ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Prosody based Attitude Recognition with Feature Selection and Its Application to Spoken Dialog System as Para-Linguistic Information

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

[1] School of Science and Engineering, Waseda University, Japan
[2] School of Human Sciences, Waseda University, Japan

10.21437/Interspeech.2004-732

Abstract

In this paper, prosody-based attitude recognition and its application to a spoken dialog system are proposed. Para-linguistic information plays a important role in the human communication. We aimed to recognize the user's attitude by prosody, and apply it to a spoken dialog system as para-linguistic information. In order to find important features to recognize the attitude from automatically extracted features, we applied some feature selection methods. Experimental results show the stepwise method, a combination of the forward selection method and the backward selection method, achieved the best recognition rate. Finally, the dialog system using the recognition results as para-linguistic information is shown.

1. Introduction

Humans exchange information in many ways. Particularly, in a spoken conversation, linguistic information carried by a speech is thought to be one of the most important information. However, we cannot communicate each other effectively enough without any sorts of information that helps the transmission of linguistic information. We call these sorts of information "para-linguistic information." In order to make a spoken dialog system talk with human more naturally, it is indispensable to introduce para-linguistic information. We focused on the speaker's attitude transmitted in the utterances.

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 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 dialog. They used the corpus of human-computer dialog developed under the DARPA Communicator Project. They introduced language model features as well as the prosodic features.

These works extracted the user's real emotion from utterances. The recognition results are important for estimating user's "real" emotional state. But the real emotion is not intuitively applicable for the dialog strategy, because people don't seem to estimate the partner's emotion for each utterance in the conversation. The important information for the dialog strategy is the partner's strategy and attitude, and they often appear as para-linguistic information in the conversation.

In this study, we recognize user's attitude using prosody as para-linguistic information. In order to find important features to recognize the attitude from automatically extracted features, we apply some feature selection methods. Finally, we show a spoken dialog system understanding para-linguistic information implemented on a humanoid robot.

2. Prosody-based Attitude Recognition

2.1. 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 vari-

Table 1: The variety of the recorded utterances.

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

ety of the recorded utterances are seen in Table 1. In the PHRASE, 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.

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.

2.2. Feature Extraction

The features are extracted from F0 pattern and phoneme alignment of the utterance. F0 pattern is extracted with *get_F0* module packaged in ESPS/Waves+, also provided as a component of an open source sound manipulation library *Snack*[5]. Phoneme alignment is obtained from the speech recognition result of a decoder, *Skood*, developed in our group.

13 features, F0 average, F0 median, F0 maximum, F0 minimum, F0 minimum of last mora, F0 range, first mora duration, last mora duration, mean of mora durations, F0 gradient of first mora, F0 gradient of last mora, F0 gradient maximum and F0 gradient minimum are chosen as basic features.

2.3. Feature Selection

In order to find important features to recognize the attitude, we applied some feature selection methods to our basic features. These methods select or eliminate the feature one by one according to the result of an experiment with previously selected feature set.

In each experiment, Gaussian Mixture Model(GMM) is learned for each category(positive/negative) and the cross validation test (750 utterances for learning set, 250 utterances for test set) is done by the Gaussian classifier with these GMMs.

Forward Selection Method(FS) is a method of incremental feature selection. It divides the basic features into one dimension features, and experiments are done for each feature. The feature that marks the most highest recognition rate, is selected as the first dimension of the feature vector. It requires experiments with the feature vectors, where the selected dimensions are fixed and the last dimension is one of the remaining features to determine the next dimension feature.

Backward Elimination Method(BE) does an experiment with basic feature vector at first. Then, the next experiments are done with feature sets, that each one dimension is eliminated from basic feature vector. The feature set that marks most highest recognition rate, is selected as the $N-1$ dimension feature vector, where N is the dimension of basic feature set. In the same way, features are eliminated in order.

Stepwise Method(S) is a combination method of FS and BE. At first, FS is done until the recognition rate is saturated. Then, BE is done with the selected feature vector by then until the rate is saturated. FS and BE are done by turns in this way.

3. Experiments and Results

3.1. Feature Selection

Table 2–4 show the results of forward selection, backward elimination and stepwise method respectively. The boldface features represent the features chosen for recognition with each method.

Table 2: Result of forward selection.

+/−	feature	recog. rate	mix.
+	**F0 grad. min.**	76.1%	8
+	**last mora dur.**	82.2%	64
+	**last mora F0 grad.**	84.2%	64
+	**F0 max.**	87.5%	64
+	F0 min.	87.0%	16
+	mora dur. mean	87.0%	16
+	first mora F0 grad.	87.4%	16
+	F0 grad. max.	87.4%	64
+	F0 mean	87.0%	16
+	F0 range	86.4%	8
+	last mora F0 min.	85.1%	8
+	F0 median	85.4%	32
+	first mora dur.	83.9%	32

Table 3: Result of backward elimination.

+/−	feature	recog. rate	mix.
		83.9%	32
−	F0 min.	85.6%	16
−	F0 median	86.3%	64
−	last mora F0 grad.	86.2%	64
−	F0 max.	86.8%	16
−	**last mora F0 min.**	86.2%	16
−	**first mora dur.**	86.0%	16
−	**first mora F0 grad.**	86.4%	64
−	**F0 grad. max.**	86.3%	16
−	**F0 grad. min.**	85.3%	8
−	**mora dur. mean**	84.2%	8
−	**F0 mean**	81.2%	32
−	**F0 range**	75.5%	32
−	**last mora dur.**		

Table 4: Result of stepwise.

+/−	feature	recog. rate	mix.
+	**F0 grad. min.**	76.1%	8
+	**last mora dur.**	82.2%	64
+	**last mora F0 grad.**	84.2%	64
+	**F0 max.**	87.5%	64
+	F0 min.	87.0%	16
+	**mora dur. mean**	87.0%	16
+	**first mora F0 grad.**	87.4%	16
−	F0 min.	87.9%	16
+	**F0 grad. min.**	88.3%	64
+	F0 median	87.3%	8
+	F0 grad. max.	86.8%	32
+	first mora dur.	86.2%	32
+	F0 range	85.5%	8
+	F0 mean	83.6%	16

3.2. Comparison with Human

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 subjects (A–E) listened to the utterances in random sequence and determined 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 dialog 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 results. The minimum, maximum and average of the rates with others was calculated for each person. The result is seen in Table 5. All three methods are the same value in the average. But the stepwise method is the best in order to obtain the highest minimum value.

Table 5: Calculation results of Cohen's κ

	min	max	average
FS	**0.48**	**0.74**	**0.64**
BE	**0.40**	**0.84**	**0.64**
S	**0.55**	**0.70**	**0.64**
A	0.52	0.85	0.71
B	0.65	0.85	0.76
C	0.52	0.79	0.65
D	0.65	0.80	0.74
E	0.70	0.79	0.75

FS: Forward Selection, BE: Backword Elimination, S:Stepwise, A–E: Subjects

4. Spoken Dialog System

4.1. Dialog Strategy

We aimed to adopt the results of the prosody based attitude recognition to a spoken dialog system as para-linguistic information. In our previous study, head gesture recognition system has been implemented [7]. This system is able to recognize several kinds 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 kinds of head gestures express the utterer's attitude well. We introduce results of head gesture recognition to a spoken dialog system as para-linguistic information.

So we must combine the both results to estimate the user's correct attitude. Particularly, when the results are contrary with each other, one is positive while the other is negative, it's serious problem to decide which result is plausible. In such case, it is recognized as "in thought", which represents that the user cannot decide clearly.

The dialog strategy, when the system receives the user's response to its suggestion about a category, is shown in Figure 1. When it is positive, it can provide detail about the category. In our task, the system suggests the restaurant in the category. When it is negative, it can provide other candidates.

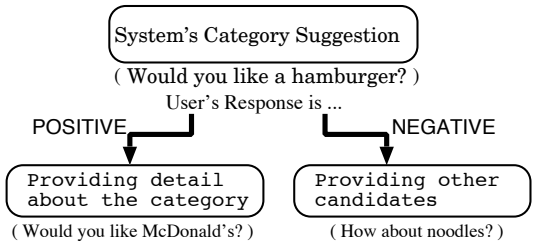

Figure 1: Dialog strategy

4.2. Proto-type spoken dialog system

We implemented a proto-type spoken dialog system with para-linguistic information on a humanoid robot ROBISUKE (Figure 2). 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. ROBISUKE is the most recently developed humanoid robot in our group.

In order to recognize user's attitude in real-time, we adopted the F0 extraction to the sound input module, and phoneme alignment to the speech recognition module. The attitude recognizer can recognize user's attitude as his utterance finishes according to the inputs from these modules(Figure 3).

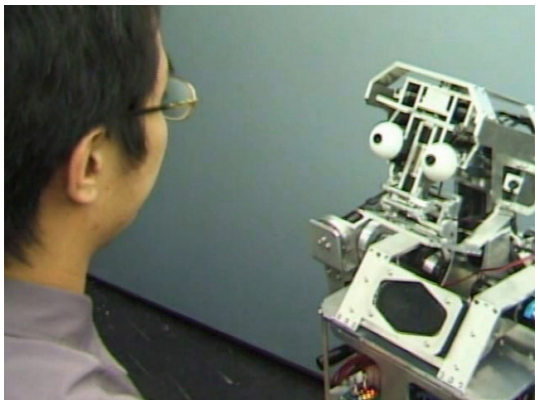

Figure 2: Humanoid robot ROBISUKE in conversation.

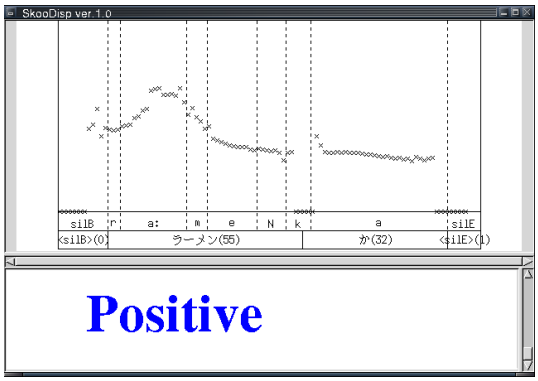

Figure 3: Prosody-based attitude recognizer.

4.3. Example

An example dialog is seen in Figure 4.

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 Figure 4, 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.

5. Conclusion

In this paper, we applied the feature selection method to the prosody-based attitude recognition. Experimental results show that the stepwise method is the best from the point of view of the recognition rate. However, even if a feature combination marks high recognition rate, not all such combination marks high agreement value. We show a spoken dialog system with para-linguistic information, and these sorts of information influence the progress of a dialog effectively.

Evaluation of the spoken dialog system, particularly on the dialog quality, is the most important issue.

`U:` Ohiru gohan nandakedo, doko ka ii tokoro naikana
(I'd like going to lunch. Do you have some 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 hamburger, there's McDonald's nearby)

Figure 4: An example dialog
User's utterances and ROBISUKE's utterances are preceded by `U:` and `R:` respectively.

6. References

[1] Dellaert, F., Polzin, T. and Waibel, A., "Recognizing Emotion in Speech," in Proceedings of ICSLP '96, 1996.

[2] McGilloway, S., Cowie, R., Cowie, E. D., Gielen, S., Westerdijk, M. and Stroeve, S., "Approaching Automatic Recognition of Emotion From Voice: A Rough Benchmark," in Proceedings of ISCA workshop on Speech and Emotion, 2000.

[3] Lee, C. M., Narayan, S. and Pieraccini, R., "Recognition of Negative Emotions from the Speech Signal," in Proceedings of IEEE ASRU2001, 2001.

[4] Ang, J., Dhillon, R., Krupski, A., Shriberg, E. and Stolcke, A., "Prosody-based Automatic Detection Of Annoyance And Frustration in Human-Computer Dialog," in Proceedings of ICSLP2002, pp.2037–2040, 2002.

[5] `http://www.speech.kth.se/snack/`

[6] Cohen, J., "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960.

[7] Ejiri, Y., Nakajima, K., Fujie, S., Kobayashi, T., "Head Gesture Recognition for the Moving Conversation Robot," IEICE Tech. Rep., Vol. 103, No.453, PRMU2003–153, pp.13–18, Nov., 2003. (in Japanese)

[8] Fujie, S., Ejiri, Y., Matsusaka, Y., Kikuchi, H. and Kobayashi, T., "Recognition of Para-Linguistic Information and Its Application to Spoken Dialogue System," in Proceedings of IEEE ASRU2003, pp.231–236, 2003.