Paasonen, Juhani; Pulkki, Ville

Short-range rendering of virtual sources for multichannel loudspeaker setups

Published in:
149th AES Convention

Published: 22/10/2020

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Paasonen, J., & Pulkki, V. (2020). Short-range rendering of virtual sources for multichannel loudspeaker setups. In *149th AES Convention* Article 10401 Curran Associates Inc.. https://www.aes.org/e-lib/browse.cfm?elib=20938

Audio Engineering Society

Convention Paper 10401

Presented at the 149th Convention
Online, 2020 October 27-30

Short-range rendering of virtual sources for multichannel loudspeaker setups

Juhani Paasonen[1] and Ville Pulkki *AES Fellow*[1]

[1]*Aalto University*

Correspondence should be addressed to Juhani Paasonen (juhani.paasonen@aalto.fi)

ABSTRACT

A method for rendering virtual sources is proposed, where the sources are perceived to be closer than the loudspeaker array radius. The development is based on an informal finding about closer perception of the range of virtual sources rendered coherently over multiple loudspeakers. To avoid sound quality issues inherent with coherent rendering, the input is split to two streams, one with more transients and the other with smoother temporal envelope. The transient stream is rendered over coherent reproduction, and the continuous stream is processed with time-frequency-domain spreading technique. The results from localization tests with moving sources show that the proposed method produces the perception of closer distances on both sweet-spot and off-sweet-spot listening.

1 Introduction

In object-based audio a number of audio objects is provided as monophonic audio channels with metadata, that are rendered independently over the loudspeaker setup in listening scenario. The rendering system targets to create such perceptual properties to the virtual sources as defined by the metadata, such as direction, distance, and loudness. The rendering of distance is typically performed with such methods as controlling the level and/or direct-to-reverberant ratio of audio objects. However, in practise the methods have shortcomings in rendering the distance to be closer than the distance to the loudspeakers in listening room. This issue is addressed in this article, and a method is suggested for the task.

The development of the method was based on earlier work in the group, where it was found that increasing the amount of coherent sound in several loudspeakers brought the virtual sources closer the listener [1]. The earlier method, however, produced the desired effect only in limited listening area, i.e. sweet spot, produced some timbral colouring issues, and had also limited range of operation.

The development of the method proposed in this article is based on trial-and-error development of audio rendering methods. The requirements set for the method were that a consistent perception of sources closer to the loudspeaker should be achieved in large listening area, while the sound quality should not suffer from processing.

2 Background

2.1 Perception of distance

Zahorik et al. [2] reviewed a collection of articles on perception of distance. Five acoustic cues were identified for perception of distance. First, the sound pressure

level is considered the primary cue. In general, the farther the sound source, the lower the amplitude as the level of direct sound decreases by 6 dB for each doubling of the distance. However, this is valid only in free field, as the room response masks the effect outside the reverberation radius. The level of reverberant sound field does not depend on observation position in a room in ideal conditions. Consequently, the direct-to-reverberant ratio of sound decreases with increasing distance, and it has been found that the perceived direction is affected by the ratio. The third identified cue is the spectral effect of atmosphere on sound. Air absorption affects more prominently high frequencies, and at distances greater than 15 m this may be audible as a low-pass effect in sound. In addition, the perception of the direction of close-by sources binaural cues have been suggested to be conveyed by the near-field effects of ITD and ILD cues.

The perceived distance is also related to the sound emitted by the source itself: it is difficult to judge the distance of a static sound source when the emitted sound is not familiar. However, often the listener has some familiarity about the loudness produced by common sound sources at different distances, and the judged distance can be based on memorized distances. This implies that distances of unnatural sounds are more difficult to judge than those of "natural" sounds. Moreover, the movement of the sound source makes the task of perceiving the distance easier. The relative changes in distances are more pronounced, and the expected result is thus easier to judge.

2.2 Rendering of distance

2.2.1 Simulation of distance attenuation

As mentioned in Sec. 2.1, the loudness is one of major cues in distance perception. This can be relatively simply simulated using distance attenuation law. Often the source in virtual reality is assumed to be point-like, emitting a spherical wave. The sound pressure $p(r)$ in the wave is inversely proportional to the distance r from the mid-point (symmetry point) of the source,

$$p(r) \propto q/r, \tag{1}$$

where q is volume velocity of the source. Often the sound pressure at certain distance d_0 is assumed to have a certain value p_0, and the target is to compute pressure p_1 at another distance d_1. Using the previous equation, it holds

$$p_1 = \frac{d_0}{d_1} p_0. \tag{2}$$

This yields that the ratio $\frac{d_0}{d_1}$ can be used as a gain factor when simulating the effect of distance attenuation on sound.

2.2.2 Vector base amplitude panning

Vector base amplitude panning (VBAP) [3] is a method to render virtual sources to user-defined directions, and it does not per se control the perceived distance. However, some mixing systems use a combination of distance attenuation and VBAP (or any other panning law), which already produces relatively plausible perception of virtual sound source distance, at least when the virtual source moves on a continuous trajectory.

In 2D VBAP a loudspeaker pair is specified with two vectors, the unit-length vectors $\mathbf{l}_m$ and $\mathbf{l}_n$ point from the listening position to the loudspeakers. The intended direction of the virtual source (panning direction) is presented with a unit-length vector $\mathbf{p}$. Vector $\mathbf{p}$ is expressed as a linear weighted sum of the loudspeaker vectors

$$\mathbf{p} = g_m \mathbf{l}_m + g_n \mathbf{l}_n. \tag{3}$$

Here g_m, and g_n are called gain factors of respective loudspeakers. The gain factors can be solved as

$$\mathbf{g} = \mathbf{p}^T \mathbf{L}_{mn}^{-1}, \tag{4}$$

where $\mathbf{g} = [g_m \; g_n]^T$ and $\mathbf{L}_{mn} = [\mathbf{l}_m \; \mathbf{l}_n]$. The calculated factors are used in amplitude panning as gain factors of the signals applied to respective loudspeakers after suitable normalization, e.g. $||\mathbf{g}|| = 1$.

2.2.3 Distance-based amplitude panning

Distance-Based Amplitude Panning (DBAP) [4] is an amplitude panning technique, which at least in principle renders also the distance of virtual sound source. The 3D positions of loudspeakers are measured, and the virtual source positions are defined to the same space. The loudspeaker gains are computed based upon the Euclidian distance d_i from the targeted position of virtual source to the real position of loudspeaker i

$$d_i = \sqrt{d_i^2 + r_s^2} \tag{5}$$

where r_s is an optional spatial blur that avoids the application of all sound of a single loudspeaker when the source position and loudspeaker position are equal. Loudspeaker gains are calculated by

$$g_i = \frac{k}{d_i a}, \tag{6}$$

where a is a distance attenuation factor

$$a = \frac{R}{20\log_{10} 2}, \tag{7}$$

and R is a free parameter that varies between 3 dB corresponding to power normalization, and 6 dB as amplitude normalization. Coefficient k enforces a unit power sum

$$k = \frac{1}{\sum_{i=1}^{M} 1/d_i^{2a}} \tag{8}$$

For $r_s > 0$, all loudspeakers are active for any virtual source position.

2.2.4 Auditory distance rendering

The previous work in this field conducted in our group was reported in [1]. The work was based on informal finding, that when coherent sound signals were applied to several loudspeakers, the listener perceived the virtual source to be closer than the loudspeakers. In the proposed method, two additional coherent amplitude-panned sources were added to fixed angles around the original sound source. The level of additional virtual sources is controlled by heuristic rules based on informal listening. It is shown, that when the level of the additional sources is increased, the perceived virtual sound sources may be brought to about 25–50 % closer distance than the distance of the loudspeakers according to formal tests. The processing was thus designed to produce coherent loudspeaker signals, which indeed brought the virtual sources closer. However, the application of coherent sound into several loudspeakers causes similar artifacts as in low-order Ambisonics playback [5], namely sound coloration caused by comb-filter effects.

The psychoacoustical mechanisms causing the virtual sources to be perceived closed than the loudspeaker distance in the technique proposed in [1] are not clear, as the added coherence on loudspeaker reproduction does not directly correspond to any of the distance cues mentioned in the previous section. However, although the psychoacoustic knowledge can not be used to explain the phenomenon, the subjective tests conducted show that the rendering is possible with the method.

3 Proposed method

The basic idea in the proposed method is to use frequency-domain rendering techniques to distribute sound into several loudspeakers, which should avoid the comb-filter effects, yet potentially creating the perception of sound source in closer distance. The processing applied in parametric spatial audio [6] may be used for distribution, where the sound is divided into frequency bands, and different bands are applied to different directions. However, the process often distorts the temporal structure of signal, and audible artifacts appear. To avoid this, we propose a to divide the input signal into two streams, one containing more transients, and the other more the continuous parts of signal. The motivation is, that different distribution methods for transient stream and continuous stream can be used, which minimize quality defects in audio, but still maximize the accuracy of distance rendering.

3.1 Division of input sound into transient and continuous parts

The processing uses a local crest factor to separate transients from temporally smooth continuous signals. The processing could be implemented for broad-band input signal. However, it is clear that the transients may occur at limited frequency areas, such as the onset of a high-pitched bell does not evoke transients at low frequencies. The processing should thus be conducted in frequency bands, and two-octave bands were chosen based on [7] that shows that phase effects are local on human hearing on about two-octave range. The block diagram for the processing is shown in Fig. 1.

The system was designed to be relatively conservative, meaning that the separation between transients and ongoing signals is not very stark, meaning that quite prominent amount of transients leaks to continuous stream and vice versa. In the first step, the input audio is split to two-octave bands with a perfect-reconstruction filter bank. We use afSTFT [8] for the filtering. The filter gains are generated with SDM Toolbox for Matlab [9]. In the second step, a moving crest factor parameter is computed. For this, we calculate two moving window sums for each sample in each subband, a short one $S(t,k)$ and a long one $L(t,k)$.

$$S(t,k) = \sum_{i=t-(w_S(k)/2)}^{i=t+(w_S(k)/2)} |x(t,k)|, \text{and} \tag{9}$$

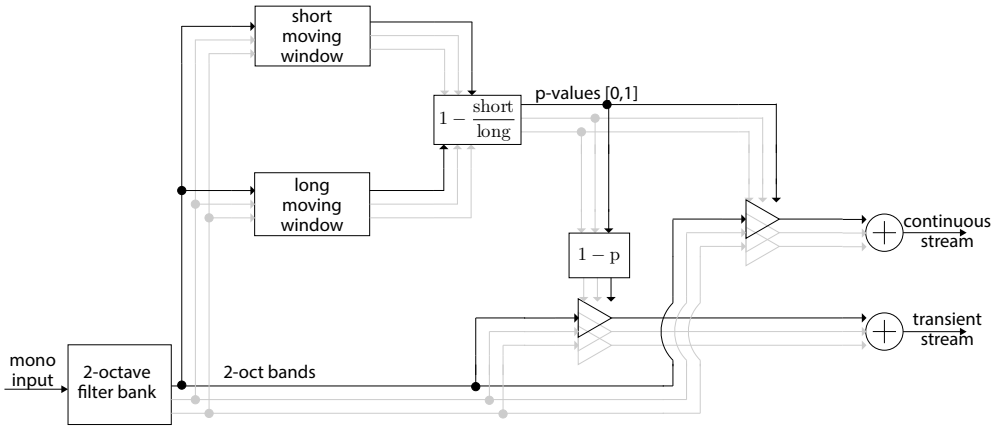

Fig. 1: Block diagram for separation of input signal into transient and continuous streams.

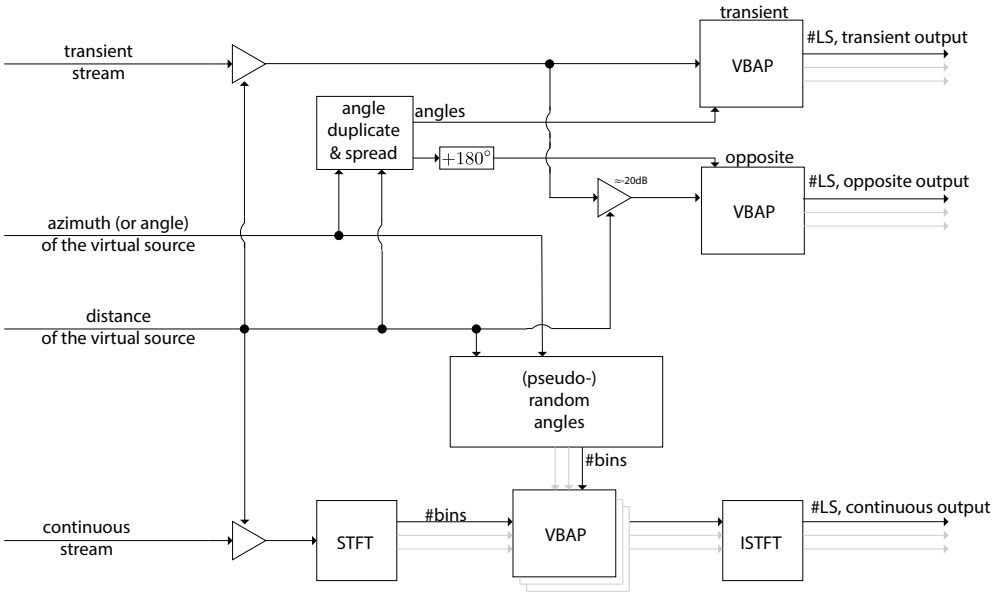

Fig. 2: Block diagram for distance rendering

$$L(t,k) = \sum_{i=t-(w_L(k)/2)}^{i=t+(w_L(k)/2)} |x(t,k)|, \tag{10}$$

where $x(t,k)$ is the output of the filter bank, t is the sample index, k is the subband index, and $w_S(k) = w_L(k)/25$ are lengths of the rectangular windows. The window lengths depend on k in such a way that at each crossover frequency of the filter bank five full cycles of the crossover frequency equals to the value of $w_S(k)$

In the third step, the local crest factor values $p(t,k)$ are calculated as

$$p(t,k) = 1 - \frac{S(t,k)}{L(t,k)} NB, \tag{11}$$

where N is the normalizing factor compensating the different window lengths, and B is a balancing factor that is decided based on informal tuning. In this case, $N = 1/25$ and $B = 1/6$. Unfortunately, p formed in this way is not guaranteed to be bounded within 0..1. Therefore p is manually forced to stay within that range:

$$p(t,k) = \begin{cases} 0, & p(t,k) < 0 \\ 1, & p(t,k) > 1 \\ p(t,k) & \text{otherwise.} \end{cases} \tag{12}$$

Finally, the filter bank output is multiplied samplewise and bandwise with the matching $p(t,k)$. Thus, the output for continuous stream is $c(t) = \sum_k p(t,k)x(t,k)$ and for the transient stream $t(t) = \sum_k (1-p(t,k))x(t,k)$, where $x(t,k)$ is the output of the filter bank. If the two streams are summed together, the result is the original input.

3.2 Distance rendering

This section presents the methods suggested for synthesizing the perceived distance for transient and continuous streams. The block diagram is shown in Fig. 2.

3.2.1 Transient stream distance rendering

The transient stream is rendered with similar approach as was done in [1], where coherent virtual sources were applied to several directions, with issues in sound quality due to comb-filter effects. We assume, that this approach is manageable with transient stream, since it optimally contains only very short onsets of sound, and colorations may be less annoying than with ongoing signals since human hearing requires relatively long

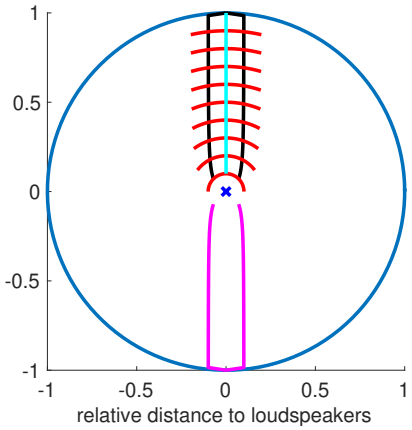

Fig. 3: The positions of the virtual sound sources pictured from above. The blue circle depicts the loudspeaker ring and the blue cross marks the listener position. The cyan line depicts the position of the actual virtual sound source at different distances in front of the listener. The black lines depict the paths of two virtual sources used to reproduce transient stream. The red lines depict the spatial spread of the continuous stream at quantized distances. The magenta lines depict the positions of two virtual sources used to reproduce the opposite transient stream. At one time, the virtual sources of transient and continuous streams are positioned to the same distance from the listener as is the actual virtual source.

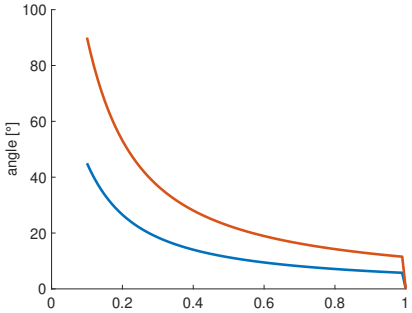

Fig. 4: Apparent widths of the different streams from listener point of view. The blue line is the width of the transient and opposite streams at different distances from the listener. The red line is the maximum spread of the continuous stream at different distances from the listener.

stimulus to perceive accurately the timbre of sound [10]. On the other hand, any decorrelation technique would cause audible defects in sound quality, since human listeners are sensitive to the temporal structure of transients [11].

A relatively simple heuristic method was designed, where the transient stream is applied to two virtual sources in equal angular distance around the actual panning direction in the horizontal plane. The constant $r = 0.1$ defines the distance between the virtual sources on the arc of the unit circle. The angular separation between the virtual sources of transient stream α_w is computed as

$$\alpha_w = \arctan(r/d), \tag{13}$$

where $d \in [0.1, 1]$ is the distance of the virtual sound source from the center of the unit circle. This equation was motivated by the idea of the virtual source having a physical extent, and then tuned by trial-and-error experimentation. In other words, the apparent width of the virtual sound source is narrower as it is close to the loudspeakers, and wide up to 90°as it is close to the listener. This is illustrated in Fig. 3.

In the same instance, we take the azimuth direction of the virtual source θ into account

$$\alpha = \arctan(\theta \pm r/d). \tag{14}$$

Transient stream attenuated by 6 dB is applied to the virtual sources in the resulting two directions using VBAP [3].

If the virtual source comes closer the listener, and continues the route to the other side of the listener, it might be beneficial to fade in the virtual source on the direction where it is going. To accomplish this, the transient stream is also reproduced with lower level on the opposite side, which is called as the opposite stream. The opposite stream consists also of two virtual sources, mirrored to the other side of the listener. In informal listening, it was noted that such duplication increases the plausibility of the reproduction. The gain of the opposite stream depends on the distance of the sound source, and it is defined by

$$g_{opp} = \frac{1-d}{\sqrt{6}}, \tag{15}$$

which, equals to zero at the distance of loudspeakers. The level of the opposite stream is thus kept low, and based on informal listening it is difficult or impossible to perceive the opposite stream as an independent sound event. In other words, the opposite stream is barely audible. The positioning of the virtual sources of opposite stream is shown in Fig. 3.

3.2.2 Continuous stream distance rendering

The method proposed in [1] applied coherent sound to a broad distribution of loudspeakers, which created the perception of closer distances, with known coloration issues. In the present work we hypothesized, that also the time-frequency-domain methods to synthesize the perception of spatially wide virtual sources [12] could also be effective also for the perception of distance. Such frequency-domain methods potentially have less issues on coloration.

In the method, the continuous stream signal is converted into STFT domain, and the frequency bins are distributed spatially using a pseudo-random algorithm [6] on a continuous arc with left and right end points. The angle between the left and right points is double of the value of the angle between the two virtual sources for the transient stream, as shown in Eq. 14. When reproduced in this way, the continuous stream is perceived plausibly to be a wide source, which seems to generate the perception of source in closer distance, at least in the tested cases.

3.3 Controlling the levels of virtual sources

The base gain controls the overall sound pressure level, and it is defined by

$$g = k + \frac{1-k}{\max(0.1, d)}, \tag{16}$$

where $k \in [0, 1]$ is a chosen constant and d is the relative distance of the source relative to the loudspeakers. In other words, the gain is always unity at loudspeaker distance. k can be changed to control how aggressively the gain increases as d decreases, while maintaining the inverse-r shape for the gain wrt distance. In our work, we chose $k = 0.3$ based on informal listening. To prevent the very high values for g, the smallest possible value for denominator is 0.1.

A further attenuation is added in continuous stream, to simulate increase in direct-to-reverberant ratio, which

increases when a real source is closer to a receiver. This is conducted by

$$g_{cont} = g\left((1-f)d+f\right), \tag{17}$$

where f is the coefficient. There is no extra attenuation at the distance of loudspeakers, and at zero distance the gain will be multiplied with f. We have chosen $f = 1/\sqrt{2}$, this makes the continuous stream 3 dB softer than the transient stream on very short distances.

4 Listening experiment

A formal listening experiment was conducted, where the target was to measure if the proposed method produced virtual sources to close distances, and how does the performance compare with amplitude-panning based techniques.

4.1 Listening experiment design

A set of moving virtual source samples was designed, where the source moved from a position at the distance of the loudspeakers to another with a path. The task for the listeners was to indicate the geometry of the path using a graphical user interface. The users used the system on a tablet, and the curve was changed by dragging a finger on the interface. The curve was always an arc of a circle, defined by the start and end position, and the third point defined by the user. The third point was forced to lay halfway between the start and end points. A typical response is expected to be closer than the start and end points, thus controlling the shortest distance of the arc to the sweet spot. The start and end positions lay at the loudspeaker distance. Three different curves with different shortest distances to the sweep spot were used in the test. On each trial, each sample was played twice. The subject could not skip nor replay samples. They could start to input their answer after the sample had played once, but they could only move forward after the two plays had finished. The user interface is presented in Fig. 6.

Three methods for virtual source positioning were compared in the listening experiment. The proposed method (DR) was used as described in Ch. 3, distance-based amplitude panning (DBAP) as in Sec. 2.2.3 and vector-base amplitude panning (VBAP) as described in Sec. 2.2.2. To match the gains of the outputs, first Eq. 16 was used to make all the methods have similar changes for the path of the virtual sound source. Each

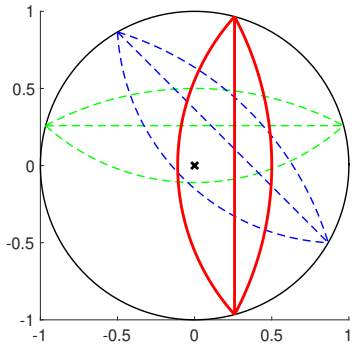

Fig. 5: The paths of the virtual sound sources in the listening experiment.

trial with similar conditions was then gain-matched by listening-based manual tuning, as there exists no trivial automated method for perceptually match the gains automatically. The test was conducted in an ITU BS-1116.1 compliant room of size 6.4x5.6x3.0 m with nine Genelec 8260A as loudspeakers at 2.0 m distance from the sweet spot. In total nine loudspeakers were positioned on angles of 0, $\pm$30, $\pm$70, $\pm$110, and $\pm$160 degrees.

Two audio samples were used in the test, a male speaking English, and a musical sample with a guitar and castanets. The first was chosen since speech is a very familiar signal for human listeners. The music excerpt was chosen since it has both a repeating transient element in the castanets and a somewhat smoother sound from the guitar. There were three different paths for the virtual sound sources, each repeated in total of three different orientations. The paths are depicted in Fig. 5.

In total, each subject responded to 54 trials for sweet spot and another 54 trials for the off-sweet spot seat, together 108 answers from each subject. The off-sweet-spot sitting position was at 45° front left from the sweet spot at 0.26 of loudspeaker distance, or 50 cm, towards the loudspeakers. This way, in each rotation of the virtual source paths, the off-sweet-spot position had different distance to the paths. The order of trials was randomized, and half of the subjects started seated on the sweet spot and the other started seated on the off-sweet spot. The test system was run on Max 7 graphical audio programming tool.

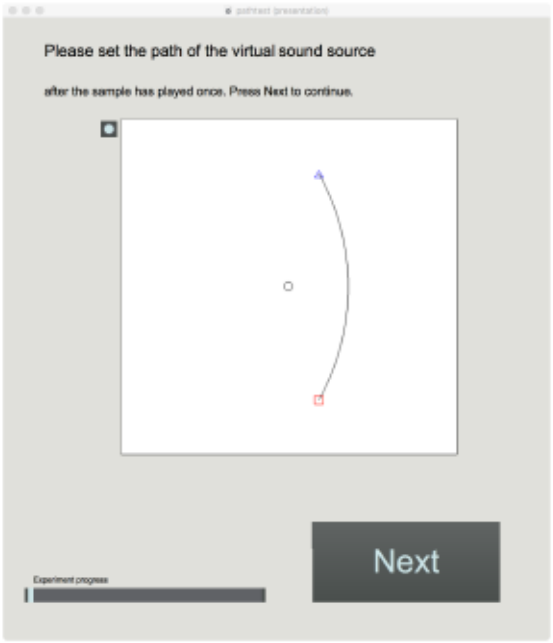

Fig. 6: User interface for the listening experiment. The subject controls the arc between the triangle and the square, which are the start and end point of the path of the virtual sound source, respectively. The seat of the user is marked with the black circle. In this case, the subject is seated on the sweet spot.

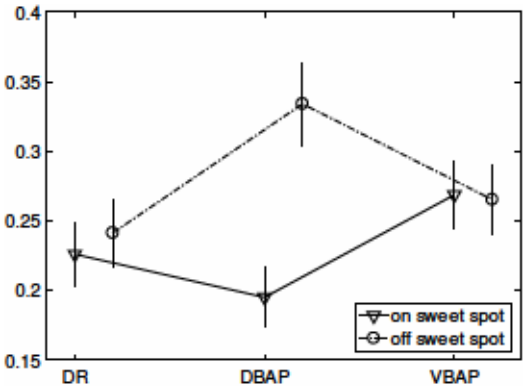

Fig. 7: Overall results of each method in both on and off sweet spot positions, represented by triangles and circles, respectively, and the 95 % confidence intervals. The number scale is the absolute value of the error, measured in distance normalized with the loudspeaker distance.

4.2 Listening test results

A total of 16 subjects, 12 male and 4 female, aged between 21 and 38, participated the listening test. All of the subjects were lab members with extensive listening experience.

The subjects were controlling the shortest distance between the arc and the subject d_{subject}, which was the output of each trial. The output was compared to the corresponding distance for the reference path $d_{\text{reference}}$, by computing the difference measure

$$d_{\text{difference}} = d_{\text{subject}} - d_{\text{reference}}. \tag{18}$$

The sign of the differences thus tells on which side of the correct virtual source path the subject placed their answer. Absolute values of the differences $d_{\text{absdif}} = |d_{\text{difference}}|$ were used in ANOVA and the most part of statistical analysis, to quantify the accuracy achieved with the methods. If the absolute values were not used, answers on opposite sides of the correct value would cancel each other due to different signs.

The results were subjected to ANOVA analysis assuming fixed effects with two-factor interactions over the independent variables *method*, *sample*, *start position*, and *path* separately for sweet spot and off-sweet-spot results. The analysis showed a number of significant effects, however, the effect of renderer is deemed most important in the scope of the article. When the subjects were seated on the sweet spot, the main effect of method was clearly significant $F(2, 838) = 12.4$, $p < 0.0001$ for sweet spot listening and $F(2, 838) = 17.6$, $p < 0.0001$ for off-sweet spot listening. The effects are shown in Fig. 7 with 95 % confidence intervals.

For on-sweet-spot listening the results with DR and DBAP are within the confidence interval showing no significant difference, but VBAP performed significantly worse. When the subjects were sitting off the sweet spot, DR performed significantly better than DBAP. However, in this case there was no significant difference between DR and VBAP. DR and VBAP showed no significant difference with regard to listening position, unlike DBAP, with which the subjects' performance deteriorated when seated off the sweet spot.

An interesting finding is that the rotation of the sound path has a significant effect on the subjects' perfor-

mance. The subjects performed the best, when the virtual sound path was rotated 45°, where the virtual sound source moved from front left to back right. In this position, even the off-sweet spot results mostly outperformed the two different rotations with all of the techniques. The virtual path rotated by 90°, travelling from front left to front right of the subject performed equally well with all the renderers. The unrotated position, travelling from front right to back right had the most variance between on and off sweet spot seats.

Many subjects reported that specifically in off sweet spot experiment the sound paths often seemed to jump around inexplicably. Many subjects also reported that they perceived some samples inside their head. We suspect these were mostly DBAP samples during the part of the path with the shortest distance to the subject, as in these cases all the loudspeakers output coherent signal with nearly equal level. Some subjects reported that they had drawn the arc through the head of the listener in the UI in these cases. This was analyzed from the listening test data and the results are shown in Fig. 8. For on-sweet spot position, the arc was drawn through the head in total 3, 39, and 1 times for DR, DBAP, and VBAP, respectively. For off-sweet spot seat, the corresponding numbers were 11, 26, and 6. This suggests that DBAP system produces more often the perception of virtual source inside the head.

In informal comments it was reported that the proposed method causes the virtual source to be extended in volume, which means that the source was perceived to occupy a considerable volume between the loudspeaker and the listener. The timbre produced by the method was perceived not to have severe coloring.

5 Discussion

It is an interesting question why do virtual sources appear to such distances as they do in the proposed method. Intuitively it is not clear why a closer perception of a virtual source is obtained when the apparent width is increased. The only analogous case comes with large sources, where larger angular range occurs with closer distances. However, for example human voice, which was used in the tests, can not be considered as a large source. An analysis of ear canal signals, and their relation to known distance cues would be an interesting research topic, which is left for future studies.

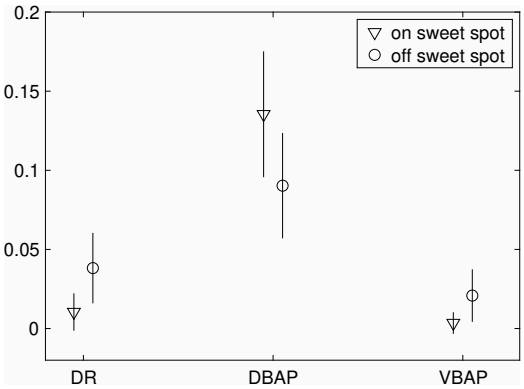

Fig. 8: Proportion of answers for which the subject had marked the path to intersect with the avatar in the GUI (Fig. 6) of the listening experiment.

The effect of the proposed method on timbre is now discussed. The measurement of the timbre of moving sources is not straightforward. The comparison of static close-distance virtual sources reproduced using different methods would have been a more straight-forward task. However, we did not find static sources relevant for the study, since the synthesized distance of a virtual source is perceived with high saliency only when it is in movement. Measuring the timbre quality for snapshots of virtual sources along the path of movement was neither seen meaningful, since the timbre of the proposed technique is in fast random change all the time, and a singular snapshot would not reveal the actual quality. Likewise, the timbral effect on DBAP depends on the current distance of the virtual source.

Nevertheless, we can compare the timbral artifacts produced by proposed techniques already by existing information. Both VBAP and DBAP produce comb-filter effects, where the spectral notches travel in frequency up and down depending on the contribution of individual loudspeakers. The proposed method in turn produces such spectral notches with lower saliency, since the loudspeaker signals are relatively incoherent, and such artifacts are largely missing. It should be noted, though, that with sparse loudspeaker arrays the proposed method produces more coherent output. Further, the time-frequency-domain processing may cause some more or less audible changes in the temporal structure of the signal, which may be audible as sound coloration. Subjective measurement of the colorations on relatively long paths of virtual sources seems ini-

tially a topic where relatively complicated listening tests should be conducted, which is left as a topic for future research. However, as an informal finding by the developers and the subjective tests was that the timbral artifacts were very low or missing with the proposed technique.

6 Conclusions

A method to control the distance of a virtual source is proposed. The system is designed for horizontal loudspeaker setups, all having an equal distance from the sweet spot. The monophonic input signal is divided into two streams, one with higher dominance of transients and the other with higher dominance of smooth ongoing parts of signal. This approach makes it possible to have different strategies for distance rendering for the streams, and consequently the artifacts caused by processing are mitigated. The methods to control the perceived distance are based on trial-and-error testing in laboratory conditions where a heuristic technique was formed. In the resulting technique the transient stream is split to two directions around the desired angle of the sound source, and it is also duplicated in the opposite direction and attenuated by 15-20 dB. The continuous stream is distributed frequency-wise, and the width of the spreading and the gain of each stream depends on the distance of the virtual sound source with a set of heuristic rules.

In a listening experiment we compared the proposed method (DR) to distance-based amplitude panning (DBAP) and vector-base amplitude panning (VBAP). All the methods were gain-matched to exclude the sound level as a distance cue. The listening experiment shows that the proposed method was on par with DBAP, when the subject was seated on the sweet spot, with VBAP performing worse. When seated off the sweet spot the proposed method was on par with VBAP, while DBAP performed the worst.

7 Acknowledgment

This project has been supported by Fraunhofer IIS and the Academy of Finland.

References

[1] Laitinen, M.-V., Walther, A., Plogsties, J., and Pulkki, V., "Auditory distance rendering using a standard 5.1 loudspeaker layout," in *Audio Engineering Society Convention 139*, Audio Engineering Society, 2015.

[2] Zahorik, P., Brungart, D. S., and Bronkhorst, A. W., "Auditory distance perception in humans: A summary of past and present research," *ACTA Acustica united with Acustica*, 91(3), pp. 409–420, 2005.

[3] Pulkki, V., "Virtual sound source positioning using vector base amplitude panning," *Journal of the audio engineering society*, 45(6), pp. 456–466, 1997.

[4] Lossius, T., Baltazar, P., and de la Hogue, T., "DBAP–distance-based amplitude panning," in *ICMC*, 2009.

[5] Solvang, A., "Spectral impairment of two-dimensional higher order ambisonics," *Journal of the Audio engineering Society*, 56(4), pp. 267–279, 2008.

[6] Pihlajamäki, T., Santala, O., and Pulkki, V., "Synthesis of spatially extended virtual source with time-frequency decomposition of mono signals," *Journal of the Audio Engineering Society*, 62(7/8), pp. 467–484, 2014.

[7] Laitinen, M.-V., Disch, S., and Pulkki, V., "Sensitivity of human hearing to changes in phase spectrum," *Journal of the Audio Engineering Society*, 61(11), pp. 860–877, 2013.

[8] Vilkamo, J., *Alias-free short-time Fourier transform*, https://github.com/jvilkamo/afSTFT, 2015.

[9] Tervo, S., *SDM Toolbox*, https://se.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox, 2018.

[10] Moore, B. C., *Hearing*, Academic Press, 1995.

[11] Laitinen, M.-V., Kuech, F., Disch, S., and Pulkki, V., "Reproducing applause-type signals with directional audio coding," *Journal of the Audio Engineering Society*, 59(1/2), pp. 29–43, 2011.

[12] Pulkki, V., Politis, A., Pihlajamäki, T., and Laitinen, M.-V., "Spatial Sound Scene Synthesis and Manipulation for Virtual Reality and Audio Effects," *Parametric Time-Frequency Domain Spatial Audio*, 2017.