EFFICIENT MERGING OF MULTIPLE AUDIO STREAMS FOR SPATIAL SOUND REPRODUCTION IN DIRECTIONAL AUDIO CODING

Giovanni Del Galdo, Fabian Kuech, Markus Kallinger, and Richard Schultz-Amling

Fraunhofer IIS, Audio Department, Am Wolfsmantel 33, 91058 Erlangen, Germany

giovanni.delgaldo@iis.fraunhofer.de

ABSTRACT

Directional Audio Coding (DirAC) is an efficient technique to capture and reproduce spatial sound. The analysis step outputs a mono DirAC stream, comprising an omnidirectional microphone pressure signal and side information, i.e., direction of arrival and diffuseness of the sound field expressed in time-frequency domain. This contribution proposes a method to merge two or more mono DirAC streams for a joint playback at the reproduction side. This problem arises in applications such as immersive spatial audio teleconferencing. With respect to a trivial direct merging, the proposed method is more efficient as it does not require the synthesis step. From this follows the benefit that the loudspeaker setup at the reproduction side does not have to be known in advance. Simulations and informal listening tests confirm the absence of any artifacts and that the proposed method is practically indistinguishable from the ideal merging.

Index Terms— Spatial audio coding, spatial audio processing

1. INTRODUCTION

Spatial audio processing is becoming more important as the variety of possible applications for multichannel audio is constantly increasing. An efficient approach to the analysis and reproduction of spatial sound is the Directional Audio Coding (DirAC) technique [1]. DirAC uses a parametric representation of sound fields based on the features which are relevant for the perception of spatial sound, namely direction of arrival (DOA) and diffuseness expressed in frequency subbands. Together with the pressure reading of an omnidirectional microphone they form a so-called *mono DirAC stream*. On the reproduction side, the signals of the loudspeaker channels are determined as a function of the DirAC parameters so that an accurate spatial rendering can be achieved at a desired listening position for arbitrary loudspeaker setups.

Note that there are substantial differences between DirAC and parametric multichannel audio coding, such as MPEG Surround [2], in that MPEG Surround is based on a time-frequency analysis of the different loudspeaker channels whereas DirAC takes as input the channels of coincident microphones, which effectively describe the sound field in one point.

This paper deals with the problem of merging two or more mono DirAC streams for a subsequent joint playback based on the conventional DirAC synthesis. One realistic scenario in which this problem arises is a spatial audio teleconferencing application with more than two parties. It has been shown that the introduction of spatial information of different speakers increases their intelligibility [3, 4]. For the considered application, we assume that the desired spatial information is included by representing each participating party by a corresponding mono DirAC stream. Let party A communicate with parties B and C via a multipoint control unit (MCU). The mono DirAC streams generated at the locations of B and C are transmitted to the MCU. The MCU needs to merge these two separate streams and send the resulting stream to A, which carries out the reproduction with the conventional DirAC synthesis technique.

The proposed scheme can also be used to provide *interactive* spatial sound in gaming and teleconferencing applications. In fact, the DirAC streams to be merged can be generated synthetically, meaning that proper side information can be added to a mono audio signal. In the teleconferencing example mentioned above, party A might receive two audio streams from B and C without any side information (e.g. due to legacy systems). It is then possible to assign each stream a certain direction and diffuseness, thus adding the side information needed to construct the DirAC streams, which are then merged before the synthesis step.

It should be mentioned that similar application scenarios are considered in the context of MPEG Spatial Audio Object Coding (SAOC) [5]. SAOC builds upon the rendering engine of MPEG Surround and treats different sound sources as objects, as opposed to the directional representation of DirAC.

A trivial solution to the problem of merging consists in adding the loudspeaker signals computed independently for each mono DirAC stream. This direct approach suffers from the severe drawback that the loudspeaker setup needs to be known in advance. In contrast, our method operates on the DirAC streams, thus avoiding in toto the synthesis step. By doing so, it follows a lower computational complexity as well as more flexibility, as no knowledge on the reproduction side is required.

The paper is organized as follows: Section 2 reviews the DirAC principle of operation and derives the proposed method. Section 3 proves the validity of our proposal on the basis of simulations, whereas Section 4 draws the conclusions.

2. PHYSICAL QUANTITIES AND DIRAC

As already mentioned in the introduction, DirAC operates in transformed time-frequency domain. Let $P(k,n)$ be the pressure reading of an omnidirectional microphone after a short time Fourier transform (STFT), where k and n are the frequency and time indices, respectively. Let further $\boldsymbol{U}(k,n) = [U_x(k,n),\, U_y(k,n),\, U_z(k,n)]^{\mathrm{T}}$ be the complex particle velocity vector for the time-frequency tile identified by (k,n). The knowledge of $P(k,n)$ and $\boldsymbol{U}(k,n)$ is necessary to compute the DirAC parameters, namely DOA and diffuse-

Part of the research leading to this paper has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2007-214793. The authors gratefully acknowledge Ville Pulkki and his colleagues from the Laboratory of Acoustics and Signal Processing, Helsinki University of Technology, TKK, for the helpful comments and discussions on DirAC.

ness. There exist different ways to obtain them, such as via a B-format microphone or via omnidirectional microphone arrays [6, 7]. The way how the DirAC streams are generated can be found in [1]. As it is not relevant for the matter dealt in this publication, it will be omitted.

The active intensity vector $\boldsymbol{I}_{\mathrm{a}}(k,n)$ is defined as [8]

$$\boldsymbol{I}_{\mathrm{a}}(k,n) = \frac{1}{2}\,\Re\left\{P(k,n)\cdot\overline{\boldsymbol{U}(k,n)}\right\}, \tag{1}$$

where $\overline{(\cdot)}$ denotes complex conjugation. This vector describes the direction and the magnitude of the net flow of energy characterizing the sound field in a specific point in space, namely where $P(k,n)$ and $\boldsymbol{U}(k,n)$ were computed.

Let c denote the speed of sound and E the sound field energy defined as [8]

$$E(k,n) = \frac{\rho_0}{4}\left\|\boldsymbol{U}(k,n)\right\|^2 + \frac{1}{4\rho_0 c^2}\left|P(k,n)\right|^2. \tag{2}$$

The *diffuseness* $\Psi(k,n)$ of the field is defined as

$$\Psi(k,n) = 1 - \frac{\left\|\mathrm{E}\{\boldsymbol{I}_{\mathrm{a}}(k,n)\}\right\|}{c\,\mathrm{E}\{E(k,n)\}}, \tag{3}$$

where $\mathrm{E}\{\cdot\}$ denotes the temporal averaging operator.

The *direction of arrival* (DOA) is expressed by means of the unit vector $\boldsymbol{e}_{\mathrm{DOA}}(k,n)$, defined as

$$\boldsymbol{e}_{\mathrm{DOA}}(k,n) = -\boldsymbol{e}_{\boldsymbol{I}}(k,n) = -\frac{\boldsymbol{I}_{\mathrm{a}}(k,n)}{\left\|\boldsymbol{I}_{\mathrm{a}}(k,n)\right\|}. \tag{4}$$

The mono DirAC stream carries the following quantities: $P(k,n)$, $\boldsymbol{e}_{\mathrm{DOA}}(k,n)$, and $\Psi(k,n)$.

2.1. DirAC for More Sound Sources

We now derive the parameters of the mono DirAC stream obtained from the ideal merging of N mono DirAC streams, i.e., we compute the relevant physical quantities (pressure and particle velocity) which we would measure if the N sources were to play at the same time in the same environment. Let $P^{(i)}(k,n)$ and $\boldsymbol{U}^{(i)}(k,n)$ be the pressure and particle velocity which would have been recorded for the i-th source, if it was to play alone. Given the linearity of the medium, when N sources play together, the observed pressure $P(k,n)$ and particle velocity $\boldsymbol{U}(k,n)$ are

$$P(k,n) = \sum_{i=1}^{N} P^{(i)}(k,n) \tag{5}$$

$$\boldsymbol{U}(k,n) = \sum_{i=1}^{N} \boldsymbol{U}^{(i)}(k,n). \tag{6}$$

By substituting $P(k,n)$ and $\boldsymbol{U}(k,n)$ into (1) and (2), we can compute diffuseness and DOA from (3) and (4), which, together with (5), represent the ideal mono DirAC parameters of the merged stream.

This shows that if both pressure and particle velocity were known for each source, obtaining the merged mono DirAC stream would be straightforward. Unfortunately, such a trivial merging is not possible for mono DirAC streams, for which the particle velocity $\boldsymbol{U}^{(i)}(k,n)$ is not transmitted.

2.2. Estimating Pressure and Particle Velocity

The novel method proposed in this paper aims at carrying out the merging without knowing the particle velocity of each source. To do so, we introduce the assumption that the field of each source consists of a plane wave summed to an ideal diffuse field. The pressure and particle velocity for the i-th source are modeled as

$$P^{(i)}(k,n) = P^{(i)}_{\mathrm{PW}}(k,n) + P^{(i)}_{\mathrm{diff}}(k,n) \tag{7}$$

$$\boldsymbol{U}^{(i)}(k,n) = \boldsymbol{U}^{(i)}_{\mathrm{PW}}(k,n) + \boldsymbol{U}^{(i)}_{\mathrm{diff}}(k,n). \tag{8}$$

We first derive estimators for $P^{(i)}_{\mathrm{PW}}(k,n)$ and $\boldsymbol{U}^{(i)}_{\mathrm{PW}}(k,n)$, denoted by $\widehat{P}^{(i)}_{\mathrm{PW}}$ and $\widehat{\boldsymbol{U}}^{(i)}_{\mathrm{PW}}$, and then, in the next section, we show how from these we can estimate the DirAC parameters of the merged stream. This approach, namely to estimate pressure and particle velocity of the underlying plane wave, is motivated by the fact that when the plane wave dominates the field, the overall pressure $P^{(i)}(k,n)$ approximates $P^{(i)}_{\mathrm{PW}}(k,n)$. The same happens for the particle velocity vector which can be derived directly from $P^{(i)}_{\mathrm{PW}}(k,n)$. In fact, the relationship between these two quantities is constant and known for plane waves [8]. On the contrary, when the diffuse field dominates the field, the mono DirAC stream does not carry any useful information which can be exploited to estimate magnitude and direction of $\boldsymbol{U}^{(i)}(k,n)$, which is then set to $\boldsymbol{0}$.

Setting the air density ρ_0 to 1, and dropping the functional dependency (k,n) for simplicity, we can write

$$\begin{aligned}
\Psi^{(i)} &= 1 - \frac{\left\|\mathrm{E}\left\{\boldsymbol{I}^{(i)}_{\mathrm{a}}\right\}\right\|}{c\,\mathrm{E}\{E^{(i)}\}} = 1 - \frac{\mathcal{A}}{\mathcal{B}} \\
\mathcal{A} &= \Big\|\mathrm{E}\Big\{\Re\Big\{P^{(i)}_{\mathrm{PW}}\overline{\boldsymbol{U}^{(i)}_{\mathrm{PW}}}\Big\}\Big\} + \mathrm{E}\Big\{\Re\Big\{P^{(i)}_{\mathrm{PW}}\overline{\boldsymbol{U}^{(i)}_{\mathrm{diff}}}\Big\}\Big\} + \\
&\quad + \mathrm{E}\Big\{\Re\Big\{P^{(i)}_{\mathrm{diff}}\overline{\boldsymbol{U}^{(i)}_{\mathrm{PW}}}\Big\}\Big\} + \mathrm{E}\Big\{\Re\Big\{P^{(i)}_{\mathrm{diff}}\overline{\boldsymbol{U}^{(i)}_{\mathrm{diff}}}\Big\}\Big\}\Big\| \\
\mathcal{B} &= \frac{c}{2}\Big\{\mathrm{E}\Big\{|P^{(i)}_{\mathrm{PW}}|^2\Big\} + \mathrm{E}\Big\{|P^{(i)}_{\mathrm{diff}}|^2\Big\} + \mathrm{E}\Big\{P^{(i)}_{\mathrm{PW}}\overline{P^{(i)}_{\mathrm{diff}}}\Big\} + \\
&+ \mathrm{E}\Big\{P^{(i)}_{\mathrm{diff}}\overline{P^{(i)}_{\mathrm{PW}}}\Big\} + \mathrm{E}\Big\{\|\boldsymbol{U}^{(i)}_{\mathrm{PW}}\|^2\Big\} + \mathrm{E}\Big\{\|\boldsymbol{U}^{(i)}_{\mathrm{diff}}\|^2\Big\} + \\
&+ \mathrm{E}\Big\{\Big(\boldsymbol{U}^{(i)}_{\mathrm{PW}}\Big)^{\mathrm{H}}\boldsymbol{U}^{(i)}_{\mathrm{diff}}\Big\} + \mathrm{E}\Big\{\Big(\boldsymbol{U}^{(i)}_{\mathrm{diff}}\Big)^{\mathrm{H}}\boldsymbol{U}^{(i)}_{\mathrm{PW}}\Big\}\Big\}.
\end{aligned}$$

Exploiting the statistical independence between the planar wave and the diffuse field we obtain

$$\Psi^{(i)} = 1 - \frac{\mathrm{E}\Big\{|P^{(i)}_{\mathrm{PW}}|^2\Big\}}{\mathrm{E}\Big\{|P^{(i)}_{\mathrm{PW}}|^2\Big\} + 2c^2\mathrm{E}\{E_{\mathrm{diff}}\}}. \tag{9}$$

Considering the statistical properties of ideal diffuse fields (see [8]), we introduce the approximation

$$\mathrm{E}\Big\{|P^{(i)}_{\mathrm{PW}}|^2\Big\} + 2c^2\mathrm{E}\{E_{\mathrm{diff}}\} \approx \mathrm{E}\Big\{|P^{(i)}|^2\Big\}, \tag{10}$$

leading to the estimator

$$\mathrm{E}\Big\{|P^{(i)}_{\mathrm{PW}}|\Big\} \approx \mathrm{E}\Big\{|\widehat{P}^{(i)}_{\mathrm{PW}}|\Big\} = \sqrt{1-\Psi^{(i)}}\,\mathrm{E}\Big\{|P^{(i)}|\Big\}. \tag{11}$$

To obtain instantaneous estimates (i.e., for each time-frequency tile), we remove the expectation operators, obtaining

$$\widehat{P}^{(i)}_{\mathrm{PW}}(k,n) = \sqrt{1-\Psi^{(i)}(k,n)}\,P^{(i)}(k,n). \tag{12}$$

By exploiting the planar wave assumption, we can derive the estimate for the particle velocity directly

$$\widehat{\boldsymbol{U}}^{(i)}_{\mathrm{PW}}(k,n) = -\frac{1}{\rho_0 c}\,\widehat{P}^{(i)}_{\mathrm{PW}}(k,n)\cdot \boldsymbol{e}^{(i)}_{\mathrm{DOA}}(k,n). \tag{13}$$

The desired estimates are then

$$\begin{aligned}\widehat{P}_{\mathrm{PW}}(k,n) &= \sum_{i=1}^{N}\widehat{P}^{(i)}_{\mathrm{PW}}(k,n)\\ \widehat{\boldsymbol{U}}_{\mathrm{PW}}(k,n) &= \sum_{i=1}^{N}\widehat{\boldsymbol{U}}^{(i)}_{\mathrm{PW}}(k,n).\end{aligned} \tag{14}$$

2.3. Estimating the DirAC Parameters

Once estimators have been defined for the pressure and particle velocity vectors, as presented in the previous section, deriving estimators for the diffuseness and direction of the merged stream is straightforward.

The estimates for the direction of arrival for the overall active field and for the diffuseness, denoted by $\widehat{e}_{\mathrm{DOA}}(k,n)$ and $\widehat{\Psi}(k,n)$, respectively, are

$$\begin{aligned}\widehat{e}_{\mathrm{DOA}}(k,n) &= -\frac{\widehat{\boldsymbol{I}}_{\mathrm{a}}(k,n)}{\left\|\widehat{\boldsymbol{I}}_{\mathrm{a}}(k,n)\right\|}\\ \widehat{\Psi}(k,n) &= 1-\frac{\left\|\mathrm{E}\left\{\widehat{\boldsymbol{I}}_{\mathrm{a}}(k,n)\right\}\right\|}{\mathrm{E}\left\{\left\|\widehat{\boldsymbol{I}}_{\mathrm{a}}(k,n)\right\|+\frac{1}{2c}\sum_{i=1}^{N}\Psi^{(i)}\cdot|P^{(i)}|^2\right\}},\end{aligned} \tag{15}$$

which, together with (5), represent the estimated mono DirAC stream obtained from the N mono DirAC streams.

3. REALISTIC SIMULATION RESULTS

In this section we present simulation results for the merging approach proposed in the previous section for a realistic teleconferencing scenario. In a room of $5\times5\times3$ m, we simulate two talkers, positioned at 1 meter from a B-format microphone. The talkers are characterized by azimuth and elevation angles of $(-60^\circ,\ 0^\circ)$ and $(45^\circ,\ 0^\circ)$ respectively, relative to the microphone. The frequency dependent reverberation time has a mean of approximately 100 ms. Microphone self-noise is modeled as white Gaussian noise with an SNR of 35 dB.

Plots (a) and (b) of Figures 1 and 2 show respectively diffuseness and azimuth for the two talkers separately. The merging carried out ideally (i.e., with perfect knowledge of all physical quantities) is shown in Plot (c) whereas Plot (d) depicts the merging carried out with the proposed method. Plot (e) shows the error affecting the estimators. The time-frequency tiles which possess energy less than 50 dB from the overall maximum are colored in white, as they are anyway irrelevant for the synthesis. As the direct sound dominates over both the echoes and self noise, we can observe a rather low diffuseness for (a) and (b) in Figure 1. When the two talkers are active at the same time, the diffuseness can raise up to approximately 0.4 in Plot (c). For the DOA we can observe that the sources are still distinguishable in both (c) and (d) of Figure 2. The reason is that it is unlikely that both sources allocate the same time-frequency tile with comparable energy and the DOA computed as in (4) corresponds to the source with the largest energy.

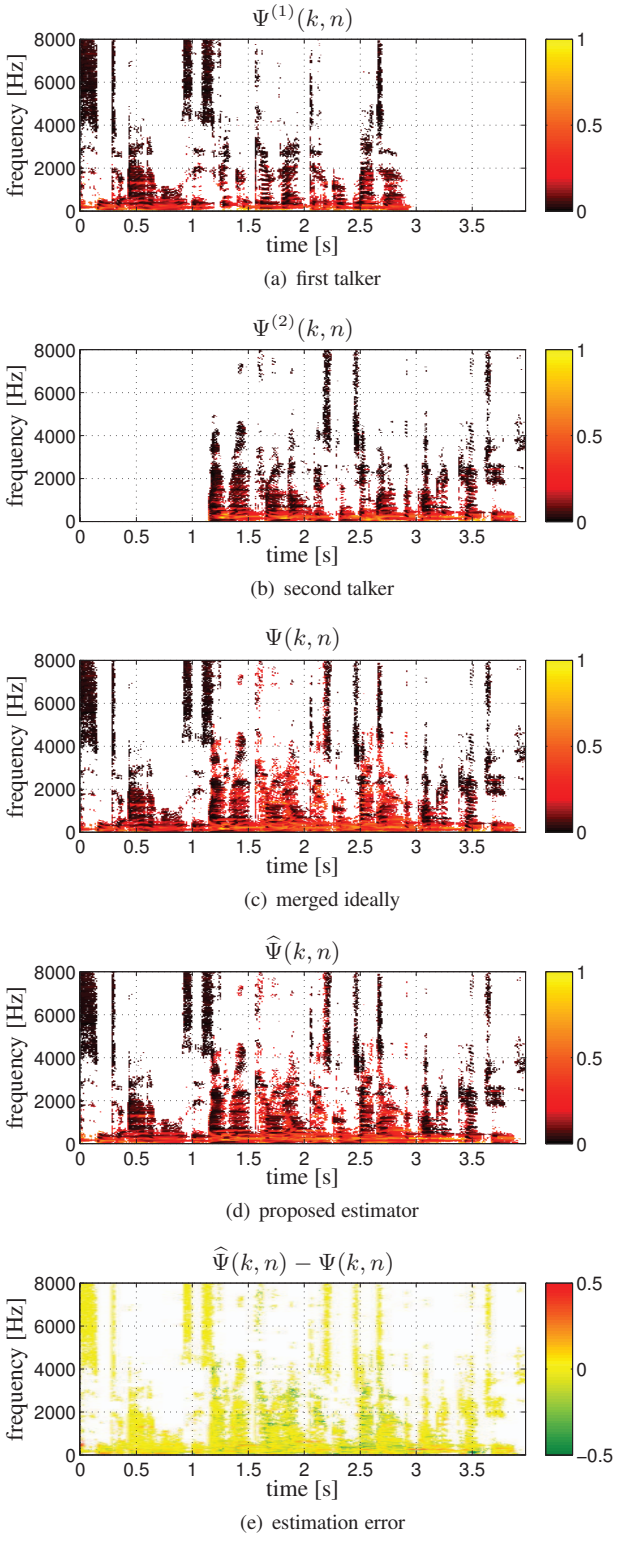
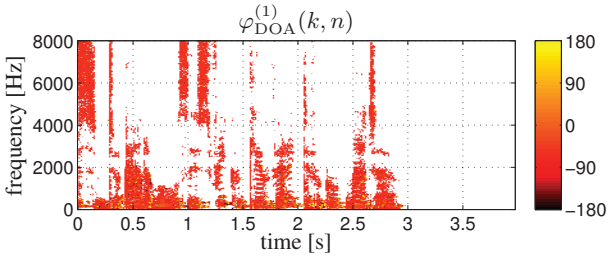
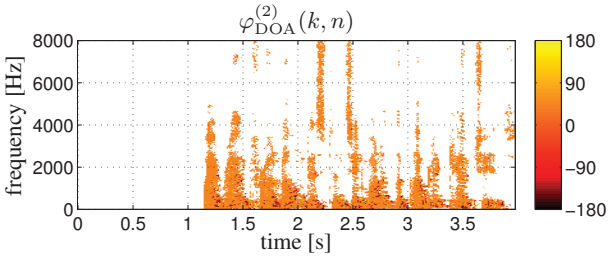


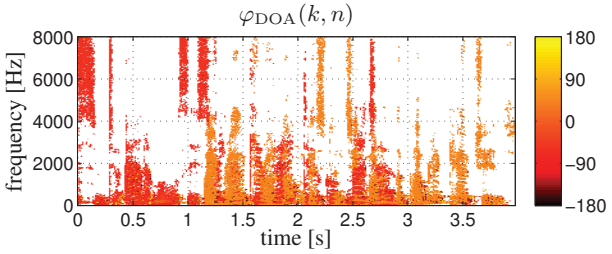
Fig. 1. Plot (a) and (b) show the diffuseness for two distinct mono DirAC streams. Plots (c) shows the diffuseness for the mono stream merged ideally whereas plot (d) shows the merging with the proposed method. Plot (e) depicts the estimation error. Time-frequency tiles with little or no energy are colored in white, as they are anyway irrelevant for the synthesis.



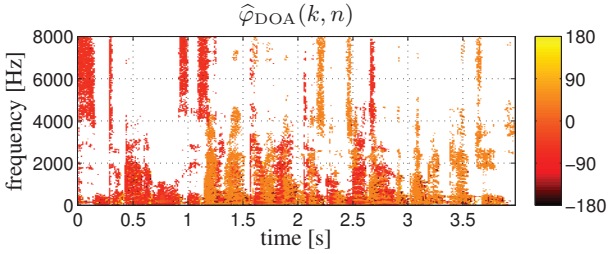
(a) first talker



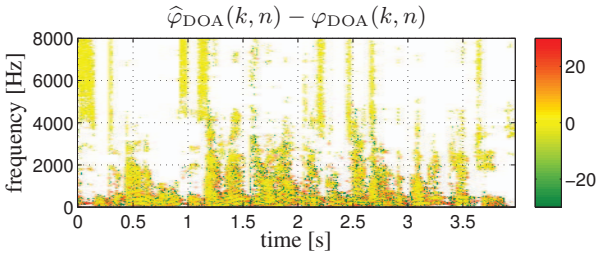
(b) second talker



(c) merged ideally



(d) proposed estimator



(e) estimation error

Fig. 2. Plot (a) and (b) show the azimuth of the direction of arrival (DOA) for two distinct mono DirAC streams. Plot (c) shows the DOA for the mono stream merged ideally. Plot (d) shows the merging with the proposed method. Plot (e) depicts the estimation error. All plots are expressed in degrees. The DOA for the time-frequency tiles which possess little or no energy is colored in white, as it is anyway irrelevant for the synthesis.

The plots indicate good agreement of the estimators with the ideal merging. In particular, the Root Mean Square Error (RMSE) on the diffuseness is .06, whereas the RMSE for the azimuth is 23°. As the diffuseness ranges between 0 and 1, an RMSE of .06 is clearly very small. On the other hand, the accuracy of the DOA might seem large. To correctly interpret this number we need first to consider that the DOA estimate improves rapidly with the energy. For instance, if we considered the time-frequency tiles with an energy larger than -20 dB (assuming 0 dB to be the maximum) the RMSE would be 15°. Such an accuracy is sufficient to correctly recreate the spatial scene at the reproduction side. In fact, informal listening tests on several scenarios indicate that the merging performed with the proposed method is practically indistinguishable from the ideal merging, i.e., the one carried out with perfect knowledge of all physical quantities. Moreover, the tests confirm the absence of any artifacts.

4. CONCLUSIONS

In this contribution we have proposed an efficient method to merge two or more mono DirAC streams. The problem of merging arises as information on the particle velocity vectors is required but is not transmitted in the mono streams. By modeling the sound field as a plane wave summed to an ideal diffuse field we are able to estimate the particle velocity vectors and perform the merging. Simulations and informal listening tests indicate that the proposed method is practically indistinguishable from an ideal merging and confirm the absence of any artifacts.

5. REFERENCES

[1] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audo Engineering Society*, vol. 55, no. 6, pp. 503–516, June 2007.

[2] L. Villemoes, J. Herre, J. Breebaart, G. Hotho, S. Disch, H. Purnhagen, and K. Kjrling, "MPEG surround: The forthcoming ISO standard for spatial audio coding," in *AES 28th International Conference, Pitea, Sweden*, June 2006.

[3] J. Ahonen, V. Pulkki, F. Kuech abd M. Kallinger, and R. Schultz-Amling, "Directional analysis of sound field with linear microphone array and applications in sound reproduction," in *124th AES Convention, May 17-20, 2008, Amsterdam, The Netherlands*, 2008.

[4] M.A. Ericson and R.L. McKinley, *The intelligibility of multiple talkers separated spatially in noise*, Mahwah: Lawrence Erlbaum Associates, 1997.

[5] J. Engdegard, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen, "Spatial audio object coding (SAOC) the upcoming MPEG standard on parametric object based audio coding," in *124th AES Convention, May 17-20, 2008, Amsterdam, The Neatherlands*, 2008.

[6] J. Merimaa, "Applications of a 3-D microphone array," in *112th AES Convention,* Paper 5501, Munich, May 2002.

[7] J. Ahonen, V. Pulkki, and T. Lokki, "Teleconference application and B-format microphone array for directional audio coding," in *30th AES International Conference*.

[8] F. J. Fahy, *Sound Intensity*, Essex: Elsevier Science Publishers Ltd., 1989.