A Novel Combined System of Direction Estimation and Sound Zooming of Multiple Speakers

Hasan KHADDOUR [1], *Jiri SCHIMMEL* [1], *Frantisek RUND* [2]

[1] Dept. of Telecommunications, Brno University of Technology, Technicka 12, 616 00 Brno, Czech Republic
[2] Dept. of Radioelectronics, Czech Technical University in Prague, Technicka 2, 166 27 Prague 6, Czech Republic

xkhadd00@stud.feec.vutbr.cz, schimmel@feec.vutbr.cz, xrund@fel.cvut.cz

Abstract. *This article presents a new system for estimating the direction of multiple speakers and zooming the sound of one of them at a time. The proposed system is a combination of two levels; namely, sound source direction estimation, and acoustic zooming. The sound source direction estimation uses the so-called energetic analysis method for estimating the direction of multiple speakers, whereas the acoustic zooming is based on modifying the parameters of the directional audio coding (DirAC) in order to zoom the sound of a selected speaker among the others. Both listening tests and objective assessments are performed to evaluate this system using different time-frequency transforms.*

Keywords

Acoustic zooming, vector base amplitude panning, sound source localization, energetic analysis method, directional audio coding

1. Introduction

Sound source direction estimation techniques can be used with several applications such as in video conferencing systems for automatic camera pointing. When the directions of the sound sources are estimated, the camera can be turned to the direction of one of them. When multiple speakers talk simultaneously, an acoustic zooming method can be used to zoom the sound of one of them and attenuate other sounds. Such acoustic zooming method can be used in several applications, for instance, in a video conferencing system to zoom the sound of one of the speakers, or it can be used to process a recorded audio file in order to hear the sound of one selected speaker clearly and attenuate the sound of other speakers. From this point of view, acoustic zooming can be compared to the blind source separation [1], but with a special method of the sound pick-up.

In this article, we introduce a compatible system for both sound source direction estimation and acoustic zooming. This system uses energetic analysis method for estimating the direction of the speakers [2], and it is based on directional audio coding (DirAC) in order to zoom the sound of one speaker and render the resulted spatial sound file [3].

This article is organized as follows: The next section briefly introduces the state of the art in this area. The third section shortly introduces directional audio coding. Section 4 describes the original energetic analysis method. The proposed system is presented in Section 5. Section 6 describes the experiments. The listening test results are presented in Section 7. Section 8 presents the objective tests of this system, and this paper is concluded in the last section.

2. State of the Art

Many sound source localization techniques have been invented during the last decades. They can be divided into categories depending on the criteria they use to localize the sound sources. The number of sound sources that can be localized by some methods cannot exceed the number of used microphones [4]. Other methods overcome this problem by using binary time frequency masks for blind separation of speech mixtures [5]. Some techniques are designed especially to work with video camera streaming [6], [7].

Some exciting systems provide the possibility of tracking the active speaker using time delay of arrival such as in [8] and [9]. However, these systems do not support the possibility of zooming the sound and they suppose the existence of only one active speaker at the same time. An algorithm for audiovisual capture applications was proposed in [10]. This algorithm achieved acoustic zooming by manipulating the signals captured by an array of a small number of low-cost microphones. The article presented in [11] studied the possibility of modifying the parameters of DirAC in order to zoom the sound of one speaker. To our knowledge, there is no another article dealing with acoustic zooming by modifying DirAC parameters.

The proposed system combines the effort of two disciplines; namely, sound source localization and acoustic zooming in order to achieve a compatible system which can estimate the direction of multiple active speakers in the same time and zoom the sound of one of them. Even more, our system provides the possibility of using two time-frequency transforms, which ensures obtaining better results depending on achieving the best resolution in time-frequency plane.

3. Directional Audio Coding

Directional audio coding (DirAC) is a method for spatial sound representation, which was invented by Pulkki [12]. The input signals of DirAC are B-format signals, i.e., $x(t)$, $y(t)$ and $w(t)$ in two-dimensional scenario and with additional $z(t)$ in the three-dimensional situation [13].

DirAC can be divided into three parts; namely, analysis, transmission and synthesis. It can be used with different sound rendering methods, e.g., Ambisonic [14] and vector base amplitude panning (VBAP) [15]. In the analysis part, DirAC computes the diffuseness and the direction of arrival of the sound signal, which are then transmitted along with $w(t)$ signal or all B-format signals to the synthesis part. In the synthesis part, DirAC divides the sound signal into diffuse and non-diffuse streams. These two streams are then processed separately. Whereas the gains for the non-diffuse stream are calculated using a rendering technique, the diffuse stream is correlated and sent to the loudspeakers array. Different time-frequency transforms can be used with DirAC, for instance, short time Fourier transform (STFT) [16] and filter banks [17]. In this article we use both STFT and Gabor transform [18].

4. Energetic Analysis Method

Energetic analysis method is a technique for multiple sound source direction estimations, which was inspired by DirAC [3]. The principle of this method relies on analyzing the acoustic intensity in the sound field recalling that the intensity vector points to the region of increase of the energy density [19].

In case of B-format signals, the acoustic intensity is expressed as [12]

$$\begin{aligned} I_x(t,f) &= \frac{1}{\sqrt{2}Z_0}\mathrm{Re}(X(t,f)W^*(t,f)),\\ I_y(t,f) &= \frac{1}{\sqrt{2}Z_0}\mathrm{Re}(Y(t,f)W^*(t,f)),\\ I_z(t,f) &= \frac{1}{\sqrt{2}Z_0}\mathrm{Re}(Z(t,f)W^*(t,f)), \end{aligned} \tag{1}$$

where $I_x(t,f)$, $I_y(t,f)$, $I_z(t,f)$ are the components of the intensity vector, Z_0 is the acoustic impedance of the air, $X(t,f)$, $Y(t,f)$, $Z(t,f)$ and $W(t,f)$ are the coefficients of the short-time Fourier transform of the B-format signals.

The direction of arrival in horizontal plane is then estimated for each frequency bin in each time frame as [12]

$$\alpha(t,f) = \begin{cases} \arctan\left[\dfrac{-I_y(t,f)}{-I_x(t,f)}\right] & \text{if } I_y(t,f) \geq 0,\\ \arctan\left[\dfrac{-I_y(t,f)}{-I_x(t,f)}\right] - 180^\circ & \text{if } I_y(t,f) < 0. \end{cases} \tag{2}$$

The direction of the sound source is then derived as

$$\alpha_{\text{est}} = \arg\max_{\alpha} F(\alpha), \tag{3}$$

where $F(\alpha)$ represents the number of the frequency bins pointing to the direction α and it is calculated for each angle as

$$F(\alpha) = \sum_{k=0}^{K} (\alpha(t,k)|\alpha), \tag{4}$$

where $\alpha \in [-180^\circ, 180^\circ]$ is the azimuth, K is the number of the frequency bins, t represents the time and $\alpha(t,k)|\alpha$ gathers the cases where the function $\alpha(t,k)$ points to the direction α. For more details about this method, the reader is referred to [2] and [20].

5. Description of the Proposed System

The proposed system depends on DirAC. It modifies the parameters of DirAC depending on the information coming from the sound source localization unit. Although the system can be modified to work in the three dimensional plane, we are interested in only two-dimensional plane in this paper since the teleconferencing usually works in the horizontal plane.

This system can be divided into four units; namely, DirAC analysis unit, sound source localization unit, zooming and synthesis unit and rendering unit. Figure 1 shows the diagram of this system when used in two-dimensional plane.

Instead of zooming the sound of all speakers, the proposed system aims at zooming the sound of one speaker and attenuating the other sounds, giving the possibility to a listener to listen to one speaker. This technique can be useful in many applications, for instance, in the teleconferencing where the listeners are interested in listening to one speaker.

DirAC analysis unit is explained shortly in the third section, so it is not explained here any deeper. The readers are referred to [3] for more details about DirAC. In the following, we describe the other units.

5.1 Sound Source Localization Unit

The sound source localization unit depends on the so-called energetic analysis method presented in [2]. The original energetic analysis method is discussed in the fourth section. However, several steps were added to the original method to improve its accuracy. These steps were designed to exploit the features of the human voice and the propagation properties in the closed room, see Fig. 2. In the following, these steps are explained.

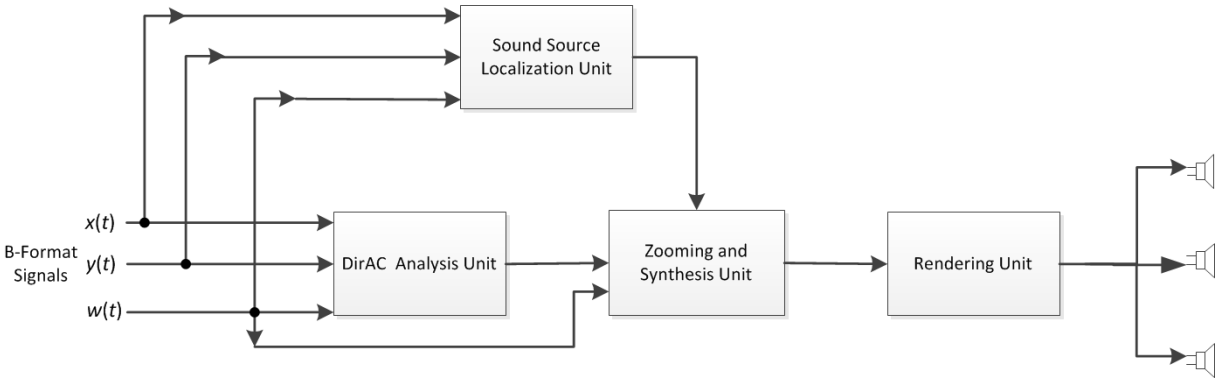

Fig. 1. The zooming system in two-dimensional plane.

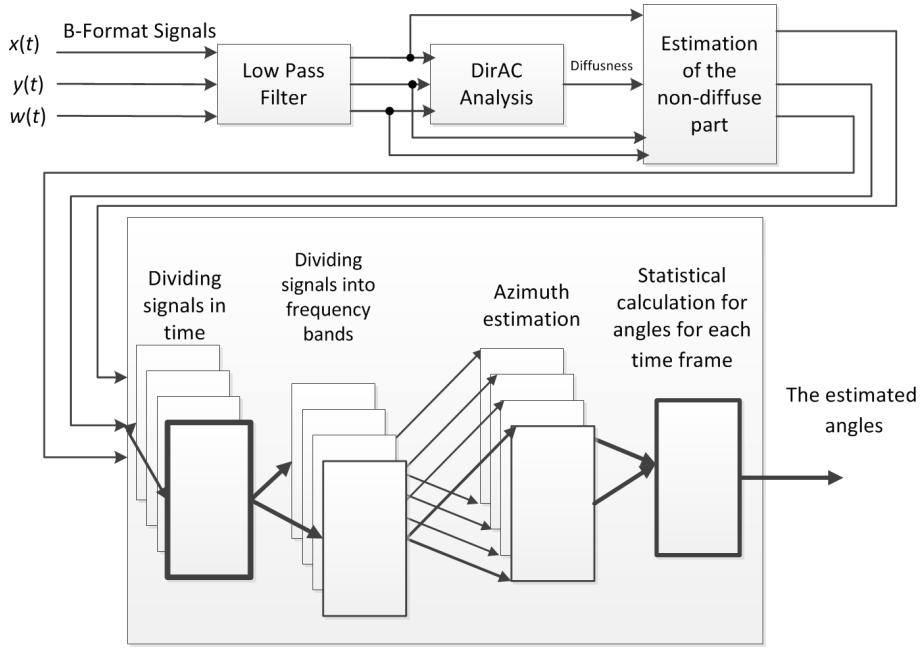

Fig. 2. Sound source localization unit in two-dimensional plane.

Step One - Filters: The input signals of this unit are B-format signals. The idea of using a low pass filter comes from the fact that we want to estimate the direction of a human speech source. The speech spectrum can be divided into two parts, the first part is flat and it contains the frequencies up to 500 Hz, whereas the second part has a slope of -10 dB/octave, and it is applied to the frequencies higher than 500 Hz [21], [22].

Applying a low pass filter to the input signals suppresses the additional interference caused by higher frequency, which belongs to the noise signals. Therefore, we applied a low pass FIR filter with cut-off frequency equal to 3500 Hz. We also applied a high pass filter with cut-off frequency equal to 100 Hz in order to minimize the effect of unevenly distributed sound energy below the critical frequency of the laboratory. It was seen that adding these filters improves the accuracy of the energetic analysis method.

Step Two - DirAC Analysis: The goal of this step in the sound source localization unit is to obtain the diffuseness parameter, which can be used to divide the sound signal into diffuse and non-diffuse part. The input signals of this step are the resulted filtered signals from the previous step. The signals are then divided in time and frequency, and the DirAC parameters are calculated [3]. The diffuseness parameter is then estimated to be applied in the next step.

Step Three - Estimation of the Non-Diffuse Part: The sound signals are first separated into diffuse and non-diffuse streams using the diffuseness parameter [3]. Then the non-diffuse part can be used to improve the accuracy of this unit by eliminating the diffusing sound, which results from the reverberant sound. The non-diffuse part is then transmitted to the time domain using inverse STFT or Gabor transform.

After processing the above mentioned steps, the original energetic analysis method is applied normally to the resulted signals. The results are in this case more accurate because of suppressing the interference caused by diffuse sound and reverberant signals.

The absolute angle error of this method with and without the mentioned steps is illustrated in Fig. 3 using boxplot. The boxes have lines at lower quartile, median, and upper quartile values. The whiskers show the extent of the rest of the data. The outliers are presented by cross outside the whiskers. As can be clearly seen, the absolute angle error was reduced when the filters were applied.

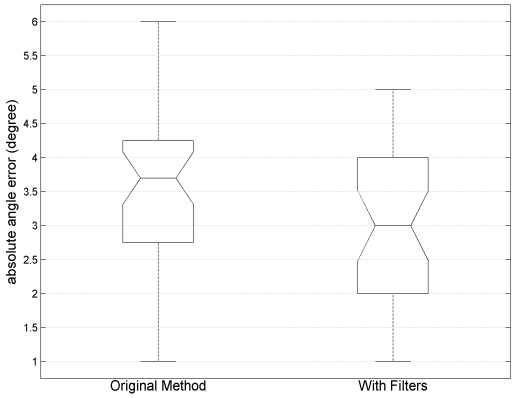

Fig. 3. The absolute angle error of the original and the modified energetic analysis method.

5.2 Zooming and Synthesis Unit

The input signals for this unit are the omni-directional B-format signal ($w(t)$), the parameters estimated from the DirAC analysis unit and the information about the directions of the speakers, which were obtained from the sound source localization unit.

The sound signal is first transmitted into frequency domain, and it is then divided into diffuse and non-diffuse part depending on the diffuseness we estimated from the DirAC analysis unit. A gain factor is then applied to the non-diffuse part, and it is calculated as

$$g(m,n) = \begin{cases} g_{\max} & \text{if DOA}(m,n) \in [\gamma+\vartheta, \gamma-\vartheta] \\ \\ g_{\min} & \text{if DOA}(m,n) \notin [\gamma+\vartheta, \gamma-\vartheta] \end{cases} \tag{5}$$

where $g(m,n)$ is the gain applied to the frequency bin number m in the time sample number n, $g_{\max}$ is the maximum gain applied to the sound we want to zoom, $g_{\min}$ is the attenuation factor, $\text{DOA}(m,n)$ is the direction of arrival estimated from DirAC analysis, γ is the direction of the speaker whose sound we want to emphasize, and it is estimated from the sound source localization unit and ϑ is the half of the angle in which we zoom the sound and it differs in each scenario. ϑ was chosen to be 5 degrees in our experiments. It was chosen depending on the length of the arc (space) that the normal-size person can occupy when he is 2 m far from the microphones.

The zooming factor impacts the quality of the sound. When a large zooming factor is used, an audible distortion occurs to the sound file, which affects the quality of the reproduced sound. Using a smoothing method improves the quality of the sound, and minimizes the distortion of the sound.

5.3 Rendering Unit

When the sound is transmitted to the time domain, it can be rendered to a set of loudspeakers, or to headphones [3]. However, a prior knowledge about the distribution of the loudspeakers should be taken into account when the rendering method is applied. In our system, we chose VBAP as a suitable method for rendering the sound since it has better localization accuracy over first-order Ambisonic [23].

6. Description of the Experiments

The experiments were designed to evaluate the ability of zooming the sound, the resolution of the zooming technique and the precision of the mentioned system. They can be divided into three stages; namely, recording the sound, processing the sound and listening stage. It should be noted that all experiments were carried out in the horizontal plane.

6.1 Recording the Sound

The recording was carried out in the acoustic laboratory at Department of Telecommunications FEEC, Brno University of Technology that meets the ITU-R BS.1116-2 requirements for the listening conditions and reproduction devices [24]. The laboratory provides semi-diffuse field with reverberation time $RT60$ around 0.3 s for one-third octave bands from 125 Hz, see Fig. 4.

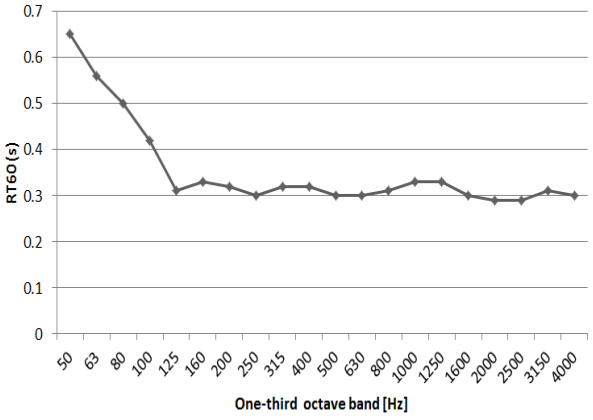

Fig. 4. $RT60$ measured in the laboratory.

A SPS200 Soundfield microphone [25] was used to record the sound of four speakers (three men and one woman). The listeners spoke simultaneously. A short English sentence was chosen as a test sentence. The duration of the speech was about 5 seconds. All speakers said the same sentence simultaneously, which ensures the most difficult situation for the system. The microphone was placed at the center of the laboratory, and the speakers stood at different positions around it at six different combinations. The sound signals were recorded as A-format signals, and then they were transmitted into B-format signals using the equations [26]

$$\begin{aligned} x(t) &= 0.5((lf(t) - lb(t) + (rf(t) - rb(t)), \\ y(t) &= 0.5((lf(t) - rb(t) - (rf(t) - lb(t)), \\ z(t) &= 0.5((lf(t) - lb(t) + (rf(t) - rb(t)), \\ w(t) &= 0.5((lf(t) + lb(t) + (rf(t) + rb(t)), \end{aligned} \tag{6}$$

where $x(t)$, $y(t)$, $z(t)$ and $w(t)$ are B-format signals, and $lf(t)$, $rf(t)$, $lb(t)$ and $rb(t)$ correspond to the signals recorded by the capsules left-front, right-front, left-back and right-back respectively.

Another recording was carried out to measure the resolution of the system. In this scenario, two speakers said simultaneously the same English sentences at different positions. The speakers came closer to each other in each new recording. The purpose of this step is to measure the smallest distance between the speakers at which the system is still able to zoom the sound of one speaker.

6.2 Processing the Sound

The mentioned system was applied to the recorded sound files in the previous paragraph. It was built using Matlab. Two time-frequency transforms were used; namely, short-time Fourier transform (STFT) and Gabor transform. The direction of the speakers was first estimated and then the zooming method was applied to each speaker of the four speakers separately. The same zooming factors were applied when both Gabor and STFT were used.

In order to achieve the best resolution in both time and frequency domains simultaneously, a compromise between time localization and frequency localization should be done. Therefore, we chose both Gabor and STFT as time-frequency transformations to study their effects on the quality of the resulted sound.

When STFT was used, a square-root Hanning window was applied, the length of this window was chosen to be 512 samples, the overlaps were chosen to be 256 points, the number of sampling points to calculate the discrete Fourier transform was 256 points, and the sampling frequency was 44100 Hz. A square-root Gaussian window was used when Gabor transform was applied. However, a similar window length and sampling frequency were used in both cases. The parameters were chosen depending on preliminary experiments, where the sound, processed using this parameters, was with the highest subjective quality.

6.3 Listening Test

In order to evaluate the zooming system, a listening test was carried out. The listening test compared the original sound rendered using DirAC and the zoomed sound using both STFT and Gabor transform. The test was performed in the acoustic laboratory described in 6.1 as follows: six loudspeakers were located in the vertices of a regular hexagon with distance of vertices from the sweet spot of 2.5 meters. For this test, ten listeners were used. The listeners have been chosen without any hearing impairment, at the age from 25 to 35 years. Five listeners have a good experience in the procedure of listening tests. For others, the procedures were explained carefully. The listeners included four women and six men. Each listener was seated at the position of the sweet spot of the loudspeaker setup. The listeners were asked to give an evaluation of the quality of the sound and of the loudness of the loudest speaker compared to the others. They were told to write their evaluation on a sheet of paper, which had the questions and a scale for each question. Five scales were available to describe the quality of the sound based on mean opinion score (MOS) [27]. The available options according to MOS are presented in Tab. 1.

Quality of the speech	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Tab. 1. Listening-quality scale (MOS).

Another scale was used to describe the loudness ratio of the speakers to each other. The available options that describe the ratio of the loudness of the speakers in this case are shown in Tab. 2.

The loudness ratio of the loudest speaker	Score
I cannot hear the others	5
Very high	4
Higher	3
Slightly higher	2
The speakers have the same loudness	1

Tab. 2. Loudness ratio.

The listening tests were also used to measure the precision of the system. The listeners were asked to localize the sound sources. A mobile loudspeaker was used as a reference sound [23]. The same sentence was rendered via the mobile loudspeaker and the original loudspeaker array alternately. The mobile loudspeaker was moved around the sweet spot in the same distance as the loudspeakers of the array till the listener said that the sound coming from it and the sound rendered via the original loudspeaker array have the same

direction. This step was applied to each of the four speakers in each audio file and only to the zoomed speaker in the zoomed files.

In order to study the relation between the value of the zooming factor and the degradation of the quality of the sound, a listening test was designed, where the same sound file with the same zooming area was processed with different zooming factors. In this listening test we used the degradation mean opinion score (DMOS) which was described in Annex D of ITU-T Recommendation P.800 [27]. The scales for DMOS are presented in Tab. 3

Degradation of the sound quality	Score
inaudible	5
audible but not annoying	4
slightly annoying	3
annoying	2
very annoying	1

Tab. 3. Degradation category scale (DMOS).

It should be noted that the duration of each test did not exceed 30 minutes, during which each listener evaluated three sound files.

7. Experimental Results

Depending on our listening test's results, the best ratio between $g_{\max}$ and $g_{\min}$ in (5) is between 13 and 15 because of the ability of zooming the sound and keeping an acceptable quality of it. Therefore, we chose the ratio 15 as a suitable value to be applied in the next listening tests. To estimate this ratio, the audio files were processed using different zooming factors as it was explained in the previous paragraph. It was seen that when small ratio between $g_{\max}$ and $g_{\min}$ is used, the zooming was not audible enough, whereas bigger ratio between $g_{\max}$ and $g_{\min}$ caused some distortion to the sound. Figure 5 shows the results regarding Tab. 2 and Tab. 3.

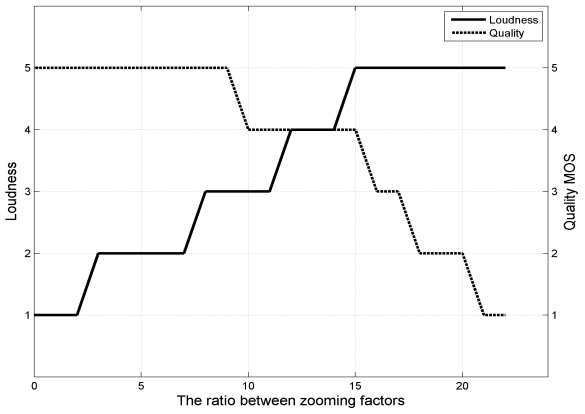

Fig. 5. The relation between the zooming ratio and the quality of the sound.

The resolution of the system was measured in a subjective way. According to our measurements, the smallest angle between the speakers at which the system was still able to zoom the sound of one speaker and attenuate the second one is 15°. When the angle between the speakers was bigger than 15°, the system worked correctly. However, when the speakers were closer to each other, the system zoomed the sound of both speakers.

A part of our experiments attended to measure the localization blur of this system, and the influence of the zooming system on this blur. In our experiments, most of the listeners explained the sound localization as "*easier*" when the zooming was applied. However, it was noticed that the listeners attended to match the sound source with the visible loudspeakers when the sound source was near them. In the original sound files i.e. without zooming, the listeners were asked to localize the four speakers, whereas they were asked to localize only the zoomed sound when the zooming was applied. The results showed that the median blur for the system was about 18°, and it was decreased a little bit when the zooming sound was applied. This little improvement in precision is mostly because of attenuating of the other sounds, which can be seen as a distraction when the listener focuses his attention on one speaker, see Fig. 6.

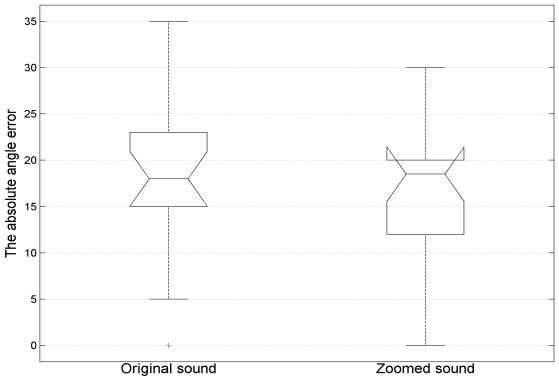

Fig. 6. The localization blur for both original and zoomed sound.

The results of the loudness of the sound are presented in both Fig. 7 and Fig. 8. The results were computed for each audio file as the average score of the evaluation given by the listeners who listened to the sound file. The results are illustrated in the graphs regarding the scales presented in Tab. 2. As seen in the previous paragraph, the zooming was applied to each speaker of the four speakers in our recordings. However, the loudness of the sound of each person differs from the others. Though, the intensity of the sound is different as well. It was seen in our experiments that zooming the sound of the loudest person achieved the best quality. When the sound of one speaker was almost inaudible in the original recording, the zoomed sound of this person achieved the worst results. The results of the sound quality are presented in both Fig. 9 and Fig. 10 regarding MOS score presented in Tab. 1.

Depending on our results, the experienced listeners, who had taken a part in listening tests before, felt the difference between the quality of the sound files when Gabor or STFT was applied to the zooming system more than the inexperienced listeners. Fig. 11 and Fig. 12 compare the results when Gabor and STFT are used using boxplot. As can be clearly seen, Gabor achieved better results. The perception of the loudness of the sound was better and it kept better sound quality.

8. Objective Measurement

For the objective assessment of quality of extracting the signal of the zoomed sound source from a given mixture we used the PEASS algorithm [28] which is designed specifically for these purposes. The algorithm [28] is based on decomposing the estimation error into three components (target distortion, interference and artifacts components), assessing the salience of each component via PEMO-Q [29] quality metric and combining these saliences via trained nonlinear mappings. The algorithm outputs are overall perceptual score (OPS), target-related perceptual score (TPS), interference-related perceptual score (IPS) and artifacts-related perceptual score (APS).

We used the sound of four speakers as the source sounds of the zooming system. The sound of the speakers was recorded in the anechoic room (reverberation time 50 ± 10 ms in octave bands from 250 Hz to 8 kHz). The sampling frequency of the recordings was 44.1 kHz and the recordings were synchronized in time. In order to align the loudness of the sound sources, their level was adjusted to RMS value of -20 dBFS with maximum peak values of -3 dBFS using the Steinberg Wavelab loudness normalizer. These recordings were rendered using four loudspeakers in the same laboratory where the subjective tests were performed. The loudspeakers were placed in the same distance from the sweet spot and in the same angles as the speakers when the recordings for the subjective tests were carried out.

At first, an omnidirectional microphone was placed in the sweet spot of the loudspeaker array and the single speakers were recorded. The recordings were carried out synchronously with the playback of given speaker. The used microphone with the recording system conforms the IEC 61672 class 1. The recordings of the individual speakers were used as the reference signals for the PEASS algorithm.

In the second step, the sound of the four speakers, which was rendered using four loudspeakers simultaneously, was recorded using the SoundField microphone, where this microphone was placed at the sweet spot of the loudspeaker as well. Recording using the omnidirectional microphone was performed as well to compare the results. The sound field recorded using SoundField microphone was then processed using the DirAC without zoom and the sound of one selected speaker was zoomed in using our system with STFT and Gabor transform. The processed sound files were rendered at the same conditions used in the subjective listening tests. The same omnidirectional microphone was placed in the sweet spot of the loudspeaker array and its signal was recorded synchronously with the playback signal of the loudspeaker array. The recorded signals in the three cases (DirAC, zoomed sound using STFT and Gabor) were then used as a test signal for the PEASS algorithm.

The results of the objective assessment of the speech quality are shown in Tab. 4. As it can be seen from the PEASS results, the overall perceptual score of the zoomed speaker is definitely better than the score of all four speakers played back simultaneously (OPS = 8) and also better than the score of the sound field of all four speakers rendered using DirAC without zooming (OPS = 19). The results are almost the same when Gabor and STFT are used for the zooming. A more detailed analysis shows that a greater suppression of the other speakers (IPS) occurs using the Gabor transformation than the STFT.

Tested signal	Original sound (four speakers)	DirAC without zoom	Zoom using Gabor	Zoom using STFT
OPS	8 %	19 %	38 %	38 %
TPS	81 %	38 %	44 %	44 %
IPS	1 %	15 %	55 %	52%
APS	87 %	54 %	44 %	45 %

Tab. 4. The average results of the speech quality assessment using the PEASS algorithm.

Absolute values of the assessment for the zooming algorithms are relatively low, but the quality improvements compared to the situation without using the zoom is clear. For the correct interpretation of the results of the PEASS algorithm it should be noted that the OPS is only 53 when we compare the recording of one speaker captured using the omnidirectional microphone in the room where the test was performed, with the recording of the speaker in the anechoic chamber, even those two recordings differ from each other only in the natural reverberation of the room. This demonstrates high sensitivity of the PEASS algorithm to any signal change. So it is necessary to take the output values of the objective assessment algorithms as the approximate values. In this case, the results of the subjective tests are primary.

9. Conclusion

A new system for estimation the direction of the speakers and zooming the sound of one of them was introduced. This system depends on the energetic analysis method for estimation the direction of the speakers, and on modifying the DirAC parameters for zooming the sound. Two time-frequency transforms are used, namely, STFT and Gabor transform. Several listening tests have been carried out to evaluate the effect of the zooming ratio on the quality of the sound, the precision (localization blur), the quality of the

sound, and the performance of the acoustic zooming system. The listening tests were mostly designed depending on ITU recommendations. The subjective experiments showed that Gabor transform achieved better results than STFT. It also showed that the resolution of this system is about 15°, and the precision (localization blur) is almost 18°.

Objective tests were done as well. The objective tests were in conformity with the subjective tests. PEASS algorithm evaluated the attenuation of other speakers (IPS) to be over 50%, whereas the ratio of the loudness of the zoomed speaker is about 3.5 till 4 from 5 point on MOS scale according to the results of the subjective tests. The comparison of the results of quality assessment is more complicated because the listeners were not told if they have to evaluate the quality degradation of the zoomed sound (equivalent to the TPS) or the quality degradation of the sound due to artifacts (equivalent to the APS). A closer analysis of each evaluation of PEASS algorithm shows that the zoomed speaker is more separated from other speakers when Gabor transform is used than when STFT is used, but other artifacts occur.

Future work will focus on improving the system to be able to work in real time, as well as on investigating the subjective and objective methods for quality assessment of this system.

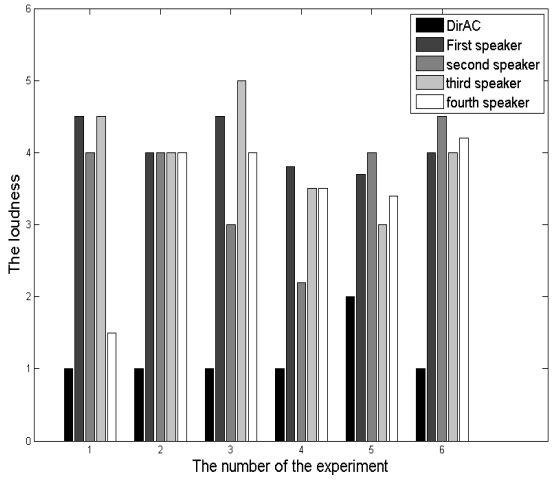

Fig. 7. The loudness ratio between the sound of the speakers when STFT was used.

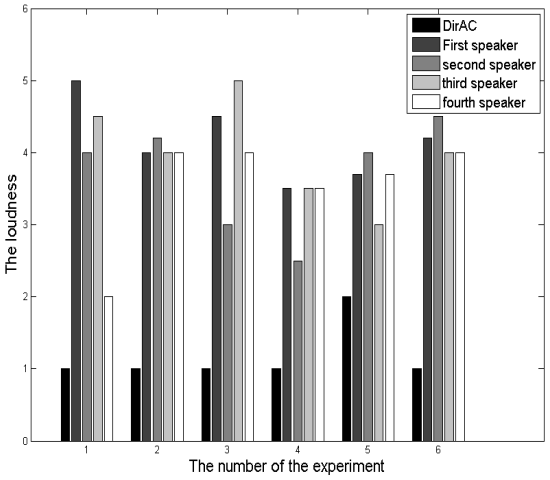

Fig. 8. The loudness ratio between the sound of the speakers when Gabor was used.

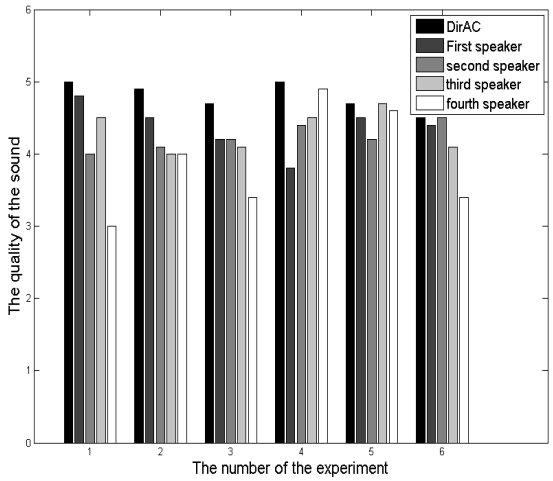

Fig. 9. The quality of the sound files according to MOS scale when STFT was used.

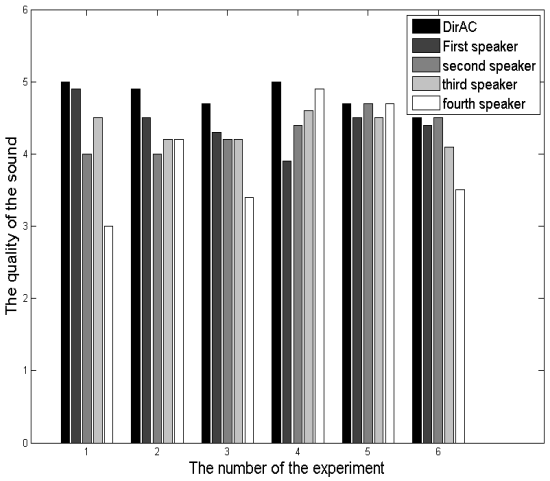

Fig. 10. The quality of the sound files according to MOS scales when Gabor was used.

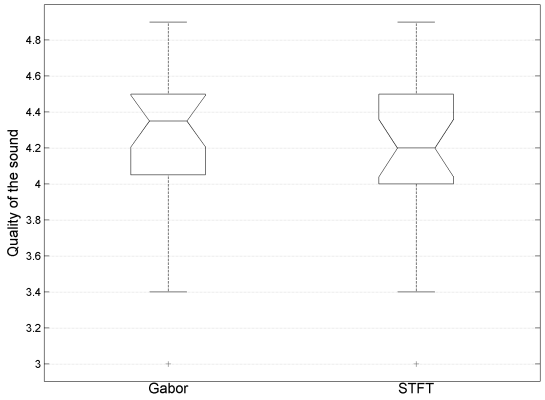

Fig. 11. The quality of the sound when Gabor and STFT are

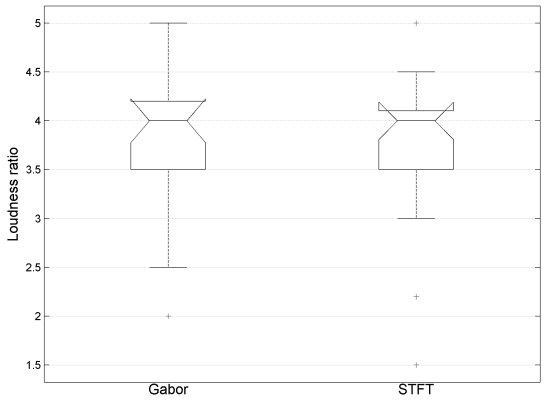

Fig. 12. The loudness ratio between the sound of the speakers when Gabor and STFT are used.

Acknowledgements

Research described in this paper was financed by the National Sustainability Program under grant LO1401. For the research, infrastructure of the SIX Center was used.

References

[1] MOHAMMED, A., BALLAL, T., GRBIC, N. Blind source separation using time-frequency masking. *Radioengineering*, 2007, vol. 16, no. 4, p. 96–100.

[2] KHADDOUR, H., SCHIMMEL, J., TRZOS, M. Estimation of direction of arrival of multiple sound sources in 3D space using B-format. *International Journal of Advances in Telecommunications, Electrotechnics, Signals and Systems*, 2013, vol. 2, no. 2, p. 63–67.

[3] PULKKI, V. Applications of directional audio coding in audio. In *Proceedings of the 19th International Congress of Acoustics.* Madrid (Spain), 2007.

[4] ITO, N., VINCENT, E., ONO, N., GRIBONVAL, R., SAGAYAMA, S. Crystal-MUSIC: Accurate localization of multiple sources in diffuse noise environments using crystal-shaped microphone arrays. In *Proceedings of the 9th International Conference on Latent Variable Analysis and Signal Separation*. Saint-Malo (France), 2010, p. 81–88. DOI: 10.1007/978-3-642-15995-4_11

[5] YILMAZ, O., RICKARD, S. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 2004, vol. 52, no. 7, p. 1830–1847. DOI: 10.1109/TSP.2004.828896

[6] HUANG,Y., BENESTY, J., ELKO, G. Passive acoustic source localization for video camera steering. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00).* Istanbul (Turkey), 2000, vol. 2, p. II909–II912. DOI: 10.1109/ICASSP.2000.859108

[7] HUANG, Y., BENESTY, J., ELKO, G., MERSEREATI, R. Real-time passive source localization: A practical linear-correction least-squares approach. *IEEE Transactions on Speech and Audio Processing*, 2001, vol. 9, no. 8, p. 943–956. DOI: 10.1109/89.966097

[8] HILDIN, J. *Voice-Following Video System*. US Patent No. 5,844,599. December 1998.

[9] WANG, H., CHU, P. Voice source localization for automatic camera pointing system in videoconferencing. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-97).* Munich (Germany), 1997, vol. 1, p. 187–190. DOI: 10.1109/ICASSP.1997.599595

[10] VAN WATERSCHOOT, T., TIRRY, W. J., MOONEN, M. Acoustic zooming by multi-microphone sound scene manipulation. *Journal of the Audio Engineering Society*, 2013, vol. 61, no. 7/8, p. 489–507.

[11] SCHULTYZ-AMLING, R., KUECH, F., THIERGART, O., KALLINGER, M. Acoustical zooming based on a parametric sound field representation. In *Proceedings of 128th Audio Engineering Society Convention*. London (UK), 2010.

[12] PULKKI,V. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 2007, vol. 55, no. 6, p. 503–516.

[13] BENJAMIN, E., HELLER, A., LEE, R. Localization in horizontal-only ambisonic systems. In *Proceedings of 121th Audio Engineering Society Convention*. San Francisco (USA), 2006.

[14] BENJAMIN, E., HELLER, A., LEE, R. Design of ambisonic decoders for irregular arrays of loudspeakers by non-linear optimization. In *Proceedings of 129th Audio Engineering Society Convention*. San Francisco (USA), 2010.

[15] VILKAMO, J., LOKKI, T., PULKKI, V. Directional audio coding: Virtual microphone-based synthesis and subjective evaluation. *Journal of the Audio Engineering Society*, 2009, vol. 57, no. 9, p. 709–724.

[16] TURI NAGY, M., ROZINAJ, G. An analysis/synthesis system of audio signal with utilization of an SN model. *Radioengineering*, 2004, vol. 13, no. 4, p. 51–57.

[17] AMBEDE, A., SMITHA, K. G., VINOD, A. P. A new low complexity uniform filter bank based on the improved coefficient decimation method. *Radioengineering*, 2013, vol. 22, no. 1, p. 34–43.

[18] ANDRAS, L., CHMURNY, J. Image compression by Gabor expansion. *Radioengineering*, 2001, vol. 10, no. 2, p. 5–8.

[19] WILLIAMS, E. *Fourier Acoustics: Sound Radiation and Near Field Acoustical Holography*. Cambridge (UK): Academic Press, 1999.

[20] KHADDOUR, H., KURC, D. Impact of applied transform on accuracy of energetic analysis method. In *Proceedings of 36th International Conference on Telecommunications and Signal Processing (TSP)*. Rome (Italy), 2013, p. 464–468.

[21] WITTENBERG, N. *Understanding Voice over IP Technology*. Cengage Learning, 2008.

[22] FURUI, S. *Digital Speech Processing, Synthesis, and Recognition*, vol. 7. CRC Press, 2001.

[23] TRZOS, M., KHADDOUR, H. Representation of sound field using ambisonic. *Elektrorevue*, 2010, vol. 41, p. 1–7.

[24] International Telecommunication Union. *Recommendation BS.1116-2: Methods for the Subjective Assessment of Small Impairments in Audio Systems*. 2014.

[25] TSL Proffessional Produsts Ltd. *SPS200 Software Controlled Microphone*. [Online] Cited 2014-2-4. Available at: http://www.tslproducts.com/soundfield /soundfieldsps200-software-controlled-microphone.

[26] RUMSEY, F., MCCORMICK, T. *Sound and Recording*. Elsevier& Focal, 2009.

[27] International Telecommunication Union. *Recommendation P.800: Methods for Subjective Determination of Transmission Quality*. 1996.

[28] VINCENT, E. Improved perceptual metrics for the evaluation of audio source separation. In *Proceedings of Latent Variable Analysis and Signal Separation*. 2012, p. 430–437. DOI: 10.1007/978-3-642-28551-6_53

[29] HUBER, R., KOLLMEIER, B. PEMO-Q – A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, vol. 14, no. 6, p. 1902–1911. DOI: 10.1109/TASL.2006.883259

About the Authors ...

Hasan KHADDOUR received his Eng. title from the Department of Telecommunications and Electronics, Faculty of Mechanical and Electrical Engineering, Tishreen University, Syria, in 2007. Since 2009, he is a Ph.D. candidate at the Department of Telecommunications, Faculty of Electrical Engineering, Brno University of Technology (BUT), Czech Republic. His current research focuses on sound source localization, acoustic zooming, and sound rendering methods.

Jiri SCHIMMEL was born in Brno, Czech Republic, in 1976. He received his M.Sc. and Ph.D. degrees in Electronics and Communications in 1999 and in Teleinformatics in 2006. He is currently an assistant professor at the Department of Telecommunications of the Faculty of Electrical Engineering and Communication of Brno University of Technology, Czech Republic. His research focuses on acoustics, multichannel digital audio signal processing, and software and hardware development for real-time audio signal processing systems. He is a member of the AES.

Frantisek RUND received his MSc. degree in Radioelectronics and his Ph.D. degree in Acoustics from Czech Technical University in Prague in 2000 and 2004, respectively. Currently he is a member of Multimedia Technology Group, Department of Radioelectronics, Czech Technical University in Prague. His research is focused on acoustics, psychoacoustic, auditory models and objective audio quality assessment. He is member of the AES.