RoboCup Asai-Pacific 2021 Symposium
Aichi, Japan
November 29, 2021

RCAP2021-04-PM1-2

A Digital Speech Watermarking Method Using Spread Spectrum for Secret Communication in RoboCup Soccer

Kazuya Tsubokura[1], Takashi Kuboya[1], Takuma Tachi[1], Akihisa Sanae[1], and Kunikazu Kobayashi[1]

Aichi Prefectural University, 1522-3 Ibaragabasama, Nagakute, Aichi 480-1198, JAPAN.
{im212008, im211004, is181040}@cis,aichi-pu.ac.jp,
kobayashi@ist.aichi-pu.ac.jp
http://www.ist.aichi-pu.ac.jp/lab/robocup-spl/index.html

Abstract. At present, in the RoboCup Soccer Standard Platform League (SPL), wireless communication is mainly used as a means of communication between robots. However, in order to become closer to human soccer, it is necessary to develop a communication method to replace wireless communication. In this paper, we investigate voice communication as an alternative to wireless communication. Specifically, we propose a method of embedding hidden data into the synthesized speech of natural language using the spread spectrum method. We then compared and verified the bit error rate when the distance between the robots and the size of the watermark information were changed.

Keywords: Soccer Standard Platform League · Humanoid Robot · Secret Communication · Digital Speech Watermarking · Spread Spectrum

1 Introduction

By 2050, RoboCup aims to create a robot soccer team that can beat the world champions in the Soccer World Cup [1].

The RoboCup Standard Platform League (SPL) is a league in which all teams compete for superiority in software systems by using the same robot [2]. Currently, a humanoid robot NAO developed by SoftBank Robotics is used for the standard robot.

At present, RoboCup uses wireless communication for inter-robot communication. However, in order to treat the same conditions as humans in the future, an alternative communication method is required. In particular, SPL carries out technical challenges such as wireless communications with limited packet sizes and sound localization using whistles. In this way, SPL requires the development of inter-robot communication technology that uses sound rather than wireless communication.

In general, communication using sound can be thought of as coded sounds such as Morse code or utterances in natural language. The former is difficult for humans to discern. Considering the entertainment aspect, it is difficult to create the interaction between players and spectators as seen in human soccer, and it is feared that the spectators will be left behind. On the other hand, Communication using natural language enables the spectators to understand the communication between robots and to guess what the robots are thinking and moving. However, there is a danger that the amount of information transmitted per unit time will be small and that the content of the dialogue will be transmitted to the other team.

In this paper, we propose a method to embed secret data in natural language utterances. Specifically, by embedding bit strings in synthesized speech using the spread spectrum method, the spectators can understand the interaction between robots from the synthesized speech, and the robot can execute commands and communications from the embedded bit strings. Furthermore, it is also possible to ensure confidentiality so that bits are not read by the other team.

2 Related work

2.1 Spread Spectrum

A Spread Spectrum (SS) method is a modulation method of signal, and is used for wireless communication such as cellular phone and wireless LAN, and digital audio watermarking [3–5]. In the SS method, communication is performed using a signal with a bandwidth much wider than the bandwidth required for normal information transmission. This method has excellent characteristics such as confidentiality and noise resistance in conducting communications.

The SS method includes Direct Sequence (DS)/SS and Frequency Hopping (FH)/SS. The DS/SS method is more secure than that the FH/SS method because signal strength can always be kept weak. In this paper, we consider speech communication using the DS/SS method, because communication between robots requires high secrecy.

Figure 1 shows the data flow in the DS/SS transmitter and receiver configuration. Figure 2 shows the data in each state. At first, the data is modulated using Binary Phase Shift Keying (BPSK) or Quadrature Phase Shift Keying

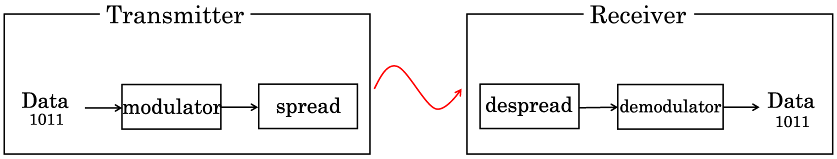

Fig. 1. Data flow in the DS/SS method.

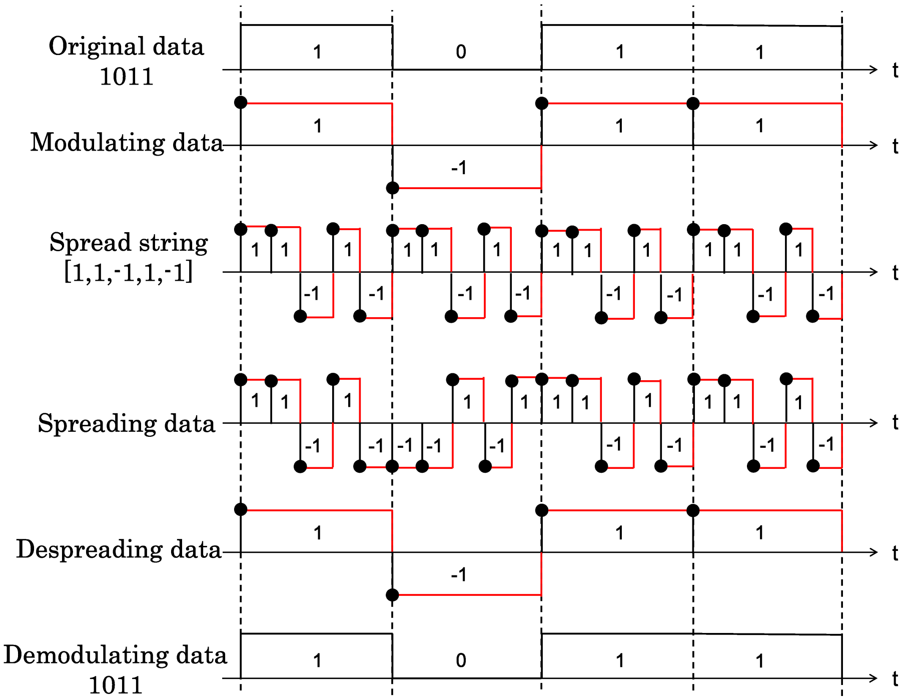

Fig. 2. Data in each state in the DS/SS method.

(QPSK). The BPSK modulation is applied in this study because it is not suitable for high-speed communication but transmission quality does not deteriorate easily. Next, data is spread using a spread string, and the data is transmitted from the transmitter to the receiver. The data before spread can be obtained by despreading the received data. Finally, data can be restored by demodulating the data with a demodulator.

Despreading In despreading, a correlator is generated with a spreading string as a weighting factor as shown in Fig. 3, and the correlation with the received signal is obtained. The higher the correlation, the larger the absolute value of the output value becomes. Therefore, by each peak of the correlator output is extracted as shown in Fig. 4, the modulated data before spreading can be restored.

Rake Combining The spread spectrum method can solve the problem of selective fading, in which the waveform of the transmitted signal becomes distorted. However, this method is affected by flat fading, which reduces the amplitude of the transmitted signal waveform, so rake combining is necessary [4].

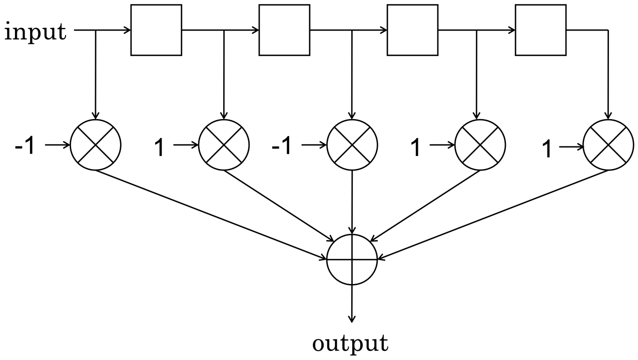

Fig. 3. Correlator Input and output images.

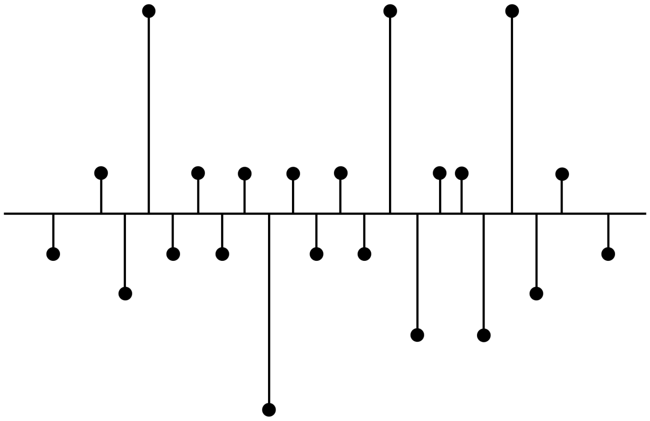

Fig. 4. Example of correlator output.

In rake combining, the multipath signal separated by the spread spectrum method is subjected to maximum ratio combining so that the SN ratio is maximized, so that the received transmitted signal becomes large and the effect of fading can be mitigated. In this case, the maximum ratio combining is a method of combining the phases by multiplying each transmission path by a weighting factor. This weighting factor is called the channel coefficient, and the signal transmitted by spreading 1 is converted using a correlator.

2.2 Linear Predictive Analysis

Linear prediction analysis is a model in which the current output sample value x_t is predicted by a linear combination of the previous N sample values. This method is widely used in speech analysis such as estimation of vocal tract filters [6, 7].

Vocal tract filter$H(z)$ is approximated by the following equation.

$$H(z) = \frac{1}{1 + \sum_{n=1}^{N} \alpha_n z^{-n}},$$

where α is called a linear prediction coefficient, and this value is adjusted to approximate the vocal tract filter.

3 Proposed Method

In this paper, a bit string to be transmitted was spectrally spread and embedded into a synthesized speech. This voice was transmitted from a robot's speaker, and the bit string was recovered from the voice received by another robot. The flow from voice generation to bit recovery in the proposed method is shown in Fig. 5. This method enables the embedding of confidential information into a voice that is easily understandable by humans.

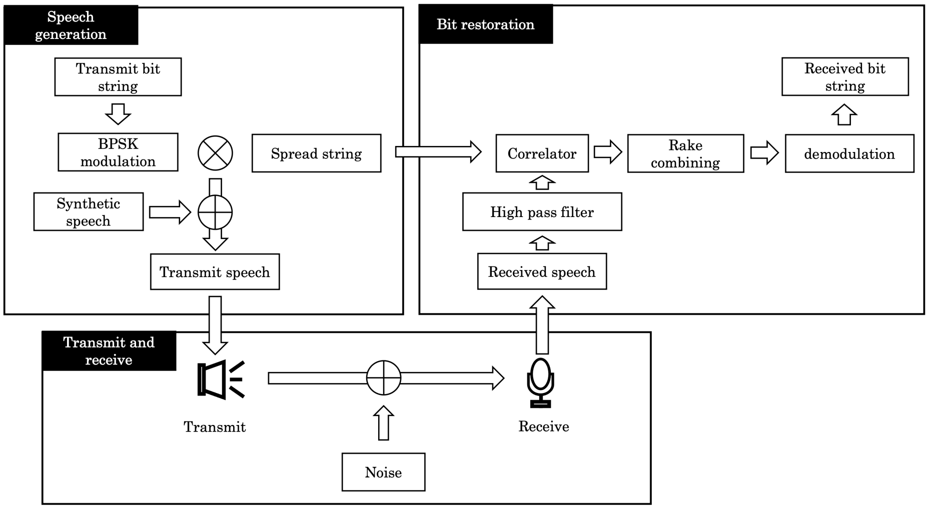

Fig. 5. Flow of the proposed method from speech generation to bit recovery.

3.1 Embedding Bit String into Audio

In the speech generation part, BPSK modulation of the bit string to be transmitted at first. For example, if the data to be transmitted is {1, 0, 1, 0}, it becomes {1, -1, 1, -1} after modulation by BPSK (0 is converted to -1). This data and the spreading string are multiplied bit by bit to obtain a spreading bit string. To adjust the sound pressure level of the spreading bit string, the synthesized speech is multiplied by a weighting factor to amplify the power and then added to the spreading bit string to obtain the voice to be played back.

3.2 Restoring Bit String

In the bit restoration section, the recorded voice is subjected to a high-pass filter at first, which reduces the influence of the synthesized speech and eliminates low-frequency noise. Next, the self-correlation of the high-pass filtered recorded voice is determined using the spreading string used to generate the spread bit string.

When recording in real environment, there is a problem that delayed waves are generated by reflecting on walls and obstacles. To solve this problem, the peak of the preceding and delayed waves are synthesized by rake combining, and then BPSK demodulation is performed to restore the bit string.

4 Experiment

In this paper, we embedded bit strings in synthesized speech using the method proposed in the previous section, and conducted transmission and reception in a real environment. In the experiment, the performance of Bit Error Rate (BER) is evaluated by changing the Signal-to-Noise Ratio (SNR) and the distance between robots.

4.1 Experimental Settings

The experiment was conducted at an indoor arena of Robotics Institute for the Next Generation (RING), Aichi Prefectural University. The environmental noise magnitude during the experiment was about $40 \sim 45$ [dB], and only the operating noise of the ventilation fan was audible.

The NAO V6 (Fig. 6 and 7) manufactured by SoftBank Robotics was used to transmit and receive audio signals. The specifications of NAO V6 is shown in Table 1 [8]. When transmitting, audio signal was output from both right and left speakers. When restoring the bit string, only the information from one of

Table 1. NAO V6 Specifications

Item	Details
OS	NAOqi 2.8
hight	57.4 [cm]
speaker	two speakers (right and left of the head)
microphone	four directional microphones

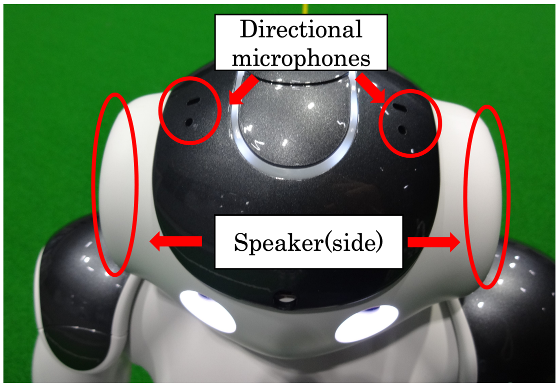

Fig. 6. NAO's frontal region

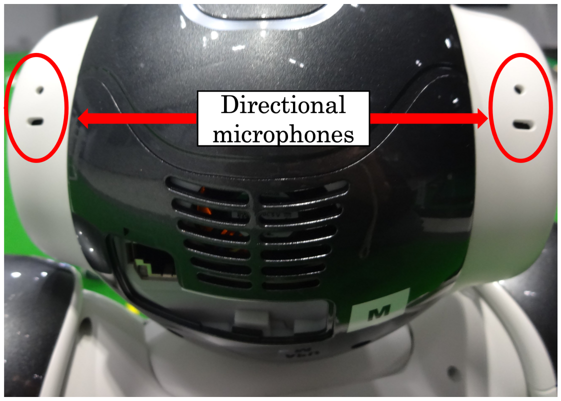

Fig. 7. NAO's occipital region

four microphones was used. The sampling frequency at the time of transmission and reception was set to 48,000 [Hz], which can be recorded in NAO.
The distance between the NAO transmitting speech and the NAO receiving speech was measured in 6 patterns of 1.0, 3.0, 5.0, 7.0, 9.0, and 10.8 [m]. The NAO was arranged to face each other as shown in Fig. 8. The NAO on the transmitting is upright and the NAO on the receiving is squatting down. The size of the field currently used in the SPL is 6.0 [m] × 9.0 [m]. Therefore, if it is possible to transmit information without errors up to 10.8 [m], it can be said that the field was sufficiently covered when actually using the technology.

4.2 Experimental Speech Generation

First, we generated a speech to be transmitted and received to conduct the experiment. The synthesized speech is generated as shown in Table 2. The vocal tract source was generated by applying a second-order filter with a resonant

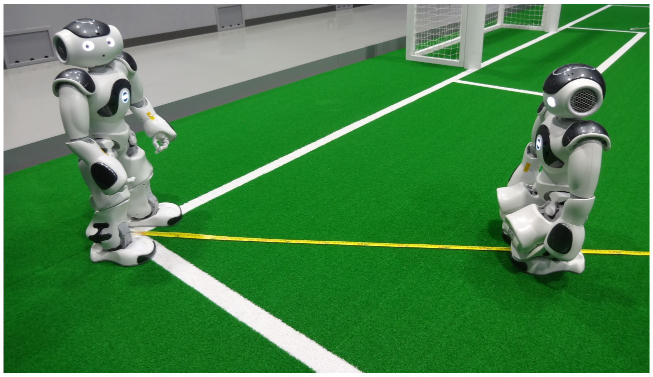

Fig. 8. Arrangement of NAO during the experiment

Table 2. Generated synthesized speech

Item	Details
speech content	"ganbare ganbare" (It means "fight fight")
speech time	7.0 [sec] (The last 0.2 [sec] is a silence interval.)
sampling frequency	48,000 [Hz]
fundamental frequency	250 [Hz]

frequency of 1,000 [Hz] to a sound source with a fundamental frequency of 250 [Hz], and then random noise was added. For the vocal tract information, linear prediction coefficients extracted by linear prediction analysis (20th order) were used and filtered to a vocal tract source in units of frames (frame length: 1,024 points). The spreading bit string is created as shown in Table 3. For the spreading sequence, we used a sequence in which 1 and -1 appeared randomly.

We generated a spreading bit string as shown in Table 3. For the spread string, we used random strings in which 1 and -1 appear randomly.

Table 3. Generated spreading bits

Item	Details
transmitted bit string	$\{1, 1, 0, 1, 0, 0, 1\}$
transmission rate	1[bps]
spread string length	48,000/bit
modulation method	BPSK

If the synthesized speech and the spread bit string are added together, the spread bit string will be too loud. Therefore, the synthesized speech was weighted in 3 patterns (500, 900, and 1,600) and then normalized by adding to the spreading bit string. Table 4 shows the SNR of the synthesized speech multiplied by the weighting factor and the spread bit string. The SNR is calculated using the spread bit string as the signal and the synthesized speech as the noise component. According to Table 4, as the weighting factor increases, the SNR decreases. As a result, a noise of the generated speech is reduced as the weighting factor increases. This makes the synthesized speech easier to hear as the weighting factor increases. when it is heard by a human

Table 4. Weight factor and SNR correspondence

Weight factor	SNR[dB]
500	-40.8
900	-45.9
1,600	-50.9

We added each weight to the synthesized speech and observed the BER when it was transmitted and received. Although 7 bits are transmitted in total, since the first bit is used as test data to determine the channel coefficient, the bit error rate is the error rate in 6 bits.

4.3 Restoring Bit String

The NAO was used for speech transmission and reception, but due to the high computational cost, a computer was used to restore the bit string. The high-pass filter processed the received speech under the conditions shown in Table 5. We used autocorrelation for the first second of the received speech as the channel coefficient for the rake combinig.

Table 5. High pass filter settings

Item	Details
software	MATLAB R2019b
function	highpass
bandpass frequency	4,000 [Hz]

4.4 Discussion

Figure 9 shows the transition of BER when the distance between robots is changed for each weighting factor of the synthesized speech. The BER is the average value obtained when five transmissions are performed for each combination of the weighting factor and the distance between robots. When the weighting factor is 500, the BER was 0 [%] for all the distances between robots.

In an experimental environment of environmental sound -45 ∼ -40 [dB], if the weighting factor was 500 (SNR = -40.8 [dB]), information could be transmitted

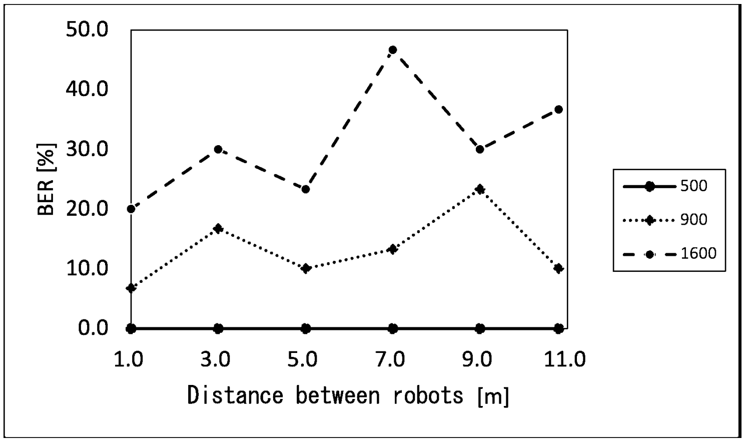

Fig. 9. BER corresponding to the distance between robots for each weight factor.

over the entire SPL field without error. However, since the SNR is -40.8 [dB] when the weighting factor is 500 and the strength of the hidden data is very weak, it is considered that three factors such as the high pass filter, the correlator, and the rake combining are the major contribution to the BER of 0 [%]. If the weighting factors were 900 and 1,600, since the speech magnitude of the spreading bit string would be too weak, speech signal cannot be separated from environmental sound.

When the weighting factor were 900 and 1,600, the error rate did not tend to increase even if the distance between the robots increased. Since the positional relation between the robot and the wall was always fixed in five transmissions, we plan to record the positional relationship between the robot and the wall by changing it in future work to investigate this tendency.

5 Conclusion

In this paper, we embedded a bit string of spread-spectrum bits in the synthesized speech and transmitted and received it in a real environment. As a result, we found that if the environmental sound is -45 ∼ -40 [dB] the speech with an SNR of -40.8 [dB] can transmit information over the entire SPL field without error. As future issues, we will consider how to cope with the situation where the environment noise is louder than and instantaneous noise can be mixed. We will also explore the limits of SNR that are comfortable for the spectators and ways to increase transmission speed.

In the experiment, the recording was performed without moving the robot. However, we should carry out the experiment even in the situation in which the robot moves in order to close to the real RoboCup game.

Acknowledgements

We would like to thank Assoc. Prof. Yukihiro Kamiya for his valuable advices on the spread spectrum method, and the members of Camellia Dragons, which is an AI × Robotics Project Team in Aichi Prefectural University for their wonderful discussions.

References

1. Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., and Osawa, E.: RoboCup: The Robot World Cup Initiative. 1995. `https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.7511`
2. RoboCup Soccer Standard Platform League. `https://spl.robocup.org/` (Last accessed on August 9, 2021).
3. Kamiya, Y.: Digital Wireless Communication Technology Using MATLAB. Corona Publishing Co.,Ltd., 2008 (in Japanese).
4. Kamiya, Y.: RF World No.31: Special Issue on Spread Spectrum Technology (Section 6). CQ Publishing Co.,Ltd., pp.46-56, 2015 (in Japanese).
5. Takehana, S., Kondo, K., and Nakagawa, K.: A Basic Study of Watermarking for Acoustic Signals Using Spread Spectrum. Journal of Tohoku-section Joint Convention of Institutes of Electrical Engineering, pp.344, 2003 (in Japanese).
6. H Banno, I: How to Master Spectrum Analysis of Speech. Journal of the Acoustical Society of Japan, Vol.68, No.4, pp.188-194, 2012 (in Japanese).
7. Itahashi, S.: Speech Engineering. Morikita Publishing Co.,Ltd., 2005 (in Japanese).
8. SoftBank Robotics: NAO - Documentation — Aldebaran 2.8.6.23e documentation. `http://doc.aldebaran.com/2-8/home_nao.html` (Last accessed on August 9, 2021).