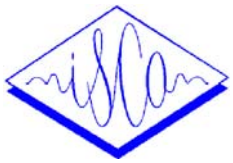

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

INTEGRATING PROSODIC FEATURES IN DIALOGUE UNDERSTANDING

Masafumi TAMOTO, Masahito Kawamori, Takeshi KAWABATA
NTT Laboratories
3-1 Morinosato-Wakamiya, Atsugi 243-0198, Japan
URL: http://www.brl.ntt.co.jp/info/dug/
mailto: tamoto@idea.brl.ntt.co.jp

ABSTRACT

We report our studies on functions of prosodic information in dialogue and on the result of our experiment in dialogue using a speech understanding system that incorporates a discrimination schema for illocutionary acts using prosodic features obtained from human-human dialogs. For constructing a speech understanding system with an 'effortless' interface, it is necessary to model coordination in dialogue. This model needs to capture such sequential constraints of illocutionary acts as answers following questions, acceptance or rejections following requests, acknowledgements following assertions, and so on. However, recognizing these speech acts by simple, superficial analysis of dialogue is often difficult because of such disfluencies as omission and interruption that abound in spontaneous dialogs. Prosodic features are important in this respect because they often contribute to identifying speech acts of utterance when explicit linguistic information is missing. In order to investigate how prosodic information is utilized, along with other linguistic information, to identify speech acts, we performed a series of experiments. We collected task oriented spontaneous dialogs between human subjects, and extracted sentences that represent dialogue control structure. A different set of subjects were chosen for an experiment in which they were asked to identify the sentence type and intonation contour of these sentences. Given the transcription of these extracted sentences with contextual information, the subjects were able to identify the speech act types of about 85% the 290 sentences. The subjects were then asked to identify the intonation contour of the same sentences by listening to the utterance modified in such a way that all voiced sounds were replaced by sinusoid so that only the fundamental frequency of the utterance can be heard. We observed syntactic and prosodic properties of those utterances Speech acts were represented as three basic categories, the illocutions of assertion, question and request. Similarly, sentence types are represented as declarative, interrogative and imperative. Intonations are classified into rise-up, fall-down and neutral pitch contour. We then made a simulation task of dialogue understanding system to incorporate the results of the human-human dialogue experiment. A sentence type was identified using human subjects. An intonation contour was identified using an algorithm that calculates the range and slope of the upper and lower bounds of unwarped segmental contour, and matches these against predefined contour templates.

10.21437/Eurospeech.1999-13

1. INTRODUCTION

Speech conversation is our usual communication method and is most comfortable man-machine interface. For constructing an effortless speech conversation system, it is necessary to implement the coordination mechanism for dialogs.

Prosodic information contributes to identifying speech acts of utterance when linguistic information is missing due to omission or obscure utterance. We aim to construct a system that successfully incorporates prosody, and to analyze the dialog coordination in human-machine conversation. The most commonly known contribution of prosody to speech communication is at the sentential pragmatic and intentional level. That is, analyzing syntactic and intonational properties of utterances, and relationship among sentence type, pitch contour and illocutionary act, intonation can be effectively used for disambiguation in mapping an utterance to these three types of illocutionary acts.

Several simplifications are introduced in the course of our experiment, speech acts are represented as three basic categories, the illocutions of assertion, question and request. Similarly, sentence types are represented as declarative, interrogative and imperative. Intonations are classified into rise-up, fall-down and neutral pitch contour. For investigating how prosodic information incorporates with linguistic information to identify speech acts, we performed a series of experiments.

1. We propose a new algorithm that classify pitch contour to three intonation types for the purpose of automated speech act identification. These intonations are largely realized in the utterance final boundary tone. An intonation contour is identified using an algorithm that calculates the range and slope of the upper and lower bounds of unwarped segmental contour of the last one mora of the utterance, and matches these against predefined contour templates. Problems, however, remain in the course of this process, such as reliable extraction of an intonation contour processing.
2. With this automated intonation contour classification, a simulation of dialogue act prediction through human-human dialogs that evaluates precision/recall of subsequent speech act prediction

based on the automated pitch contour classification and sentence identification by human subjects is performed.

2. PROBLEM SETTING

It can be assumed that three basic sentence types, interrogative, imperative, and declarative express the illocutionary acts of questioning, requesting and asserting, respectively, and a successful mapping of these three sentence types might be expected to discriminate the majority of illocutionary acts.

We consider an illocutionary act set, containing the illocution of assertion, question and request; and combination of a sentence type set, consisting three basic sentence types of declarative, interrogative and imperative, and a pitch contour type set, consisting of neutral, rising and falling.

Likewise, we find the feature that efficiently classify pitch contours into a pitch contour type of neutral, rising and falling. Our research addresses this problem and uses the schema to identify illocutionary act of the given utterance, and to make a dialogue understanding system to incorporate the results of the human-human dialogue experiment.

3. ILLOCUTIONARY ACT IDENTIFICATION WITH AUTOMATED PITCH CONTOUR CLASSIFICATION

In this section, we report an experiment for evaluating an automated pitch contour classification method and the illocutionary act identification proposed above.

3.1. Method and Procedure

Based on the pitch contour classification criteria performed by human subjects, the automated pitch contour classification criteria was produced following way.

3.2. Speech data

The speech data for the experiment was produced in the following way. Two participants perform two coordinative tasks. One is the "map task". One participant informs a given route on map to the other, exchanging information of their map and route. The other task is "maze task". Each of the two participants has one half of a maze divided into two halves. They have to find a passage of the maze through exchanging information of their piece of maze with each other. These dialogs are 30 minutes length totally, sampled at 12kHz and transcribed and divided into chunks. All chunks had to form complete meaningful utterances.

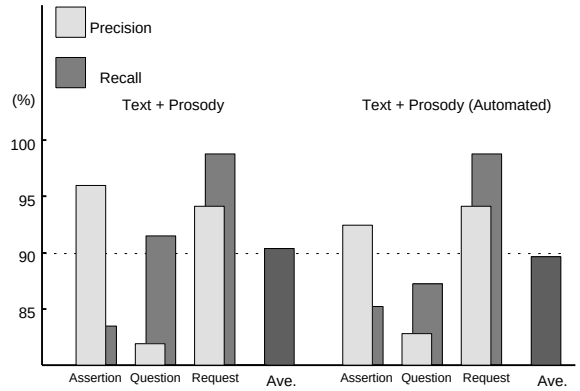

Figure 2. Estimation efficiency of Speech act by sentence type and automated intonation type discrimination

Pitch contour extraction

1. Using pitch tracking software that is based loosely on an algorithm by Medan, Yair and Chazan and also utilizes dynamic programming, rough pitch contour is extracted.

2. Post-processing to determine voiced/unvoiced decision and to omit glitches including half-tone and overtone. The method to omit glitches is as follows: when square power of input speech becomes less than the threshold, output pitch frequency is regarded as unreliable; when the pitch frequency changes rapidly more than the specified rate, pitch frequency is recalculated as less than over-tone and higher than half-tone of recent frequency.

Pitch contour approximation

1. Dividing unwarped contour into monotonous segments from utterance end. Deviations that cannot affect to pitch contour determination are ignored. Average pitch slopes are calculated.

2. Repeat the above procedure until vowel boundary appears.

3.3. Results

To find the efficient combinations that identify illocutionary act from sentence type a pitch contour type, we used a measurement of precision/recall.

Figure 1 shows precision/recall rates of the illocutionary acts identification obtained for each methods of classification with automated pitch contour classification. The pitch deviation threshold between rise-up and neutral is

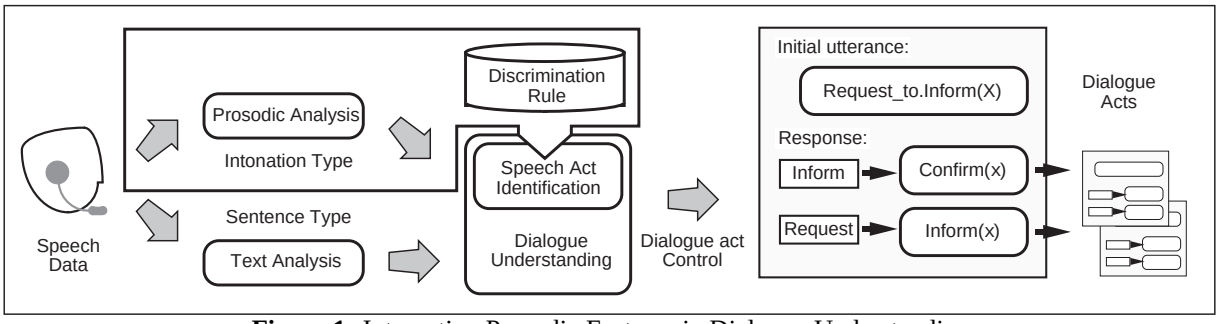

Figure 1. Integrating Prosodic Features in Dialogue Understanding

$0.38 octave/sec$, and threshold between fall-down and neutral is $-1.43 octave/sec$, to fit the classification that human subjects performed. Using automated pitch contour classification, the error rate drops slightly, less than 1%, against pitch contour classification by human subjects.

This figure shows the recall rates of identifying the illocution of question decrease due to classification errors between neutral and fall-down pitch contour. This is considered due to the fact that the slow but long fall-down pitch contour is classified as neutral pitch contour, whereas fall-down pitch contour with sustained power is recognized as neutral pitch contour by human subjects.

4. INTEGRATING PROSODIC FEATURES IN DIALOGUE UNDERSTANDING

In order to implement this algorithm into a speech understanding system, we have to make a module based on the result of these experiments that determines the type of a sentence by analyzing its syntactic and intonational properties.

4.1. Dialogue Act Representation

We prepared a set of dialogue act schema that represent subsequent response to the initial utterance.

Initial Utterance: Request.Inform(X)
Response:

Inform(a)	$\rightarrow$	Confirm(a)
Request.Inform(a)	$\rightarrow$	Inform(eval(a)).

For each line in response representation, an arrow denotes the sequential constraints of illocutionary acts of following utterances. Such as answers following questions, acceptance or rejections following requests, acknowledgements following assertions, and so on.

For example, subsequent response is expected at a request of choosing "day of the week".

A	What day of the week ?	Request.Inform(*Day*)
B	Tuesday	Inform(*Day*)
A	Tuesday	Confirm(*Day*)

Occasionary, query appered as subsequent utterance instead of response.

A	What day of the week ?	Request.Inform(*Day*)
B	This week ?	Request.Inform(*Week*)
A	This week	Inform(eval(*Week*))

Dialogue understanding system may adopt one of subsequent speech acts listed in the dialogue act schema. Figure 2. shows the speech understanding system based on dialogue act schema. It determines its dialogue act by the speech act of the utterance, and transit subsequent dialogue act. Prosodic features are applied in case of syntactically ambiguous utterance.

4.2. Results

We performed an experiment of predicting subsequent speech act on 121 dialogue acts extracted from human-human dialogue. A speech act of initial utterance and sentence type was identified using human subjects. Then speech act of complementary utterance was determined by the proposed method, and predict speech act of subsequent utterance. This results 105(86%) of complementary utterance was correctly classified, and 39(32%) of subsequent utterance was predicted correctly.

5. SUMMARY AND FUTURE WORK

We have proposed a new discrimination schema for illocutionary acts using prosodic features based on experimental results. An intonation contour is identified using an algorithm that calculates the range and slope of the upper and lower bounds of unwarped segmental contour, and matches these against predefined contour templates. This algorithm could correctly recognize 78% of the pitch contour types in the utterances. Furthermore, by this automated intonation contour classification, nearly 90% of speech acts could be correctly identified. Then we made a simulation task of dialogue understanding system to incorporate the results of the automated intonation contour classification. A sentence type was identified using human subjects.

There are a number of problems that need to be solved in the future. First, in order to reduce the number of intonation contour identification errors due to inappropriate analysis, a new method is expected with an algorithm that explicitly uses pitch contours in a parametric form (e.g., by integration into distance scores) for more precise identification, and that calculates the cross correlation factor between two neighboring frames so that the probability density of fundamental frequency may be derived.

6. ACKNOWLEDGEMENT

Acknowledgement is made to Dr. Hironori Hagita, the Executive Manager of the Information Science Research Laboratory of NTT Basic Research Laboratories, for his support. The authors also acknowledge their debt to the other members of the Dialog Understanding research group.

REFERENCES

[1] A. Black. Predicting the intonation of discourse segments from examples in dialogue speech. In Y. Sagisaka, N. Campbell, and N. Higuchi, editors, *Computing Prosody*, pages 117–128. Springer-Verlag, 1997.

[2] B. Grosz, J. Hirschberg, and C. Nakatani. Some intonational characteristics of discourse structure. In *Proceedings of ICSLP*, 1992.

[3] N. Kaiki and Y. Sagisaka. Pause characteristics and local phrase-dependency structure in Japanese. In *Proceedings of ICSLP*, pages 357–360, 1992.

[4] Willem J. Levelt. From intention to articulation. In *Speaking*. The MIT Press, 1989.

[5] Stephen C. Levinson. *Pragmatics*. Press Syndicate of the University of Cambridge, 1983.

[6] Y. Medan and E. Yair. Pitch synchronous spectral analysis scheme for voiced speech. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-37(9):1321, 1989.

[7] N Minematsu and K Hirose. Role of prosodic features in the human process of perceiving spoken words and sentences in Japanese. *THE JOURNAL of the Acoustical Society of Japan (E)*, pages 311–320, 1995.

[8] Shin'ya Nakajima and J. F. Allen. Prosody as a cue for discourse structure. In *Proceedings of ICSLP*, pages 425–428, 1992.

[9] J Pierrehumbert and J Hirschberg. The meaning of intonational contours in the interpretation of discourse. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 271–312. The MIT Press, 1989.

[10] J. Pitrelli, E. Beckman, and J. Hirschberg. Evaluation of prosodic transcription labelling reliability in the ToBI framework. In *Proceedings of ICSLP*, pages 123–126, 1995.

[11] L. R. Rabiner, M. J. Cheng, and A. E. Rosenberg. A comparative performance study of several pitch detection. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 24, pages 399–418. IEEE, 1976.

[12] Marc Swerts and Mari Ostendorf. Prosodic and lexical indications of discourse structure in human-machine interactions. *Speech Communication*, 22:25–41, 1997.

[13] Masafumi. Tamoto and Takeshi Kawabata. A schema for illocutionary act identification with prosodic feature. In *Proceedings of ICSLP*, volume 3, pages 687–690, 1998.

[14] David Traum and James Allen. *A Computatioinal Theory of Grounding in Natural Language Conversation*. Department of Computer Science, Unversity of Rochester, 1994.

[15] Alex Waibel. *Prosody and Speech Recognition*. Morgan Kaufmann, 1988. .