Supplementary Document

Anonymous WACV Algorithms Track submission

Paper ID 515

1. More Visual Results

In Fig. 1 we present more results obtained with the three methods. In addition, Fig. 2 and Fig. 3 show the results of hiding different length of audio data in our base single-layer system for the LJ Speech database and the CMU-ARCTIC database [1]. The iterative hiding results of the double nested hidden architecture is shown in Fig. 4.

2. Details of User Study

The rating results by each user are shown in Tab. 1∼Tab. 3, which represent the scores of the results obtained by baseline (raw audio data), STFT and compression respectively.

3. Decoded Video and Audio Results

We show the video and audio results of each experiment in the folder named **Results**, in which the folder named **single-layer** and **double-layer** denote the results for the base system and the double nested hidden architecture, respectively. For fair comparison, we chose the same facial image as input for each experiment. The file names in **single-layer** are: **the results form_the range_the format of the method**, *e.g.* video_0-10_mel.mp4, audio_0-10_mel.wav and container image_0-10_mel.png are the video result, the audio result and the corresponding container image. Similarly, the files in **double-layer** are: **the results form_the range_the N layer**, *e.g.* video_0-20_first-layer.mp4, video_0-20_second-layer.mp4, audio_0-20_first-layer.wav, audio_0-20_second-layer.wav and container image_0-20.png are the video, the audio and the corresponding container image for both layers.

References

[1] John Kominek and Alan W Black. The CMU arctic speech databases. In *Fifth ISCA workshop on speech synthesis*, 2004. 1, 3

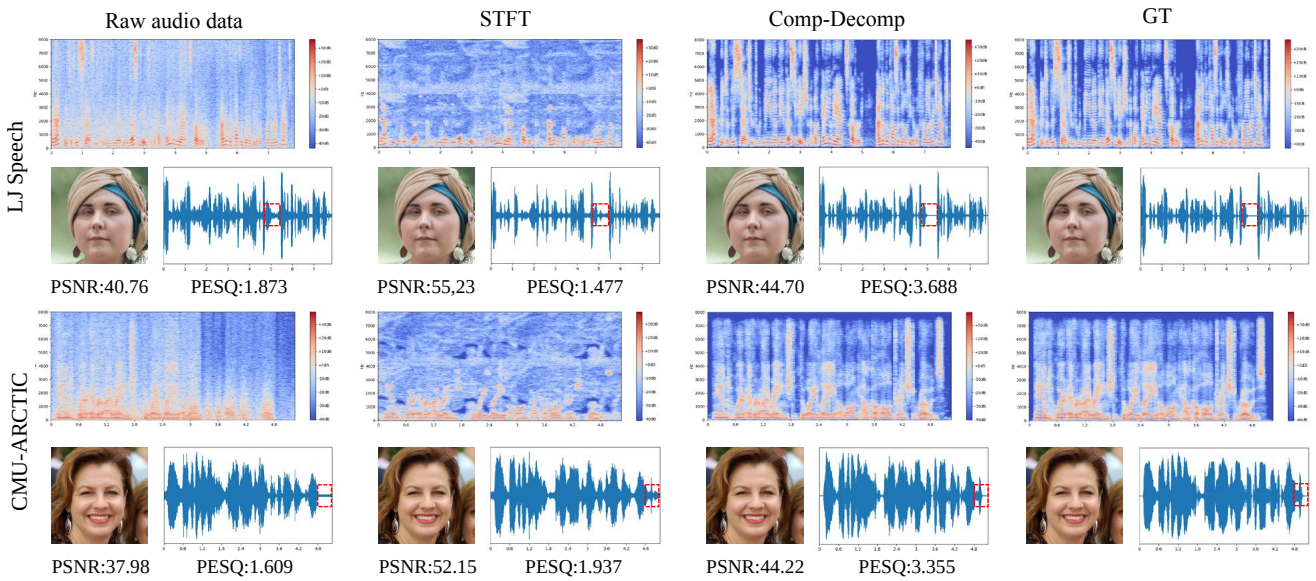

Figure 1. The visual results of container images and recovered audio from different methods.

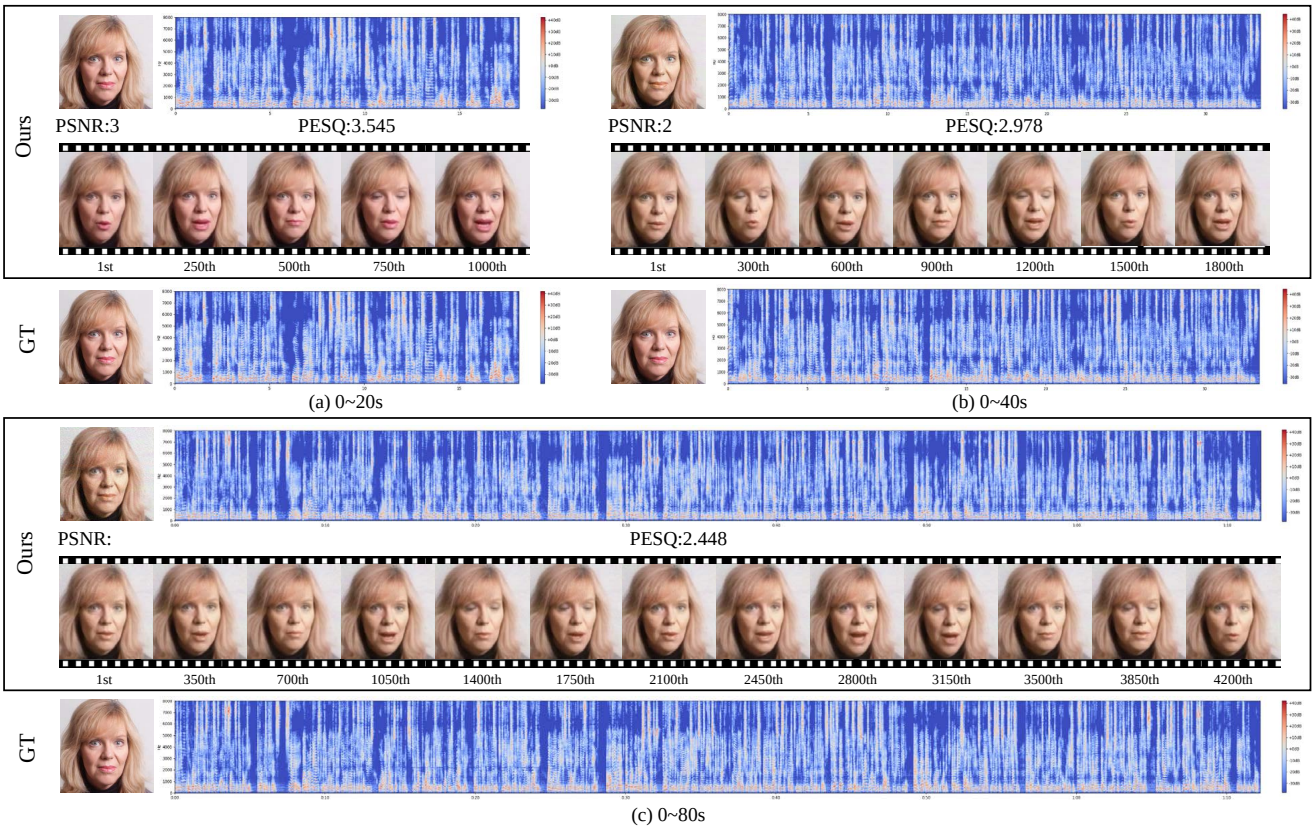

Figure 2. The decoded audio and video frames corresponding to different audio lengths in the LJ Speech database.

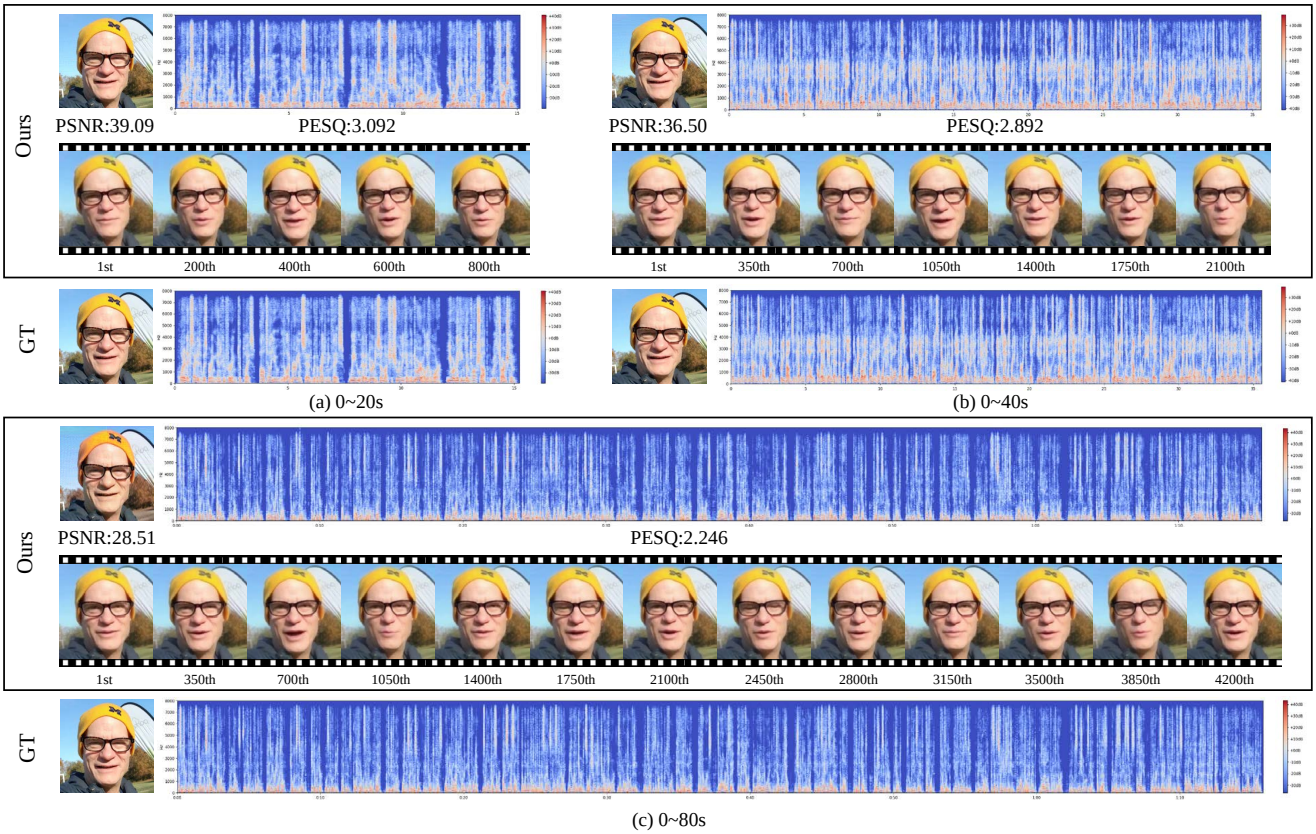

Figure 3. The decoded audio and video frames corresponding to different audio lengths in the CMU-ARCTIC database [1].

Figure 4. Visual results of the double-layer nested hidden architecture in hiding different ranges of audio lengths.

V / U	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19	v20
u1	4.0	3.0	4.0	4.0	3.5	4.0	4.0	4.0	3.0	3.0	4.0	2.5	4.0	4.0	4.0	4.0	4.0	3.5	4.0	3.5
u2	2.5	3.0	3.0	3.0	3.0	3.0	2.5	3.0	2.5	3.0	3.0	2.0	3.5	3.0	3.0	3.0	3.0	2.5	3.0	2.5
u3	3.0	3.5	2.5	3.0	2.5	3.0	3.0	3.5	3.0	3.0	4.0	0.5	4.0	3.0	3.5	4.0	4.0	3.5	4.0	4.0
u4	3.5	3.0	3.0	3.5	3.5	3.5	3.5	4.0	1.0	3.5	3.5	0.0	4.5	3.5	3.0	4.0	4.5	3.0	4.0	2.0
u5	1.0	0.0	0.5	1.0	0.0	1.0	0.0	1.0	0.5	1.0	3.0	0.0	4.0	1.0	1.0	0.0	1.0	0.0	1.0	2.0
u6	3.5	3.0	3.5	3.0	4.0	3.0	3.5	2.5	4.0	3.5	2.5	3.5	3.5	3.0	4.0	4.0	3.5	3.5	3.5	4.0
u7	4.0	3.5	3.5	3.0	3.5	3.5	3.0	4.0	2.5	3.5	3.5	2.5	4.0	3.5	3.5	3.5	3.5	3.0	3.5	3.5
u8	2.0	1.5	2.0	2.0	2.0	1.5	1.5	1.5	1.0	2.0	1.5	1.0	1.5	1.5	1.5	2.0	1.5	2.0	1.5	1.5
u9	2.5	2.0	2.0	2.0	1.5	2.0	1.5	2.5	2.0	3.0	3.0	1.5	3.0	3.0	2.0	2.0	2.0	3.0	1.0	3.5
u10	2.0	2.0	2.5	3.0	3.0	2.0	1.5	3.0	1.5	2.0	2.0	1.5	4.0	3.0	3.0	4.0	2.0	1.5	2.5	2.0
u11	2.0	1.0	1.0	1.0	1.5	1.0	1.0	2.0	1.5	3.0	3.0	1.0	3.0	2.5	1.0	3.0	1.0	2.5	3.0	3.0
u12	2.0	1.5	2.0	2.5	3.0	2.5	1.5	1.5	1.0	1.5	2.0	1.0	3.5	2.0	1.0	2.5	1.5	1.0	2.0	2.0
u13	0.5	0.0	0.0	0.5	1.0	0.5	0.0	0.5	0.0	1.0	1.0	0.0	1.5	1.5	0.5	0.5	0.5	0.0	0.5	0.0
u14	3.0	2.0	3.0	3.0	3.0	3.0	2.0	3.0	2.0	3.0	3.0	1.0	3.0	2.0	3.5	2.0	3.0	2.5	3.5	3.0
u15	1.5	1.0	1.0	1.5	2.0	1.5	1.0	1.0	0.5	1.5	1.5	0.5	2.5	2.0	1.0	1.5	1.0	0.5	1.5	1.0
u16	2.5	1.5	1.5	3.0	3.0	3.0	2.0	2.5	0.5	1.0	1.5	0.5	2.5	1.5	1.5	3.5	1.5	1.0	3.0	2.0
u17	2.0	2.0	1.0	2.0	1.0	2.0	3.0	2.0	1.0	1.0	2.0	2.0	2.0	1.0	1.0	2.0	1.0	1.0	2.0	1.0
u18	2.0	1.0	2.0	3.0	3.0	2.0	1.0	1.0	1.0	2.0	2.0	1.0	3.0	2.0	1.0	2.0	2.0	1.0	2.0	2.0
u19	4.0	4.0	3.5	3.5	4.0	4.0	4.0	4.0	2.5	3.5	4.0	2.0	3.5	3.5	3.5	3.5	2.5	1.5	2.5	2.0
u20	3.0	2.0	3.0	3.0	3.0	3.5	3.0	3.5	2.0	3.0	3.5	2.0	3.5	3.0	2.5	3.0	2.0	1.0	3.0	2.5
u21	3.5	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.5	4.5	4.5	4.5	4.0	4.5	4.0	4.5	4.5	4.0	4.5	4.5
u22	3.0	2.0	2.0	3.5	3.0	2.0	1.5	2.0	1.5	3.0	2.5	1.5	3.0	2.0	3.0	2.5	2.0	2.0	3.0	2.5
u23	3.0	1.5	1.0	2.0	2.5	2.5	0.5	0.5	0.5	0.5	2.5	0.5	2.0	2.5	2.0	4.0	2.0	1.0	2.5	2.5
u24	3.5	3.0	3.5	3.5	3.5	3.5	3.0	3.5	2.5	3.0	3.0	2.5	3.5	2.5	3.0	3.5	3.0	2.5	3.5	3.5
u25	3.0	2.5	3.0	3.5	3.0	3.0	2.5	2.0	2.5	3.0	2.5	2.0	3.0	3.5	2.5	4.0	3.0	2.5	3.0	2.5
u26	3.5	3.0	2.5	2.5	3.0	3.0	1.5	3.0	2.5	3.0	2.5	1.5	3.0	3.5	4.0	4.5	1.5	1.0	2.5	2.5
u27	3.5	2.5	3.0	3.5	3.0	4.0	3.5	3.5	1.0	3.0	3.0	2.0	2.0	3.0	2.0	3.5	3.0	1.0	3.5	3.5
u28	3.0	2.0	3.0	3.0	4.0	3.0	3.0	3.0	3.0	4.0	4.0	2.0	4.0	3.0	3.0	2.0	2.0	3.0	3.0	4.0
u29	3.5	3.0	3.5	3.5	3.5	3.5	3.0	3.0	2.5	3.0	3.5	2.0	3.0	2.5	2.5	2.5	2.5	2.5	2.5	3.0
u30	2.0	1.0	1.5	1.5	2.0	2.0	2.5	1.5	3.0	1.0	3.0	1.0	2.0	2.5	2.0	2.0	1.0	2.0	3.0	2.5

Table 1. The scores for the generated audio-visual results of the baseline method. "v1∼v20" are the 20 videos, and "u1∼u30" represent the 30 users.

U \ V	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19	v20
u1	2.5	2.5	3.5	4.0	3.5	4.0	4.0	3.5	3.0	2.5	3.5	2.5	4.0	4.0	4.0	3.0	3.0	3.0	3.0	4.0
u2	3.0	2.5	2.5	3.0	2.5	2.5	2.5	3.0	2.5	2.5	3.0	2.5	3.5	3.0	3.0	3.5	3.0	3.0	3.0	3.0
u3	2.0	1.5	1.5	0.5	2.0	1.0	2.0	4.0	2.5	2.5	4.5	1.0	3.5	2.5	3.0	3.5	3.5	3.0	3.5	3.5
u4	3.0	2.0	1.5	3.0	3.0	3.0	3.0	4.0	3.0	3.5	4.0	0.5	4.0	3.0	2.5	4.0	4.0	3.5	4.0	3.0
u5	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.5	2.0	0.0	3.0	0.0	4.0	1.0	1.0	0.0	1.0	1.0	0.0	1.0
u6	3.0	2.5	3.0	2.5	3.5	3.0	3.5	2.0	3.5	3.0	2.0	3.5	2.5	2.5	3.0	3.0	3.5	3.0	3.0	4.0
u7	2.5	2.0	2.0	2.5	3.0	2.5	2.5	3.5	2.5	3.0	3.5	2.5	3.0	3.0	2.5	3.5	3.0	2.5	3.0	3.5
u8	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	1.5	2.0	2.0	1.0	2.0	2.0	2.0	2.0	1.5	2.0	2.0	2.0
u9	2.0	2.5	2.0	2.5	2.0	2.0	2.0	3.0	2.0	2.0	3.0	2.5	2.5	2.5	2.5	3.5	2.0	2.0	3.0	3.0
u10	1.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1.0	1.0	1.0	0.5	1.5	0.0	0.5	0.5	0.5	0.5	0.5	0.5
u11	3.0	2.0	2.5	2.0	2.5	3.0	2.0	4.0	3.0	3.0	4.0	3.0	4.0	3.5	3.0	3.0	2.5	3.5	4.0	3.5
u12	3.0	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.0	1.5	2.0	1.0	3.5	2.0	1.5	2.5	1.0	1.5	2.0	2.0
u13	1.0	0.5	1.0	1.0	1.5	1.0	1.0	1.5	0.5	0.5	2.0	1.5	3.5	2.0	1.5	1.5	1.5	1.0	1.5	1.5
u14	3.0	2.0	2.0	2.0	2.0	2.0	2.0	3.0	2.0	2.0	2.0	1.0	3.0	2.0	3.5	3.0	2.5	2.0	3.0	3.0
u15	2.0	1.0	1.5	1.5	1.5	1.5	1.5	1.5	1.0	1.0	2.0	1.5	3.5	2.0	1.5	2.0	1.5	1.5	2.0	2.0
u16	3.0	1.0	1.0	2.5	2.0	2.5	2.0	2.0	1.0	2.0	2.0	2.0	3.0	2.0	2.5	3.5	2.0	2.0	3.0	3.0
u17	3.0	3.5	3.0	3.0	2.0	3.0	4.0	3.0	3.0	2.0	3.0	2.0	2.0	3.0	2.0	2.0	3.0	2.0	2.0	2.0
u18	2.0	1.0	1.0	2.0	1.0	1.0	2.0	1.0	1.0	2.0	2.0	1.0	3.0	2.0	2.0	2.0	1.0	1.0	2.0	2.0
u19	3.0	3.0	3.0	3.5	2.0	3.5	4.0	4.0	2.5	3.5	4.5	2.5	4.0	3.5	3.5	4.0	2.5	2.0	2.5	3.0
u20	2.5	3.0	2.0	1.5	1.5	2.0	1.5	2.5	1.5	2.0	3.0	1.0	2.5	2.0	2.5	2.5	3.0	2.0	2.5	3.5
u21	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.5	4.5	4.0	4.0	4.5	4.5	4.0	4.5	4.5
u22	3.5	2.5	3.0	3.5	3.0	3.0	3.0	3.0	2.5	3.0	3.5	3.0	4.0	3.5	3.5	3.5	2.5	3.5	4.0	3.5
u23	4.0	2.5	3.0	3.0	3.5	2.5	2.0	3.0	1.5	1.0	3.0	0.5	3.0	3.0	3.0	4.5	2.5	2.0	3.0	3.0
u24	3.5	3.0	3.0	3.0	3.0	3.0	3.5	3.5	2.5	2.5	3.0	2.0	3.0	2.5	2.5	3.0	2.5	2.0	3.0	3.0
u25	3.0	2.5	3.0	3.0	2.5	2.5	2.5	2.0	2.5	3.5	3.0	2.0	3.0	3.0	3.0	3.5	2.5	2.5	2.5	2.5
u26	3.0	2.5	2.0	3.0	3.5	2,5	2.0	3.5	3.0	2.5	3.5	1.0	2.5	2.0	3.0	3.5	2.0	3.5	1.0	2.0
u27	2.0	3.0	3.0	4.5	2.0	3.0	2.5	4.0	2.5	4.0	3.5	2.0	3.0	2.0	3.5	3.0	4.0	3.0	2.5	2.5
u28	4.0	2.0	4.0	2.0	3.0	2.0	3.0	4.0	3.0	3.0	4.0	2.0	4.0	2.0	3.0	2.0	3.0	2.0	2.0	3.0
u29	4.0	3.5	3.5	3.0	3.0	3.0	3.0	3.0	2.5	3.0	3.5	2.0	3.0	2.5	2.5	2.5	2.5	2.5	2.5	3.0
u30	3.0	2.0	2.0	1.0	2.5	2.0	2.0	1.5	2.0	1.0	2.5	1.0	1.5	2.0	1.0	1.5	1.5	2.0	2.5	3.5

Table 2. The scores for the audio-visual results of the STFT method.

U \ V	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19	v20
u1	5.0	5.0	5.0	3.5	5.0	5.0	5.0	3.5	5.0	5.0	5.0	5.0	5.0	4.0	4.5	5.0	5.0	4.5	5.0	5.0
u2	5.0	5.0	3.5	5.0	5.0	5.0	4.5	5.0	5.0	4.0	5.0	4.5	5.0	5.0	5.0	5.0	4.0	5.0	5.0	3.5
u3	5.0	4.5	5.0	5.0	5.0	4.5	4.0	5.0	4.5	4.5	5.0	3.5	4.0	4.0	4.0	4.0	5.0	4.0	5.0	4.5
u4	5.0	4.5	4.5	5.0	4.5	5.0	4.5	4.5	4.0	4.5	5.0	4.5	4.5	4.5	4.5	4.5	5.0	5.0	5.0	5.0
u5	2.0	2.5	3.0	4.0	2.0	4.0	1.0	4.0	4.0	3.0	3.5	2.0	5.0	4.0	3.0	1.0	4.0	4.5	2.0	3.0
u6	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
u7	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
u8	3.0	3.0	5.0	3.5	3.0	3.0	3.5	5.0	3.0	3.0	3.5	2.0	3.0	2.5	3.0	3.5	2.0	2.5	3.0	3.5
u9	4.5	4.0	4.0	5.0	3.0	4.0	3.5	4.5	5.0	5.0	3.5	3.0	3.0	5.0	3.0	5.0	3.0	4.5	3.5	4.5
u10	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
u11	4.0	3.5	4.5	3.0	4.0	4.5	4.5	5.0	4.0	4.0	4.5	4.0	5.0	5.0	4.5	3.5	3.0	4.5	4.5	4.0
u12	4.5	4.5	5.0	5.0	5.0	4.5	4.0	3.5	3.5	4.5	4.0	3.5	5.0	4.5	4.5	5.0	5.0	4.5	5.0	5.0
u13	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
u14	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
u15	5.0	5.0	5.0	5.0	5.0	5.0	4.5	4.5	4.5	5.0	4.5	4.5	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
u16	5.0	5.0	4.5	4.5	5.0	5.0	4.0	5.0	3.5	4.0	5.0	3.0	5.0	4.5	4.0	4.5	4.0	5.0	5.0	5.0
u17	4.5	4.5	4.0	4.0	4.0	4.5	5.0	3.5	4.5	4.0	4.5	4.0	4.5	5.0	4.0	4.5	5.0	4.5	4.5	4.0
u18	4.0	3.0	4.0	4.0	4.0	4.0	4.0	3.0	3.0	5.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
u19	5.0	4.5	4.0	5.0	4.5	4.5	5.0	4.5	5.0	4.0	5.0	5.0	4.0	4.0	4.0	5.0	5.0	4.0	4.0	4.5
u20	4.0	4.5	4.5	5.0	4.5	4.0	4.5	5.0	4.5	4.0	5.0	4.5	5.0	4.0	3.5	5.0	4.0	3.5	3.5	4.0
u21	5.0	5.0	5.0	5.0	5.0	5.0	5.0	4.0	5.0	5.0	5.0	4.5	5.0	5.0	5.0	5.0	5.0	5.0	4.5	5.0
u22	4.5	4.5	5.0	5.0	4.5	4.5	4.5	4.0	4.0	4.5	4.5	5.0	5.0	5.0	4.5	4.5	4.5	4.5	5.0	5.0
u23	5.0	5.0	4.5	5.0	4.5	5.0	5.0	4.5	5.0	4.5	5.0	4.0	5.0	5.0	4.0	5.0	4.5	5.0	5.0	5.0
u24	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.0	3.5	5.0	4.0	3.5	4.0	4.0	4.0	4.5	3.5	4.0	4.0	4.5
u25	4.0	3.5	4.0	4.5	4.0	4.0	4.0	3.5	4.0	4.5	4.0	3.5	3.5	4.0	3.5	4.5	3.5	4.0	4.0	5.0
u26	4.5	3.0	2.0	4.0	4.0	4.0	3.5	4.5	4.0	4.5	4.5	3.0	4.0	4.5	5.0	5.0	4.5	4.5	5.0	4.5
u27	4.0	4.0	4.0	5.0	3.5	3.5	4.5	4.5	5.0	4.0	5.0	4.0	4.0	4.0	3.5	4.0	5.0	5.0	5.0	5.0
u28	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	4.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
u29	3.5	4.5	4.0	5.0	4.0	4.5	4.5	3.5	3.5	4.0	4.0	2.5	3.0	3.5	3.0	3.5	3.5	4.0	3.0	3.5
u30	3.5	3.0	3.0	3.0	3.0	2.5	1.5	3.0	4.0	3.5	2.0	3.5	4.0	3.5	4.0	4.5	3.5	3.5	3.5	4.0

Table 3. The scores for the audio-visual results of the compression-decompression method.