applied sciences

MDPI

Article

Generating Robotic Speech Prosody for Human Robot Interaction: A Preliminary Study

Jaeryoung Lee

Department of Robotic Science and Technology, Chubu University, Kasugai 487-8501, Japan; jaeryounglee@isc.chubu.ac.jp

Abstract: The use of affective speech in robotic applications has increased in recent years, especially regarding the developments or studies of emotional prosody for a specific group of people. The current work proposes a prosody-based communication system that considers the limited parameters found in speech recognition for the elderly, for example. This work explored what types of voices were more effective for understanding presented information, and if the affects of robot voices reflected on the emotional states of listeners. By using functions of a small humanoid robot, two different experiments conducted to find out comprehension level and the affective reflection respectively. University students participated in both tests. The results showed that affective voices helped the users understand the information, as well as that they felt corresponding negative emotions in conversations with negative voices.

Keywords: speech prosody; human robot interaction; emotions; elderly care; social robot

check for updates

Citation: Lee, J. Generating Robotic Speech Prosody for Human Robot Interaction: A Preliminary Study. *Appl. Sci.* **2021**, *11*, 3468. https://doi.org/10.3390/app11083468

Academic Editor: Keun-Chang Kwak

Received: 28 February 2021
Accepted: 10 April 2021
Published: 13 April 2021

1. Introduction

The use of robots for applications in the daily lives of elderly people has increased in recent years for entertainment [1], security [2], and healthcare [3]. The development of assistive systems involving robots that interact with elderly people has the potential to increase their quality of life [4], and studies have already shown positive results [5,6]. For example, interacting with robots in the homes or care facilities of the elderly can help alleviate depression [7]. While numerous studies focus on elderly support [8,9], several aspects still lack sufficient study to provide proper human–robot interaction for target users. Verbal communication is challenging in human–robot interaction [10], and it is a crucial aspect of natural interaction [11], particularly for supporting elderly people, as non-verbal communication is insufficient to meet their preferences [12]. For elderly care, the services provided by robots are generally initiated with a conversation that requires several social aspects, including naturalness [11], an understanding of interpersonal relationships, and cultural backgrounds [13]. Ishi et al. investigated attention-drawing speech and argued that the way that attention is drawn through speech includes some cultural background [14]. Most potential scenarios for elderly support by robots include applications for reminding users about medicine or providing weather information. However, the ability of the robot to gain attention from the user is a crucial aspect for the interaction be successful. A robotic system with flaws in drawing attention from the user cannot be properly implemented in a support context.

1.1. Voice via Intelligent Systems

Recent studies have shown that people prefer to communicate with robots using voice [15]. Synthetic speech has been developed for natural communication between humans and robots [16], and, by applying emotion to synthetic speech, the conversation can be made much more natural [17]. Additionally, voice signals can transmit emotional information by vocal prosody, which is an important aspect regarding the communication between humans and intelligent systems [18]. Prosody is the part of linguistics that studies the intonation, rhythm, and accent (intensity, height, duration) of spoken language and other related attributes in speech. Prosody describes all acoustic properties of speech that cannot be predicted by orthographic transcription [10]. Natural language recognition has been handled as an important issue to control the humanoid robots in human robot interaction areas [19].

Although the use of emotional speech in applications of robotics, i.e., the ability of intelligent systems to generate prosodic vocal communication, can lead to the improvement of system performance and reduce training time [18], there are few approaches that focus on the development or study of emotional prosody for a specific group of people. This aspect, which is lacking in development, could be important because we may use voice tones to different degrees among different individuals, depending on social context or relationships [20]. Furthermore, there is a significant difference in perception for the elderly regarding the pitch and timing in speech when compared to younger individuals. (Here, the pitch represents the perceptual part of sound related with the frequency. Timing means the speed and pauses in the speech. The detailed explanation is described in Section 2). A system intended to interact with the elderly using prosodic vocal communication should take these parameters into account.

1.2. Novel Method for Specific Group of People

Because of the problems of generating prosodic vocal speech that is perceptible to listening limitations of the elderly, the present work proposes a method capable of generating verbal communication, while accounting for the pitch and timing perception of the elderly. The method is based on the Hidden Markov Model (HMM), but it uses correlation analysis between the raw voice present in the database with its correspondent emotional characteristics represented by a markup language (such as SSML or EmotionML). The correlation works as a classifier, clustering emotional information and voices with weak pitch contours. A filter is run on the selected data to increase the pitch contour but maintains the original emotional characteristics. The filtered data, along with the other voices with strong pitch contours, are used as input data to calibrate the HMM parameters. The output of the HMM should provide prosodic vocal speech without weak pitch contours, which are difficult to be identified by the elderly.

1.3. Contributions and Paper Overview

The study raised this issue because, as the rapid changes of technology in society occurs, the users (e.g., elderly people) have difficulties in using the new devices, and those using specific support (e.g., healthcare robots) also need more effective functions of technologies. This paper presents a study aimed at exploration of an adequate way and structure of robotic speech during human robot interaction. To find them, two independent experiments are conducted in the context of interacting with social robot. As this work is a preliminary study, the participants are recruited as the university students instead of age diverse groups. This study provides insights into the better understanding of robotic actions, while the human robot interaction works in our real life in terms of robotic prosody and the structure of interaction. To find the solution, the comprehension level and the understanding the prosody are explored during the verbal communication between human and robot by the questionnaires. In addition, those are assessed when users are in more real-life interaction, if they feel the emotions of robotic voice, and are reflected, and those are evaluated by emotional recognition tool.

The remainder of this paper is organized as the follows. Section 2 addresses the literature reviews in terms of speech features related to speech prosody and robot. Section 3 describes the experiments to explore the prosody with different voices of robot and interpretation of the emotional intent. Section 4 indicates the experimental results. Section 5 discusses the limitations and future work. Section 6 provides the conclusion of this work.

2. Literature Reviews on Speech Prosody and Robot

To express emotion using prosodic vocal speech, it is important to analyze which features present in direct verbal communication will be generated by the model. Machine learning techniques can provide a better understanding regarding which features of speech signals can be used to classify emotional speech [10]. Pitch, timing (speaking rate in Reference [21]), and intensity (volume in Reference [21]) are the features of speech that are usually related to the uses of emotion through vocal prosody. The literature indicates the big three of vocal prosody, namely pitch, timing, and loudness [22]. Pitch refers to how the human ear perceives the fundamental frequency of the sounds. Low frequencies are perceived as bass sounds and the highest are perceived as treble sounds, although human speech is more complex than a signal composed by one particular periodic sinusoid. Timing is related to speed of the speech and the pauses found between words. The speed of someone speaking is typically measured in words per second, whereas the pauses can be measured in seconds. Intensity is related to how loud or quiet a person's voice sounds, and can be seen as the power level of the voice [10]. Specific combinations of these three factors can represent some group of emotions. For example, rapid speech can indicate that the speaker is happy or angry. A slow rate of speech typically indicates that the speaker is sad [10]. The specific case of dynamic pitch, also known as intonation, has an important role in providing information regarding emotion [23], and pitch can also be used to infer interpersonal relationships between individuals [21]. Changes in dynamic pitch alone can express the degrees of friendliness and authority, for example [23].

Several robots with different way of speech have been suggested in recent studies. Crumpton and Bethel explored the vocal prosody of a speech synthesizer called MARY which was embedded in a robot used in an interactive experiment [24]. In a series of interactions with two NAO robots, it was observed that children display more expressions when interacting with a robot that displays emotion and adapts its expressions [25]. In this specific experiment, the voice of the robot was influenced by its arousal, where the robot would speak louder under higher arousal. Some studies also mention the affect used in conversation. Ito et al. revealed that the feeling of a robot's speech being natural and comfortable varies depending on the user [16].

In case of the robots with verbal communication for specific uses (e.g., elderly care), it is challenging. The communication between robots and elderly people can be difficult as a result of being too vague because elderly people are likely to forget the terms used by a robot and may not use precise delivery in their speech [26]. Studies have shown that the elderly tend to have less ability to process dynamic pitch [23], particularly when dealing with weak pitch contours with rising pitch patterns. Moreover, these studies suggest that this problem may be prominent when the dynamic pitch is carried by natural speech and when the pitch contour is not strong [23]. Making changes in speech to highlight the pitch contours could improve the ability of the elderly to perceive speech accurately when interacting with robotic systems.

3. Method

Sections 3.3.1 and 3.3.2 will describe two different experiments, as shown in Figure 1. Both experiments used robot and measured the emotional states of participants. The different parts are the shape of interaction and the method of measuring the affect of humans. In Section 3.3.1 (experiment A), the robot acts as the talker and human is listener. Emotional states are measured by the questionnaire. The other experiment in Section 3.3.2 (experiment B) is conducted wherein both robot and human perform as talker and listener in a natural interaction in real life. The emotional states recognition system is used as method of assessment.

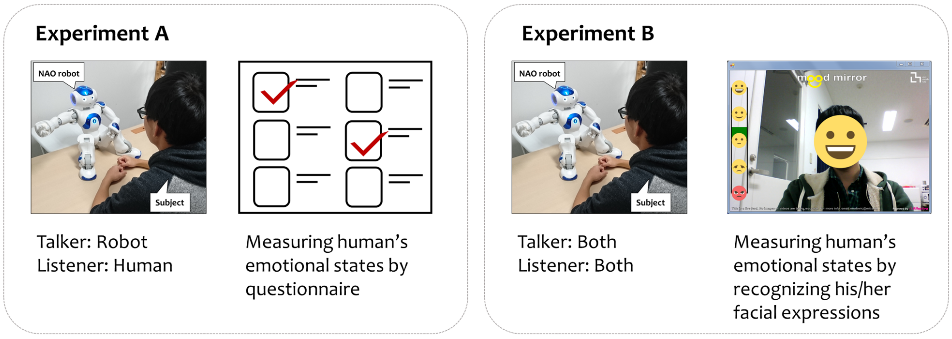
Experiment A
NAO robot
Subject
Talker: Robot
Listener: Human
Measuring human's emotional states by questionnaire
Experiment B
NAO robot
Subject
mood mirror
Talker: Both
Listener: Both
Measuring human's emotional states by recognizing his/her facial expressions

Figure 1. The experiments of Section 3.3.1 (experiment A) and Section 3.3.2 (experiment B). The robot mainly talks, and human subjects try understanding what it says in experiment A, while robot and human have a conversation in the other experiment. The tools for measuring the emotional states are different in the two experiments: questionnaire in experiment A and autonomous recognition in experiment B.

3.1. Hypothesis

In previous study, voice has had significantly more impact on the ability for listeners to infer feelings relative to the contents of conversation (i.e., the words used) [27]. Kitayama addressed that East Asian cultures have high-context language, where contextual and non-verbal cues are much more important than the contents of a conversation, and that the affect of a speaker's voice has a significant effect on the conversation [28]. Based on the above studies, we hypothesize as follows: **H1**—Participants will show better understanding when they talk to a robot that has either a positive or negative voice than one with a neutral voice. **H2**—The affect of a robot's voice will reflect on the emotional states of listeners.

3.2. Materials

A small humanoid robot NAO was used in the experiment (see Figure 1). The robot has a main software for controlling, called NAOqi. It is a programming framework and capable of commuicating between multi modules. The *ALTextToSpeech* module was used to command the NAO to speak by using the text-to-speech engine, and the value for pitch in NAOqi was coded as *vct=value*. NAO uses the AlTalk (This software is powered by AI Japan: https://www.ai-j.jp/english.) speech synthesizer for Japanese language and has values ranging from 50 to 200. To transform the values into pitch Hertz values, the Equation (1) was used. Here, P_{value} is the value of NAO pitch in NAOqi, and P_{pitch} is the pitch in Hertz. To define Equation (1), NAO speaking was recorded and converted using WaveSurfer (WaveSurfer was developed at KTH in Stockholm, Sweden).

$$P_{pitch} = 2.864 P_{value} + 4.6727. \tag{1}$$

The pitch values were taken referring to the values of *FO* and *mel scale* from the study of Suzuki and Tamura [29] as Equation (2) and were applied to three different voices: neutral, joy, and sadness, as in Table 1.

$$mel = (1000/log2)log(f/1000+1). \tag{2}$$

Table 1. The pitch, timing, and loudness values in each session.

	Neutral	Positive (Joy)	Negative (Sadness)
Pitch (Hz)	292	314	251
Timing (%)	100	100	100
Loudness (%)	100	100	100

The robotic voices in both experiments were validated. We have extracted the features by using openSMILE [30]. A part of IS09-emotion.conf [31] was used for the feature set. The original IS09-emotion.conf has 384 features that contained 32 speech and music-related features (Signal energy, Loudness, Mel-/Bark-/Octave-spectra, MFCC, PLP-CC, Pitch, Voice quality, Formants, LPC, Line Spectral Pairs, Spectral Shape descriptors), and 12 statistical functionals (Means, Extremes, Moments, Segments, Samples, Peaks, Linear and quadratic regression, Percentiles, Durations, Onsets, DCT coefficients, Zero-crossings, Modulation spectrum), but, here, we used the most important 11 features: △RMSenergy-max, △ MFCC[9]-linregc2, RMSenergy-linregc2, △voiceProb-maxPos, MFCC[1]-linregc2, △ MFCC[12]-linregc2, △ MFCC[10]-max, △ MFCC[4]-linregc1, △ MFCC[10]-min, ZCR-amean, and △ MFCC[8]-linregc1. These features were extracted through the FSS (Forward Stepwise Selection) method used by Nhat el al. [32]. After extracting the features, we trained the robotic voice dataset by using the Weka software, which is a machine learning workbench [33]. For classification, we used sequential minimal optimization (SMO) [34]. k-fold cross validation was used for the validation method in Weka. The value k in the dataset used in the experiment A was 13, and that for the dataset used in the experiment B was 10.

In both datasets (datasets from the experiment A and B), the distribution is highly skewed, which means that the neutral emotion resulted in higher recognition. As shown in Figure 2, when the robot spoke about the weather forecast with some natural small talk, this provided a better prediction result (Figure 2B) than when the robot only talked about the weather forecast(Figure 2A). In case of Negative prediction, both speeches moderately same results. The result in Table 2 also indicates that the robot speech with small talk showed better recognition than only delivering information. As shown in Figure 2, neutral was predicted higher than other two emotions in both datasets.

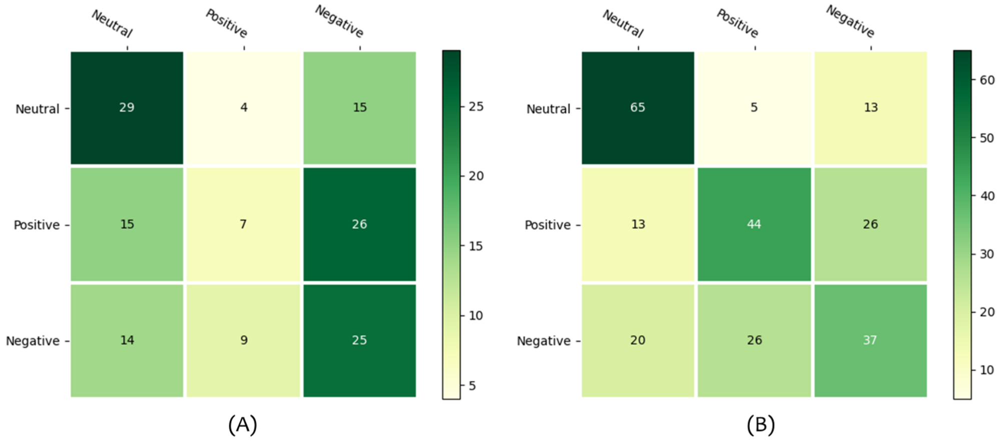

Figure 2. Prediction results of each dataset; (**A**) dataset of only the weather forecast, (**B**) dataset of forecast with both the weather forecast and small talk.

Table 2. F-measure of both datasets. Dataset in the experiment of Section 3.3.1 includes only the weather forecast. Dataset in the experiment of Section 3.3.2 includes both the weather forecast and small talk.

	Dataset from Section 3.3.1			**Dataset from Section 3.3.2**		
	Precision	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
Neutral	0.500	0.604	0.547	0.663	0.783	0.718
Positive	0.350	0.146	0.206	0.587	0.530	0.557
Negative	0.410	0.424	0.397	0.487	0.446	0.465

3.3. Protocols

3.3.1. Experiment A

We recruited the participants in the University, and 21 undergraduate students (1 female), aged from 20 to 22 (M:21.10, SD: 1.41), participated in the 5-min interaction session. The data was collected in accordance with the Chubu University ethical guidelines for research (approved number of study on contacting with NAO: 270099). Informed consent was obtained from all subjects involved in the study. All students are familiar enough with NAO robots to avoid the apprehension of a new environment. Each participant entered a room with a NAO placed on a table and was instructed to sit on a chair at the table. The subject enters the experiment room to find the robot called NAO sitting on a table. The user sits on the chair, and the experiment starts with the experimenter explaining the session, saying "This robot is called NAO, and it will tell you some stories. You can note the stories by using this pen and paper." The participants were allowed to note the information from NAO in order to avoid the additional burden of memorization. The participants engaged in 3 sessions: (1) neutral(default) session, (2) positive(joy) session, (3) negative(sadness) session. All participants started from the neutral session, and subsequent positive and negative sessions proceeded in a randomized order.

The pitch of the NAO speech was varied. The neutral voice of NAO used the default value of 291.58 Hz. The pitch values of joy and sadness were 314.46 Hz and 251.36 Hz, respectively, as noted in Table 1.

The contents of the NAO speech were as follows:

> *I will tell you tomorrow's weather. Hokkaido will be snowy due to low pressure. Tohoku will be snowy due to cold weather. Kanto will be cloudy with the influence of cold. Chubu will be covered with high pressure and it will be sunny. Kinki will be rainy due to the influence of an atmospheric pressure valley. Chugoku will be rainy due to cold air. Shikoku will be covered with high pressure and it will be sunny. Kyushu will be rainy due to moist air. Okinawa will be cloudy with the influence of moist air. It was the weather forecast.*

The information given was different in each session in order to avoid the learning effect. After listening to NAO, participants answered a questionnaire that included questions, such as "What is the weather of Hokkaido?", "Was NAO's speech easy to understand? If it wasn't, why?", and "Does it sound like emotional speech? If it does, what kinds of emotions did it sound like?" A yes/no and open-ended questions were employed. This test was to see if there was a difference in the understanding of the participants between positive and negative sessions, and whether the default value could be used as the neutral voice of NAO.

3.3.2. Experiment B

In Section 3.3.1, the contents of NAO speech were only about the weather forecast and focused on delivering information to the users. Considering the applications of this study and the target users being elderly people, we applied realistic scenarios to the contents of the conversation by adding greetings, self-introduction, small talk, and non-verbal cues [35,36] apart from the weather forecast, as shown in Figure 3.

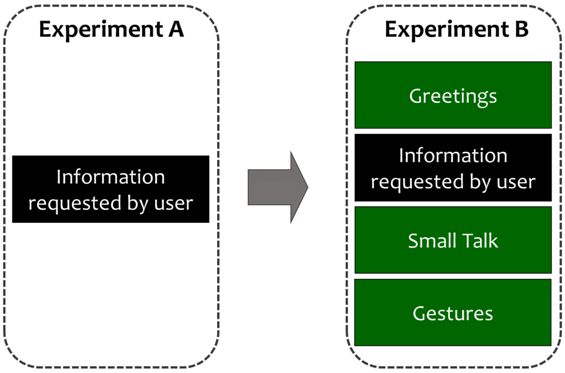

Figure 3. The contents of interactions with NAO in experiment A (Section 3.3.1) and experiment B (Section 3.3.2).

The interaction time is approximately 5 min, and, by adding longer pause between the sentences, NAO speaks more slowly than in the experiment in Section 3.3.1. An emotion detection software called Mood mirror (Mood mirror was developed by Dr. Ognjen Rudovic of MIT Media Lab) was used to recognize if the participants felt emotions (anger, sadness, fear, joy). This software employed the validated toolkit, affectiva SDK, for recognizing the facial expressions of humans [37]. This toolkit is one of the most used toolkits that is a cross-platform to recognize multi-face expression in real-time by using the facial action coding system (FACS). It has been showing a high accuracy and reliable system for several application [38–40]. The length of time for each recognized emotions was calculated and analyzed to determine if the emotions presented in the robot's voice reflected on the emotional states of users. Note that the loudness was 80% instead of 100% because a number of participants commented the voice of robot was too loud considering the physical distance between them in the experiment of Section 3.3.1. However, this revision did not influence on the studies because the same number was used to all sessions (neutral voice session, positive and negative ones).

Three participants (all males, aged 22) took part in interaction sessions in which NAO spoke in three different voices (i.e., neutral, positive, and negative voices). In these sessions, NAO started the conversation with small talk for ice-breaking and natural conversation. Each participant engaged in three sessions and was recorded by a small web-camera in front of the NAO. The video data were used for analyzing the emotional states of the participants. The participants interacted with the NAO robot for approximately 5 min, mainly by verbal communication, as follows:

> *Hello. My name is Nao. Can you tell me your name? I am a humanoid robot. Have you ever seen me somewhere? Recently, there are some friends who started working at the hotel. There are some friends playing soccer. Don't you think that this figure looks like a Pepper robot? Actually, I was built by the same company with Pepper robot. Pepper is like a brother. Pepper is also the same robot as I am talking to everyone in various places, such as a company and nursing home. I also like chatting with everyone. That's it. I want to know more about you. Now, I will ask you a couple of questions, so please answer to my questions. Thank you. Well then, what's your favorite color? (pause a little bit) I like that color! In addition, I like blue. I think that is cool. Alright, I will ask another question. What's your favorite food? (pause a little bit) Alright. I have not eaten yet, but I'd like to eat hot pot. By the way, I saw the weather forecast for tomorrow on today's news. Let tell you. The weather tomorrow is cloudy, the probability of precipitation is 30 percent. If you go out, I think it would be better to have an umbrella. In addition, it seems to be tomorrow, so it is better to keep your cold measures steady. It is about time to have the questionnaire answered. Today, thank you for chatting with me. Then, please answer the questionnaire from now.*

After interacting with NAO in each session, participants answered a questionnaire with questions, such as "Was it easy to understand?" and "Did you feel the emotions of robot?"

4. Results

4.1. Experiment A

The rate of correct answers in the sessions corresponding to the neutral, positive, and negative voices are given in Figure 4. The participants showed a better understanding when talking to NAO with either a positive (314 Hz) or negative voice (251 Hz) than with a neutral voice (292 Hz). The percentage of correct answer was simply calculated that the total number of questions divided the number of correct answers. The result was 60.48% when the participants listened to the neutral voice of robot. 88.57% was indicated for interacting with the positive voice. During listening to the robot with negative voice, the percentage of correct answer was 93.81%. The results showed a statistically significant difference between groups as determined by one-way ANOVA (F(2,60) = 14.681, $p < 0.001$). A Tukey post-hoc test revealed that the listening score with the neutral voice was significantly lower than that with the positive voice ($p = 0.001$) or negative voice ($p < 0.001$).

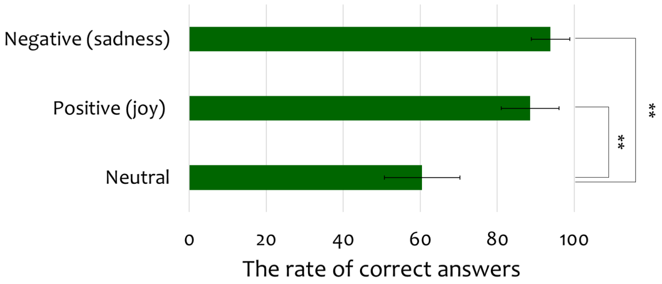

Figure 4. The average percentage of correct answers to the questionnaire. The rate in session interacting with neutral voice of robot showed 60.48%; the one with positive voice was 88.75%; the one with negative voice was 93.81%; ** means $p < 0.001$.

The participants answered the questionnaires, and the populations are indicated in Figure 5. More than 70% of the participants in all sessions responded that the NAO speech was easy to understand. Less than 40% of the populations felt the emotional feeling during their conversions.

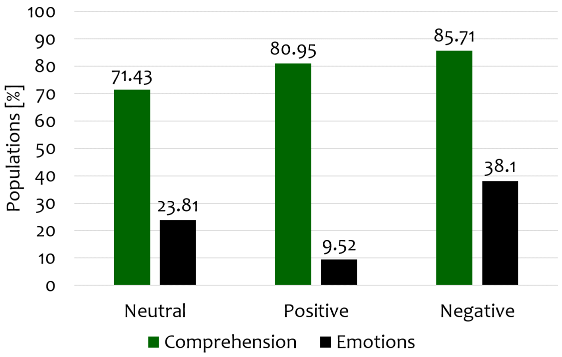

Figure 5. The population of positive answers to understanding and emotion in each robotic speech. The green bar shows whether it is understandable, and the black one indicated the feeling of emotion into the robot.

4.2. Experiment B

All participants correctly answered 100% of the weather information questions.

As shown in Figure 6 and Table 3, subject 1 showed sadness for 56% of the time and joy for 44% of the time when interacting with NAO using the neutral voice, and negative faces for more than 95% of the time using emotional voices (positive and negative). The result for subject 2 indicated negative faces for more than 70% of the time in prosody sessions. Subject 3 did not make any joy faces during any of the sessions. Overall, the participants mostly reflected negative emotions during the sessions.

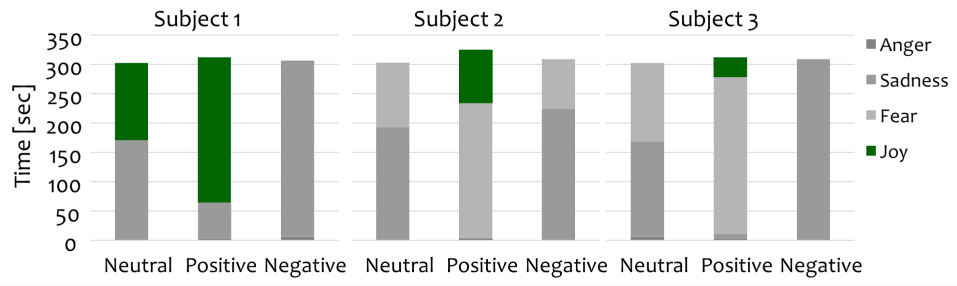

Figure 6. Durations of the subjects' emotional states.

Table 3. The Negative face (anger, sadness, fear) and Positive face (joy) time (seconds) of each participant.

Subject No.	Anger	Sadness	Fear	Joy
Subject 1	2.246	175.426	81.646	43.972
Subject 2	0	22.494	166.242	124.071
Subject 3	2.277	277.787	28.271	0

Figure 7 showed a significant result that communication with a negative voice reflects a negative emotion in the listeners. Table 4 shows average times for Negative and Positive faces for each voice.

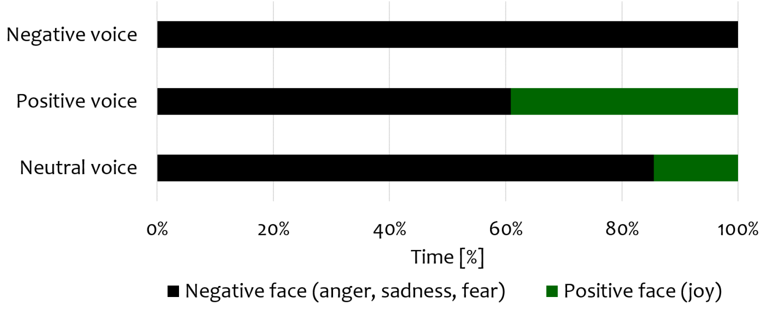

Figure 7. The average times of three participants for Negative face (anger, sadness, fear) and Positive face (joy).

Table 4. The average times (seconds) of three participants for Negative face (anger, sadness, fear) and Positive face (joy).

	Negative Face	Positive Face
Neutral voice	259	44
Positive voice	193	124
Negative voice	308	0

5. Discussion

Considering the importance of prosodic vocal speech in robotic applications, and based on the low capability of elderly people to perceive weak pitch contours, this work suggests a method that is able to filter information into a more perceivable pitch without changing its original emotional characteristics, as well as presents the results of preliminary experiments regarding the prosodic vocal speech of robots. This study generated robotic speech for two cases; in one case, only information is delivered, and, in the other case, certain information is delivered in addition to natural small talk. Three emotions were used in the generated voices, and we have validated the robotic voices. The results showed that the speech with small talk provided better predictions. The user-study in this paper introduced two experimental results for investigating the impact of prosody in human robot interactions. The result of experiment A showed that users understood the information better when NAO spoke with a positive or negative voice than with a neutral voice, which supported **H1**. The second result, given in experiment A, implied that the information can be easy to understand with all voices, but did not reflect emotions to the users, and thus **H2** was not proved by the experiment. Another result described in experiment B partially supported **H2**, where the negative voice reflected on the emotional states of listeners. **H1** was not revealed in the experiment. One of limitations in this study is the small sample size. The experiment B had three participants, which is insufficient to statistically validate the result. Moreover, as a preliminary study, the subjects in experiment A and B were only university students, and it will be necessary to collect data using larger age range of people as participants.

The findings in this paper could be applied to a speech prosody model in the future. To alter the pitch contour to be perceived more easily by the elderly while still maintaining the original emotional characteristics, first, a voice database is used, such as that used in Reference [41]. Each set of the database should be related to an emotional markup language. This can be done using information clustering as proposed in Reference [22] or by manual tagging. In this way, each piece of data is related to a specific emotional state. Although this structure is presented as a combination of two sets of information, it can also be implemented as one dataset of emotional speech. The basic idea for the system is to find a set of prosody in a voice database that retains the emotional value with voice parameters under the limitation presented by elderly users. This specific group of elderly-oriented speech has its characteristics used by a filter in order to create new and equally successful speech prosody, even with the above cited limitations in the voice characteristics. The voice information is sent to a classifier, which will correlate the speech features according to emotional markup information. Component analysis is used as a classifier technique to separate and analyze aspects inside the original data that are difficult to observe with normal correlation or density analysis. The classifier clusters the voices present in the database according to its recognition level, taking into consideration the hearing limitation presented by the elderly. It is also possible to use different degrees of limitations. This clustered information is used by a filter which learns what kinds of changes are necessary for the voice parameters to express a specific input motion and text. The filter component is based on an HMM speech synthesizer, which will create prosody with strong pitch contours while maintaining the original emotional characteristics. During the training stage of the filter, this process can be performed several times, changing the characteristics of the classifier and the filter in order to achieve a final desired speech output. The optimum values and details for the filter are under development. Moreover, as future works, we intend to evaluate the method comparing the engagement of elderly speech when talking to a robot in two different scenarios, one using a normal speech synthesizer and the other one using the proposed method. It is believed that such analysis will contribute to a deeper understating of which speech features can be better used for robot interaction with the elderly.

6. Conclusions

This work tested the effect of robotic speech prosody during human robot interaction. Although the work in this paper indicated the experimental results from university students, if considering the specific situations, such as elderly care by social robots, the natural and more adequate interaction is required for successful human robot interaction. In this study, it has been shown that personalized speech prosody is needed on human robot interaction when the robot interacts with elderly people because it is a new and complicated device to use for them. The findings of this study showed that the speech prosody of robot helped people understand the information provided by the robot and, particularly, negative voices of the robot emotionally reflected to the listener. These results could lead to the conclusion that the robot has to avoid the negative prosody when it is used for elderly support. Taken together, the questions raised and the results of this work show that the speech prosody of robot should be more studied for effective human robot interaction, and the findings of this study will increase the research in that field.

Funding: The part of this work was supported by the EU and Japan coordinated R&D project on "Culture Aware Robots and Environmental Sensor Systems for Elderly Support" commissioned by the Ministry of Internal Affairs and Communications of Japan and EC Horizon 2020 No 737858. This work was also supported in part by Chubu University research grant.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of Chubu University (270099).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: The author would like to thank Nicholas Melo, Ryusei Otani, Masayoshi Iwase for their research assistance.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Ahn, H.S.; Lee, M.H.; Broadbent, E.; MacDonald, B.A. Is Entertainment Services of a Healthcare Service Robot for Older People Useful to Young People? In Proceedings of the IEEE International Conference on Robotic Computing (IRC), Taichung, Taiwan, 10–12 April 2017; pp. 330–335.
2. Joh, E.E. Private Security Robots, Artificial Intelligence, and Deadly Force. *UCDL Rev.* **2017**, *51*, 569.
3. Matarić, M.J. Socially assistive robotics: Human augmentation versus automation. *Sci. Robot.* **2017**, *2*, eaam5410. [CrossRef]
4. Moyle, W.; Arnautovska, U.; Ownsworth, T.; Jones, C. Potential of telepresence robots to enhance social connectedness in older adults with dementia: an integrative review of feasibility. *Int. Psychogeriatr.* **2017**, *29*, 1951–1964. [CrossRef]
5. Cudd, P.; De Witte, L. Robots for Elderly Care: Their Level of Social Interactions and the Targeted End User. *Harnessing Power Technol. Improv. Lives* **2017**, *242*, 472.
6. Bedaf, S.M. The Future is Now: The Potential of Service Robots in Elderly Care. Ph.D. Thesis, Maastricht University, Maastricht, The Netherlands, 2017.
7. Wada, K.; Shibata, T.; Saito, T.; Tanie, K. Psychological and social effects of robot assisted activity to elderly people who stay at a health service facility for the aged. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'03), Barcelona, Spain, 18–22 April 2003; Volume 3, pp. 3996–4001.
8. Das, R.; Tuna, A.; Demirel, S.; Netas, A.S.K.; Yurdakul, M.K. A Survey on the Internet of Things Solutions for the Elderly and Disabled: Applications, Prospects, and Challenges. *Int. J. Comput. Netw. Appl. (IJCNA)* **2017**, *4*, 84–92. [CrossRef]
9. Lewis, L.; Metzler, T.; Cook, L. Evaluating Human-Robot Interaction Using a Robot Exercise Instructor at a Senior Living Community. In Proceedings of the International Conference on Intelligent Robotics and Applications (ICIRA 2016), Tokyo, Japan, 22–24 August 2016; pp. 15–25.
10. Crumpton, J.; Bethel, C.L. A survey of using vocal prosody to convey emotion in robot speech. *Int. J. Soc. Robot.* **2016**, *8*, 271–285. [CrossRef]
11. Christensen, H.I.; Okamura, A.; Mataric, M.; Kumar, V.; Hager, G.; Choset, H. Next generation robotics. *arXiv* **2016**, arXiv:1606.09205.
12. Hammer, S.; Kirchner, K.; André, E.; Lugrin, B. Touch or Talk: Comparing Social Robots and Tablet PCs for an Elderly Assistant Recommender System. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI 2017), Vienna, Austria, 6–9 March 2017; pp. 129–130.

13. Ishiguro, N. Care robots in Japanese elderly care. In *The Routledge Handbook of Social Care Work around the World*; Taylor & Francis Group: London, UK, 2017; p. 256.
14. Ishi, C.; Arai, J.; Hagita, N. Prosodic analysis of attention-drawing speech. In Proceedings of the 2017 Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 909–913.
15. Ray, C.; Mondada, F.; Siegwart, R. What do people expect from robots? In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008 (IROS 2008), Nice, France, 22–26 September 2008; pp. 3816–3821.
16. Ito, R.; Komatani, K.; Kawahara, T.; Okuno, H.G. Analysis and Detection of Emotional States in Spoken Dialogue with Robot. *Inf. Process. Soc. Jpn. SLP* **2003**, *2003*, 107–112. (In Japanese)
17. Kimura, H.; Tomita, Y.; Honda, S. Synthesis of emotional voice by changing the parameters in the characteristics of vocal cords and vocal tract. *Jpn. J. Ergon.* **1996**, *32*, 319–325. (In Japanese)
18. Mitchell, R.L.; Xu, Y. What is the Value of Embedding Artificial Emotional Prosody in Human–Computer Interactions? Implications for Theory and Design in Psychological Science. *Front. Psychol.* **2015**, *6*, 1750. [CrossRef] [PubMed]
19. Recupero, D.R.; Spiga, F. Knowledge acquisition from parsing natural language expressions for humanoid robot action commands. *Inf. Process. Manag.* **2020**, *57*, 102094. [CrossRef]
20. Pullin, G.; Cook, A. The value of visualizing tone of voice. *Logop. Phoniatr. Vocol.* **2013**, *38*, 105–114. [CrossRef] [PubMed]
21. Moriyama, T.; Mori, S.; Ozawa, S. A Synthesis Method of Emotional Speech Using Subspace Constraints in Prosody. *J. Inf. Process. Soc. Jpn.* **2009**, *50*, 1181–1191.
22. Vinciarelli, A.; Pantic, M.; Bourlard, H.; Pentland, A. Social signal processing: state-of-the-art and future perspectives of an emerging domain. In Proceedings of the 16th ACM International Conference on Multimedia, Vancouver, BC, Canada, 27–31 October 2008; pp. 1061–1070.
23. Clinard, C.G.; Cotter, C.M. Neural representation of dynamic frequency is degraded in older adults. *Hear. Res.* **2015**, *323*, 91–98. [CrossRef]
24. Crumpton, J.; Bethel, C.L. Validation of vocal prosody modifications to communicate emotion in robot speech. In Proceedings of the 2015 International Conference on Collaboration Technologies and Systems (CTS), Atlanta, GA, USA, 1–5 June 2015; pp. 39–46.
25. Tielman, M.; Neerincx, M.; Meyer, J.J.; Looije, R. Adaptive emotional expression in robot-child interaction. In Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, Bielefeld, Germany, 3–6 March 2014; pp. 407–414.
26. Tejima, N. Rehabilitation Robots for the Elderly-Trend and Futre. *J. JSPE* **1999**, *65*, 507–511. (In Japanese)
27. Sperber, D.; Wilson, D. Précis of relevance: Communication and cognition. *Behav. Brain Sci.* **1987**, *10*, 697–710. [CrossRef]
28. Kitayama, S.; Ishii, K. Word and voice: Spontaneous attention to emotional utterances in two languages. *Cogn. Emot.* **2002**, *16*, 29–59. [CrossRef]
29. Suzuki, T.; Tamura, N. Features of emotional voices: Focus in differences between expression and recognition. *Jpn. J. Psychol.* **2006**, *77*, 149–156. (In Japanese) [CrossRef]
30. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
31. Schuller, B.; Steidl, S.; Batliner, A. The interspeech 2009 emotion challenge. In Proceedings of the 10th Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009.
32. Nhat, T.B.; Mera, K.; Kurosawa, Y.; Takezawa, T. Natural Language Dialogue System considering Emotion: Guessed from Acoustic Features. In Proceedings of the Human-Agent Interaction Symposium 2014 (HAI'14), Tsukuba, Japan, 28–31 October 2014.
33. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [CrossRef]
34. Platt, J. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*; Microsoft: Redmond, WA, USA, 1998.
35. Hold, B.; Schleidt, M. The importance of human odour in non-verbal communication. *Ethology* **1977**, *43*, 225–238. [CrossRef] [PubMed]
36. Breazeal, C.; Kidd, C.D.; Thomaz, A.L.; Hoffman, G.; Berlin, M. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005.(IROS 2005), Edmonton, AB, Canada, 2–6 August 2005; pp. 708–713.
37. McDuff, D.; Mahmoud, A.; Mavadati, M.; Amr, M.; Turcot, J.; Kaliouby, R.E. AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, San Jose, CA, USA, 7–12 May 2016; pp. 3723–3726.
38. Magdin, M.; Benko, L.; Koprda, Š. A case study of facial emotion classification using affdex. *Sensors* **2019**, *19*, 2140. [CrossRef] [PubMed]
39. Lopez-Rincon, A. Emotion recognition using facial expressions in children using the NAO Robot. In Proceedings of the 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP 2019), Cholula, Mexico, 27 February–1 March 2019; pp. 146–153.
40. Dupré, D.; Krumhuber, E.G.; Küster, D.; McKeown, G.J. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PLoS ONE* **2020**, *15*, e0231968. [CrossRef] [PubMed]
41. Kominek, J.; Black, A.W. The CMU Arctic speech databases. In Proceedings of the Fifth ISCA Workshop on Speech Synthesis, Pittsburgh, PA, USA, 14–16 June 2004.