Special Paper
방송공학회논문지 제29권 제7호, 2024년 12월 (JBE Vol. 29, No. 7, December 2024)
https://doi.org/10.5909/JBE.2024.29.7.1178
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

# Audio Compression Technique for Low Delay and High Efficiency Using Complex Audio Data

Seungkwon Beack[a)‡], Byeongho Jo[a)], Wootaek Lim[a)], Jungwon Kang[a)]

Abstract

Audio compression techniques have been developed to improve compression ratios with the goal of enhancing audio quality for one-way streaming service. However, these one-way compression techniques focus on enhancing compression efficiency without imposing constraints on latency, making them unsuitable for interactive communication. In this paper, we introduce an audio compression technique that reduces latency while maintaining compression efficiency. To achieve this, we propose using complex data to compress audio signals within a short, fixed frame. A novel method was applied to reduce frequency domain information using efficient quantization on complex values. We demonstrate that compression efficiency can be enhanced ven within short time intervals. As a result, we have developed a low-latency, high-quality acoustic compression technique with less than 50 milliseconds of latency for 96 kbps stereo signals with MOS 4.5 quality.

Keyword : Audio coding, Audio compression, Low delay, Complex audio data

## I. Introduction

Audio compression technology is designed to efficiently compress audio signals, representing them with a lower amount of information compared to the original signal, and converting them into a bitstream format. Conventional audio compression technologies were primarily developed for use in one-way streaming services, playing a significant role in the transmission of content such as broadcasting and media services. By leveraging these technologies, the storage and transmission efficiency of audio data is maximized, contributing to the reduction of network bandwidth usage and lowering user data consumption costs. Audio compression technology has been developed through the MPEG international standards organization. The latest audio compression standard codec is MPEG-H 3D Audio (3DA) encoding technology[1], which is currently used in UHD broadcasting services, among others[2]. The 3DA

a) Electronics and Telecommunications Research Institute, Media Research Division, Media Coding Research Team
‡ Corresponding Author : Seungkwon Beack
E-mail: skbeack@etri.re.kr
Tel: +82-42-860-1745
ORCID: https://orcid.org/0000-0002-7594-0828
※ This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [24ZC1100, The research of the basic media·contents technologies].
· Manuscript November 8, 2024; Revised November 21, 2024; Accepted November 22, 2024.

technology was designed as an audio compression standard codec for one-way broadcasting and streaming services, with minimal consideration for latency constraints, focusing instead on the compression and reproduction of multi-channel audio signals. Consequently, it enables immersive audio transmission and reproduction for multi-channel and object-based audio signals[3]. However, 3DA is limited to one-way services even though it support new functionalities, allowing delays of up to several hundred milliseconds (msec) during the compression and reproduction process.

The 3DA method essentially adopts a modified form of the USAC (Unified Speech and Audio Coding) compression method, incorporating structural changes to support low complexity rather than improving compression efficiency. However, it did not significantly contribute to enhancing compression rates. The main reasons for latency in audio compression technologies like 3DA can be summarized into three points. First, the process of switching between compression tools requires look-ahead information for predictive processing. When analyzing and quantizing arbitrary signals, selecting the optimal compression tool requires predictive processing of future signals, which leads to latency. Second, future signal analysis is necessary during the predictive processing to determine the optimal frame length. Lastly, in the compression process defined by each tool, latency close to the least common multiple of the samples is needed for time sample alignment. Therefore, currently commercialized high-quality audio compression standard technologies are not suitable for real-time interactive audio transmission and require improvements in latency. This cannot be solved simply by reducing latency; low-latency audio compression technology that guarantees the same audio quality at the same compression rate must be developed to be used as an audio compression codec suitable for real-time sound communication[4].

This paper proposes an audio compression technology that can reduce latency while maintaining high quality. The proposed audio compression technology performs quantization for fixed-frame audio compression to achieve low latency and utilizes integrated bit reduction technology through frequency-domain quantization. Therefore, future signal analysis for predictive processing is not required. To address these constraints and improve quantization efficiency, the proposed method introduces a novel encoding strategy that analyzes and quantizes audio signals in the complex domain[5]. While conventional technologies perform quantization based on real-domain transformation analysis such as MDCT (Modified Discrete Cosine Transform), the newly proposed technology aims to enhance encoding efficiency by analyzing and quantizing audio frequency coefficients in the complex domain using methods like MCLT (Modulated Complex Lapped Transform) and DFT (Discrete Fourier Transform). This paper is organized as follows: Section 2 describes the proposed complex frequency-domain audio compression technology, Section 3 verifies the performance of the proposed audio compression technology through experiments, and Section 4 concludes the paper.

## Ⅱ. Audio compression with complex frequency data

In this paper, we propose a method to compress audio signals by increasing the encoding efficiency within a fixed-length frame to reduce latency. Existing techniques have been unable to use a fixed single frame due to the lack of an optimal algorithm that can simultaneously encode time and frequency information. Predicting the amount of change in the time domain is difficult without changing the frame size, and reducing the amount of information in the frequency domain should require processing in the frequency domain. In this section, we propose a method to perform the encoding in the complex frequency domain in order to do those simultaneously. However, if the amount of change in the time domain with-

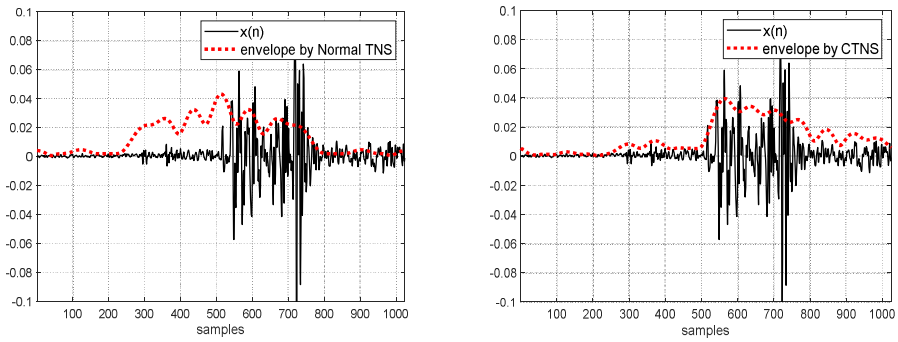
in a frame is not large, the encoding is normally performed by using the real transform domain. In this paper, the DFT method is adopted as the complex frequency domain conversion method, but MCLT, whose real part is MDCT, can be used equally well[5]. As a real frequency transform method, MDCT, which is used in conventional sound compression, is selected.

### 1. Complex LPC based audio compression

To compress an audio signal, information reduction can be achieved by quantization process. Information reduction reduces the number of bits by performing the quantization with controlled quantized noise within a range of allowed distortion that is not perceptually degraded by humans. The quantization noise can be introduced in the frequency domain and the time domain. The quantization noise in the frequency domain can be controlled by using psychoacoustic models to perform information reduction for sound compression[6], and by estimating spectral envelope information in the frequency domain using time-domain linear prediction coefficients[7]. Time domain information volume estimation and quantization is mainly performed by adjusting the time domain analysis interval and changing an analysis window to a short interval. In particular, the amount of information in the time domain can be represented by the amount of signal variation over time, and if quantization is performed using a long analysis window in a section with a large amount of time variation, sound quality distortion such as pre-echo occurs[8]. Therefore, to reduce time domain compression distortion, one alternative is to perform quantization with short analysis windows and frames. This conversion process requires future sample information for predictive processing to determine the analysis window and results in additional latency. It also reduces compression efficiency, as relatively more bits need to be allocated to short bins to minimize time base quantization distortion such as pre-echo. To improve this, a temporal noise shaping technique, Temporal Noise Shaping (TNS), has been introduced[6]. TNS is a technique that utilizes linear prediction techniques in the frequency domain to reduce the amount of time base information. The scheme of linear prediction in frequency is based on the duality properties of the linear prediction coefficient (LPC), which enables the time envelope information in the time domain to be obtained from the linear prediction information in the frequency domain. However, the conventional technique based on the real transform has the disadvantage that the envelope in the time domain cannot be accurately predicted by performing linear prediction on the real frequency coefficients obtained by the MDCT conversion method due to incomplete phase information, and it still requires a short window and cannot significantly improve the latency[8]. In this paper, the complex linear prediction technique adopted er is complex TNS (CTNS), which performs linear prediction in the complex frequency domain[6]. Complex linear prediction is performed in the complex frequency domain. In general, an arbitrary signal can be represented in the time domain as an analytic signal from in the complex domain[7].

$$x_a(n) = x(n) + jx_h(n). \tag{1}$$

Here, the signal $x_a(n)$ can be represented as an analytic signal, including the imaginary part $x_h(n)$, which is the Hilbert Transform of the original signal. The change in the amount of information on the time axis can be obtained from the envelope information $|x(n)+jx_h(n)|$, which is the absolute value of the analysed signal. Therefore, a prediction method that can accurately estimate the envelope information of $x_a(n)$ is required, and to do this, it is necessary to apply CTNS technology that can perform filtering in the complex domain[9]. Therefore, for CTNS implementation, linear prediction coefficients for complex filtering are required, and a complex number of LPCs are extracted from the DFT domain and used as filter coefficients. We define these as complex LPCs (CLPCs).




(a) Envelope by Common TNS (b) Envelope by Complex TNS

Fig. 1. CTNS vs. TNS with Complex Linear Prediction Coefficients

Fig. 1 shows a comparison of the time domain envelope estimation results of the adopted CTNS technique compared to conventional TNS. Since conventional TNS estimates the envelope based on the linear prediction coefficients obtained from MDCT, it cannot predict the exact envelope of the original signal due to the aliasing distortion that occurs in the time domain. In contrast, CTNS utilizing CLPC extracted from the complex domain shows that the envelope information of the predicted time base is a good estimate of the envelope of the original signal. When CTNS is applied to perform quantization, the complex linear prediction coefficient CLPC is also information that needs to be transmitted and converted into bit information through quantization[9,10].

Unlike the conventional LPCs, the complex linear prediction coefficients CLPCs do not have roots in the form of conjugate complex numbers. Therefore, a new quantization method is required for the conversion process so that they can be expressed correspondingly on the unit circle when expressed as P(z) and Q(z)[9]. In this paper, a method of quantizing and transmitting the roots of the complex linear prediction coefficients directly is applied for the transmission technique of the complex linear prediction coefficients. The roots represented by the complex linear prediction coefficients all exist in a unit circle in the z-plane, and the magnitude and phase values of each root can be separated. Each set of the separated magnitude and phase vectors can be subjected to VQ (Vector Quantization). The VQ for each of the magnitude and phase vectors was performed hierarchically in two stages, with each stage designed to have a codebook of 10 bits for the 8th-order complex linear prediction coefficients. Thus, 40 bits of information about the complex linear prediction coefficients are transmitted per frame (4 stages are required for encoding absolute values and phases of CLPs). However, if the amount of temporal information does not change significantly, it is efficient to quantize the audio data without performing CTNS. For this purpose, the proposed compression technique is designed with a dual quantization structure, so that if the time domain information volume reduction is not required, the encoding is performed by MDCT conversion as before. A schematic block diagram is shown in Fig. 2.

First, let's briefly review the process of complex quantization: If the time domain information reduction is deemed effective and the quantization is performed in the complex domain, DFT is performed on the input signal $\boldsymbol{x}(b)$ with frame index '$b$' for complex frequency domain conversion. Extract CLPC from $\boldsymbol{x}_f(b)$ with complex frequency coefficients and perform quantisation to obtain $\boldsymbol{clpc}_q(b)$ to be transmitted. Using $\boldsymbol{clpc}_q(b)$, perform filtering on the complex coefficients according to conventional LPC filtering

to obtain the residual signal $\boldsymbol{x}_{c,f}(b)$. Finally, the complex residual signal $\boldsymbol{x}_{c,f}(b)$ is quantised to transmit the bit information about $\boldsymbol{x}_{c,q}(b)$. The method of quantising $\boldsymbol{x}_{c,f}(b)$ performs quantization on the real and imaginary parts with a single scale-factor based on the absolute values of the complex coefficients. The quantization method follows a similar approach to the conventional MPEG audio codec, with the difference that the quantization is performed by dividing the residual signal into multi-bands[11]. Similarly, when MDCT is selected, the real part quantization method performed is the same as that of conventional MPEG audio codecs. The criterion for switching between the two modes is the predicted gain of the residual signal, which can be expressed in general terms as Eq (2).

$$Complex\_gain = gain\_metric\{\boldsymbol{x}_f(b), \boldsymbol{x}_{c,f}(b)\} \quad (2)$$

In this paper, SNR (Signal-to-Noise Ratio) is used as the $gain\_metric$, and complex domain coding is performed when the SNR difference is more than 6 dB regarding $\boldsymbol{x}_f(b)$ is considered as the original signal.

Fig. 3 shows the predicted gain of CLPC for frames actually selected with the complex value coding mode and the real value coding mode. In Fig. 3(a), it can be seen that the amount of information in the residual signal generated after applying CLPC with the complex value coding mode is significantly reduced compared with original ones. This is not only on the time domain in the upper part of (a), but also in the frequency domain plot in the lower part of (a). In Fig. 3(b), we can see that the prediction gain for CLPC is negligible. The dual-structure encoder is adopted to increase the compression efficiency, and the decoding process requires frame-to-frame compensation in the overlapping region due to the dual structure. In the case of MDCT, the phase information of the time signals is distorted, which can cause interference signals due to aliasing after inverse transform. To cancel the aliasing, MDCT information from the previous decoded frame is required. If

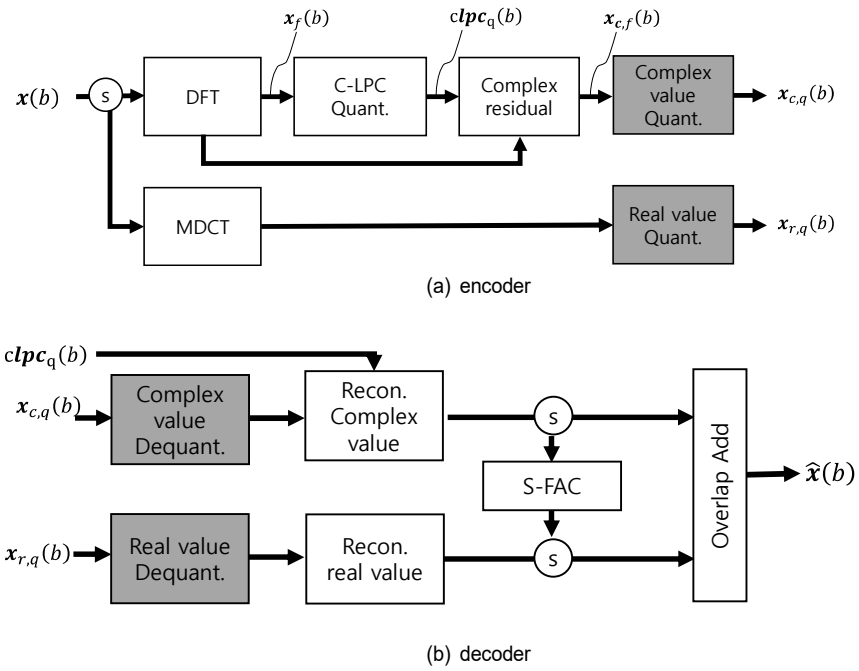
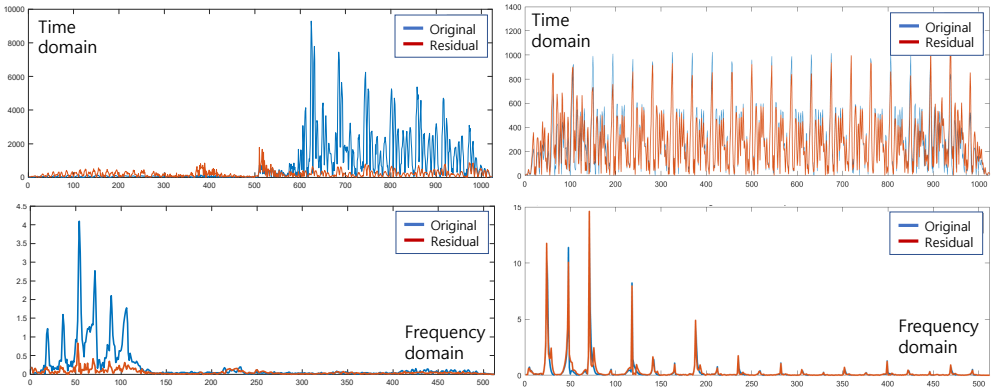



Fig. 2. Audio encoder/decoder Conceptual Block Diagram with Complex Linear Prediction Coefficients




(a) Complex value coding mode (b) Real value coding mode

Fig. 3. CLPC Predicted Gain Example with respect to quantization mode

a frame-to-frame coding scheme switch occurs during encoding, switching from a DFT complex value quantized frame to an MDCT quantized frame, the decoding process can artificially synthesize the signal of the previous MDCT quantized signals using the time-axis information obtained from the DFT complex coefficients.

This synthesized signal serves to cancel out the aliasing, and the aliasing can be compensated for by an overlap-add operation. This artificially synthesized MDCT aliasing correction signal is defined as Synthesized Forward Aliasing Cancellation (S-FAC) in Fig. 2. This refers to the block that performs the operation of superposition summation of the S-FAC signal with respect to the MDCT-based reconstructed real signal. By default, both the complex signal and real signal reconstruction schemes perform a 50% overlap summation operation, which is intended to increase the quantization efficiency in the frequency domain by using a symmetrically complete analysis and synthesis window[12].

## 2. Channel coding on complex domain

Most audio content is stereo by default, and to increase the efficiency of sound compression techniques, channel coding scheme must be applied to the audio channel signals. Channel coding based on the information reduction between audio channels reduces redundant information due to the correlation between the channels. The basic principle of the channel coding is to convert the highly correlated part between two channels into a principal component channel, and the remaining signal excluding the principal component into a minor component channel. The principal component channel signal is normally called the mid signal, and the remained component channel signal is called the side signal. A simple way to obtain the principal component and subcomponents from the channels is to obtain the signals of the sum and difference between the two channel signals[13]. The sum signal is defined as the mid signal and the difference signal as the side signal, and bits are allocated according to the size of the information amount, and quantization is performed within the allocated bits for each channel. Therefore, the larger the difference between the information amount of the main and side signals, the more efficient the bit allocation is achieved, which can increase the encoding efficiency. The simplest and most common stereo channel encoding technique is the M/S (mid side) encoding technique. M/S encoding is a very simple channel encoding technique, but its encoding efficiency is not high. Therefore, a more dynamic and active conversion method for the main and sub signals is required,

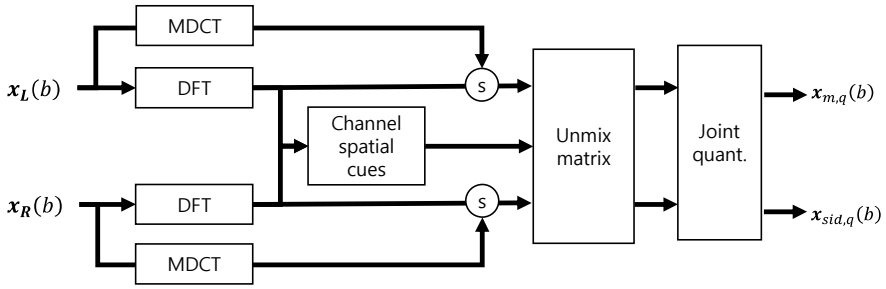



Fig. 4. Block diagram of complex data-based channel encoding

and the MPS (MPEG Surround) channel encoding technique meets this requirement[14]. However, the MPS technique applies a QMF (Quadrature Mirror Filter) frequency conversion method for dynamic main/sub conversion. The QMF conversion performs a filtering operation, which requires additional analysis time and introduces significant latency.

In this paper, we apply a spatial cue-based channel encoding technique similar to MPS, and its conceptual block diagram is shown in Fig. 4. Basically, the role of unmixing the main/sub signals using spatial cues is the same as that of MPS, but the proposed method is characterized by obtaining the spatial cues in the DFT domain and then applying a unmixing matrix (unmix matrix) to generate the main/sub channels based on them. The unmixing matrix is applied in the complex or real frequency domain depending on the coding scheme. The spatial cues required to generate this unmixing matrix are the Channel Level Difference (CLD) and Inter-Channel Coherence (ICC), which are used to generate the unmixing matrix in the same way as MPS. However, the proposed audio compression technique already includes a quantization process in the complex transform domain, which has the structural advantage of facilitating the extraction of spatial cues from complex data. Thus, the unmixing matrix for channel coding can be constructed by analyzing the spatial cues without any additional delay. An example of the predicted gain of the main/side signals obtained through the dynamic unmix matrix is shown in Fig. 5, where it can be observed that the dynamic unmix matrix-based side signal (‘dynamic matrix side’) of the proposed technique has a larger information difference compared to the main signal (‘mid’), regarding the side signal (‘M/S side’) of the passive M/S method performed in the conventional MDCT domain. This difference between the main and sub-channel signals can be exploited to achieve improved encoding efficiency and sound quality by flexibly adjusting the bit allocation through joint quantization. The key strategy of a common quantizer described in Fig. 2 is to allocate a relatively large number of bits to the major component channel (‘mid’) for quantization process, and to use the residual bits for the residual component (‘side’). In this paper, we apply an eight-stage bit allocation scheme that selects the bit allocation mode based on the amount of information difference between the mid and side signals.

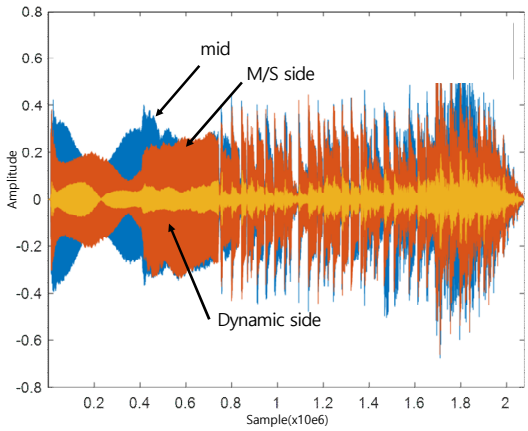



Fig. 5. M/S comparison according to the dynamic unmixing matrix

## III. Evaluation

### 1. Latency performance evaluation

The latency of an audio coding scheme is calculated as the sum of the delays incurred during encoding and decoding. Latency in signal processing-based audio compression techniques can be calculated as a fixed value by arithmetically calculating the delay when applying the algorithm. The latency of an algorithm is calculated by summing the latency of each tool and algorithm that perform the compression. The main reason of latency is the size of the sample buffer required to perform the algorithm, which is essentially reflected in the latency. Also, if the algorithm requires future samples to perform compression, the corresponding buffer size for future samples should be reflected in the latency. The case of requiring future samples is mainly related to the signal classifier used when choosing the optimal combination of tools to increase the compression efficiency. The proposed coding compression technique performs the encoding within a single fixed frame without any classifier. The analysis for selecting the quantization scheme is also based on the predicted gain within a given current frame, which does not introduce any additional delay. The channel encoding process also performs the analysis in the same complex frequency domain, which does not need any additional delay.

Table 1 shows the processing latencies occurred during the encoding and decoding of the proposed acoustic compression technique. To increase the compression efficiency, the internal sampling frequency was converted to 38 kHz, which adds an additional 0.84 msec of latency due to the filtering operation for down-sampling. The encoding of each frame is a fixed-length sample frame, which is quantized to reduce the amount of time and frequency information to increase the compression efficiency. Based on a 38 kHz internal sampling frequency, the frame latency by the encoder was 26.9 msec, using a fixed frame length. In addition, an output superposition overlap delay of 13.5 msec was performed to eliminate time base aliasing. An additional 6.8 msec was needed as a future samples analysis during the time domain linear prediction process for performing Frequency Domain Noise Shaping (FDNS)[7]. The CLPC performance is only within the transform block, so there is no additional latency. The final overall latency is 48.1 msec. An example of the latency difference between the decoded signal and the original signal is shown in Fig. 6. As a result, based on the subjective evaluation in the next section, it can be observed that at least 50 msec latency is achieved at 96 kbps while maintaining high-quality performance and compression efficiency equivalent to MOS 4.5.

Table 1. Total delay samples and its configuration for proposed system

| Latency components | Delay samples/times |
|---|---|
| Frame size | 1024 samples / 26.9 msec |
| Overlap-add buffer size | 512 samples / 13.5 msec |
| FDNS analysis look ahead | 260 samples / 6.8 msec |
| Down-sampling filter | 32 samples / 0.84 msec |
| Overall delay | 1828 samples / 48.1 msec |

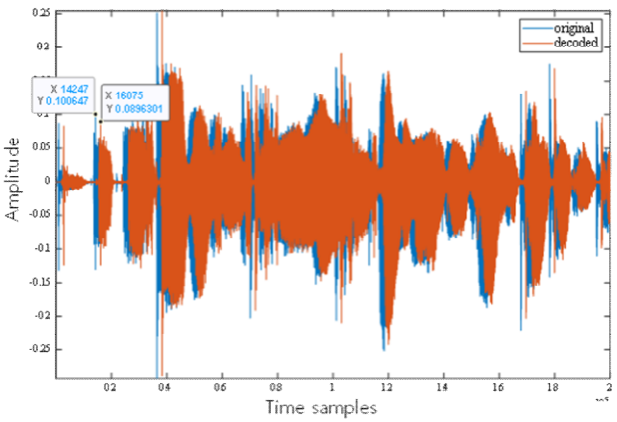



Fig. 6. Example of proposed codec delay

### 2. Subjective listening Evaluation

Subjective listening experiments were conducted to eval-

uate the performance of the proposed audio compression encoder with low latency and high efficiency. Since the performance of audio encoders is measured by subjective quality evaluation based on bit rate, it is necessary to adopt an evaluation method that allows objective statistical analysis. We selected MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) as a suitable evaluation method to assess subjective quality in a blinded manner on a scale from 0 to 100[15]. The MUSHRA evaluation includes a hidden original signals and a low-quality anchor signals to verify the reliability of the subjects' ratings, and data from subjects whose ratings did not meet the reliability criteria were excluded from the final statistical analysis. In the end, the evaluation results of 10 subjects were selected. Five systems were included in the subjective test. The proposed system (‘sys_A’) compressed and reconstructed a 95.25 kbps stereo signal, and the comparison system, MPEG-H 3D audio (‘sys_B’) operates at 128 kbps for stereo signals and has a latency of a few hundred msec. A blind original signal (‘org’) are also included, an anchor signal (‘lp70’) with a 7 kHz band, and an anchor signal (‘lp35’) with a 3.5 kHz band are included as the test system.

Fig. 7 shows the subjective evaluation results, with scores for the test items and overall mean, including 95% confidence intervals. The detailed values of the mean scores and confidence intervals are shown in Table 2. The experimental results show that the final total average value of the proposed system is above 90 points considering the 95% confidence interval, and the reference model, 3DA 128 kbps, also shows a performance above 90 points, which corresponds to approximately MOS 4.5, considering the confidence interval. In particular, the confidence interval of the subjective evaluation average score of the proposed system and the 3DA system overlaps, which can be observed that the 96 kbps of the proposed technology and the 3DA 128 kbps of the reference system are statistically equivalent performances.

Table 2. MUSHRA Mean score with confidence intervals(CI) for subjective test

| Systems | Low of CI | Mean | High of CI |
|---|---|---|---|
| Proposed (‘sys_A’) | 91.1 | 92.1 | 93.0 |
| 3DA (‘sys_B’) | 91.2 | 92.0 | 92.7 |

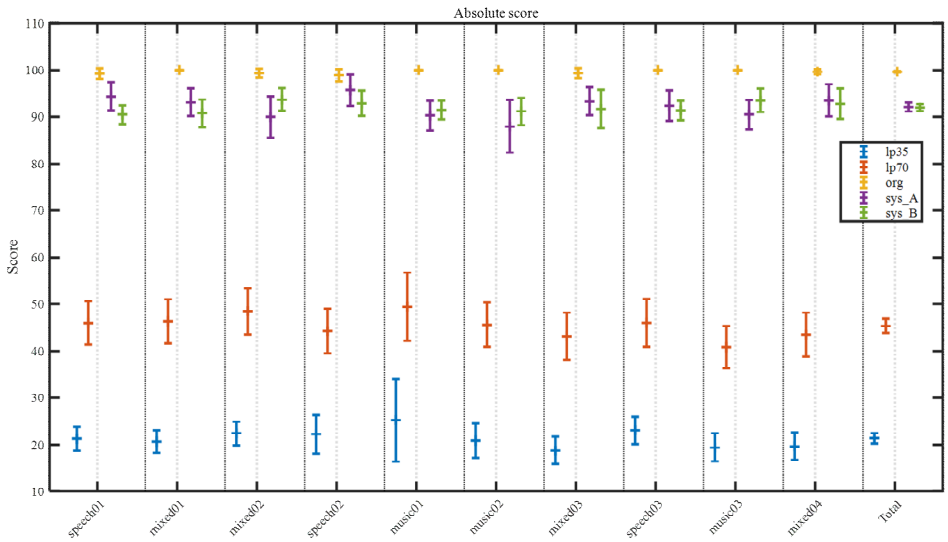



Fig. 7. Subjective listening test results

## Ⅳ. Conclusion

In this paper, a low-latency and high-efficiency audio compression technique is proposed. The proposed technique reduces latency by performing transform and quantization using a fixed single frame size without altering the frame length. Performance evaluation was conducted by comparing the system with the latest MPEG standard, 3DA, to confirm that it achieves latency below 50 ms at a bitrate of 96 kbps while maintaining high audio quality. Experimental results showed that the proposed technique achieved a compression ratio improvement of 1.33 times and a latency of 48.1 ms, while maintaining an audio quality level equivalent to a score of over 90 points, comparable to 3DA. Future work will focus on enhancing multi-channel compression methods and developing compression algorithms to improve quantization efficiency in short-segment analysis, aiming for even higher compression efficiency.

## References


[1] J Herre, J. Hilpert, A. Kuntz, and J. Plogsties, “MPEG-H audio－the new standard for universal spatial/3D audio coding,” in Proc. 137th AES Convention, 2014.

[2] C. Seo, Y. Im , S. Jeon, J. Seo, S. Choi, “A Study on Delivery Integration of UHD, Mobile HD, Digital Radio based on ATSC 3.0,” Journal of Broadcast Engineering, vol.24, no.4, pp. 643-659, Jul. 2019. doi: https://doi.org/10.5909/JBE.2019.24.4.643

[3] ISO/IEC 23008-3:2022, “Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio - Amendment 1: MPEG-H, 3D audio profile and levels,” 2015.

[4] Seungkwon Beack, Byeongho Jo, Wootaek Lim, Jungwon Kang, “Audio coding technology for supporting low-latency and high-quality,” The Korean Institute of Broadcast and Media Engineers Fall Conference, Nov. 2024. doi: https://doi.org/10.1145/351234

[5] B. Jo, and S. Beack. “Hybrid noise shaping for audio coding using perfectly overlapped window,” IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2023. doi: https://doi.org/10.1109/WASPAA58266.2023.10248116

[6] J. Herre et al., “Extending the MPEG-4 AAC codec by perceptual noise substitution,” in Proc. 104th AES Convention, 1998. doi: https://doi.org/10.17743/aesconv.2023.978-1-942220-43-5

[7] ISO/IEC 23003-3, “Information technology - MPEG audio technologies - Part 3: Unified speech and audio coding,” 2012.

[8] C. M. Liu, H. W. Hsu and W. C. Lee, “Compression artifacts in perceptual audio coding,” in IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 4, pp. 681－695, May 2008. doi: https://doi.org/10.1109/TASL.2008.918979

[9] B. Jo, and S. Beack. “Representations of the complex-valued frequency-domain LPC for audio coding.” IEEE Signal Processing Letters, Jan. 2024. doi: https://doi.org/10.1109/LSP.2024.3353162

[10] B. Jo, and S. Beack. “Efficient complex immittance spectral frequency with the perceptual-metric-based codebook search,” IEEE Signal Processing Letters, Aug. 2024. doi: https://doi.org/10.1109/LSP.2024.3466012

[11] B. Jo, S. Beack, and T. Lee, “Audio coding with unified noise shaping and phase contrast control,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023. doi: https://doi.org/10.1109/ICASSP49357.2023.10096386

[12] T. Lee, S. Beack, K. Kang, and W. Kim. Adaptive TCX Windowing Technology for Unified Structure MPEG-D USAC. ETRI Journal, 34(3):474－477, 2012. doi: https://doi.org/10.4218/etrij.12.0211.0404

[13] K. Brandenburg, “MP3 and AAC explained,” in Proc. AES 17th International Conference on High Quality Audio Coding, 1999.

[14] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, and H. Purnhagen et al., “MPEG Surround－The ISO/MPEG Standard for Efficient and Compatible Multi-channel Audio Coding,” in Proc. 122nd AES Convention, 2007.

[15] I.-R. BS.1534, “Method for the subjective assessment of intermediate sound quality (MUSHRA).”

## Introduction Authors


### Seungkwon Beack

- Ph.D. degree from the Department of Information and Communications Engineering, the Korea Advanced Institute of Science & Technology, Daejeon, South Korea, in 2005
- Principal Researcher with Media Coding Research Section at the Electronics and Telecommunications Research Institute, Daejeon, South Korea
- ORCID : https://orcid.org/0000-0002-6254-2062
- Research interests : Signal processing, audio and speech codec, and neural audio coding



### Byeongho Jo

- Ph.D degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea in 2021
- Senior Researcher with Media Coding Research Section at the Electronics and Telecommunications Research Institute, Daejeon, South Korea
- ORCID : https://orcid.org/0000-0002-3785-6406
- Research interests : Audio coding, array signal processing, digital signal processing



### Wootaek Lim

- B.S. and M.S. degrees in Electronic Engineering from Kwangwoon University, Seoul, South Korea, in 2010 and 2012, respectively
- Senior Researcher with Media Coding Research Section at the Electronics and Telecommunications Research Institute, Daejeon, South Korea
- ORCID : https://orcid.org/0009-0006-4640-4301
- Research interests : Audio signal processing, Machine learning



### Jungwon Kang

- PhD degree in electrical and computer engineering in 2003 from the Georgia Institute of Technology, Atlanta, GA, USA
- Principal Researcher with Media Coding Research Section at the Electronics and Telecommunications Research Institute, Daejeon, South Korea
- ORCID : https://orcid.org/0000-0003-4003-4638
- Research interests : Video coding, Audio coding, Multimedia processing