Intelligibility of Low Bit rate MPEG-coded Japanese Speech in Virtual 3D audio space

†Yousuke Kobayashi , ‡Kazuhiro Kondo , ‡Kiyoshi Nakagawa
†‡Graduate School of Science and Engineering, Yamagata University
4-3-16 Zyonan, Yonezawa, Yamagata, 992-8510 , Japan
† Tar46576@st.yamagata-u.ac.jp
‡ {kkondo , nakagawa}@yz.yamagata-u.ac.jp

ABSTRACT

In this paper, we investigated the influence of stereo coding on Japanese speech localized in 3-D virtual space. We encoded localized speech using Joint Stereo and Parametric Stereo modes within the HE-AAC (High-Efficiency Advanced Audio Coding) encoder at identical data rates.

First, the sound quality of the localized speech signal was checked using MUSHRA subjective tests. The result showed that the speech quality for Joint Stereo is higher than Parametric Stereo when localized at $\pm 45°$ (where $0°$ refers to localization directly in front of the listener) by 20 to 30 MUSHRA score points. The scores for Joint Stereo were relatively proportional to bit rate. However, Parametric Stereo scores were not proportional to bit rate, and remained fairly constant with bit rate.

Next, the Japanese word intelligibility tests were conducted using the Japanese Diagnostic Rhyme Tests (JDRT). Test speech was localized in front, while competing noise were localized at various angles. The result showed that speech could not be separated from the noise for Joint Stereo when the noise was in located in the frontal region, from $+45°$ to $-45°$, and intelligibility degrades significantly. However at other azimuth, the intelligibility improves dramatically. On the other hand, intelligibility with Parametric Stereo remained constant, at about 70 to 80 %.

1. INTRODUCTION

In this research, we are aiming to use sound localization to separate the primary speaker speech from the other speakers in a multi-party virtual 3D audio conferencing environment. We are using HRTF (Head-Related Transfer Function) to separate sound sources. Vocal Village [1] and Voiscape [2] are examples of such systems. These system integrates both audio and image (still and movie) in a virtual 3D environment. Avatars indicating participants and sound-generating objects are placed at arbitrary locations in this virtual space. Each participant's speech and sound objects are localized at corresponding locations. The user is free to move around in this space, and the sound image locations are altered according to relative position changes.

The focus of the Voiscape system is in the creation of a 3-D multimedia "chat" environment. However, we are focusing more on the communication networking aspects of a similar system. When speech is localized within a virtual space, they require multi-channel representation, most likely stereo if to be presented over a headphone. Thus, stereo signal processing and transmission is required for localized speech. Stereo coding is known to influence stereo sound image. For example sound image is broadened with Twin VQ coding [3]. Mid-Side Stereo and Parametric Stereo coding [4] were shown to have different sound localization azimuth dependency in terms of quality and perceived localization accuracy.

In this paper, we study the effect of stereo audio coding on localized stereo speech, specifically on sound quality and Japanese word intelligibility. Although we are aware of attempts to assess conversional (bidirectional) quality in a similar setup [5], we will only deal with listening only quality here. We used HE-AAC (High-Efficiency Advanced Audio Coding) [6] [7] which is the latest audio encoder currently available, and compared Joint Stereo [7] (which adaptively switches between Simple Stereo and Mid-Side Stereo) with Parametric Stereo [7] [8], which are both part of the HE-AAC standard codec. We only vary the stereo coding mode. The same single-channel audio coding was used (*i.e.* the AAC framework of the HE-AAC codec). Sampling rates supported by the standard are 32, 44.1 and 48 kHz. We used 32 kHz in our experiments since we mainly deal with wideband speed, which typically has bandwidth of 7kHz, and requires sampling rate at or above 16 kHz.

In previous work [9], we have shown that the speech intelligibility of target speech can be kept above 70 % if competing source is placed at azimuth of more than $45°$ from the target speech on the horizontal plane. The sound localization in this case was achieved by applying HRTF of KEMAR (Knowles Electronics Mannequin for Acoustic Research) [10] (to be noted KEMAR-HRTF or KEMAR-HRIR) to the individual sources. We will attempt the same test with stereo-coded speech in this paper.

This paper is organized as follows. In the next chapter, subjective quality listening tests as well as its results are given for the two stereo coding methods. This is followed by speech intelligibility tests when localized speech is presented with competing noise. Finally, conclusions and discussions are given.

2. SUBJECTIVE QUALITY TESTS

In this chapter, we investigated the subjective audio quality of localized speech using the MUSHRA (MUlti Stimulus test with Hidden Reference and Anchors) method [11] [12]. The coding rate as well as the speech localization azimuth was varied. The number of subjects of this test was 10, and the tests were run semi-automatically on a Windows computer.

2.1. Sound Sources

In this listening test, we used 3 read sentences from the Acoustical Society of Japan continuous speech database for research. We chose two male speakers and one female speaker. One of the male speakers had relatively high tone, while the other had a low voice. For all sound sources, the sampling

frequency was 16 kHz, and the quantization bits were 16 bits linear. The sampling rate was up-sampled to 32 kHz.

2.2. Audio Codecs Used in MUSHRA Listening Test

Table. 1 lists the codecs used in this test. We used not only the prescribed 3.5 kHz low-pass filtered audio, but also a 2.0 kHz low-pass filtered speech to use as anchors in the MUSHRA test. The reason for this is that since we are using speech, most of the spectral components are below 3.5 kHz, and thus this filter is not enough for anchors. The HE-AAC coding and decoding was done at 24, 32, 56 kbps for both Joint Stereo and Parametric Stereo coding using the aacPlus? Encoder ver.1.28 [8]. The labels shown in the Table will be used in the figures in later chapters as well.

Table.1. *Audio codecs used in this research.*

Label	Stereo cording	Data rate (kbps)	Codec
Ref.	Simple Stereo	1024	None
LPF3.5k			
LPF2.0k			
ST24	Joint Stereo	24	HE-AAC (aacPlus Encoder Ver.1.28)
ST32		32	
ST56		56	
Pa24	Parametric Stereo	24	
Pa32		32	
Pa56		56	

2.3. Sound Localized in Virtual Audio Space

The test speech was localized using the KEMAR-HRIR [10]. All speech sources were localized at an azimuth of $0°$ and $\pm 45°$.Standard MUSHRA listening tests were conducted with these localized speech as well as the reference and the two anchors. The listeners listened to the localized speech for one speaker, and rated the subjective quality on a 100-point scale. The subjects rated all test sources from the same speaker, and moved on to the next speaker. The presentation order of the speakers was randomized.

2.4. Results of MUSHRA Listening Test

Figure.2 (a) to (c) shows the results of MUSHRA listening test for 3 azimuths. The scores are average for the three tested speakers. The error bars shown are the 95 % confidence intervals.

First, for sound localized at azimuth $0°$, Joint Stereo and Parametric Stereo, the subjective quality is relatively independent of the data rate. However, Parametric Stereo is slightly better for the same data rate, as was seen in [8]. All scores were in the range of 80-100, which refers to a quality rating of "excellent," and these sources are essentially indistinguishable from the reference.

Next, for sound sources localized at azimuth $\pm 45°$, all Joint Stereo coded sound is higher than Parametric Stereo. Moreover, the Joint stereo at azimuth $-45°$ is higher by about 10 points than $+45°$ except at 24 kbps. While Joint Stereo coded quality is rate-dependent, the Parametric Stereo quality is relatively independent of the bit rate, and remains constantly below Joint Stereo at 24 kbps.

ST56 and ST32 show significantly higher scores at $-45°$ than at $+45°$. This difference is not limited to a few listeners, but is seen in at least some samples for almost all listeners. Thus the cause for this difference is still unclear, and will be investigated. Thus, in terms of subjective quality, Joint Stereo coding generally gives superior quality speech than Parametric Stereo coding, especially when speech is localized to the sides.

3. SPEECH INTELLIGIBILITY TEST

In this chapter, we tested the Japanese speech intelligibility with the two stereo coding modes when competing noise is present. We used the Japanese Diagnostic Rhyme Tests (JDRT) [14] for localized speech in 3-D virtual space, as has been investigated in [9] [15].The number of subjects of this test was 7, and tests were ran semi-automatically on a Windows PC.

3.1. The JDRT Intelligibility Test

We conducted the JDRT to measure intelligibility of Japanese. JDRT uses word-pairs that are different only an initial phoneme. In this research, we did not use words that start with vowel. Therefore, changing one initial phoneme means changing the consonant. Consonants were categorized into six attributes, and intelligibility is measured by attributes. We chose a word-pair list consisting of 120 words, or 60 word pairs, 10 word pairs per attribute. The six phonetic attributes were voicing (vocalic and non-vocalic), nasality (nasal and oral), sustention (continuant and interrupted), graveness (grave and acute) and compactness (compact and diffuse).

During the test, the subject listens to the sound of a word. Both words in the word pair are presented visually on the screen, and the subject selects one of the words as the correct word. The subject can repeatedly hear the same sound if they choose to. When the next button is selected, the following sound is presented. This procedure is repeated until the predetermined numbers of words are tested. The words are presented in random order. The selected word is recorded and processed with a PC automatically. The percentage of correct response is adjusted for chance, and is evaluated using the following expression.

$$Chance-Adjusted\ percentage\ Correct\ Response\ [\%] = \frac{(Correct\ Responses - Incorect\ Respnses)}{Total\ Number\ of\ Responses} \times 100\ [\%]$$

3.2. Sound Sources

In the DRT test, we used read speech of one female speaker. Speech samples were originally sampled 16 kHz, and the quantization bits were 16 bits linear. These samples were up-sampling to 32 kHz. We did not test all 6 phonetic attributes since it was previously shown that nasality gives good estimation of the overall average intelligibility across all attributes [16]. Table.2. shows the nasality word pairs.

Table.2. *Nasality word list.*

man	ban	mushi	bushi
nai	dai	men	ben
misu	bisu	neru	deru
miru	biru	mon	bon
muri	buri	nora	dora

3.3. Codec Used in the JDRT

We used all encoding method listed in Table 1, except the anchors, LPF3.5kHz and LPF2.0kHz. Again, the encoder used was aacPlus encoder ver.1.28.

3.4. Localized Source Position in Virtual Audio Space

Previously, we reported in [9], [15], [16] and [17] the JDRT results using the KEMAR-HRTF. We used a similar setup in this test as well. Figure.1. shows the localized position of JDRT sound sources in 3D virtual audio spaces. All sources were located on the horizontal plane. JDRT word speech was localized and presented for target speech, and multi-talker noise [18] was localized and presented as competing noise. Multi-talker noise is just a mixture of many voices, similar to what would be heard in a busy cafeteria. In all tests, the target speech was localized in front ($0°$). We localized the noise at azimuth in $15°$ increments in the frontal region between $\pm 45°$, and $45°$ increments outside this region. We located the noise on a radius relative to the distance between the target speaker and the listener. Denoting as "*a*" the normalized speaker-listener distance, noise was located on a radius with the same distance ("*a*"), twice the distance ("2.0*a*") and half the distance ("0.5*a*").

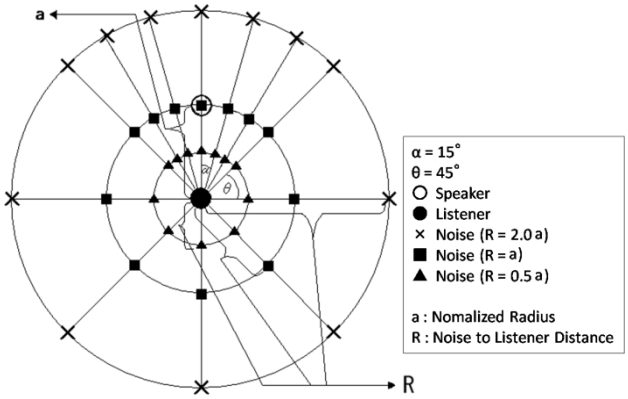


Fig.1 . *Location of the sound sources.*

3.5. JDRT Results

Figure.3. (a) to (c) shown results of JDRT for the two stereo coding modes. The intelligibility scores (Chance-Adjusted Percentage Correct, CACR) shown are the average over all localized noise azimuth and listener to noise distance. The error bars are the 95 % confidence interval.

First, the CACR shown in (a) and (b) are not significant different for both Joint Stereo and Parametric Stereo. Overall the CACR in (a) and (b) are very high, similar to the results in [13].

However, when noise is closer to the listener (c), CACR shows a wider 95 %confidence interval than (a) and (b). Accordingly, the CACR for this distance was broken down into noise azimuth vs. CACR in Fig. 4 (a) for Joint Stereo, and (b) for Parametric Stereo coding. Interestingly, the effect of noise azimuth on CACR is quite different by stereo coding. Joint Stereo shows a significant decrease in intelligibility when noise is located in front of the listener, while in Parametric Stereo, the effect is negligible. However, Joint Stereo shows a much higher CACR, by about 10% compared to Parametric Stereo, at noise azimuth beyond $\pm 90°$. Moreover Parametric Stereo tends to show similar CACR as the reference at all azimuth, ranging mostly from 70% to 80%. Larger variation of CACR with noise azimuth seems to be the cause of the large 95 % confidence interval in Fig. 3(c), especially for Joint Stereo coding.

Thus, coding rate does not have effect on the speech intelligibility. The stereo coding mode also does not have effect on the average speech intelligibility. Noise azimuths relative to the target speech affects the speech intelligibility. However, the effect of noise azimuth differs with the stereo coding mode.

4. CONCLUSIONS

We tested the influence of stereo cording on subjective audio quality and speech intelligibility using HE-AAC on the 3D localized Japanese speech. Joint Stereo and Parametric Stereo was used for stereo coding, while the basic audio coding was fixed at AAC coding.

First, it was found that the subjective audio quality for Joint Stereo compared to Parametric Stereo is higher by about 20 to 30 MUSHRA score points at azimuth $\pm 45°$. However, for sound source sets in front of listener (azimuth $0°$), Joint Stereo and Parametric Stereo did not show difference. The quality of Joint Stereo was proportional to bit rate, but the quality of Parametric Stereo was independent of the bit rate.

Next, the speech intelligibility when target speech was localized in front of the listener at a distance *a*, and the competing noise was localized on the horizontal plane at various azimuths and at a radius of *a*, 2*a* and 0.5*a* was tested using the Japanese DRT, a two-to-one selection based speech intelligibility test. Joint Stereo and Parametric Stereo did not show significant difference when noise was located at R = 2.0*a* and R = *a*. However, at R = 0.5*a*, significant difference in intelligibility was shown, by about 20%, when the noise was located in front of listener. Moreover, With Parametric Stereo coding, the speech intelligibility was relatively independent of noise source location, showing similar correct response rate of 70% to 80% as the reference speech. However, generally, intelligibility and bit rate do not have clear effect on intelligibility. Intelligibility was rather shown to be related to noise azimuth.

From these results, when encoding speech for 3-D audio displays, coding bit rates and stereo coding modes are not a critical factor. Care should be taken to localize sources so that they may be located well away from competing sources. These results suggest some guidelines when designing a 3-D audio conferencing systems using stereo audio coding in the future.

In this research, we only tested with the HE-AAC implementation from Coding Technologies, who proposed the main parts of the standard. Thus, we believe their implementation is fully compliant with their standard, including the inter-channel phase processing. Nonetheless, we would like to test with other standard implementations as well. We also would like to expand our experiments to include other locations of the target speech, including azimuths and radii.

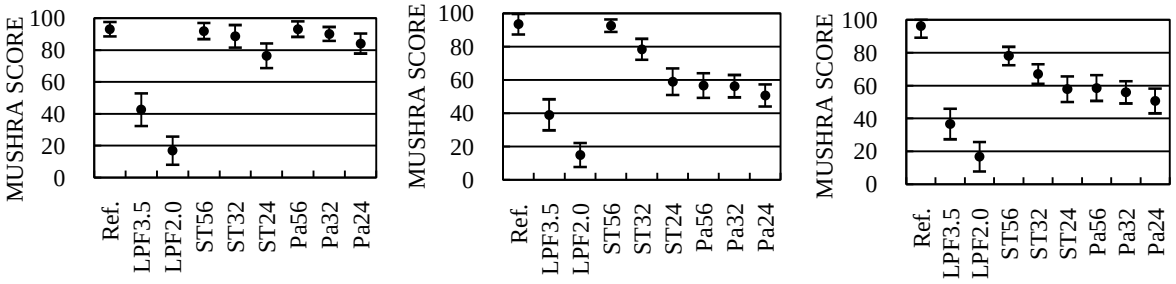
5. ACKNOWLEDGEMENTS

Additional support was provided by The NEC C&C Foundation , The Hara Research Foundation and the Yonezawa Industrial association.

6. REFERENCES

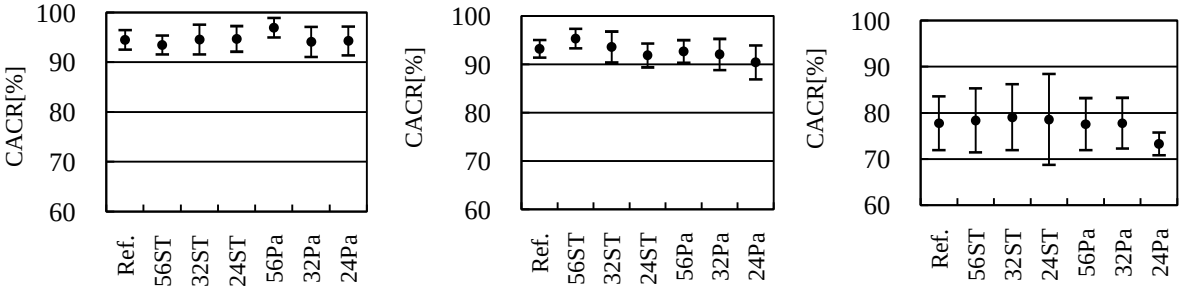
[1] Kilgore, R. *et al.*, "The Vocal Village : Enhancing Collaboration with Spatialized Audioconferencing," *Proc.*

World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education. 2004.
[2] Y.Kaneda, "Subjective Evaluation of Voiscape - A Virtual "Sound Room" Based Communication-Medium", *IEICE*, EA2007-42, Aug. 2007.
[3] J.Nakagai *et al.* "Effects of speech cording with TwinVQ on the perception of a sound image", *IEICE*, EA2000-34, Aug. 2000.
[4] Y.Kobayashi *et al.*, "The Influence of Stereo Coding on the 3D Sound Localization Accuracy", *IEICE*, EA2008-56, Aug. 2008.
[5] Raake, A. et al., "Auditory Assessment of Conversational Speech Quality of Traditional and Spatialized Teleconferences," In: *Proc. 8. ITG-Fachtagung Sprachkommunikation*, 8.-10. Oct. 2008, D-Aachen.
[6] ISO/IEC 14496-3:2003/Amd.1
[7] ISO/IEC 14496-3:2005/Amd.2
[8] Breebaart *et al.* "Parametric Coding of Stereo Audio", *EURASIP J. On Applied Signal Processing*, 9:1305-1322,2004
[9] Y.Kitashima *et al.* "Intelligibility of read Japanese words with competing noise in virtual acoustic space," *J. Acoustical Science and Technology*, vol. 29, no. 1, pp. 74-81 Jan. 2008.
[10] http://sound.media.mit.edu/KEMAR.html
[11] http://www.codingtechnologies.com/products/aacPlus.htm
[12] Recommendation ITU-R BS.1534-1,"Method for the subjective assessment of intermediate quality level coding system"(2001-2003)
[13] G.Stoll *et al.*, "EBU Report on the Subjective Listening Tests of Some Commercial" in *EBU Technical Review*, no. 28, 2000
[14] K.Kondo *et al.* "On a Two-to-one Selection Based Japanese Speech Intelligibility Test", *J. Acoust Soc. Jpn.*, 63, 196-205, 2007
[15] Y.Kobayashi *et al.* "Intelligibility of MPEG-coded read Japanese words in virtual 3D space", *Acoust. Soc. Jpn. Am.* pp. 733-734, 2008
[16] N.Yano *et al.* "The Effect of Localized Speech and Noise Distance on the Speech Intelligibility" *IPSJ-Touhoku* B-2-3, Mar. 2008
[17] Takahito Chiba *et al.* "On the influence of localized position of interference noise on the intelligibility of read Japanese words in remote conference systems," *Inter-noise 2008*, PO-2-0294, Oct. 26-29, 2008
[18] Rice University : Signal Processing Information Base (SPIB), http://spib.rice.edu/spib/select_noise.html



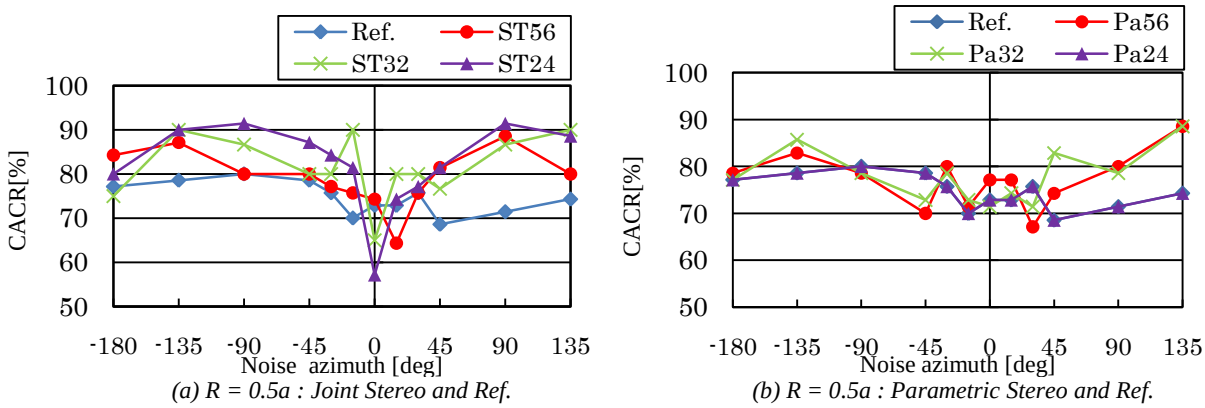
(a) Azimuth 0° *(b) Azimuth-45°* *(c) Azimut+45°*

Fig.2. *Results of the MUSHRA listening test.*



(a) R = 2.0a *(b) R = a* *(c) R = 0.5a*

Fig.3. *Results of JDRT (Average over all azimuths).*



(a) R = 0.5a : Joint Stereo and Ref. *(b) R = 0.5a : Parametric Stereo and Ref.*

Fig.4. *Results of JDRT (vs. noise azimuth).*