Audio Engineering Society

Convention Paper 8831

Presented at the 134th Convention
2013 May 4–7 Rome, Italy

Novel 5.1 downmix algorithm with improved dialogue intelligibility

Kuba Łopatka, Bartosz Kunka, and Andrzej Czy□ewski

Gda□sk University of Technology, Faculty of Electronics, Telecommunications and Informatics,
Multimedia Systems Department
Narutowicza 11/12, 80-233 Gda□sk
{klopatka,kuneck,andcz}@multimed.org

ABSTRACT

A new algorithm for 5.1 to stereo downmix is introduced, which addresses the problem of dialogue intelligibility. The algorithm utilizes proposed signal processing algorithms to enhance the intelligibility of movie dialogues, especially in difficult listening conditions or in compromised speaker setup. To account for the latter, a playback configuration utilizing a portable device, i.e. an ultrabook, is examined. The experiments are presented which confirm the efficiency of the introduced method. Both objective measurements and subjective listening tests were conducted. The new downmix algorithm is compared to the output of a standard downmix matrix method. The results of subjective tests prove that an improved dialogue intelligibility is achieved.

1. INTRODUCTION

The consumption of audio-visual media is one of the main activities of the users of consumer's electronics. One of the most popular activities is watching films – from DVD or Blu-ray optical discs, as well as streamed or downloaded from the Internet. In such cases the user often watches the movies on a portable computer device, such as laptop, netbook, ultrabook, smartphone or tablet. Such devices are usually equipped with rather poor quality electroacoustic transducers, being in most cases miniaturized cost-effective speakers (in many cases – a single speaker). The users often complain that the dialogue intelligibility in the films is too low, especially if there are loud sound effects present in the movie soundtrack or the users are present in a noisy acoustic environment (e.g. public transportation, airport, street etc.). In our research we aim to deal with this problem by means of digital signal processing application. The problem is caused by the fact that the producers of movie soundtrack for DVDs consider home theatre systems as a target platform. In a home theatre the system employs a separate speaker dedicated to the center channel, which positively influences the dialogue intelligibility. Meanwhile, whenever the soundtrack is played back on a portable device, with a limited number of speakers, the downmix operation is needed, i.e. reducing the number of channels (usually

from 6 to 2). The operation of downmixing of 5.1 soundtrack to stereo is well described in the literature and standardized. However, it does not address the issue of dialogue intelligibility, adequately.

Hence, we propose a downmix algorithm which is able to enhance the intelligibility of dialogue in movies by scaling the relevant frequency components of the center channel. To achieve this, first the analysis of the soundtrack is performed to identify the partials of the signal in the center channel which in turn are related to dialogue. Next, the identified components are amplified. Thanks to this operation, an increased intelligibility of dialogue is achieved while the rest of the soundtrack remains unchanged. The algorithm requires a soundtrack in the 5.1 format. This requirement is very often met nowadays, even in media obtained from the Internet.

The performance of the algorithm was assessed by means of objective and subjective evaluation. The subjective listening tests were conducted employing a portable computer belonging to the "ultrabook" class. Its key feature is its thinness and a small weight. Therefore, it can be expected that ultrabooks will gain in popularity in the near future. The acoustic transducers installed in the device, however, are cost- and dimensions-effective and therefore they do not produce high quality sound. Thus, it is important to evaluate the effectiveness of dialogue enhancement algorithm in such a compromised listening setup.

The remainder of the paper is organized as follows. In Section 2 we present the existing downmix methods, according to the literature review. In Section 3 we present the engineered downmix algorithm with an improved dialogue intelligibility. In the following sections we introduce the research material and the evaluation performed using this material. The conclusions, including the analysis of results, are presented in Section 6.

2. EXISTING DOWNMIX METHODS

The most popular downmix method implemented in different audio decoders has been recommended by the International Telecommunication Union [2]. The main assumption of the ITU downmix method is to simulate general image of sound scene retaining surround sound experience without any enhancement of dialogue intelligibility. The downmixing procedure consists in summing up particular channels – front left (L), front right (R), front center (C), surround left (Ls) and surround right (Rs) – with relevant gain coefficients. A pair of equations (1) presents formulas representing downmixed channel left (L') and right (R') respectively [2]:

$$\begin{aligned} L' &= 1.0 \cdot L + 0.7071 \cdot C + 0.7071 \cdot Ls \\ R' &= 1.0 \cdot R + 0.7071 \cdot C + 0.7071 \cdot Rs \end{aligned} \quad (1)$$

According to the ITU recommendation [2], utilizing the LFE channel in the downmix procedure is optional. It is assumed that an ideal acoustic level of the LFE channel should be gained of +10 dB with respect to the main channels (L, R). We decided to omit LFE channel in our research.

We can assume that the most widely available standard of encoding surround sound is Dolby Digital, also known as AC-3. In case of the downmix method Dolby Digital format introduces two elements that define relative balance of center and surround channels with respect to the left and right channels: `cmixlev` (Center Mix Level) and `surmixlev` (Surround Mix Level) [3]. Values of gain coefficients `clev` referring to `cmixlev` element are shown in Tab. 1 and `slev` coefficients corresponding to `surmixlev` element are presented in Tab. 2.

cmixlev	**clev**
'00'	0.707 (-3.0 dB)
'01'	0.595 (-4.5 dB)
'10'	0.500 (-6.0 dB)
'11'	reserved

Table 1 Gain coefficients of Center Mix Level [3]

surmixlev	**slev**
'00'	0.707 (-3.0 dB)
'01'	0.500 (-6.0 dB)
'10'	0
'11'	reserved

Table 2 Gain coefficients of Surround Mix Level [3]

Dolby Digital format provides two downmix algorithms: L_o/R_o (left only / right only) expressed by pair of equations (2) and L_t/R_t (left total / right total) – expressed by pair of equations (3). The L_t/R_t scheme is also called the Dolby Pro Logic II method [3].

$$\begin{aligned} L_o &= 1.0 \cdot L + clev \cdot C + slev \cdot Ls \\ R_o &= 1.0 \cdot R + clev \cdot C + slev \cdot Rs \end{aligned} \quad (2)$$

$$L_t = 1.0 \cdot L + 0.707 \cdot C - 0.707 \cdot Ls - 0.707 \cdot Rs$$
$$R_t = 1.0 \cdot R + 0.707 \cdot C + 0.707 \cdot Ls + 0.707 \cdot Rs \quad (3)$$

It should be stressed that there are different approaches to downmixing surround sound. There are publications considering the maintenance of spatial sound experience [3–5], as well as encoding multichannel sound stream to the two-channel stream [6–8]. Moreover, there is a possibility to change the gain coefficients of downmix equations in the audio codec settings, manually. The audio codec FFDSHOW allows increasing the relative parameter values of: “voice”, “atmosphere” and “LFE”. The existing solutions are based on setting the parameters implemented in the codec without processing and analysis of audio signals. Therefore, the developed algorithm can be regarded as a novel 5.1 downmix method to enhance the dialogue intelligibility in advanced way.

It is worth mentioning that some interesting solutions were employed in the field of downmix methods related to maintenance of spatial sound experience. Bai and Shih utilized filtering the center, the rear left and the rear right channels by the corresponding HRTFs at 0°, +110°, and −110° and feeding a Shuffler filter by rear surround channels [4]. The architecture of the Bai and Shih’s downmixing technique was presented in Fig. 1. It is worth noting that in general the HRTF-based downmixing procedures differ from the ITU downmix method in the spatial quality and experience of immersion in reproduced sound scene significantly.

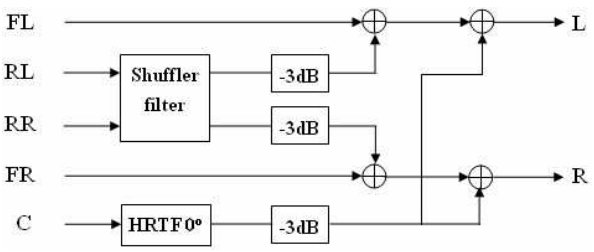


Figure 1 The architecture of the HRTF-based downmixing method [4]

Faller and Schillebeeckx proposed the method which enables to control the amount of ambience in the downmix independently of direct sound. Moreover, they defined a matrix surround downmix – formula (4). It mixes surround ambience directly to the left and right downmix channels without crosstalk and phase inversion [5].

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & \frac{1}{\sqrt{2}} & j\frac{\sqrt{3}}{2} & j\frac{1}{2} \\ 0 & 1 & \frac{1}{\sqrt{2}} & -j\frac{1}{2} & -j\frac{\sqrt{3}}{2} \end{bmatrix} \quad (4)$$

The matrix surround downmix is dedicated to the direct sound channels, whereas ITU downmix is applied to the ambient sound channels. The scheme of this concept was presented in Fig. 2. The purpose of the Faller and Schillebeeckx’s method is an enhancement of surround ambience in the downmix output.

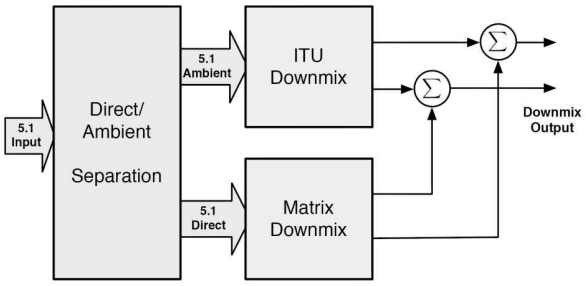


Figure 2 Proposed matrix surround downmix with different algorithms for direct and ambient sound [5]

It should be stressed that there are not any published downmix methods providing an improvement of dialogue intelligibility. Therefore, we believe our downmix algorithm can be regarded as a novel one.

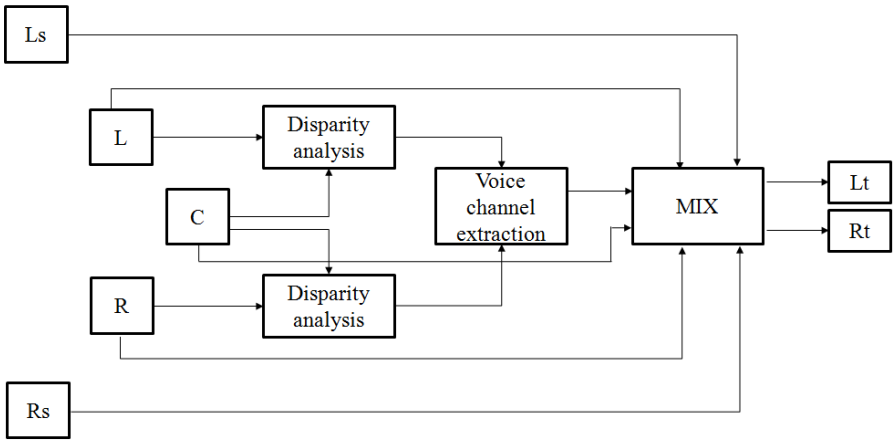


Figure 3 Block diagram of the engineered downmix algorithm

3. DIALOGUE ENHANCEMENT ALGORITHM

The general block diagram of the engineered algorithm is presented in Fig. 3. Similar to the state-of-the-art downmix methods, only 5 channels are taken into consideration: L, R, C, Ls and Rs. We can represent the downmix operation in the form of the following equation:

$$l_t[n] = l[n] + 0.707 \cdot c[n] + (d_{lev} - 1) \cdot e[n] + 0.5 \cdot l_s[n]$$
$$r_t[n] = r[n] + 0.707 \cdot c[n] + (d_{lev} - 1) \cdot e[n] + 0.5 \cdot r_s[n] \quad (5)$$

where $e[n]$ is the extracted voice signal, d_{lev} represents the dialogue level and all considered signals are represented in the digital domain, in which n denotes the sample index.

The key part of the presented algorithm is *voice channel extraction*. To achieve this feature, the *disparity analysis* of the signals in front channels (L,C,R) is essential. Extraction of the voice channel allows for controlling the level of the dialogues compared to the other sounds in the soundtrack. It is worth noting that the formula presented in Eq. 5 is merely a mathematical concept. In fact, the dialogue boost is performed in the frequency domain. The details of this operation will be given in the following subsections.

3.1. Disparity analysis

The separation of the dialogue from the other sounds in the soundtrack is achieved by means of disparity analysis between the center channel and the remaining front channels – left and right. From the study of the typical 5.1 movie soundtracks, the following assumptions were derived:

- C channel contains dialogue and (in a majority of cases) also other sounds, i.e. sound effects, illustrative music etc.
- L and R channels **do not** contain dialogue, only other sounds.

Provided that the soundtrack meets these requirements it is possible to extract the dialogue channel by analyzing the differences between the signals in the L, C and R channels in the frequency domain.

The processing flow applied to each channel is presented in Fig. 4. First, the channels are divided into OLA (OverLap and Add) frames with 50% overlap and multiplied by Hamming window function. Next, the signals $l[n]$, $c[n]$ and $r[n]$ are transformed to the frequency domain using a 4096 point FFT (Fast Fourier Transform), which yields the complex spectra of the signals - $L[k]$, $C[k]$ and $R[k]$ respectively, where k denotes the index of the spectral bin. Subsequently, the magnitude spectra are calculated and then smoothed

using moving average with the length of k = 5 spectral bins (58 Hz). As a result we obtain the smoothed magnitude spectra: $\overline{|L[k]|}$, $\overline{|C[k]|}$ and $\overline{|R[k]|}$. The comparison of example magnitude spectra of the signals in front channels is presented in Fig. 5. For a better illustration in this and in the following figures the spectra are plotted as functions of frequency, instead of spectral bin index k. The distinct harmonic components of the C channel, which are related to dialogue, can be observed. Moreover, the band above 4000 Hz is significantly more prominent in the center channel, as it contains the frequency components which positively influence speech clarity.

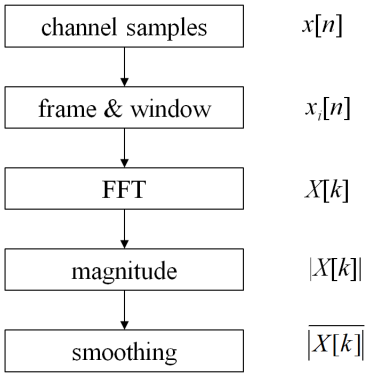


Figure 4 Calculation of the smoothed magnitude spectrum

In the next step the disparity function is calculated according to the definition in Eq. (6):

$$V[k] = \frac{\overline{|C[k]|} - \overline{|L[k]|}}{\overline{|C[k]|} + \overline{|L[k]|}} \cdot \frac{\overline{|C[k]|} - \overline{|R[k]|}}{\overline{|C[k]|} + \overline{|R[k]|}} \qquad (6)$$
$$-1 \le V[k] \le 1$$

The function $V[k]$ represents the dissimilarity of each frequency component of the signal in the center channel and the remaining front channels. To improve the effectiveness of the calculation, linear trend is removed from the spectra before computing $V[k]$. The measure is by definition constrained between -1 and 1, which allows for straightforward application of threshold. The threshold, henceforth referred to as *voice extraction threshold* is an essential parameter of the proposed dialogue intelligibility enhancement algorithm.

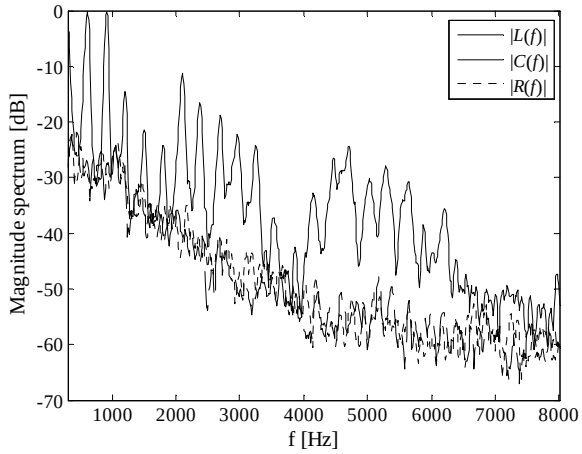


Figure 5 Comparison of the magnitude spectra of front channels

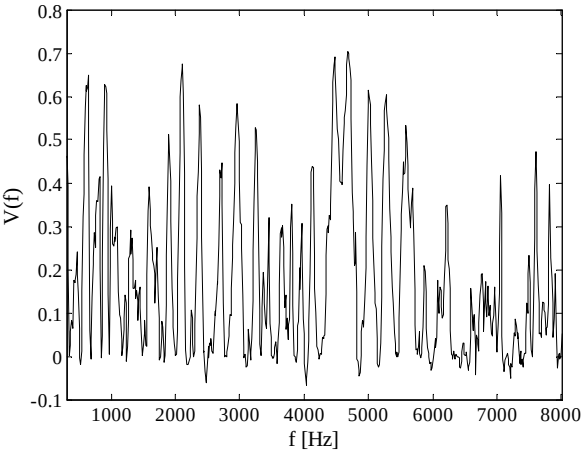


Figure 6 Example channel disparity function

3.2. Voice channel extraction

Basing on the calculated disparity function $V[k]$, the dialogue frequency components can be identified. In Fig. 6 the concept of extracting the voice components is presented. The threshold (here equal to 0.25) is applied to the dissimilarity function $V[k]$. The frequency components which are above the threshold, are considered related to dialogue. Thus, the spectrum of the extracted voice channel can be derived according to the formula in Eq. (7).

$$E[k] = \begin{cases} C[k] & if\ V[k] > t \\ 0 & if\ V[k] \le t \end{cases} \qquad (7)$$

where *t* denotes the voice extraction threshold.

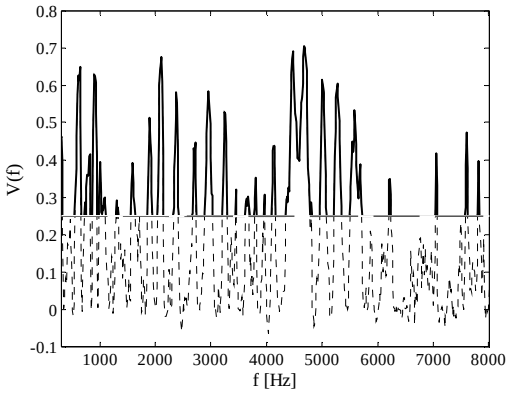


Figure 7 Identification of dialogue frequency components. The dashed horizontal line represents the voice extraction threshold.

3.3. Dialogue boosting

The next operation is boosting the level of the dialogue. The center channel is modified by selective scaling of the detected frequency components, which belong to the voice channel.

$$C[k] = C[k] + (d_{lev} - 1) \cdot E[k] \quad (8)$$

The d_{lev} coefficient represents the dialogue level in the resulting downmix. For boosting the dialogue the value of d_{lev} has to be greater than 1. The operation of scaling the detected frequency components is presented in Fig. 8, in which a 10dB dialogue boost is applied. To avoid boosting the frequency components, which do not belong to dialogue, this operation can be constrained to a given frequency band. Here, we limit the processing to the band 300-8000 Hz. Next, the modified center channel is transformed backed into the time domain using standard OLA resynthesis scheme.

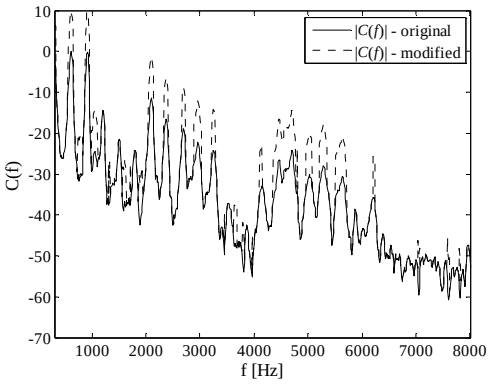


Figure 8 Scaling of the dialogue frequency components in the frequency domain

The example processing results are shown in Fig. 9. The original C channel, the extracted voice and the modified center channel are plotted. The dialogue is boosted by 10 dB. It is visible that dialogue boosting does not affect the level of the remaining part of the soundtrack.

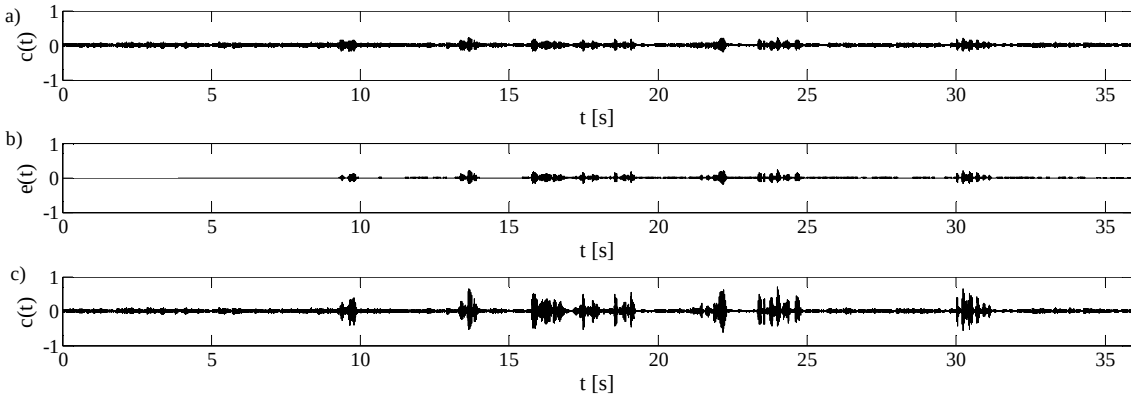


Figure 9 Results of center channel processing: a) original center channel, b) extracted voice channel, c) modified center channel

4. RESEARCH MATERIAL

Audio-video material gathered within our studies consisted of two groups: DVD samples and so-called custom samples. The first group included over 40 different movie samples with surround sound encoded in Dolby Digital format. These samples were chosen based on soundtrack content, especially on type of sound effects and illustrative music. Effectiveness of dialogue enhancement algorithm involving DVD samples was evaluated in an objective and in a subjective way. The second group – custom samples – included three samples with soundtracks prepared by the Authors. The main assumption of custom mixes was to obtain the dialogue track providing the reference signal for dialogue channel extracted by the engineered algorithm. A comparison of extracted dialogue channel with the reference one indicates the effectiveness of proposed method in an objective way.

In the experiments we utilized 8 DVD samples and 2 custom samples. A detailed description of test samples is contained in Tab. 3. It is worth noting sample No. 5 was the reference sample. It was devoid of any sound effects. We wanted to assess whether our algorithm decreases dialogue intelligibility when the dialogue track is not disturbed by other sounds in the soundtrack.

3.	*BHW_02*	„Black Hawk Down”	helicopter rotor, music
4.	*DD_02*	„Dirty Dancing”	rock music performed by music band on the stage
5. *reference*	*MDB_01*	„Million Dollar Baby”	–
6.	*SPR_01*	„Saving Private Ryan”	machine gun shots, explosions, shouts
7.	*MDB_02*	„Million Dollar Baby”	audience shouts (low level)
8.	*Avatar_02*	„Avatar”	mainly illustrative music, gunshots, shattered glass
9.	*custom_1*	-	city sounds, sounds of the surrounding nature and illustrative music
10.	*custom_2*	-	the engine’s sound and sounds of passing cars

Table 3 Movie samples utilized in the experiments

5. OBJECTIVE EVALUATION

As it was mentioned in Sec. 2, the voice extraction algorithm is the key aspect of the dialogue enhancement. The dialogue extraction process can be verified objectively. Therefore, we introduce the objective evaluation of the engineered algorithm. The defined metrics, the methodology and the obtained results will be presented in this section. We also present the results of objective evaluation using the well-known PESQ measure which was calculated using OPERA software.

Sample No.	Sample name	Movie title	Sound effects
1.	*2012_02*	„2012”	collapsing buildings, car tire
2.	*GDT_01*	„Girl with the Dragon Tattoo”	illustrative music

5.1. Methodology of objective evaluation

The employed methodology for objective evaluation of dialogue extraction is presented in Fig. 10. The front channels *l*, *r* and *c* are fed to the dialogue extraction algorithm. This operation results in extraction signal *e* which by intention contains only dialog. Next, the extraction signal is compared with the reference signal *d* and the metrics are calculated. Indexing of the reference signal is needed to determine, which part of the signal contains a dialog. The set of indices which correspond to the dialog is hereafter referred to as *Ground Truth* (GT). The key parameter of the algorithm, which most strongly affects the obtained results, is the voice extraction threshold.

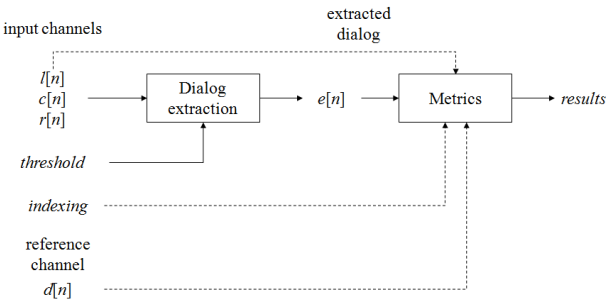


Figure 10 Objective evaluation methodology

We define two kinds of metrics for assessment of dialogue extraction: time-domain and frequency-domain ones. The time-domain metrics are:

- TDE – *True Dialog Extraction* – the ratio of energy of the extracted dialog to the energy of the reference signal calculated in the parts of the signal which are considered to contain spoken dialog.
- FDE – *False Dialog Extraction* – the ratio of energy of the extraction signal beyond the regions which contain dialog to the total energy of the extraction signal.

The TDE and FDE measures are defined by Eq. (9) and Eq. (10).

$$TDE = \frac{\sum_{n\in\{GT\}} e[n]^2}{\sum_{n\in\{GT\}} c[n]^2} \tag{9}$$

$$FDE = \frac{\sum_{n\notin\{GT\}} e[n]^2}{\sum_{n} e[n]^2} \tag{10}$$

The frequency domain metrics is mean square error (MSER). To calculate MSER the reference dialogue track has to be available. Then, the extracted signal is compared with the reference signal by calculating the MSER as defined in Eq. (11).

$$MSER = \sqrt{\frac{\sum_{k=k_1}^{k_2} \left(E(k) - D(k)\right)^2}{\sum_{k=k_1}^{k_2} D(k)^2}} \tag{11}$$

where $E(k)$, $D(k)$ are the amplitude spectra of extraction and reference dialog channel respectively and k_1, k_2 are the lower and upper spectral bin limit, corresponding to the processing band of the algorithm (300-8000Hz).

The reference channel *d* should contain only dialogue. However, as far as DVD samples are concerned, such a signal is not available. Therefore, for the evaluation of DVD downmix we used the center channel as reference. This fact influences the evaluation results. For custom soundtracks the *clean* dialogue channel was available.

5.2. Results

First, we present the evaluation of dialogue extraction from DVD samples. The DET (Detection Error Tradeoff) plots are used to depict the relation between TDE and FDE for different values of voice extraction threshold. The threshold *t* was changed in the interval [0;1]. The typical dependency is plotted in Fig. 11, obtained from sample *MDB_02*. Lowering the threshold leads to more dialogue being extracted, thus elevating the TDE measure. However, at the same time more signal is falsely treated as dialogue, which leads to an increase of the FDE metrics.

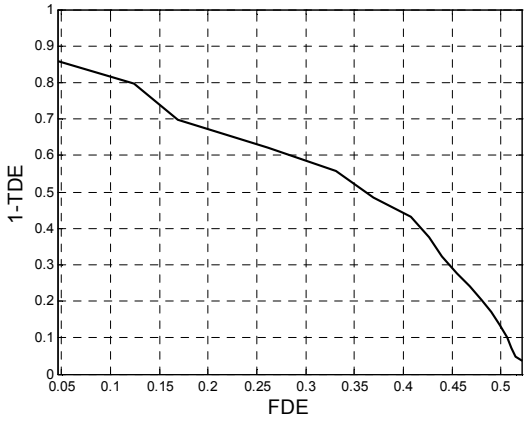


Figure 11 Objective evaluation result for sample *MDB_02*

In another example, shown in Fig. 12, obtained from sample *Avatar_02*, it can be seen, that the false negative detections (1-TDE) cannot be lowered beneath a certain lower limit – in this case the limit being equal to ca. 0.3. It demonstrates a limitation of the methodology assumed for evaluation of dialogue extraction from the DVD samples. Several factors contribute to this fact, most important of which are:

- the center channel is not appropriate for reference, since it contains other sounds than speech (sound effects, music etc.)
- the dialog extraction algorithm works in a limited frequency band (e.g. 300-8000 Hz), while the energy calculated to form the TDE and FDE measures is calculated over the entire band.
- GT indexing may be inaccurate, thus some of positive detections may be treated as false positives.

The evaluation performed with the samples of custom soundtrack is less sensitive to these inaccuracies. Since the reference dialogue channel is available, it is possible to compare the extracted voice signal with the true reference. Moreover, the MSER is calculated instead of TDE. Comparing the signal in the frequency domain, according to Eq. (11), is more accurate, since it can be limited only to the band in which the algorithm operates (i.e. 300-8000 Hz). The comparison of TDE and MSER-based evaluation is shown in Fig. 13. The metrics were calculated for the same soundtrack, which was one of the custom samples. It is visible that MSER evaluation yields more dynamics and allows for a better assessment of the voice extraction process.

An example of MSER evaluation is shown in Fig. 14. The MSER is plotted vs. voice extraction threshold. From this plot, the optimum threshold, for which the voice is most accurately extracted, can be indicated. This threshold value equals ca. 0.2. Above this threshold fewer frequency components are indentified as voice and below, whereas in turn more unwanted partials are falsely extracted.

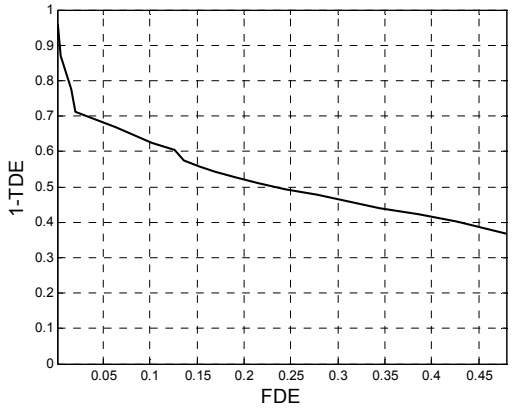


Figure 12 Objective evaluation result for sample *Avatar_02*

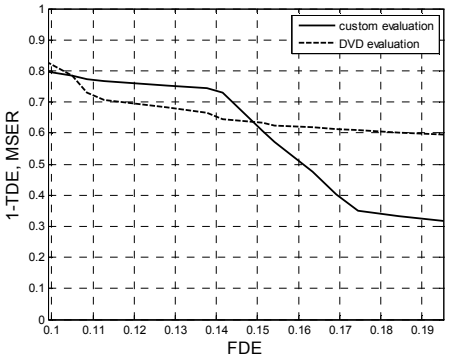


Figure 13 Comparison of MSER and TDE metrics for sample *custom_2*

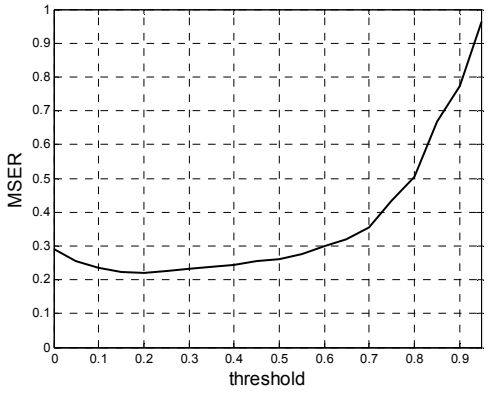


Figure 14 MSER evaluation of sample *custom_1*

5.3. PESQ evaluation

In addition to the results presented in the previous subsection, we introduce the evaluation based on PESQ metrics. The Perceptual Evaluation of Speech Quality (PESQ) measure was defined to assess the degradation of speech signals in communication channels. It is reported to have a strong correlation with speech intelligibility [9]. In this case, we treat the downmix operation as the telecommunication channel and the employed algorithms– as signal degradation. The PESQ parameter is used to asses if the speech intelligibility is improved or impaired by the employed signal processing.

The evaluation was performed using Opticom OPERA software [10], which serves as a tool for calculating several speech-related metrics (i.a. PESQ). Only the custom soundtracks were used since a clean reference signal is needed. Three types of signals were analyzed:

- original, unmodified center channel;
- modified center channel with 10 dB dialogue boost – with threshold changing from 0 to 0.9;
- extracted voice channel – with threshold changing from 0 to 0.9

The result of PESQ evaluation of the sample *custom_1* is presented in Fig. 15. The unmodified center channel received a PESQ value of 2.5. It can be understood that the other sounds added to the dialogue in the course of the soundtrack production are considered as degradation of speech intelligibility compared to the clean dialogue. Modifying the C channel by boosting the frequency components which are identified as voice by 10 dB leads to an increase of PESQ. The highest value (2.85) is achieved for the threshold equal to 0.2. The extracted voice channel also yields a higher PESQ than the original C channel, however only for thresholds remaining below 0.5. For higher thresholds too many frequency components are omitted. The important observations are that:

1. The PESQ of modified C channel is always higher than of the original, which proves that boosting the dialogue leads to an objective increase in speech intelligibility
2. The shapes of the PESQ plot and the inverted MSER plot (compare Fig. 14 and Fig. 15) are similar. This proves the correctness of the metrics employed for evaluation of the employed signal processing operations.

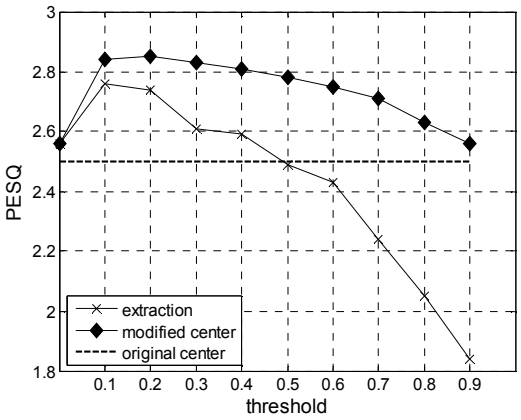


Figure 15 PESQ evaluation of sample *custom_1*

6. SUBJECTIVE EVALUATION

We assume that the engineered downmix algorithm will be applied in some audio decoders or software multimedia players in the future. Therefore, a very important aspect of the conducted studies was to assess the effectiveness of the algorithm in a subjective way.

It has to be stressed that the presented evaluation is not speech intelligibility examination as it is understood in some state-of-the-art approaches. To evaluate speech intelligibility (SI), nonsense syllables should be used and the SI factor should be calculated as the ratio of syllables correctly repeated by the listener [1]. Such approach is impossible to follow in this research since actual movie soundtracks are used as test material. Therefore, the assessed parameter is in fact subjectively perceived dialogue clarity which certainly contributes to an increase in speech intelligibility. Henceforth, we will use the term *dialogue intelligibility* understood as the listener's impression that they can understand what is being said in the movie in the presence of all other sounds.

6.1. Experiment conditions

We conducted two series of experiments in two different listening conditions involving two independent subject groups:

– professional listening room with the professional audio reproduction system (stereo basis equal to 2 m) utilizing Nexo loudspeakers;

– the room of auditory conditions close to real employing the ultrabook emitting sound.

In the second listening setup we used the ultrabook hp Folio 13. During the experiment the subject sat in the standard distance to the ultrabook which was 0.6 m. The level of sound reproduced in both configurations allowed for a comfortable listening experience.

We assumed to use the same set of test samples in both listening conditions and different groups of subjects involved in the experiments for each setup. The total number of 30 subjects participated in each configuration. Subjects were students (average age: 22 years) of the Faculty of Electronics, Telecommunications and Informatics of Gdansk University of Technology. They were not familiar with the research topic – untrained subjects.

According to our assumptions the subjects compared dialogue intelligibility in the movie soundtrack downmixed by the ITU algorithm and by engineered dialogue enhancement algorithm (DEA). Thus, the pairwise comparison test was applied for the dialogue intelligibility assessment in accordance with ITU-T Recommendation P.800 [11]. Two parameters were studied in the course of the subjective evaluation: dialogue intelligibility (7-point rating scale) and quality in the context of perceived distortions (5-point rating scale).

6.2. Results

According to ITU recommendation [11] we utilized ANOVA test to assess statistical significance of dialogue intelligibility enhancement associated with the developed algorithm.

6.2.1. Professional listening setup

The column chart presented in Fig. 16 shows the general trend of the obtained results indicating differences in the evaluation of the dialogue intelligibility.

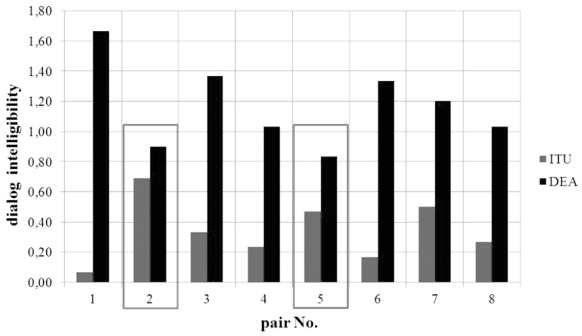


Figure 16 Average score of intelligibility in the professional listening setup

Plots presenting the range of subjective assessments with some more details were shown in Fig. 17 – pair No. 1 and 2 (example plots). Furthermore, results of ANOVA test were summarized in Tab. 4.

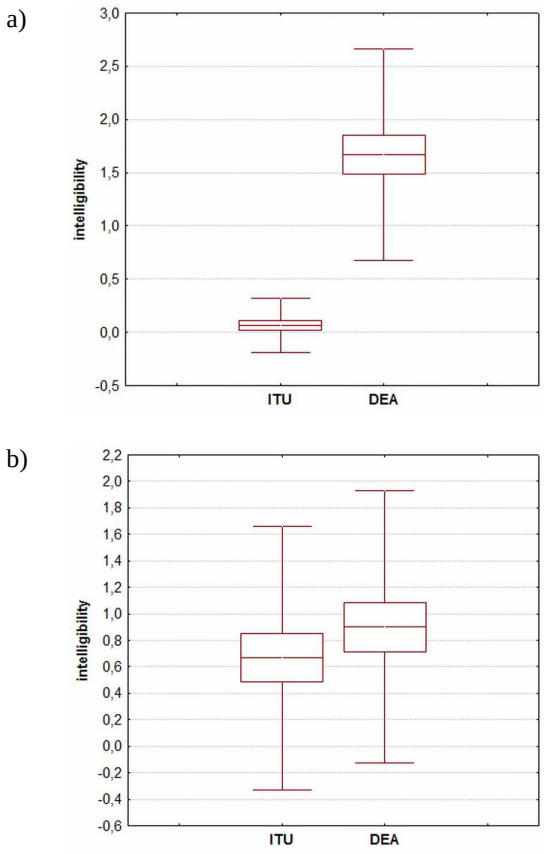


Figure 17 Box-and-whisker plots of dialogue intelligibility parameter: a) pair No. 1, b) pair No. 2

The second assessed parameter was the quality of the downmixed audio. The obtained results are collated in Fig. 18. The presented values are average quality scores given by all test participants.

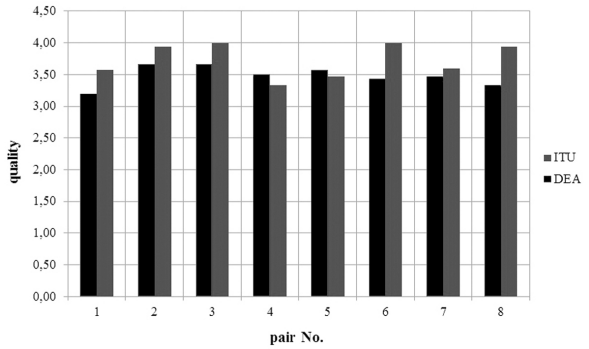


Figure 18 Quality of ITU and DEA algorithm evaluated by subjects in the professional setup

	intelligibility		quality	
pair No.	*F*	*p*	*F*	*p*
1.	72.94	0.0000	2.24	0.14
2.	0.80	0.38	1.61	0.21
3.	24.68	0.000006	2.07	0.16
4.	18.36	0.00007	0.46	0.50
5. (ref)	2.56	0.11	0.15	0.70
6.	32.15	0.0000005	4.41	0.04
7.	7.85	0.0069	0.31	0.58
8.	10.47	0.002	5.74	0.02

Table 4 Results of ANOVA test for intelligibility and quality (professional setup)

The best enhancement of dialogue intelligibility was observed for pairs: No. 1, No. 3, No. 4 and No. 6. For these samples *p*-value was very close to 0. In case of sample No. 2 a little difference between subjects' assessments was caused by relatively high intelligibility of dialogue channel for ITU and DEA. An unnoticeable difference in case of sample No. 5 means that the employment of DEA did not decrease dialog intelligibility. The underlined outcomes of intelligibility in Tab. 4 indicate that statistical significance was not met ($p>0.05$). In case of quality parameter the underlined outcomes mean that the observed differences between subjective evaluations are statistically significant ($p<0.05$). It is not desirable because the quality should remain unchanged after modifications.

According to values presented in Fig. 19 and Tab. 4 the subjects did not perceive statistically significant differences between the quality of samples processed with ITU and DEA. The exceptions are pairs No. 6 and No. 8. In these cases the observed difference was significant in a statistical sense ($p<0.05$).

6.2.2. Ultrabook

The second configuration was the ultrabook setup. The test samples were played through the ultrabook in an auditory room. The mean scores of each sample are presented in Fig. 19. It is visible that the scores of DEA samples are higher than those of ITU downmix. The exception is pair No. 5 providing the reference sample. We observe that in the case of pair No. 7 and pair No. 8 a dramatic increase in dialog intelligibility was achieved in the ultrabook setup. The effect is even more prominent than in listening room conditions.

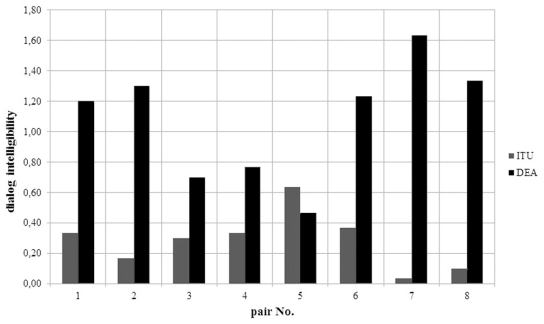
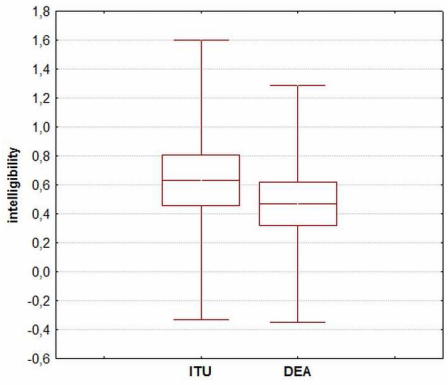


Figure 19 Average score of intelligibility in the ultrabook setup

Example plots – for the pairs No. 5 and No. 6 – presenting the range of subjective evaluations were shown in Fig. 20. The results of ANOVA test were summarized in Tab. 5.

a)



b)

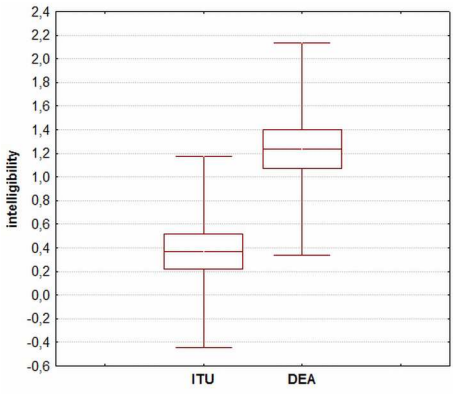


Figure 20 Box-and-whisker plots of dialogue intelligibility parameter: a) pair No. 5, b) pair No. 6

Similarly to the results obtained for the professional listening room conditions, no significant quality degradation is observed for DEA samples. The results of quality assessment in auditory conditions close to real are presented in Fig. 21.

It is worth noting that in case of ultrabook configuration all pairs of samples except reference pair (No. 5) produced a statistical significance for dialogue intelligibility parameter.

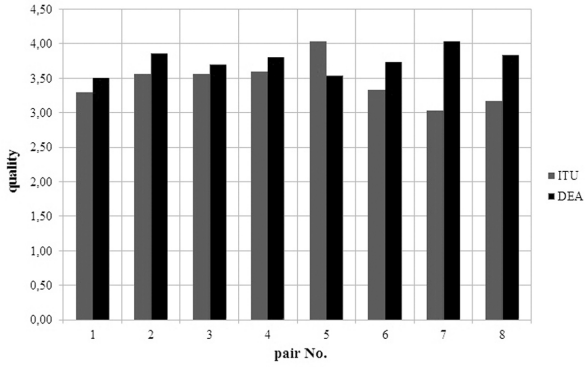


Figure 21 Quality of ITU and DEA evaluated by subjects in the ultrabook setup

	intelligibility		quality	
pair No.	*F*	*p*	*F*	*p*
1.	13.77	0.0005	0.53	0.47
2.	27.61	0.0000	1.67	0.20
3.	4.27	0.04	0.23	0.63
4.	5.44	0.02	0.87	0.35
5. (ref)	0.52	0.47	5.38	0.02
6.	15.44	0.0002	2.16	0.15
7.	65.63	0.0000	14.05	0.0004
8.	39.66	0.0000	8.35	0.0054

Table 5 Results of ANOVA test for intelligibility and quality of dialogue channel (ultrabook setup)

An interesting result was achieved as far as the quality of the reference sample is concerned. Therefore, the aspect of enhancement of dialog intelligibility and quality also for samples with undisturbed dialogue should be studied in future research.

7. CONCLUSIONS

A novel 5.1 to stereo downmix algorithm was presented, which addresses the issue of dialogue intelligibility is certain listening conditions. The details of the algorithm were presented and the objective and subjective evaluation results were shown. The results indicate that a significant increase of dialogue intelligibility was achieved employing the introduced signal processing algorithms. The objective evaluation showed that the accuracy of extracting the dialogue from the soundtrack is strongly dependent on the sensitivity, i.e. voice extraction threshold. It is difficult to determine the

optimum threshold *a priori*, since it depends, among other factors, on the type of sounds present in the soundtrack. In a practical application, the user should have the possibility to set this threshold to the most comfortable value. Moreover, the PESQ analysis of the test samples proved that the proposed dialogue enhancement method yields an objective increase in speech intelligibility.

Concluding the obtained results of subjective evaluation procedure it should be noted that the engineered dialogue enhancement algorithm improved dialogue intelligibility ensuring a statistical significance in both listening conditions. Slightly better results were achieved for the ultrabook setup. Marginally worse quality was obtained for processed samples in the absence of sound effects disturbing the dialogue channel (reference sample – pair No. 5). The subjects reported that the effectiveness of the dialogue enhancement algorithm is significantly better in professional listening conditions for continuous sound effects (music, helicopter rotor etc. – pair No. 3 and 4). The results of assessed quality parameter regarded as perceived distortions are highly correlated to the results of dialogue intelligibility evaluations. Moreover, distortions in the signal do not influence subjective evaluation of the downmixed soundtrack's quality.

In future work an effort should be made to limit the distortion introduced in the signal by the dialogue enhancement operation. It was shown that the degradation is imperceptible when there is a lot of sound effects present in the signal. However, some quality impairment was reported also by the tested subjects when the soundtrack was devoid of additional sound effects.

8. ACKNOWLEDGEMENTS

This research was funded by Intel Labs University Research Office.

9. REFERENCES

[1] J. Benesty, M. Sondhi, and Y. Huang, "Acoustical Information Required for Speech Perception," in *Speech Processing*, Springer Berlin Heidelberg, 2008, pp. 70–82.

[2] ITU, "ITU-R Recommendation BS.775-1, Multichannel stereophonic sound system with and without accompanying picture," 1994.

[3] "Digital Audio Compression Standard (AC-3, E-AC-3)," 2010.

[4] M. Bai and G. Shih, "Upmixing and Downmixing Two-channel Stereo Audio for Consumer Electronics," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 3, pp. 1011–1019, Aug. 2007.

[5] C. Faller and P. Schillebeeckx, "Improved ITU and Matrix Surround Downmixing," in *Audio Engineering Society, Convention Paper 8339*, 2011.

[6] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, and J. Engdegård, "MPEG Spatial Audio Object Coding—The ISO/MPEG Standard for Efficient Coding of Interactive Audio Scenes," *Journal of Audio Engineering Society*, vol. 60, no. 9, pp. 655–673, 2012.

[7] B. Schick, R. Maillard, and C.-C. Spenger, "First investigations on the use of manually and automatically generated stereo downmixes for spatial audio coding," in *Audio Engineering Society' Convention Paper 6448*, 2005.

[8] C.-M. Liu, S.-W. Lee, and W.-C. Lee, "Efficient Downmixing Methods for Dolby AC-3 Decoders," *IEEE Trans. on Speech And Audio Processing*, 1997.

[9] A. W. Rix, "Comparison between subjective listening quality and P.862 PESQ score," in *white paper*, 2003.

[10] "Opticom software homepage," 2013. [Online]. Available: http://www.opticom.de.

[11] ITU, "ITU-T Recommendation P.800, Methods for Subjective Determination of Transmission Quality," 1996.