/2017-DAFx-ApplauseUpmix.

3. APPLAUSE UPMIX

After having the input applause signal separated into individual claps and noise-like background, the signal parts are upmixed separately. Upmixing the noise-like background was realized by using the original background signal as the left channel and a decorrelated version $\widehat{N}(k,m)$ as the right channel of the upmix. Decorrelation was achieved by scrambling the temporal order of the original noise-like background signal. This method was originally proposed in [4], where the time signal is divided into segments which are themselves divided into overlapping subsegments. Subsegments are windowed and their temporal order is scrambled. Applying overlap-add yields the decorrelated output signal. The processing for the noise-like background signal was modified in the sense that it operates in short-time Fourier transform (STFT)-domain and with a small segment size. A segment size of 10 blocks corresponding to 13 ms was used, where each block represented a subsegment.

Upmixing of foreground claps makes use of inter-clap relations. This means, the upmixing process incorporates the assumptions that claps originating from a certain direction should sound similar, as well as that claps originating from a certain direction should exhibit some sort of periodicity.

In the beginning, arbitrary discrete directions ϕ_d within the stereo panorama are chosen, where $d = 0...D-1$ denotes the index of a distinctive direction within the direction vector $\Phi = [\phi_0, \phi_1, ..., \phi_{D-1}]$. The total number of directions is given by D. Furthermore, the mean clap spectrum for each detected clap is computed, whereby a clap is considered as a set of consecutive blocks, each of which having non-zero energy and framed by at least one block to each side containing zero energy. With the start and stop block index $\gamma_s(c)$ and $\gamma_e(c)$ of clap c, the claps' mean spectra are given by

$$\widehat{C}_c(k) = \frac{1}{\gamma_e(c) - \gamma_s(c) + 1} \sum_{m=\gamma_s(c)}^{\gamma_e(c)} |C(k,m)|^2 . \tag{8}$$

The upmix process operates on a per clap basis rather than on individual blocks. Considering the first clap to be upmixed, the target direction $\Theta(c)$ is chosen randomly from the vector of available directions $\Phi(d)$. The spectrum of the current clap is stored in the matrix $S(k,d)$ which holds the mean spectra of the last clap assigned to a distinctive direction:

$$S(k,d) = \begin{cases} \widehat{C}_c(k) & \text{if } S(k,d) = 0 \\ 0.5(S(k,d) + \widehat{C}_c(k)), & \text{else} \end{cases} \tag{9}$$

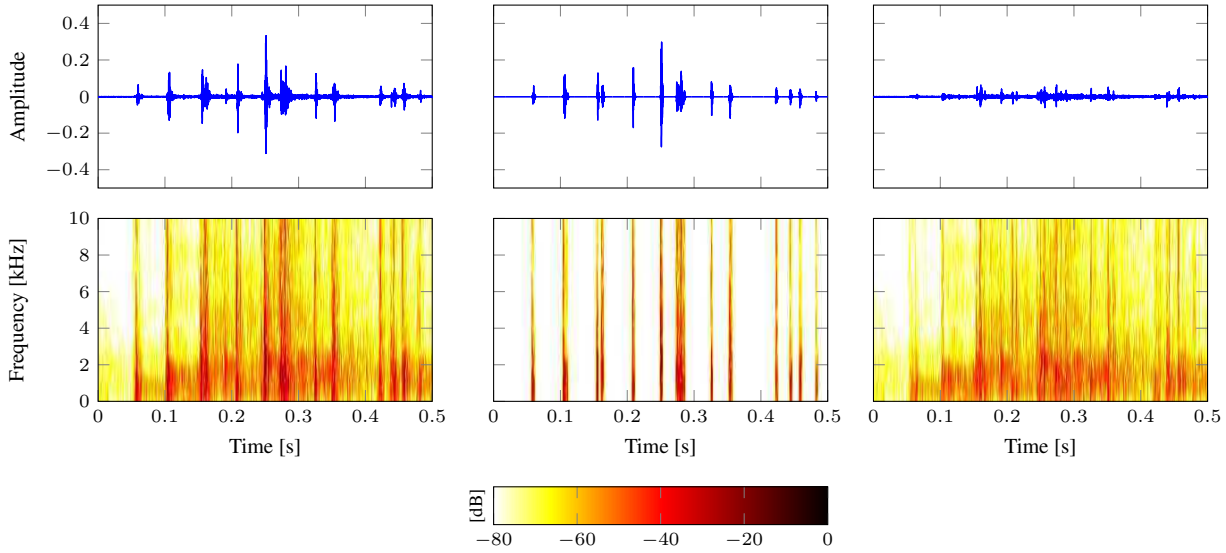

Figure 2: *Waveforms and spectrograms of an applause signal (left) and the respective separated clap signal (middle) and residual noise-like background (right). In the Figure the spectrogram is only plotted in the range of 0 to 10 kHz.*

Additionally, the start block number of the current clap is stored in a vector $T(d)$ holding the respective start block number of the last clap assigned to a direction:

$$T(d) = \gamma_s(c). \tag{10}$$

For all further claps, it is determined how well the respective current clap fits to the claps distributed previously in each direction. This is done with respect to timbral similarity as well as with respect to temporal periodicity. As a measure for timbral similarity, the log spectral distances of the current clap to the previously stored mean spectra of the claps attributed to a direction is computed according to

$$\mathrm{LSD}(d) = \sqrt{\frac{1}{K_u - K_l + 1} \sum_{k=K_l}^{K_u} \left[10 \log_{10} \left(\frac{S_d(k)}{\widehat{C}_c(k)} \right) \right]^2}, \tag{11}$$

where K_l and K_u denote the lower and upper bin of the frequency region relevant for similarity. Similarity is mostly influenced by the spectral peaks/resonances resulting form the air cavity between hands as a consequence of the positioning of hands relative to each other while clapping; these peaks are in the region up to 2 kHz [1, 7, 10]. Based on this observation and including some additional headroom, frequency bins corresponding to the frequency region between 200 Hz and 4 kHz were considered in the log spectral distance measurement.

To determine how well the current clap fits into a periodicity scheme within a certain direction, the time difference (i.e., in blocks) of the current clap to the last distributed clap in every direction is computed:

$$\Delta(d) = \gamma_s(c) - T(d). \tag{12}$$

The resulting time differences are compared to a target clap frequency. In an experiment with more than 70 participants, Neda et. al [12] found clap rates of 2 to 6 Hz with a peak at around 4 Hz. Based on this experimental data and some internal test runs, we chose a target clap rate of $\delta_t = 3$ Hz, corresponding to a target block difference of $\delta_m = 250$ blocks, with an additional tolerance scheme of ± 1 Hz. This means for the time difference, it has to be within the range of $\delta_m - 62$ and $\delta_m + 125$, i.e., 188 and 375 blocks to be considered as a periodic continuation of claps in a direction. In the case that two or more claps occur simultaneously the applause separator detects only one single clap which leads to a gap in the periodicity pattern of the directions the other (masked) claps would have belonged to. To compensate for this effect the search for periodicity is extended to multiples of the expected target block difference, specifically to $3 \cdot \delta_m$. The actually used time difference value is the one with smallest absolute difference to the considered multiples of the target frequency. If the current raw time difference is outside of the tolerance scheme, it is biased with a penalty.

In the next step, log spectral differences as well as time difference are normalized to have zero mean and a variance of one:

$$\widehat{\mathrm{LSD}}(d) = \frac{\mathrm{LSD}(d) - \mu_{\mathrm{LSD}}}{\sigma_{\mathrm{LSD}}} \tag{13}$$

$$\widehat{\Delta}(d) = \frac{\Delta(d) - \mu_{\Delta}}{\sigma_{\Delta}}. \tag{14}$$

Means μ and standard deviations σ are computed using the respective raw time differences and log spectral distances corresponding to the last 25 assigned claps. Finding the most suitable direction for the current clap can be considered as the problem of finding the direction d where the length of a vector $\begin{bmatrix} \widehat{\mathrm{LSD}}(d) \\ \widehat{\Delta}(d) \end{bmatrix}$ is minimal. The vector norm $\Lambda(d)$ for every direction is computed by

$$\Lambda(d) = \sqrt{\widehat{\mathrm{LSD}}(d)^2 + \widehat{\Delta}(d)^2}. \tag{15}$$

The direction to which the current clap fits best is determined as

the index d_{new} where $\Lambda(d)$ is minimal:

$$d_{\text{new}} = \arg\min_{d} \Lambda(d). \tag{16}$$

The current clap is then assigned to the direction $\Phi(d_{\text{new}})$ and the buffer for the respective mean and standard deviation computations as well as $S(k, d_{\text{new}})$ and $T(d_{\text{new}})$ are updated.

As long as each available direction was not assigned with at least one clap, it is additionally checked whether the vector norm $\Lambda(d_{\text{new}})$ is larger than a threshold $\tau_\Lambda = \sqrt{\left(\frac{\tau_{\text{LSD}}}{\sigma_{\text{LSD}}}\right)^2 + \left(\frac{\tau_\Delta}{\sigma_\Delta}\right)^2}$ with $\tau_{\text{LSD}} = 1.9$ and $\tau_\Delta = 5.5$. If so, the current clap is considered as too different from the claps in the so far assigned directions and, consequently, is assigned to a new or 'free' direction. The new direction is chosen randomly from the available free directions.

Finally, the blocks corresponding to the current clap $\gamma_s(c) \leq m \leq \gamma_e(c)$ are scaled with the corresponding panning coefficient $g(\phi)$ [13] to appear under the determined direction $\phi_{d_{\text{new}}}$. Left and right clap signal are obtained according to

$$C_L(k, m) = g(\phi_{d_{\text{new}}}) \cdot C(k, m) \tag{17}$$

$$C_R(k, m) = \sqrt{1 - g(\phi_{d_{\text{new}}})^2} \cdot C(k, m). \tag{18}$$

The final upmixed left and right signals are obtained by superposing the respective left and right upmixed clap and noise signals:

$$L(k, m) = C_L(k, m) + \frac{1}{\sqrt{2}} N(k, m) \tag{19}$$

$$R(k, m) = C_R(k, m) + \frac{1}{\sqrt{2}} \widehat{N}(k, m). \tag{20}$$

4. SUBJECTIVE EVALUATION

To subjectively evaluate the performance of the proposed blind upmix method, a listening test was performed, where the plausibility criteria-driven upmix was compared to a context-agnostic upmix.

4.1. Stimuli

Two sets of stimuli were used for the listening test: one set consisted of synthetically generated applause signals with controlled parameters, whereas the other set consisted of naturally recorded applause signals. All signals were sampled at a sampling rate of 48 kHz. Seven synthetic signals were generated based on the method proposed in [2] and with respective number of virtual people $\widehat{P}_\Sigma = [2, 4, 8, 16, 32, 64, 128]$. The signals had a uniform length of 5 seconds.

The set of naturally recorded signals was a subset of the stimuli used in [11]. For reasons of comparability the same stimuli numbering was applied in this paper. In particular stimuli with numbers 1, 7, 10, 11, 13, 14, and 18 were used, each of which having a uniform length of 4 seconds. Both stimulus sets covered an applause density range from very sparse to medium-high.

Stimuli of both sets were blindly upmixed using different processing schemes. In the first scheme, upmixing was done according to the above proposed processing and will be denoted as *proposedUpmix*. In the second scheme, the applause signals were also separated into claps and noise-like background but the claps were distributed in a context-agnostic manner, i.e., claps were randomly assigned to the available directions within the stereo panorama. This type of upmix is denoted as *randomUpmix*. For both upmixing schemes, 13 discrete directions were available; these were in particular $\Phi = [\pm 30, \pm 25, \pm 20, \pm 15, \pm 10, \pm 5, 0]$. To ensure equal loudness, all stimuli were loudness-normalized to -27 LUFS (loudness unit relative to full scale). Exemplary stimuli are available at https://www.audiolabs-erlangen.de/resources/2017-DAFx-ApplauseUpmix.

4.2. Procedure

The listening test procedure followed a forced-choice preference test methodology. This means, in each trial subjects are presented with two stimuli in randomized order as hidden conditions. Subjects have to listen to both versions and were asked which stimulus sounds more plausible. The order of trials was randomized, as well.

Before the test, subjects were instructed that applause can be considered as a superposition of distinctive and individually perceivable foreground claps and more noise-like background. It was furthermore stated that claps usually exhibit certain temporal, timbral, and spatial structures, e.g., claps originating from the same person do not vary considerably in spatial position and clap frequency, etc. As listening task, subjects were asked to focus on plausibility of foreground claps.

After the instructions, there was a training to firstly familiarize subjects with the concept of foreground claps and noise-like background and secondly with plausibility of foreground claps. In the first case, the same procedure as in [9, 11] was also used here: four exemplary stimuli of varying density and accompanied with additional supplementary explanations regarding foreground clap density were provided. For the second, two synthetically generated stereo stimuli with $\widehat{P}_\Sigma = 6$ were presented whereby in one of which timbre and time intervals between consecutive claps were modified to decrease plausibility. Subjects were provided with supplementary information regarding the stimuli and their expected plausibility.

It should be noted that instructing subjects on such a detailed level appears to bear a risk of biasing them into a certain direction. However, this came as a result of a pre-test where subjects were simply asked to rate naturalness of applause sounds without providing any further information. In this pre-test, it was found in interviews that subjects based their ratings of naturalness on quite different aspects of the stimuli. For example, for some subjects, naturalness was predominantly influenced by the room/ambient sound, others focused more on imperfections of the applause synthesis and did not take spatial aspects into account. The plurality of influencing factors made it impossible to obtain a reasonably consistent rating between the subjects. Thus, it was decided to focus the listening test on the notion of foreground claps. Furthermore, it emerged from the subject inter- views that asking for *plausibility* potentially puts more focus on properties of the clap sounds themselves than using the more broadly defined term *naturalness*.

The listening test was conducted in an acoustically optimized sound laboratory at the International Audio Laboratories Erlangen. Stimuli were presented via KSdigital ADM 25 loudspeakers.

4.3. Subjects

A total number of 17 subjects among which 3 female and 14 male took part in the listening test. Subjects' average age was 33.1 (SD = 8) years ranging from 23 to 53 years. All subjects were stu-

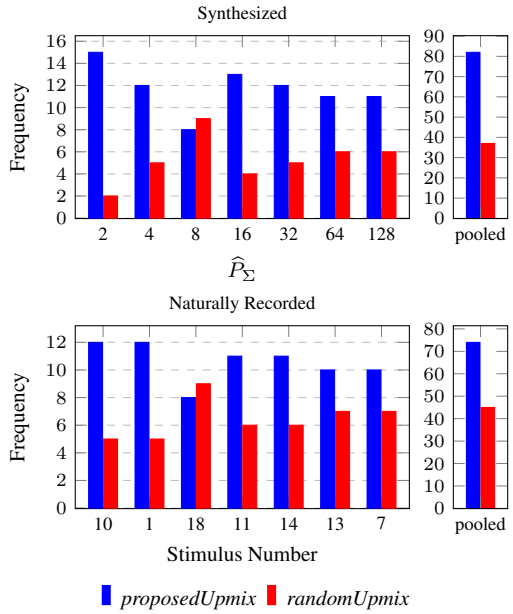

Figure 3: *Histogram of subjects' preference for synthetical generated (top plane) and naturally recorded (bottom plane) stimuli. Additionally, the pooled data across stimuli is provided.*

	Synthesized							
$\widehat{P}_\Sigma$	2	4	8	16	32	64	128	pooled
p-value	.001	.072	.685	.025	.072	.166	.166	<.001
	Recorded							
Stimulus	10	1	18	11	14	13	7	pooled
p-value	.072	.072	.685	.166	.166	.315	.315	.005

Table 1: *P-values of the statistical analysis by means of binomial testing of preference data.*

dents or employees either at Fraunhofer IIS or at the International Audio Laboratories Erlangen with varying degree of experience. In a questionnaire after the listening test, subjects reported in how many listening test they have participated in so far; a number of up to 5 was considered as low-experienced (2 subjects), a number between 6 and 14 was considered as medium-experienced (4 subjects), and anything above was considered as expert listener (11 subjects).

4.4. Results

Figure 3 depicts subjects' preferences for each stimulus where in the top plane responses for the synthetically generated and in the bottom plane responses for the naturally recorded stimuli are depicted. Stimuli are ordered according to increasing applause density. On the respective right hand sides, the pooled data across stimuli are provided.

For the synthesized signals and except for $\widehat{P}_\Sigma = 8$, subjects in total preferred *proposedUpmix*. There is even a trend of the subjects' preference recognizable: at (very) low density the *proposedUpmix* is clearly preferred over *randomUpmix* and with increasing density, subjects' preference for *proposedUpmix* decreased. An exception is $\widehat{P}_\Sigma = 8$ where both upmix methods were about equally frequently chosen. The pooled results support a general preference for *proposedUpmix*.

Also regarding the naturally recorded stimuli, there is a general preference towards *proposedUpmix* visible in the pooled results. Considering the individual stimuli, the data suggests that except for stimulus number 18, where preference for both methods was about similar, *proposedUpmix* was preferred. There is also a weak trend in the data visible indicating that preference for *proposedUpmix* decreases with increasing density. In no case of both stimulus sets, the *randomUpmix* was clearly preferred over *proposedUpmix*.

The indicated trend of preferences in both stimulus sets makes sense given that at high densities clap events occur in a more chaotic and pseudo-random fashion and the event rate gets too dense to be evaluated by the human auditory system on a per clap basis. Instead, the denser the clapping becomes, the more it can be considered as a sound texture and, in further consequence, general signal statistics gain more relevance for perception than properties of individual clap events [14].

Additionally to the visual evaluation, statistical test results are provided in Table 1. For every stimulus as well as the respective pooled data, a one-tailed binomial test was carried out which tested against the hypothesis that the upmix methods were chosen by chance, i.e., an expected relative frequency of 0.5. Considering the individual stimuli only results for $\widehat{P}_\Sigma \in [2, 16]$ of the synthesized and none of the recorded stimulus set are significant, where a significance level of $\alpha = 0.05$ was used. However, the overall results show that for both stimulus sets, *proposedUpmix* was clearly and statistically significantly preferred over *randomUpmix*.

5. CONCLUSION

A blind upmix approach for applause-like signals incorporating quasi-periodicity and timbral similarity of consecutive claps from individual spatial directions was proposed and evaluated. The input signal was firstly decomposed into distinctive and individually perceivable foreground claps and more noise-like background. While the background signal was simply decorrelated, the foreground claps were distributed amongst random positions in the stereo panorama based on timbral similarity and temporal periodicity of claps. The proposed upmix was evaluated by means of a preference test and based on synthetically generated as well as naturally recorded applause stimuli. Results showed that the perceptual quality with respect to plausibility of the spatial scene produced by the proposed upmix was clearly preferred over the one of an upmix where foreground claps were distributed in a context-agnostic manner. Statistical analysis proved subjects' overall preference to be significant.

6. REFERENCES

[1] B. H. Repp, "The Sound of two hands clapping: An exploratory study," *Journal of the Acoustical Society of America*, vol. 81, no. 4, pp. 1100–1109, 1987.

[2] A. Adami, S. Disch, G. Steba, and J. Herre, "Assessing Applause Density Perception Using Synthesized Layered Applause Signals," in *Proceedings of the 19th International Conference on Digital Audio Effects (DAFx-16)*, Brno, Czech Republic, 2016, pp. 183–189.

[3] S. Disch and A. Kuntz, "A Dedicated Decorrelator for Parametric Spatial Coding of Applause-like Audio Signals," in *Microelectronic Systems*, A. Heuberger, G. Elst, and R. Hanke, Eds. Springer-Verlag Berlin Heidelberg, 2011, pp. 363–371.

[4] G. Hotho, S. van de Par, and J. Breebaart, "Multichannel Coding of Applause Signals," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, 2008.

[5] F. Ghido, S. Disch, J. Herre, F. Reutelhuber, and A. Adami, "Coding of Fine Granular Audio Signals Using High Resolution Envelope Processing (HREP)," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017.

[6] C. Uhle, "Applause Sound Detection," *J. Audio Eng. Soc*, vol. 59, no. 4, pp. 213–224, 2011.

[7] L. Peltola, C. Erkut, P. R. Cook, and V. Välimäki, "Synthesis of Hand Clapping Sounds," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1021–1029, 2007.

[8] W. Ahmad and A. M. Kondoz, "Analysis and Synthesis of Hand Clapping Sounds Based on Adaptive Dictionary," in *Proceedings of the International Computer Music Conference*, vol. 2011, Huddersfield, UK, 2011, pp. 257–263.

[9] A. Adami, L. Brand, and J. Herre, "Investigations Towards Plausible Blind Upmixing of Applause Signals," in *142nd International Convention of the AES*, Berlin, Germany, 2017.

[10] A. Jylhä and C. Erkut, "Inferring the Hand Configuration from Hand Clapping Sounds," in *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland, 2008.

[11] A. Adami and J. Herre, "Perception and Measurement of Applause Characteristics: Wahrnehmung und Messung von Applauseigenschaften," in *Proceedings of the 29th Tonmeistertagung (TMT29)*. Cologne, Germany: Verband Deutscher Tonmeister e.V., 2016, pp. 199–206.

[12] Z. Néda, E. Ravasz, T. Vicsek, Y. Brechet, and A.-L. Barabási, "Physics of the rhythmic applause," *Phys. Rev. E*, vol. 61, no. 6, pp. 6987–6992, 2000.

[13] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997.

[14] J. H. McDermott and E. P. Simoncelli, "Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.