Body-conductive acoustic sensors in human-robot communication

Panikos Heracleous, Carlos T. Ishi, Takahiro Miyashita, and Norihiro Hagita

Intelligent Robotics and Communication Laboratories
Advanced Telecommunications Research Institute International
2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto-fu 619-0288, Japan
{panikos,carlos,miyasita,hagita}@atr.jp

Abstract

In this study, the use of alternative acoustic sensors in human-robot communication is investigated. In particular, a Non-Audible Murmur (NAM) microphone was applied in teleoperating Geminoid HI-1 robot in noisy environments. The current study introduces the methodology and the results of speech intelligibility subjective tests when a NAM microphone was used in comparison with using a standard microphone. The results show the advantage of using NAM microphone when the operation takes place in adverse environmental conditions. In addition, the effect of Geminoid's lip movements on speech intelligibility is also investigated. Subjective speech intelligibility tests show that the operator's speech can be perceived with higher intelligibility scores when operator's audio speech is perceived along with the lip movements of robots.

Keywords: Geminoid, NAM microphone, Subjective evaluation

1. Introduction

To date, many studies have addressed the problem of human-robot interaction (Stiefelhagen et al., 2007). Components or modalities such as speech, gestures, gaze, and others have been used in order to facilitate a natural human-robot interaction. Particular efforts have been focused on designing and developing human-like robots (Kanda et al., 2004). The Geminoid HI-1 robot, which was developed at the ATR, Intelligent Robotics and Communication Laboratories, Japan (Nishio et al., 2007; Backer-Asano et al., 2010) is a teleoperated anthropomorphic robot, which is a duplicate of an existing person. Since speech is the most natural modality for human-human communication, in human-Geminoid interaction speech communication also plays an important role. In addition to audio speech, communication is also performed by lip movements of Geminoid.

In this study, results of subjective speech intelligibility tests conducted to evaluate the importance and the effect of Geminoid's lip movements on speech intelligibility are reported. Several subjects have been employed in order to evaluate the intelligibility of speech perceived during interaction, and under clean and noisy conditions.

Moreover, in this study the use of alternative acoustic sensors is also investigated. In particular, speech uttered by an operator while using a NAM microphone to teleoperate Geminoid was subjectively evaluated and compared with speech uttered while using a standard microphone.

The remainder of this study is organized as follows: In section 2, the Geminoid system and its characteristics are introduced. In Section 3, we introduce the NAM microphone. The differences between speech perceived by a normal microphone and a NAM microphone are also described. Section 4 introduces the Diagnostic Rhyme Test, which was used in this study for evaluation of speech intelligibility. Section 5 describes the methodology and the experimental setup, and Section 6 reports the results obtained. In Section 7 the current work is discussed, and Section 8 concludes the work.

(a) (b)

Figure 1: (a) and (b): The Geminoid HI-1 (left) and its master person (right).

2. The Geminoid HI-1 teleoperated android

Figure 1a and Figure 1b show the Geminoid HI-1 robot with its master person. Geminoid HI-1 is a duplicate of its creator. A geminoid is an android, designed to look exactly as real human. Its terminology is originated from the Latin word "geminus", which means twin and "oides" meaning similarity. Anthropomorphic robots such as Geminoid, are designed to be very similar to real humans with features such as artificial skin and hair, and are controlled through a computer system that replicates the facial movements of the operator in the robot. Humanoid robots belong to another robot family, and includes robots designed for human-robot interaction. Humanoid robots, however, do not have the appearance of real humans.

In Geminoid HI-1, the robotic element has identical structure as previous androids (Ishiguro, 2005). Particular efforts focused on designing a robot to be a copy of the master person. Silicone skin was molded by a cast taken from the original person; shape adjustments and skin textures were painted manually based on MRI scans and photographs. Fifty pneumatic actuators drive the robot to generate smooth and quiet movements. The 50 actuators were

This work was supported by KAKENHI (21118001).

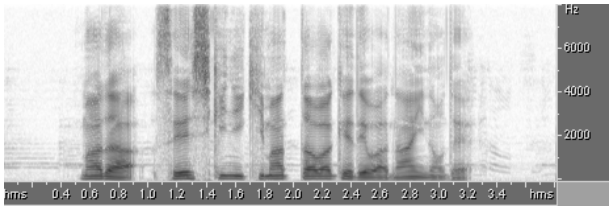

Figure 2: Spectrogram of an audible utterance received by a close-talking microphone.

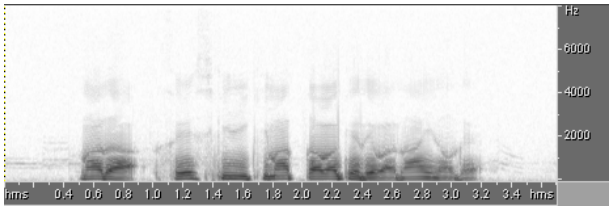

Figure 3: Spectrogram of an audible utterance received by a NAM microphone.

determined to effectively show the movements necessary for human interaction, and also express the master's personality traits. Thirteen actuators are embedded in the face, 15 in the torso, and the remaining 22 move the arms and legs.

3. Non-Audible Murmur (NAM)

Non-Audible Murmur (NAM) refers to a very softly uttered speech received through the body tissue. A special acoustic sensor (i.e., the NAM microphone) is attached behind the talker's ear. This receives very soft sounds that are inaudible to other listeners who are in close proximity to the talker.

The first NAM microphone was based on stethoscopes used by medical doctors to examine patients, and was called the stethoscopic microphone (Nakajima et al., 2003). Stethoscopic microphones were used by the first author for the automatic recognition of NAM speech (Heracleous et al., 2004). The silicon NAM microphone is a more advanced version of the NAM microphone (Nakajima et al., 2005). The silicon NAM microphone is a highly sensitive microphone wrapped in silicon; silicon is used because its impedance is similar to that of human skin. Silicon NAM microphones have been employed for automatic recognition of NAM speech as well as for NAM-to-speech conversion (Toda and Shikano, 2005). Similar approaches have been introduced for speech enhancement or speech recognition (Zheng et al., 2003).

The speech received by a NAM microphone has different spectral characteristics in comparison to normal speech. In particular, the NAM speech shows limited high-frequency contents because of body transmission. Frequency components above the 3500-4000 Hz range are not included in NAM speech. The NAM microphone can also be used to receive audible speech directly from the body [Body Transmitted Ordinary Speech (BTOS)]. This enables automatic speech recognition in a conventional way while taking advantage of the robustness of NAM against noise. Figure 2 shows the spectrogram of an audible utterance received by a close-talking microphone and Figure 3 shows the spectrogram of the same utterance received by a NAM microphone. As is shown in the figure, only low frequency components are included in the NAM speech.

Figure 4: Old version (left) and new version (right) of NAM microphone.

Previously, the first author of the current paper reported experiments for NAM speech automatic recognition that produced very promising results. A word accuracy of 93.9% was achieved for a 20k Japanese vocabulary dictation task when a small amount of training data from a single speaker was used (Heracleous et al., 2004). The HMM distances of NAM sounds in comparison with the HMM distances of normal speech were also investigated, which indicated distance reduction when NAM sounds were concerned (Heracleous et al., 2010).

In this study, a new version of NAM microphone is used to receive audible speech through the body tissue (i.e., BTOS) at the operator's side. The study aims at investigating the use of NAM microphones in Geminoid-human interaction under noisy conditions. However, very often Geminoid or similar robots are operated in noisy environments, such as conferences, malls, etc. Since NAM microphones show increased robustness against noise, using a NAM microphone instead of a standard microphone in operating Geminoid might be advantageous in adverse environmental conditions.

Figure 4 shows the old and the new versions of the NAM microphone. As is shown, the new version of NAM microphone is smaller than the previous one, and can be attached to the talker more easily. The new version is made by a special material and – to some extent – can be attached to the talker without using any supporting device.

4. Diagnostic Rhyme Test

Speech can be evaluated based on intelligibility, naturalness, and suitability for a specific application. Speech intelligibility is a measure of how well speech can be understood, and is different from speech naturalness. Depending on the application, intelligibility or naturalness appear to

Table 1: Description of the JDRT

Characteristics	Description	Examples
Voicing	voiced - unvoiced	zai - sai
Nasality	nasal - oral	nai - dai
Sustension	sustained - interrupted	hata - kata
Sibilation	sibilated - unsibilated	jamu - gamu
Graveness	grave - acute	waku - raku
Compactness	compact - diffuse	kai - pai

be the most important. For instance, in reading machines for blind, speech intelligibility with high scores is more important than speech naturalness. In contrast, other applications (e.g., multimedia applications) require speech with high rates of naturalness.

Speech intelligibility can be evaluated subjectively or objectively. In the case of subjective evaluation, speech intelligibility is measured by subjective listening tests based on human perceptions. The response sets are usually syllables, words, or sentences. The test sets usually focus on consonants, because consonants have more important role in speech understanding, than vowels.

Among other subjective tests, Diagnostic Rhyme Test (DRT) is very widely used to evaluate speech intelligibility. In this case, a set of word pairs is used to test speech intelligibility. A pair consists of two words, which differ by a single phonetic characteristic in the initial consonant. Specifically, voicing, nasality, sustension, sibilation, graveness, and compactness phonetic characteristics are evaluated by DRT.

In this study, the Japanese Diagnostic Rhyme Test (JDRT) was used to evaluate speech intelligibility (Fujimori et al., 2006). JDRT consists of 60 word pairs, which evaluate the six phonetic characteristics (i.e., 10 word pairs for each characteristic). Table 1 shows the description of the JDRT.

5. Controlling the lip movements of Geminoid

The proposed method is divided in two parts; one is the operator side, while the other is the robot side. In the operator side, formant extraction is firstly conducted on the input speech signal. Then, the origin of the coordinates of the formant space given by the first and second formants (F1 and F2) are translated to the center of the vowel space of the speaker (adjusted by a graphic user interface), for accounting for differences in the vowel space depending on the speaker's feature, such as gender, age and height. A rotation of 25 degrees is realized on the new coordinates, so that the F1 axis has better matching with the lip height. Although lip width can also be estimated from the normalized vowel space, only lip height is controlled in the present work due to physical limitations of the robot. The estimated lip height is then converted to the actuator commands by a linear scaling. Audio packets and actuator commands are sent to a remote robot in intervals of 20 ms. In the robot side, the audio packets and the lip motion actuator commands are received, the actuator commands are sent to the robot for moving the lip actuators, and a delay is controlled for playing the received audio packets, for synchronizing the two streams.

6. Methods

6.1. Procedure for evaluation of Geminoid's lip movements

The experiments were conducted in the Geminoid's room, in a reverberant environment with 38 dB(A) background noise. The environment was chosen to be as much as similar to the environment where human-Geminoid interaction occurs.

For this subjective evaluation, ten subjects (i.e., 6 males and 4 females) were employed. The subjects were normal-hearing, undergraduate students, and were paid to participate the experiments. Their age was between 20 and 24 years old. Before the experiments, they did not meet in real life the Geminoid robot.

The subjects were seat in front of Geminoid at a distance of about 1.5m. They were instructed to also watch the Geminoid's lips/face while listening to a word. The stimulus consisted of the 120 words of the JDRT, which were pre-recorded and played back one by one during the experiment. Each subject was provided with a list, which included the correct uttered word and its pair in the JDRT. The subjects were instructed to definitely choose one of the two written words. For example, if the uttered word was zai, the subjects had to choose between zai and sai.

The experiment consisted of four sessions. Specifically, speech intelligibility was evaluated under the following conditions: speech with lip movements, speech without lip movements, speech with lip movements and background babble noise of 70 dB(A) level played back through a loud speaker at the subject's place, and speech without lip movements with the same babble noise at the subject's place. The order of the four sessions was randomly selected in order to avoid memorizing the correct words by the subjects.

6.2. Procedure for evaluation of NAM microphone

Intelligibility of speech received by a NAM microphone and speech received by a standard microphone were evaluated in noisy conditions. This experiment corresponds to the case when the operator is located in a noisy environment and uses NAM microphone to teleoperate Geminoid.

In this experiment, the uttered words for intelligibility evaluation included noise. To simulate this situation, babble noises at 70 dB(A) and 80 dB(A) levels were played back through a loud speaker, and were recorded simultaneously by a NAM and a standard microphone. The recorded noises were then superimposed onto the clean words to simulate the noisy stimuli.

For this evaluation, four subjects were used (i.e., two males and two females). The subjects were students and employees working in the laboratory. All of them were normal-hearing. While listening the stimulus, the subjects were instructed to watch the Geminoid's lips/face movements. The same word lists as in the previous experiments were used.

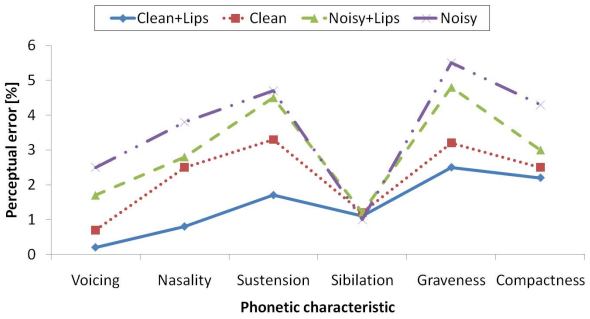

Figure 5: Evaluation of speech intelligibility with respect to Geminoid's lip movements.

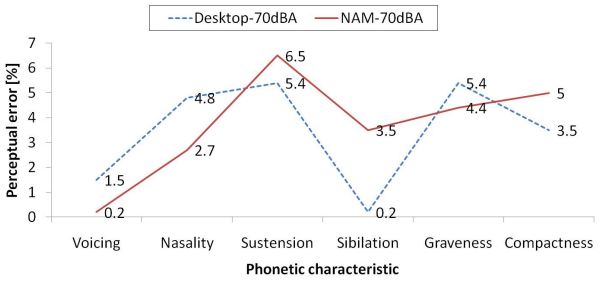

Figure 6: Evaluation of speech intelligibility with noisy stimuli [70dB(A)].

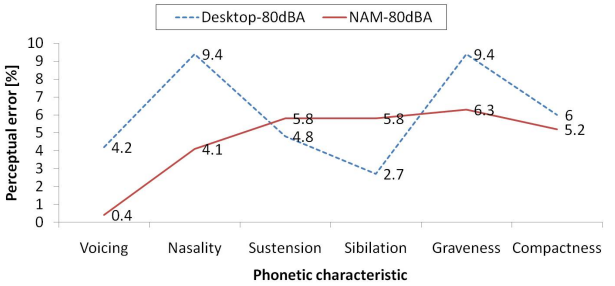

Figure 7: Evaluation of speech intelligibility with noisy stimuli [80dB(A)].

7. Results

7.1. The effect of lip movements on speech intelligibility

Figure 5 shows the results obtained in the experiments. The results show, that lip movements of Geminoid have an effect on perceptual scores in both clean and noisy environments. The highest overall score was achieved when speech intelligibility was evaluated in a clean environment with lip movements. In this case, the evaluation score was 92.2%. The second highest score was obtained when the evaluation was performed in clean environment, but without lip movements. In this case, the score was 87.5%. When the speech was evaluated in noisy environment with lip movements, the score was 82.2%. Finally, in the case of evaluation in noisy environment without lip movements the score was as low as 79.7%.

The results clearly show that when interacting with Geminoid in both clean and noisy environment, the operator's clean speech can be better understood when also watching Geminoid's face and lip movements. This phenomenon is very similar to the one appearing in human-human communication. Since speech includes both audio and visual modalities, audiovisual speech perception results in higher speech intelligibility rates.

7.2. Speech intelligibility using NAM and desktop microphones

Figure 6 shows the results when noisy stimulus of 70 dB(A) was used. The results show that in the case of sustension, sibilation, and compactness phonetic characteristics, lower error rates were obtained when using standard microphone. In all the other cases, the NAM microphone achieved lower error rates. The differences in the error rates might be explained by the limited frequency band of the NAM microphone. When, however, the uttered word has initial consonant with rich information in the higher frequency band, this word may be confused resulting in lower intelligibility scores. The overall perceptual score was 79.2% when using the standard microphone and 77.7% when the NAM microphone was used. The scores are closely comparable, and the difference was statistically not significant.

Figure 7 shows the results when noisy stimulus of 80 dB(A) was used. As is shown, using a NAM microphone significantly lower error rates were obtained in most of the cases. Only in the cases of sustension and sibilation, desktop microphone performed slightly better. The overall intelligibility score for the standard microphone was 63.5% and for the NAM microphone 72.4%. The difference in intelligibility scores was statistically significant.

Table 2 shows the p-values of multiple paired two-tailed t-tests (Box, 1987). As is shown in the Table, in the case of NAM microphones, the difference between 70 dB(A) and 80 dB(A)noise levels was statistically not significant. This observation indicates that NAM microphones are more robust against noise compared to standard microphones. In the case of standard microphones, the difference between 70 dB(A) and 80 dB(A) noise levels was statistically significant.

8. Discussion

This study reports the results of subjective speech intelligibility tests aiming at evaluating several aspects in human-robot interaction. Experiments were conducted to evaluate the effect of Geminoid's lip movements on speech intelligibility. The results obtained justify the effectiveness of Geminoid's lip movements while interacting with humans in both noisy and clean environments. Also, the achieved results show the effectiveness of the method used to control the lip movements. It might be possible, however, that lower intelligibility scores are achieved if the speech was not accurately synchronized with lip motions, or if the lip motions were not accurately synthesized according to the uttered speech. This would result in a similar effect to the McGurk effect (McGurk and MacDonald, 1976).

Table 2: p-values of multiple paired two-tailed t-tests

Speech	Speech			
	NAM 70dB	NAM 80dB	Standard 70dB	Standard 80dB
NAM 70dB	-	0.3451	0.6952	0.0354
NAM 80dB	-	-	0.0154	0.0267
Standard 70dB	-	-	-	0.0084
Standard 80dB	-	-	-	-

Anthropomorphic robots such as Geminoid HI-1 are designed and developed to interact with humans in different environments where also noise might be present. For more effective operation of Geminoid, the use of alternative acoustic sensors was also investigated in this study. Previously, the authors conducted experiments using a NAM microphone and demonstrated its robustness against noise. For this reason, the use of NAM microphone by the operator was also investigated when Geminoid was teleoperated in noisy environments. The results obtained show that NAM microphone might be very useful in highly noisy environments.

Although at this stage an automatic speech recognition module is not included in the Geminoid, in the future we plan to investigate the possibility of decreasing the efforts of the operator by using automatic speech recognition while interacting with Geminoid. When the interaction takes place in adverse environments, users can use NAM microphones to operate the speech recognition engine.

9. Conclusions

In this study, the effect of Geminoid's lip movements on speech intelligibility was investigated by conducting subjective tests. The results showed that with lip movements, the intelligibility rates increase. This study also compares speech intelligibility using a NAM microphone and a standard microphone. The achieved results show the effectiveness of using NAM microphone in adverse environmental conditions.

Acknowledgements

The authors thank Dr. Christian Becker-Asano for providing supporting materials, and also Tomoko Honda for assisting in the experiments.

10. References

C. Backer-Asano, K. Ogawa, S. Nishio, and H. Ishiguro. 2010. Exploring the uncanny valley with geminoid hi-1 in a real-world application. *IADIS Intl. Conf. Interfaces and Human Computer Interaction*, pages 121–128.

J. F. Box. 1987. Guinness, gosset, fisher, and small samples. *Statistical Science*, 2:61–66.

M. Fujimori, K. Kondo, K. Takano, and K. Nakagawa. 2006. On a revised word-pair list for the japanese intelligibility test. *In Proc. of International Workshop Frontiers in Speech and Hearing Research*, pages 103–108.

P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano. 2004. Non-audible murmur (nam) recognition using a stethoscopic nam microphone. *In Proc. of Interspeech2004-ICSLP*, pages 1469–1472.

P. Heracleous, V.A. Tran, T. Nagai, and K. Shikano. 2010. Analysis and recognition of nam speech using hmm distances and visual information. *IEEE Transactions on Audio, Speech, and Language Processing*, 8 Issue 6:1528–1538.

H. Ishiguro. 2005. Android science: Toward a new cross-disciplinary framework. *In Proc. of Toward Social Mechanisms of Android Science: A CogSci 2005 Workshop*, pages 1–6.

T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro. 2004. Interactive robots as social partners and peer tutors for children: A field trial. *Human Computer Interaction (Special issues on human-robot interaction*, 19:61–84.

H. McGurk and J. MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. 2003. Non-audible murmur recognition. *In Proc. of EUROSPEECH*, pages 2601–2604.

Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. 2005. Remodeling of the sensor for non-audible murmur (nam). *In Proc. of Interspeech2005-EUROSPEECH*, pages 389–392.

S. Nishio, H. Ishiguro, and N. Hagita. 2007. Geminoid: Teleoperated android of an existing person. *in Humanoid Robots: New Developments, de Pina Filho, A.C. eds., I-Tech Education and Publishing*.

R. Stiefelhagen, H. K. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel. 2007. Enabling multimodal human-robot interaction for the karlsrune humanoid robot. *IEEE Transactions on Robotics*, 23, No 5:840–851.

T. Toda and K. Shikano. 2005. Nam-to-speech conversion with gaussian mixture models. *In Proc. of Interspeech2005*, pages 1957–1960.

Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, , and Z. Huang. 2003. Air- and bone-conductive integrated microphones for robust speech detection and enhancement. *in Proc. of ASRU*, pages 249–253.