Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Development and Evaluation of Bone-conducted Ultrasonic Hearing-aid Regarding Transmission of Speaker Emotion: Comparison of DSB-TC and DSB-SC Amplitude Modulation Method

Takayuki Kagomiya[1], *Seiji Nakagawa*[2]

[1]National Institute for Japanese Language and Linguistics, Japan
[2]National Institute of Advanced Industrial Science and Technology (AIST), Japan
t-kagomiya@ninjal.ac.jp, s-nakagawa@aist.go.jp

Abstract

Human listeners can perceive speech signals in a voice-modulated ultrasonic carrier from a bone-conduction stimulator even for sensorineural hearing loss patients. Considering this fact, we have developed a bone-conducted ultrasonic hearing aid (BCUHA). However, there remains considerable scope for improvement, particularly in terms of sound quality. Voice-modulated BCU is accompanied by a strong high-pitched tone and some distortion depending on the amplitude modulation method. In this study, the sound quality of a BCUHA with double-sideband transmitted carrier (DSB-TC) modulation and double-sideband suppressed carrier (DSB-SC) modulation methods was examined. The assessment was conducted by examining the transmission of the emotional state of speakers. The evaluation used emotion-identification experiments. Types of emotion included Ekman's basic six emotions ("anger," "disgust," "fear," "joy," "sadness," and "surprise") and "neutral." In addition, a series of subjective evaluations regarding "voice clarity," "comfortableness," and "preference" was conducted. The results showed that although DSB-SC sound was superior in "comfortableness" and "preference," voice emotion transmission was more effective in DSB-TC conditions.

Index Terms: hearing-aid, ultrasound, bone-conduction, amplitude-modulation, speaker emotion, multi-dimensional scaling

1. Introduction

We have developed a bone-conducted ultrasonic hearing aid (BCUHA) for sensorineural hearing-impaired patients [1] . A BCUHA consists of 2 components: an amplitude-modulated ultrasound processor and a bone-conduction vibrator (Figure 1).

Ultrasound is defined as sound with a frequency higher than the limitation of human perception (about 15 kHz). However, humans can perceive sound transmitted as ultrasound through a bone-conduction vibrator (bone-conducted ultrasound, BCU). Moreover, if the ultrasound is amplitude modulated by speech sounds, original speech sounds can be perceived besides the carrier sound, and this amplitude-modulated BCU can be perceived by not only normal-hearing (NH) listeners but also hearing-impaired patients. We based the development of our BCUHA on these observations.

As mentioned above, the BCUHA provide amplitude-modulated ultrasound through bone-conduction vibrator. Thus sound quality of BCUHA depend on amplitude modulation (AM) method to some extent. For current prototype of BCUHA, double-sideband transmitted carrier (DSB-TC) method was adopted because BCUHA with DSB-TC showed high performance at transmission of linguistic messages in our previous studies [1, 2] . A series of listening tests by profoundly hearing impaired patients revealed that 40 % of participants were able to perceive some sounds and 17 % were able to recognize words [1] .

Figure 1: *Ceramic vibrator of the BCUHA attached to the mastoid with a hair band-like device*

Although these results demonstrated potential of the BCUHA, however, further development and improvement for practical use are required. Particularly, there is room for improvement in sound quality. As mentioned above, AM method for the current prototype of BCUHA is DSB-TC. In the DSB-TC method, modulated amplitude envelope corresponds to original sound signal. However, DSB-TC method transmit not only original sound information, but also carrier signal; in our prototype BCUHA, 30 kHz sinusoid. This carrier-peak power is perceived as a strong high pitch tone; about 15-17 kHz. In our previous study, BCUHA listeners reported that this high pitch prevents speech listening.

To reduce this high pitch tone, we have proposed other AM methods. In this study, we reports evaluation of double-sideband suppressed carrier (DSB-SC) method. The DSB-SC method can suppress peak power at carrier frequency, thus it is expected that this AM method has advantage in reducing high pitch tone. On the other hand, envelope of DSB-SC waveform does not correspond to the original signal. The pitch accompanied by the envelope is almost twice as high as that of the original signal, and contains some distortion.

As mentioned above, mono-syllable or word intelligibility scores BCUHA with DSB-TC were superior to that of DSB-SC [1, 2, 3] . This inferiority of DSB-SC in transmission of linguistic message can be explained by its sound distortion. Linguistic messages are mainly connected with segmental aspects of spoken language. Segmental phones are short term events, thus it is

 doi: 10.21437/SpeechProsody.2016-244

Table 1: *Selected words*

num. mora	initial acc.	N-1 acc.	no acc
3	midori (green)	naname (slant)	nagame (scenery)
4	arawani (revealed)	arayuru (every)	omonaga (long-faced)
5	naniyorimo (first of all)	yawarageru (soften)	amarimono (remains)

inferred that segmental information are fragile for sound distortion. However, oral speech does not only convey linguistic messages but also paralinguistic messages: emotion, gender, age of speaker, etc. Many researches pointed out that such paralinguistic messages are connected with prosodic aspects of speech [4, 5, 6] . Prosodic events are relatively long term and global, thus it is expected robust for sound distortion of DSB-SC.

To prove this hypothesis and investigate the situation where DSB-SC is effective, a series of listening experiments was conducted.

2. Method

2.1. Stimuli

The experiment was designed using emotion detection tasks. Stimuli were selected according to the following procedures.

2.1.1. Types of emotions

For this task, 7 emotion types were selected. Six types out of the 7 were Ekman's basic emotions [7] : "anger," "disgust," "fear," "joy," "sadness," and "surprise." In addition to these emotions, "neutral" was selected.

2.2. Selection of sounds

Half of the sounds were obtained from the Keio Emotional Speech Database (Keio-ESD). This corpus was developed by the Keio University, and contains 20 words or phrases spoken by a male in his 30s with 47 emotions [8] .

From the Keio-ESD corpus, 9 accentual phrases each spoken with the 7 emotions described above were selected according to accent types. The Japanese language has tonal or pitch accent. Manner of accent is one of the major cues used to convey emotional state or other paralinguistic messages [4] . Accent types selected were "initial accent," "N-1 accent," and "no accent." "Initial accent" words are enunciated with an accent on the first mora. "N-1 accent" words have a penultimate accent. In this study, "no accent" words included 2 types: words with no lexical accent and words with a final-mora accent; both these types of words have no accent in isolated speech. The number of moras was also considered in the selection of words, and all words had 3-5 moras. Following these procedures, the 9 words listed in Table 1 were selected.

As mentioned above, the speech sounds contained in the Keio-ESD were spoken by one male speaker. Thus, a female voice was recorded to balance the gender of speakers. A female speaker in her 20s participated in recording. She was an amateur actress and majored in opera singing in a university of music. The words and emotions were identical to those selected from the Keio-ESD.

As a result of these processes, 126 stimuli were generated (7 emotions $\times$ 9 words $\times$ 2 genders).

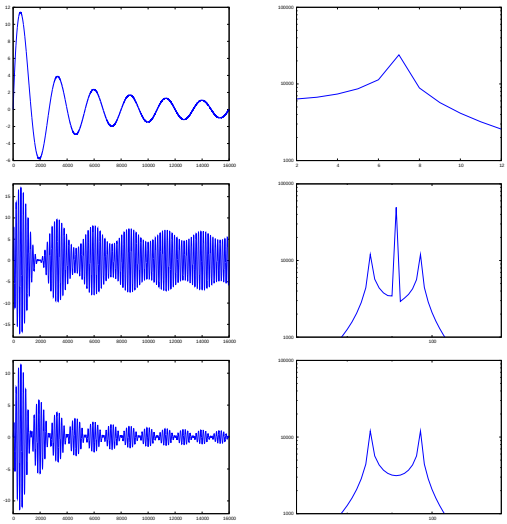

Figure 2: *Waveform and spectrum patterns of DSB-TC and DSB-SC. Upper: original, middle: DSB-TC, bottom: DSB-SC, left: waveform, right: spectrum*

2.3. Bone-conducted ultrasound

The stimuli were converted to BCU stimuli in the form of amplitude-modulated 30-kHz sinusoid waves.

In the DSB-TC modulation, an AM signal $U_{\mathrm{tc}}(t)$ at a point of time t is represented as equation (1)

$$U_{\mathrm{tc}}(t) = [S(t) - S_{\mathrm{min}}] \times \sin(2\pi f_c t) \quad (1)$$

where $S(t)$ represents original signal, S_{min} is minimal value in $S(t)$, f_c indicates frequency of carrier signal (30kHz).

While in the DSB-SC modulation, an AM signal $U_{\mathrm{sc}}(t)$ is represented as following equation

$$U_{\mathrm{sc}}(t) = S(t) \times \sin(2\pi f_c t) \quad (2)$$

Example of both AM modulation are shown in Figure 2. A damped wave signal (upper) is modulated by DSB-TC (middle) and DSB-SC (bottom) respectively. As shown in Figure 2, a DSB-TC modulation signal has a substantial peak power at carrier frequency, it is accompanied by a strong high pitch tone. On the other hand, a DSB-SC signal lacks such a strong peak power at carrier frequency, however the pitch accompanied by the envelope is almost twice as high as that of the original signal.

2.4. Experiments

2.4.1. Participants

Nine native Japanese speakers (7 males and 2 females) with no reported hearing or speaking defects participated in the experiments. Their ages were in the range 19-42 years.

2.4.2. Procedures

The BCU stimuli were presented using a custom-made ceramic vibrator (Figure 1). Bone-conducted ultrasound can be perceived when it is applied to various parts of the body, and the mastoids are among the locations where such perception is high.

Table 2: *Mean values and results of Wilcoxon signed-rank tests for the subjective evaluation scores*

	voice clarity	comfortableness	preference
DSB-TC	3.222	2.667	2.889
DSB-SC	3.889	3.889	3.889
p	p =0.280	p =0.019	p =0.018

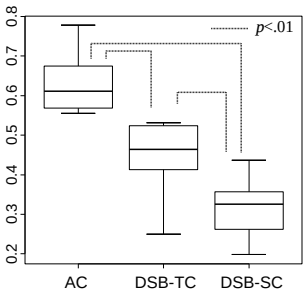

Figure 3: *Correct perceived ratio in each condition*

Therefore, we applied the vibrator to the left or right mastoid of the subject by using a hair band-like device (Figure 1).

Besides the BCU conditions, air-conduction (AC) condition experiments were conducted as a baseline condition. For AC condition, the sound stimuli were presented through a headphone (Sennheiser HD200).

Participants were presented with the stimuli, and they were requested to identify the emotion from 7 choices: the 7 emotion types.

The order of the stimuli was counterbalanced. All experiments were conducted in a soundproof chamber, and the sound levels of the stimuli were adjusted to the most comfortable level for each participant.

2.5. Subjective evaluation

Besides main experiments, sevens-scale subjective evaluations were conducted. Participants were requested to judge subjective values about "voice clarity," "comfortableness" and "preference" of the each AM-BCU sound in seven-scales after main experiments.

3. Results and Analysis

3.1. Subjective evaluation

Table 2 shows mean value of each subjective values for both AM methods. As shown in Table 2, DSB-SC sound showed higher scores in all items. A series of Wilcoxson signed-rank tests showed the differences were significant in "comfortableness" and "preference" at 5 % level.

3.2. Confusion Pattern of the Emotions

Table 3 shows the confusion patterns of each emotion. Responses of all participants were pooled. Rows represent stimuli and columns show responses. Correct perception ratios of each emotion are shown as diagonal elements.

Further, as shown in Table 3, overall correct perceived ratio of the AC condition was 0.643, the DSB-TC condition was 0.443 and DSB-SC was 0.319. An ANOVA and post-hoc test (pairwise t test with Holm 's adjust method) revealed that the differences between each condition were significant at 1% level (Figure 3).

Table 3: *Confusion matrices of each emotions*

AC	anger	disgust	fear	joy	neutral	sad	surprise
anger	0.758	0.126	0.000	0.011	0.022	0.005	0.066
disgust	0.148	0.379	0.000	0.110	0.324	0.005	0.038
fear	0.011	0.016	0.308	0.000	0.000	0.484	0.170
joy	0.038	0.093	0.000	0.604	0.225	0.000	0.033
neutral	0.027	0.044	0.000	0.016	0.901	0.000	0.000
sad	0.005	0.022	0.242	0.005	0.011	0.703	0.000
surprise	0.137	0.005	0.044	0.060	0.005	0.033	0.714
	correct perceived ratio: 0.628						
DSB-TC	anger	disgust	fear	joy	neutral	sad	surprise
anger	0.389	0.161	0.044	0.072	0.189	0.050	0.106
disgust	0.117	0.256	0.078	0.050	0.394	0.050	0.056
fear	0.044	0.067	0.289	0.039	0.056	0.383	0.128
joy	0.100	0.061	0.033	0.467	0.283	0.028	0.033
neutral	0.078	0.100	0.006	0.044	0.744	0.017	0.011
sad	0.011	0.100	0.283	0.028	0.044	0.511	0.022
surprise	0.222	0.028	0.083	0.072	0.039	0.100	0.456
	correct perceived ratio: 0.443						
DSB-SC	anger	disgust	fear	joy	neutral	sad	surprise
anger	0.378	0.161	0.067	0.033	0.194	0.067	0.100
disgust	0.178	0.278	0.050	0.072	0.306	0.072	0.050
fear	0.144	0.061	0.183	0.100	0.122	0.294	0.094
joy	0.233	0.089	0.044	0.261	0.222	0.050	0.100
neutral	0.061	0.183	0.044	0.072	0.544	0.061	0.033
sad	0.089	0.089	0.217	0.106	0.117	0.344	0.039
surprise	0.311	0.056	0.061	0.144	0.117	0.067	0.244
	correct perceived ratio: 0.319						

3.3. Multi-Dimensional Scaling Analysis

To obtain more information from the confusion matrices, a series of Kruskal's multi-dimensional scaling (MDS) analyses [9] were conducted. MDS analyses were applied for each confusion matrix shown in Table 3. Stress scores were checked and all scores were below 0.15 in the 2-dimensional condition (AC, 0.007; DSB-TC, 0.07; DSB-SC, 0.14). These scores indicated that the fitness of the MDS models to the confusion matrices was "good." Figure 4 shows distribution of each emotion according to the results of MDS.

In the AC condition, each emotion was well separated. On dimension 1, "fear" and "sad" located edge of positive area, while "disgust" located opposite side. On dimension 2, "anger" and "joy" were located at both extremities. From these tendencies, the distribution of each emotion generally can be explained by Schlosberg's two dimensional model [10]; dimension 1 can be interpreted as "attention-reject" dimension, while dimension 2 can be regarded as "pleasantness-unpleasantness" dimension.

While in the DSB-TC condition diagram, although "fear" and "sadness" had relatively small areas of distribution and nearly overlapped, the other emotions were relatively well separated. In dimension 1, "joy" was located positive area, "fear" and "sad" were distributed in negative area. From these tendencies, dimension 1 can be also regarded as "pleasant-unpleasant" dimension. In dimension 2, "surprise" and "anger" were located in positive area, "disgust" and "neutral" lay negative area. Therefore this dimension can be interpreted as "attention-reject" dimension. Consequently, broad tendency of the distribution under DSB-TC condition is similar to that of AC condition.

Although dimension 2 was inverted, a similar tendency was

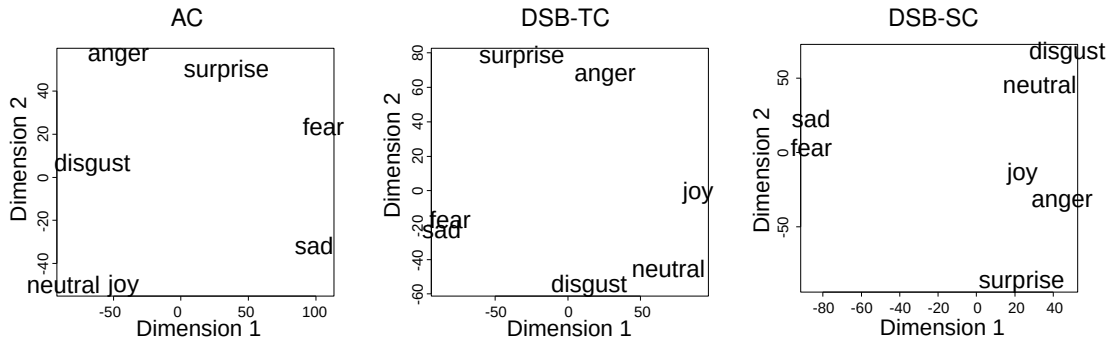

Figure 4: *Scatter diagram of each emotion by the results of MDS*

Table 4: *Acoustic parameters applied for multiple regression analysis*

F0 avg	mean value of F0 (standardized within speakers)
F0 SD	standard deviation of F0 (standardized within speakers)
RMS avg	mean value of RMS power
RMS SD	standard deviation of RMS power
Duration	duration of utterance
Speed	mora/second

Table 5: Result of multiple regression analysis

	β	Std. β	p
intercept	-132.067		$p < 0,001$
F0 avg	0.716	0.715	$p < 0,001$
F0 SD	-0.886	-0.290	$p < 0,001$
adjusted R^2: 0.431			

also observed in the DSB-SC condition; "fear" and "sadness" were also closely distributed, "joy" and "sad" or "fear" were opposite to each other on dimension 1, "surprise" and "disgust" or "neutral" were in opposition on dimension 2. Thus, dimension 1 can be interpreted as "pleasant-unpleasant" dimension, dimension 2 can be regarded as "attention-reject" dimension.

However, difference between DSB-TC and DSB-SC distribution was also observed. Under DSB-SC condition, "disgust," "neutral," "joy," "anger," and "surprise" distribute in narrow range on dimension 1, while these emotions scattered on dimension 1 under DSB-TC condition. Thus, it is inferred that DSB-SC listeners have difficulty in perception of degree of "pleasant."

3.4. Factors to predict "pleasant" score

As described above, the results of MDS suggest that DSB-SC listeners have difficulty in perceiving acoustic cues that help in "pleasant" degree. In this respect, the DSB-TC is superior to the DSB-SC. Therefore, determining which cues contribute to prediction of "pleasant" scores facilitates understanding of the differences between DSB-TC and DSB-SC performances.

To determine the parameters used to prediction of "pleasant" scores, a series of multiple regression analyses and variable selection were conducted. Acoustic parameters listed in Table 4 were used as predictor variables, and all parameters were estimated for each stimulus. Akaike 's information criterion (AIC) was adopted for choosing the best predictor subsets.

As a result of these procedures, a predictor subset consisting of "F0 avg" and "F0 SD" was selected. Table 5 shows the result of the multiple regression analysis. From this analysis, it is inferred that the main factor that helps in transmission of "pleasant" degree is the F0 value. In other words, under the DSB-SC condition, listeners may have had difficulties in perceiving differences in average F0 value and levels of F0 movement between the stimuli.

4. General Discussion

First, the results of this study reveal that emotional state is transmitted more efficiently with the DSB-TC than DSB-SC. From the results of MDS and multiple regression analysis, this advantage of the DSB-TC is mainly due to differences in transmission of F0 information. The reason of this inferiority in F0 transmission of DSB-SC can be accounted for by its sound distortion and twofold pitch envelope.

However, DSB-TC listeners have discontent about "comfortableness" and "preference." The subjective evaluation score revealed DSB-SC coding can reduce this discomforness. Thus it may be efficient in practical use that switching DSB-TC and DSB-SC according to situation: switch to DSB-TC when listening human speech, in other situation, switch to DSB-SC.

In addition, the results of this study demonstrate the effectiveness of applying emotion transmission experiments for the evaluation of hearing aids. The inferiority in F0 transmission is hard to comprehend word or monosyllable intelligibility tests.

5. Conclusion

To determine best AM method for the BCUHA, speaker emotion identification experiments were conducted. The evaluation was conducted by comparison of AC, DSB-TC and DSB-SC conditions. The results showed that although subjective preference is better than DSB-TC, DSB-SC coding was inferior to DSB-TC in speaker emotion transmission. This result indicates that the best AM method for BCUHA in this respect is DSB-TC.

6. Acknowledgements

This research was supported by a Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (24500675, 25280063, 2628130, 15K01495, 15K00245 and 26560320).

7. References

[1] S. Nakagawa, Y. Okamoto, and Y. Fujisaka, "Development of a bone-conducted ultrasonic hearing aid for the profoundly sensorineural deaf," *Transactions of the Japanese Society for Medical and Biological Engineering : BME*, vol. 44, no. 1, pp. 184–189, 2006.

[2] Y. Okamoto, S. Nakagawa, K. Fujimoto, and M. Tonoike, "Intelligibility of bone-conducted ultrasonic speech," *Hearing Research*, vol. 208, pp. 107–113, 2005.

[3] S. Nakagawa, C. Fujiyuki, and T. Kagomiya, "Development of a bone-conducted ultrasonic hearing aid for the profoundly deaf: Evaluation of sound quality using a semantic differential method," *Japanese Journal of Applied Physics*, vol. 52, no. 7, pp. 07HF06:1–6, 2013.

[4] K. Maekawa, "Production and perception of 'paralinguistic' information," in *Proceedings of Speech Prosody 2004*, 2004, pp. 367–374.

[5] H. Helfrich, "Age markers in speech," in *Social Markers in Speech*, K. R. Scherer and H. Giles, Eds. London, New York and Melbourne: Cambridge University Press, 1979, ch. 3, pp. 63–107.

[6] P. M. Smith, "Sex markers in speech," in *Social Markers in Speech*, K. R. Scherer and H. Giles, Eds. London, New York and Melbourne: Cambridge University Press, 1979, ch. 4, pp. 109–146.

[7] P. Ekman and W. V. Friesen, *Unmasking the face: a guide to recognizing emotions from facial clues*. Oxford, England: Prentice-Hall, 1975.

[8] T. Moriyama, S. Mori, and S. Ozawa, "A synthesis method of emotional speech using subspace constraints in prosody," *Journal of Information Processing Society of Japan*, vol. 50, no. 3, pp. 1181–1191, 2009, (in Japanese).

[9] J. B. Kruskal, "Nonmetric multidimensional scaling: a numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, 1964.

[10] H. Schlosberg, "The description of facial expression in terms of two dimensions," *Journal of Experimental Psychology*, vol. 44, pp. 229–237, 1952.