Digital Signal Processing 48 (2016) 40–49

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Improving listeners' experience for movie playback through enhancing dialogue clarity in soundtracks

Kuba Lopatka [a,*], Andrzej Czyzewski [a], Bozena Kostek [b]

[a] Multimedia Systems Department, Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Narutowicza 11/12, 80-233 Gdansk, Poland

[b] Audio Acoustics Laboratory, Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Narutowicza 11/12, 80-233 Gdansk, Poland

ARTICLE INFO

Article history:
Available online 8 September 2015

Keywords:
Dialogue clarity
Center channel extraction
Speech processing
5.1 downmix
Quality of experience

ABSTRACT

This paper presents a method for improving users' quality of experience through processing of movie soundtracks. The dialogue clarity enhancement algorithms were introduced for detecting dialogue in movie soundtrack mixes and then for amplifying the dialogue components. The front channel signals (left, right, center) are analyzed in the frequency domain. The selected partials in the center channel signal, which yield high disparity between left and right channels, are detected as dialogue. Subsequently, the dialogue frequency components are boosted to achieve an increased dialogue intelligibility. Techniques for reduction of artifacts in the processed signal are also introduced. It is done through smoothing in the time domain and in the frequency domain, applied to reduce unpleasant artifacts. The results of objective and subjective tests are provided, which prove that an increased dialogue intelligibility is achieved with the aid of the proposed algorithm. The algorithm is particularly applicable in mobile devices while listening in changing conditions and in the presence of noise.

1. Introduction

The prevailing paradigm of the current computer devices market is mobility. The devices such as ultrabooks and tablets become more and more popular and are used for various activities. One of the frequent use cases of such devices is watching movies in mobile conditions. Despite the difficulties arising from the mobility, the users' quality of experience (QoE) should be taken into account and optimized. The issue most often taken into consideration is the problem of video compression influence on quality of experience evaluation [1]. However, the quality of audio also contributes to QoE. It was shown by Beerends and DeCaluwe that the audio quality has a significant impact on perceived audiovisual quality [2]. The aspect of correlation between audio and video streams was also investigated by Kunka and Kostek [3]. Currently, most films are delivered with 5.1 soundtrack, which is originally produced for cinema or home cinema systems equipped with high-end loudspeakers. Playing a 5.1 movie soundtrack on a mobile device with stereo (2.0) speakers requires the so-called downmix operation. In such case, low intelligibility of movie dialogue is a common problem, which is particularly important in mobile listening conditions when noise disturbs the listener or when the movie is not in the listener's native language. The primary goal of this work is to abate the problem of low clarity of dialogue, thus improving the listener's experience.

In contrast to existing downmix standards [4–6], we intend to improve dialogue clarity by providing an algorithm which detects the dialogue in movie soundtrack mixes and selectively amplifies the signal components which are related to speech. The dialogue is sought in the center channel of the 5.1 soundtrack. Typically, contrary to a common misconception, the center channel does not contain only dialogue. Other sounds, including music and sound effects, are also prominent. However, it is a fair assumption, that once the dialogue is present in the center channel, it is not present in the remaining channels. Therefore, an algorithm based on channel disparity is proposed, which analyzes the spectral content of left, right and center channels and identifies the partials which are related to dialogue. Subsequently, a selective amplification of dialogue components is applied, which yields an improvement in dialogue intelligibility. This application is focused on speech, and not on singing voice. Albeit the method would also work for singing voice, we believe that the level of voice and instruments in music should not be tampered with. The target implementation is in an Audio Processing Object (APO) in Windows system's audio driver.

* Corresponding author.
E-mail address: klopatka@multimed.org (K. Lopatka).

http://dx.doi.org/10.1016/j.dsp.2015.08.015
1051-2004/

Therefore, the CPU and memory load should be kept at a minimum level.

Several approaches to improving dialogue intelligibility have been outlined by Fuchs of the Fraunhofer Institute [7]. According to Fuchs, the solution is either to transmit different mixes with different levels of dialogue, to transmit separate audio sources, to introduce a new object-based approach, or to enhance dialogue clarity at the receiver's side by means of signal processing. The proposed method falls into the last category, which obviously is the only one which can be applied to existing soundtrack mixes without changing the source data.

Some of the known works approach the problem of dialogue detection employing machine learning. Kotti et al. utilized Support Vector Machines [8] and neural networks [9]. Alatan et al. introduced Hidden Markov Models for detecting dialogue scenes using audiovisual cues [10]. The advantage of the machine learning approach is that the algorithm can learn the characteristics of the dialogue sound, such as its spectro-temporal energy distribution. Thus, the dialogue extraction process does not have to rely on where the dialogue is located in the mix. However, the drawback is that the pattern recognition algorithm requires reference data. In a recent work by Hennequin et al. this reference was provided by dubbing the movie dialogue by the user [11]. In our method we do not employ machine learning, aiming at simplicity, lower complexity and finding a general solution, which does not require establishing a signal model. Our assumptions, however, can lead to occasional errors when the dialogue is not panned in the center.

The problem of voice extraction (or center channel extraction) is known in the field of music signal analysis. Han and Chen [12] proposed a PLCA-based (Probabilistic Latent Component Analysis) algorithm for extracting the main vocal melody from stereophonic music recordings. Lee et al. [13] focused on speech extraction using Blind Source Separation technique based on ICA (Independent Component Analysis). The problem of separating sound sources from stereo recordings was also addressed by Abdipour et al. [14]. A spatial model of sources was used and a learning method with adaptation was employed for separating multiple moving sources. Our algorithm employs a much less complicated method for identification of speech signal, which does not require any statistical processing and is more suitable for online operation in the audio driver of the operating system. Barry et al. [15], on the other hand, proposed a real-time algorithm based on the frequency–azimuth analysis, which enables the separation of sound sources in the stereo recording (ADRess algorithm). The method proposed in our work is based on a different channel layout, which alters the assumptions and the processing required to detect speech in the recording. An interesting solution was also presented by Avendano and Jot [16], in which interchannel similarity of spectral components was derived from a coherence measure. It was utilized to identify ambient sounds in 2.0 mix, which were later upmixed to a 5.1 speaker configuration. The method introduced by Avendano and Jot also enabled identification of frequency components panned in the center. Hence, it could be employed for dialogue detection. The method proposed by Avendano and Jot is considered in depth in Section 2.1.

In our previous work the 5.1 to stereo downmix algorithm with an improved dialogue intelligibility was introduced for the first time [17]. In this paper we showed how we developed the dialogue detection algorithm by adding signal processing operations which improve quality of the resulting signal. The current paper is directly based on another published work [18], in which the detection of dialogue was explored in more detail. The current work is extended with respect to previous conference publications. The algorithms are presented in more detail and the methods are optimized. Also, one of the improvements of the method featured in this paper, is the ability to work with 2.0 soundtrack, i.e. the soundtrack which has already undergone the downmix operation. Moreover, new results of both objective measurements and subjective listening tests are provided.

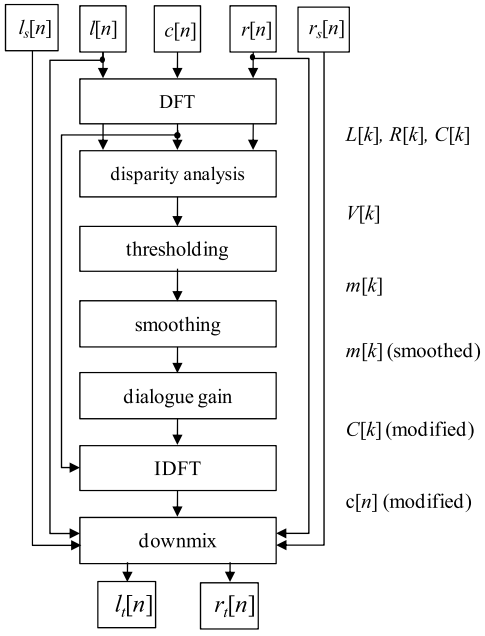

Fig. 1. Block diagram of the algorithm (symbols are explained in the text).

The paper is organized as follows. In Section 2 the principles of the algorithm are outlined. In Section 3 the evaluation by means of subjective listening tests and objective measurements is featured. The conclusions are presented in Section 4.

2. Algorithm

In the base version of the algorithm it is assumed that the input is a 5.1 channel movie soundtrack with the input signals as follows: front left channel – $l[n]$, front right channel – $r[n]$, front center channel – $c[n]$, low frequency channel – $lf[n]$, rear left channel – $l_s[n]$, rear right channel – $r_s[n]$. The low frequency channel, as in most downmix methods, is discarded. The DFT-based (Digital Fourier Transform) spectral representations of the signals in front channels are denoted $L[k]$, $R[k]$, $C[k]$. The diagram of the algorithm is presented in Fig. 1. The method is based on modifications performed in the frequency domain. The signals in front channels are transformed with DFT and analyzed to detect dialogue components. Next, the spectrum of the center channel signal is modified by selective boosting of components related to speech. The center channel signal is then reconstructed from the modified spectrum and the downmix operation is performed. The details of the signal processing are outlined in the following subsections. One of the main assumptions of the algorithm is low complexity, since it is intended for operation inside the Windows audio engine. Moreover, the method is meant for mobile devices, which frequently operate on battery. Thus, the chosen methods are computationally less expensive than some state-of-the-art algorithms featured in the literature.

2.1. Calculation of dialogue mask

As shown in Fig. 1, the most important operation in the presented algorithm is the calculation of the dialogue detection mask $m[k]$. It is achieved by spectral disparity analysis of the signals in front channels $l[n]$, $r[n]$ and $c[n]$, which are input signals of the algorithm obtained by decoding the multichannel soundtrack delivered with the film. The front channels are divided into