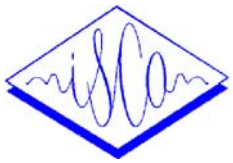


Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Spoken Dialogue System Using Prosody as Para-Linguistic Information

Shinya Fujie[1], *Daizo Yagi*[1], *Yosuke Matsusaka*[1], *Hideaki Kikuchi*[2] & *Tetsunori Kobayashi*[1]

[1]School of Science and Engineering, Waseda University, Japan
[2]School of Human Sciences, Waseda University, Japan
{fujie, yagi, yosuke, koba}@tk.elec.waseda.ac.jp, kikuchi@waseda.jp

10.21437/SpeechProsody.2004-89

Abstract

An attitude recognizer of a speaker which uses prosodic features of speech is proposed and it is successfully applied to the dialogue system aiming at agreement formation. We use not only linguistic information but also some sorts of additional information supporting linguistic information in our human communication. In agreement formation dialogues, we are often required to express our attitude (positive or negative) to conversational partners' proposals. We sometimes reply explicitly in linguistic information. We sometimes reply information ambiguously. However, even in the ambiguous case, we implicitly express our attitude using prosodic information. By realizing the abilities of catching these nuances, the dialogue system can be more sophisticated. In this paper, we implemented an attitude recognizer based on the GMM using prosodic feature parameters. The performance of the system is comparable to the human ability. We also realized a proto-type of spoken dialogue system using the recognizer. We show how these abilities contribute to efficient conversation.

1. Introduction

In a spoken dialogue-based human communication, we use not only linguistic information but also some sorts of additional information supporting linguistic information. We call these sorts of additional information "para-linguistic information."

In most conventional spoken dialogue systems a user and a system exchanges utterances in order, and the system tries to understand the user's intention with the linguistic-information in his/her utterance. In an actual spoken dialogue, however, humans exchange their intentions effectively with para-linguistic information along with linguistic information. We sometimes do not express our intention explicitly in words. Nevertheless, hearers can recognize our intention through our face or voice expressions. Therefore, it is indispensable that a natural spoken dialogue system has the ability to recognize para-linguistic information.

Delleart et al[1] recognized four emotions, such as "happy", "sad", "anger", and "fear", by prosody. In this work, 17 pitch features divided into 5 groups are introduced. The only 5 features selected in order by the performance of the cross-validation experiment, achieved better performance than the original features. The majority voting of specialists method, where the original feature space is divided into the small sub spaces and the results of the sub spaces are combined, marked better as well. McGilloway et al[2] gave a benchmark of automatic emotion recognition using features by ASSESS system by Cowie. Lee et al[3] recognized negative and non-negative emotions using prosody, in order to improve the quality of the service in call center applications. In this work a total of 10 features representing F0 and energy information are introduced. The features reduced the dimension (6 or 7) by PCA and feature selection(the same method with [1]) marked better than the original base features. Ang et al[4] used prosody for the detection of frustration and annoyance in natural human-computer dialogue. They used the corpus of human-computer dialogue developed under the DARPA Communicator Project. The prosodic features used in this work are durations, speaking rate, pause, pitch and so on. They introduced language model feature as well. For more literature of the works on emotion recognition, refer [5].

The keyword of the categories to recognize in these works is "Emotion." The recognition results are important for estimating user's "real" emotional state, but we think that the behavior of the users and the intended message included implicitly as para-linguistic information in utterance are more important for effective and smooth conversation. In addition, the real emotion is not intuitively applicable for the dialogue strategy.

In order to make human-machine conversation more effective and smooth, we incorporate some methods to recognize para-linguistic information and treat those as user's implicit messages. In this paper, we introduce a method for recognizing a user's attitude (positive or negative) from prosody. This information can be used as the user's evaluation against the system's suggestion. The spoken dialogue system that combines these sorts of information, including head gesture, and talks with human effectively is also presented.

The rest of the paper is organized as follows. Section 2 presents the methods and the experimental results of the attitude recognition by prosody. Section 3 presents how some sorts of para-linguistic information are combined and show the implemented spoken dialogue system.

2. Attitude Recognition Using Prosody

2.1. Target

The target is to recognize a user's attitude using para-linguistic information in the agreement formation dialogue, in which a user decides a plan through interaction with an advisor who suggests some plans. Please imagine the case that the advisor suggests a plan; for example he proposes a hamburger for lunch, "How about hamburger." Often the users do not express their evaluation explicitly. They simply repeat the advisors' words, say "Hamburger...". Linguistic information, the repeated words, itself do not include the users' evaluation (positive/negative) against the plan. However, the evaluation often appears in the expression of face or voice, that is called para-linguistic information. If this evaluation is automatically recognized by the system, it contributes to efficient and smooth conversation. The aim of this section is to give a method to

recognize users' evaluation (positive/negative) to the suggested plan using the prosodic information of the repeated words.

2.2. Data

We recorded the utterances of the users' responses to the system's suggestions. In order to collect a large number of utterances that include the positive or negative attitudes, the users' responses in the single turn (the system's suggestion and user's response) were recorded. The variety of the recorded utterances are seen in Table 1. In the PHRASE column of this table, the underlined part represents the repetition of the CATEGORY or RESTAURANT in the system's suggestion. The words "ka" and "ne" are post-positional particles of Japanese, which do not define positive or negative by themselves. The phrases "iinjanai" and "so-dane" are usually received as positive, but they can be received as negative depending on the use. The utterances of 20 students in our laboratory, total 2000 utterances, were recorded.

Table 1: *The variety of the recorded utterances.*

CATEGORY/RESTAURANT		PHRASE	ATTITUDE
Hamburger	McDonald's	_____ ka	positive
Noodles	Ajigen	_____ ne	negative
Lunch Box	Mumin	_____	
Curry	Hoka-ben	iinjanai	
Refectory	Soba-no-mi	so-dane	

For example, in the case that the CATEGORY is "Hamburger," the PHRASE is "____ ka" and the ATTITUDE is negative. First, the user hears the utterance "Hamburger nan te do kana", which means "What about a hamburger?". After that, the user utters "Hamburger ka" with negative attitude, and we record it.

2.3. Recognition method

We apply the fundamental frequency (F0) extraction and the phoneme alignment to the recorded utterances. As consideration of these results, we decided to use the following 3-dimensional feature $\boldsymbol{x} = (x_1, x_2, x_3)$ to distinguish the attitude.

x_1 : the gradient of F0 at the vowel part of the first mora in the utterance

x_2 : the range of F0 throughout the utterance

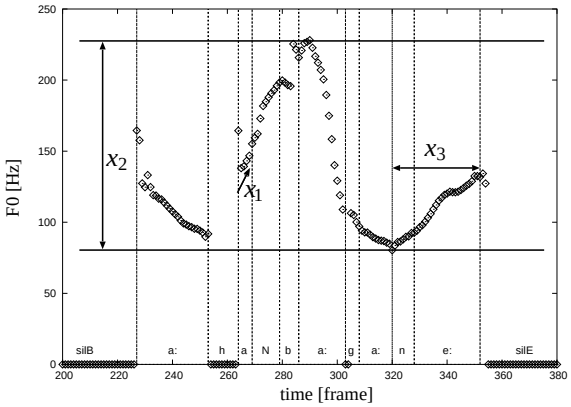
x_3 : the time length of the last mora in the utterance

x_1, the gradient is calculated with the least square method. The filler phrase at the beginning of the utterance, for example "ah", "un" and so forth, was accepted, so we ignored it in the feature extraction.

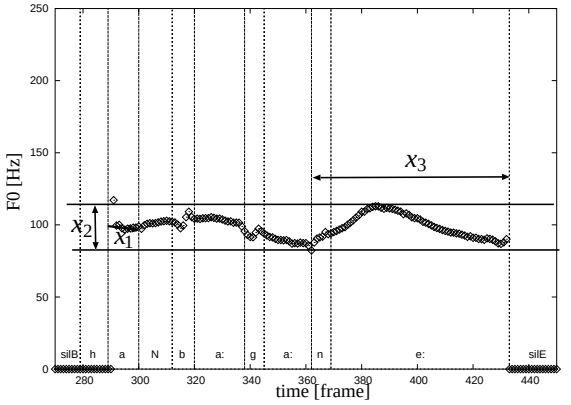
In Fig.1, examples of feature extraction are seen. In this figure, the following tendency of features can be seen.

- x_1 is a positive number if the attitude is positive, or a negative number otherwise.
- x_2 is larger if the attitude is positive than otherwise.
- x_3 is smaller if the attitude is positive than otherwise.

An utterance is recognized as positive or negative with Bayesian classifier.



(a) positive



(b) negative

Figure 1: *Examples of prosodic feature extraction.*

2.4. Experiment and results

The recorded utterances are divided into 4 sets. Each set contains 5 person's utterances, total 500 utterances. In one experiment process, we use 3 sets as a training set and the remainder as a test set. This process was performed for each set, total 4 times, and we calculated the average recognition rate of all the results.

In Fig.2, the experiment result of each person is shown.

We aimed to reduce the error caused by the differences between phrases and that between persons, by introducing Gaussian Mixture Model(GMM) as the probabilistic model. In Fig.3, the experiment result of GMM is seen.

In Fig.2, recognition rates are different for each person. While for some persons the positive result is better than the negative one, the negative result is better for others. In Fig.3, the maximum recognition rate (82.9%) is marked when the number of mixture is 16.

2.5. Comparison with recognitions by human

In the experiment in the previous section, the proper result is the attitude which the speaker implied when he spoke. In order to compare the recognition ability of our model with that of a human, we calculate a measure of agreement between results of humans' annotations and the experiment results.

Among the recorded utterances, we picked up 20 utterances from each category, positive and negative. 5 annotators (A–E)

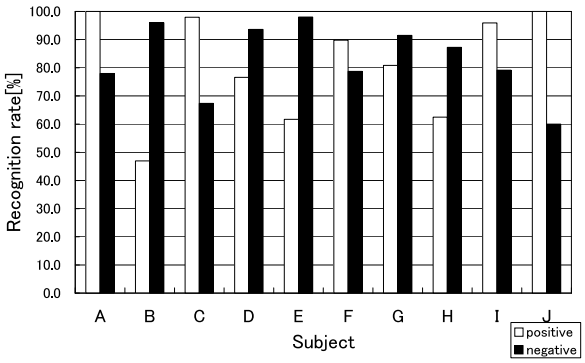


Figure 2: *Recognition results of selected 10 persons.*

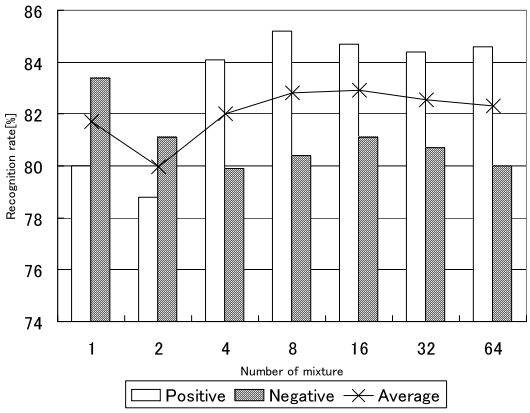


Figure 3: *Recognition results of GMM(Gaussian Mixture Model.)*

listened to the utterances in random sequence and decided the attitude, positive or negative, for each.

We introduce Cohen's κ [6] as a measure of agreement. It is often used as an evaluation measure of a dialogue corpus annotation. Since Cohen's κ is a value calculated from 2 annotations, we calculated it for each combination of the annotations including the experiment result(M). The minimum, maximum and average of the rates with others was calculated for each person. The result is seen in Table 2.

Table 2: *Calculation results of Cohen's κ.*

	min	max	average
M	**0.52**	**0.80**	**0.66**
A	0.52	0.85	0.72
B	0.52	0.79	0.62
C	0.65	0.80	0.72
D	0.65	0.85	0.76
E	0.61	0.79	0.72

Although the machine(M) is ranked fifth on average, the maximum and minimum values are similar to humans. As evaluating with Cohen's κ, the agreement is good if $0.60 < \kappa < 0.75$. In this point of view, this result is sufficient to say that the model has the recognition ability similar to human's.

3. SYSTEM

3.1. Combining the information

We aimed to adopt the results of the attitude recognition by prosody described in Section 2 to a spoken dialogue system as para-linguistic information. In our previous study, a head gesture recognition system has been implemented [7]. This system is able to recognize several types of head gestures, such as "nod", "tilt", and "shake," using optical flow as the feature and HMM as the probabilistic model. As well as prosodic information, these types of head gestures express the utterer's attitude well. We introduce results of a head gesture recognition to a spoken dialogue system as para-linguistic information. We now must combine both results to estimate the user's correct attitude. Particularly, when the results are contrary to each other, one is positive while the other is negative, it is a serious problem to decide which result is plausible.

In this study, we introduce the combination as in Table. 3. For example, If the result of the head gesture recognition is "nod" and the result of the attitude recognition by prosody is "positive", the user's attitude is recognized as "strongly positive", while if the latter is "negative", it is recognized as "in thought", which represents that the user cannot decide clearly.

Table 3: *Combining strategy of the recognition results.* (* indicates that this combination hardly occurs in a dialogue)

		head gesture		
		nod	tilt	shake
prosody	positive	strongly positive	weakly positive	strongly negative*
	negative	in thought	negative	strongly negative

3.2. Proto-type spoken dialogue system

We implemented a proto-type spoken dialog system with para-linguistic information on a humanoid robot ROBISUKE (Fig. 4). We have developed humanoid robots that interact with humans with their perceptual and expressive abilities; such as speech recognition, individual identification using face images, and facial expressions using its eyes, eyebrows and mouth[9]. ROBISUKE is the most recent humanoid robot developed in our group.

In this system, user's responses to the system's suggestions are combined with the method described in section 3.1, and the system makes the dialogue proceed according to the decisions listed as follows.

- strongly positive : the suggestion is accepted, so that the system gives the details promptly.
- weakly positive : the suggestion is almost accepted, so that the system gives the details.
- in thought : the decision is delayed, so that the system says "un(yes)" or just nods its head to receive another response by the user.
- negative or strongly negative : the suggestion is rejected, so that the system gives another suggestion if possible.

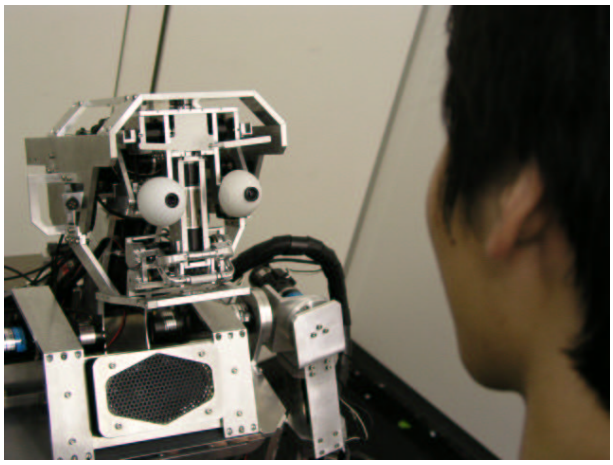


Figure 4: Conversation robot ROBISUKE.

3.3. Example

An example dialogue is seen in Fig. 5.

The system suggests another idea when the user's response is negative to the previous suggestion. It suggests the restaurant in the suggested category when the user's response is positive to the previous suggestion.

In Fig.5, the responses "Curry ka" and "Bento ne" which cannot be decided positive or negative with the only linguistic information, are appropriately interpreted using para-linguistic information.

U: Ohiru gohan nandakedo, doko ka ii tokoro naikana
(I'd like to go to lunch. Do you have an idea?)
R: Curry nante do
(What about curry?)
U: Curry ka **(negative)**
(Curry)
R: Sorejaa, hamburger nante doukana
(Well then, what about hamburger)
U: Ah, hamburger ne **(strongly positive)**
(Oh, hamburger)
R: Hamburger nara chikaku ni McDonald ga aruyo
(For a hamburger, there's a McDonald's nearby)

Figure 5: An example dialogue.
User's utterances and ROBISUKE's utterances are preceded by U: and R: respectively.

4. Conclusion

In this paper, we implemented the attitude recognition by prosodic information in utterances as the para-linguistic information in spoken dialogue. Experimental results show that the recognition ability is similar to human's. We show a spoken dialogue system with para-linguistic information, and these sorts of information influence the progress of a dialogue effectively.

We aim to develop recognition methods of more information, for example, the attitudes recognition in several kinds of utterances and the facial expression and gaze direction recognition, which are accompanied as para-linguistic information. Evaluation of the spoken dialogue system, particularly on the dialogue quality, is the most important issue.

j

5. References

[1] Dellaert, F.; Polzin, T.; Waibel, A., 1996. Recognizing emotion in speech: in *Proceedings of ICSLP'96*,

[2] McGilloway, S.; Cowie, R.; Cowie, E. D.; Gielen, S.; Westerdijk, M.; Stroeve, S., 2000. Approaching automatic recognition of emotion from voice: a rough benchmark: in *Proceedings of ISCA workshop on Speech and Emotion*.

[3] Lee, C. M.; Narayanan, S.; Pieraccini, R., 2001. Recognition of negative emotions from the speech signal: in *Proceedings of IEEE ASRU2001*, 240–243.

[4] Ang, J.; Dhillon, R.; Krupski, A.; Shriberg, E.; Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog: in *Proceedings of ICSLP2002*, 2037–2040.

[5] Cowie, R.; Cowie, E. D.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. G., 2001. Emotion recogntion in human-computer interaction: *IEEE Signal Processing Magazie*, 18(1), 32–80.

[6] Cohen, J., 1960. A coefficient of agreement for nominal scales: *Educational and Psychological Measurement*, 20(1), 37–46.

[7] Ejiri, Y.; Matsusaka, Y.; Kobayashi, T., 2002. Recognition of the head gesture under dialog: *Tehcnical Report of IEICE*, Vol.102, No.218, PRMU2002-61, 31–36. (*in Japanese*)

[8] Tojo, T.; Matsusaka, Y.; Ishii, T.; Kobayashi, T., 2000. A conversational robot utilizing facial and body expressions, in *Proceedings of 2000 IEEE International Conference on Systems, Man and Cyberntics(SMC2000)*, Vol. 2, 858–863.

[9] Matsusaka, Y.; Kobayashi, T., 2001. System software for collaborative development of interactive robot, in *Proceedings of IEEE-Humanoids2001*, 271–277.