



PROSODICAL SENTENCE STRUCTURE INFERENCE FOR NATURAL CONVERSATIONAL SPEECH UNDERSTANDING

Akio Komatsu, Eiji Oohira, and Akira Ichikawa

Central Research Laboratory, Hitachi, Ltd.
Kokubunji, Tokyo 185, JAPAN

ABSTRACT

In order to develop a system capable of understanding natural conversational speech, along with the current developments in technology for phonetic information processing, a technology must be developed that will utilize prosodic information of natural speech.

We propose here an algorithm for generating a parsing tree that represents for the semantical relationships between phrases, based on analysis of fundamental frequency contour patterns.

The feasibility and validity of our algorithm are confirmed by computer simulation experiments. The capability of prosodical sentence structure inference is confirmed using natural Japanese speech. Also, it is confirmed that our algorithm is speaker independent and task independent. Furthermore, it is confirmed that the algorithm is applicable to another language, by experiments using English speech.

INTRODUCTION

The development of a system capable of understanding spoken language is a highly desirable objective, because oral communication is the most natural form of communication, and because it can convey more information, such as nuances, emotion or intention, than written communication.

When analyzing natural conversational speech, it becomes clear that comprehension mainly involves the use of prosodic information for structural analysis, and the use of phonetic information for content analysis. Thus, along with the current developments in technology for phonetic information processing, a technology must be developed that will utilize prosodic information of natural speech.

Numerous studies show the importance of prosody in human speech perception. However, only a few systems use prosodic information for speech understanding ([1], [2], [3], [4], [5], [7]). In these systems, prosodic information is used to identify boundaries between grammatical units, or is incorporated into a weighting of a distance parameter to improve recognition score or to prune irrelevant recognition candidates. Anyhow, important aspects of prosody have not yet been fully utilized at a higher level of processing (e.g. suprasegmental level).

Here we propose an algorithm for generating a parsing tree that stands for the semantical relationships between phrase units (clause, phrase, word or some other unit), based on analysis of fundamental frequency contour patterns and inference of phrase components.

BASIC APPROACH

The basic approach of a conversational speech understanding system which is composed of prosodic, phonetic, and linguistic information processing units is shown in Fig. 1. These units, which share a common knowledge base and communicate with each other, perform their own data processing, in a cooperative problem solving framework ([7]).

As each unit can be an independent knowledge source, operatable in a parallel computational environment, a sequence of understanding processes is essentially nondeterministic, or data driven. A typical processing sequence may be as follows. From prosodical analysis, a syntactic and/or semantical structure is inferred. By linguistic processing, word hypotheses can be obtained from structural inference, and speech can be understood as a set of key words obtained by word spotting in the phonetic processing unit.

OVERVIEW OF PROSODICAL STRUCTURE INFERENCE

The structure of a spoken sentence comprises not only syntactical structure, but also semantical and emotional structure. To understand this structure, prosody provides suprasegmented cues.

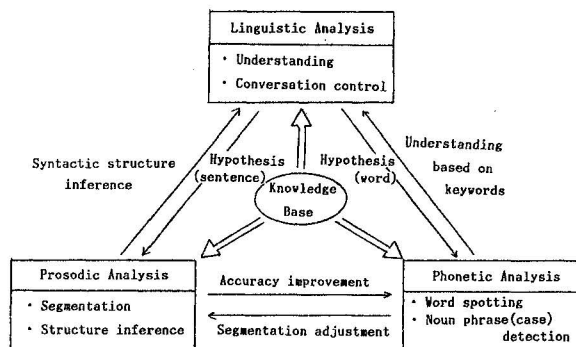


Fig. 1 Basic Approach of Conversational Speech Understanding([7]).

Fundamental frequency (F0) contour is one of the most important prosodic parameters. Therefore, by analyzing the F0 contour pattern, the structure of a spoken sentence can be inferred, and the inferred structure can be represented as a parsing tree. The process of prosodical sentence structure inference is composed of boundary detection and structure inference.

(1) Boundary Detection

From the F0 contour, the likely area of boundaries in a spoken sentence is relatively unambiguous. However, in oral communication, there are no spoken equivalents for commas or periods, and speakers are often almost indifferent to grammatical rules. Therefore, the classification of the unit between two boundaries into a grammatical unit, such as a sentence, clause, phrase or word, in the grammatical sense of written language, is ambiguous.

Thus, boundary detection does not include the identification of grammatical units, but does include the segmentation of continuous speech into a number of units (later called "phrase units"), each of which respectively represent some meaningful unit in a spoken sentence.

Boundary detection analysis is similar to morphological analysis in natural language understanding.

(2) Structure Inference

There should be a model for constructing a structure for phrase units. For this purpose, we introduce a connection rate between two adjacent phrase units, and develop rules for constructing phrase units into hierarchical structure, which can be represented as a parsing tree.

Structural inference is similar to syntactic and/or semantic analysis in natural language understanding.

EXPERIMENTAL PROCEDURE FOR STRUCTURAL INFERENCE

The general functional configuration used to implement the algorithm of prosodical structural inference is shown in Fig. 2.

Conversation Model

In order to develop practical procedures, a model of conversation is tentatively based on the tasks of a PBX telephone operator (the operator understands what the user says, and connects the user with the specified extension telephone line), and actual speech sentences in the model are analyzed.

Prosodical Information Extraction

The incoming speech is processed as follows:

- (1) A/D conversion (12 bit) sampled at 12kHz,
- (2) Prosodic feature extraction: Fundamental frequency calculation with a 60 msec rectangular window in a 20 msec step.

In order to cope with erroneous pitch detection, window size is set wider than in phonetic feature extraction (40 msec), and each F0 value is confirmed using speech power and normalized

residual power. Furthermore, by checking F0 continuity, some adjustment of F0 value is performed.

Boundary Detection

The boundaries between the phrase units of incoming speech are detected as the local minimum points of an F0 contour pattern. Irrelevant boundaries are pruned by skipping those boundaries which follow short phrase units and those at small F0 contour dips because these boundaries are probably those of stressed syllables within words.

Phrase units surrounded by detected boundaries may be sentences, clauses, phrases, words and in some instances, parts of words. The detected boundary position may differ from the actual boundary position by a few syllables. (This type of difference will be absorbed by a word spotting mechanism, not described in this paper).

Line Approximation of F0 Contour

The F0 contour pattern of a phrase unit is represented as a single approximate straight line which has the least cumulative difference between it and actual F0 values. Thus, this line is usually a declining line, and its slope may represent the phrase component of speech.

Qualitative Analysis of Connection Rate

The connection rate between consecutive phrase units is introduced as a measure of strength of the relationship between phrase units in a semantical sense rather than in a syntactical or grammatical sense. Several parameters contribute to setting the value of the connection rate:

- (1) Gap of F0 --- When the gap in the F0 value, between the tail of the leading unit and the head of the following unit, is big, a new phrase component can be implied, and the connection rate should be small.
- (2) Length of leading unit --- When the leading unit is short, it may be a word or a part of a word, and the connection rate should be large.
- (3) Declining slope --- When the declining slope of the approximation line of the unit is steep, it may not represent a phrase component, but rather an accent component within a phrase, and the connection rate should be large.

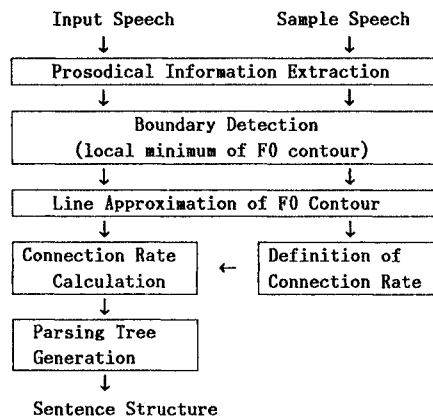


Fig. 2 Experimental Procedure for Prosodical Inference.

Definition of Connection Rate

The more parameters and rules used to define the connection rate, the greater accuracy of the rate obtained in a wider variety of speech. Here, the connection rate $R_i(A,B)$, at the point i , between two consecutive phrase units, A and B , is defined as follows (Fig. 3) :

$$R_i(A,B) = W_d * d + W_l * l + W_g * g \quad (1)$$

where,

d : decline slope of approximation line (Hz/Sec)

l : length of phrase unit (Sec)

g : gap of fundamental frequency (Hz)

W_d, W_l, W_g : weighting coefficients

The versatile weighting coefficients should be decided based on the analyses of many conversational sentences spoken by many persons. However, here, we decide the weighting coefficients based on only one speech example (of one male speaker). By analyzing the F0 contour pattern of sample speech, the relationship between weighting coefficients can be deduced, based on a number of constraints, to generate the proper parsing tree of the learning speech sentence. The value of the weighting coefficients will be set to satisfy these relationships ([7]).

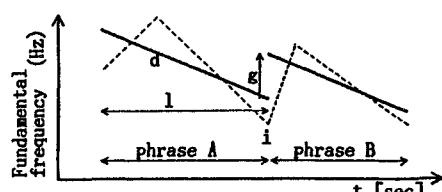
Weighting coefficients decided in this heuristic manner are applicable to a variety of speech inputs, as will be shown later from experimental results. Thus, it may be said that the general suprasegmental features of prosody will be global. Therefore, a few essential or typical data can represent the characteristics of these features.

Parsing Tree Generation

A parsing tree of a spoken sentence is obtained by comparing connection rates, and combining phrase units in accordance with the rates. Using a push down stack, input spoken sentences are processed progressively, based on so-called left-to-right parsing, and a parsing tree is generated (Fig.4).

A new phrase unit P_x of the input speech is compared with the phrase units P_{i-1} and P_i in the stack, and the two connection rates $R(P_{i-1}, P_i)$ and $R(P_i, P_x)$ are calculated.

When the preceding connection rate is greater than the following connection rate, it means that the relationship between the two preceding phrase



$$R_i(A,B) = W_d * d + W_l * l + W_g * g$$

W_d, W_l, W_g : weighting coefficients
 d : decline of approximation line (Hz/Sec)
 l : length of speech (Sec)
 g : gap of fundamental frequencies (Hz)

Fig. 3 Definition of connection rate([7]).

units (P_{i-1} and P_i) in the stack is stronger than that between the current and preceding units (P_x and P_i). Thus, in this case, the two phrase units in the stack are popped and combined into a new phrase unit which is put at the top of the stack. When the preceding connection rate is smaller, the phrase unit P_x is simply pushed into the stack. The combining process is postponed until the connection rate between two phrase units is greater than that between two succeeding ones.

At the end of the input speech, the phrase units remaining in the stack are analyzed and popped in accordance with the connection rate.

EXPERIMENTAL RESULTS

Computer simulation experiments are performed using actual speech and several kinds of sentences (Fig. 5).

Another speech sample, consisting of the same sentence used for weighting coefficient learning, is analyzed and a proper parsing tree is obtained (Fig. 5a):

[ohayougozaimasu(Good morning),
 [[[chuukeN(Cent.res.lab), 6-bu(6th dep.)],
 komatsudesuga(Komatsu speaking)],
 [8-bu(8th dep.), suzukionegai(Suzuki please)]]].

For speech in which the same sentence is spoken by another speaker (male), the same parsing tree is obtained. From another sentence in the PBX task, a reasonable parsing tree is obtained (Fig. 5b).

As ambiguous speech sentences,
 [niwaniwa(in garden), niwatori gairu(hen is)], and
 [niwa(two), [niwaniwa(in garden), tori gairu(bird is)]] are discriminated into different parsing trees (Fig.5c).

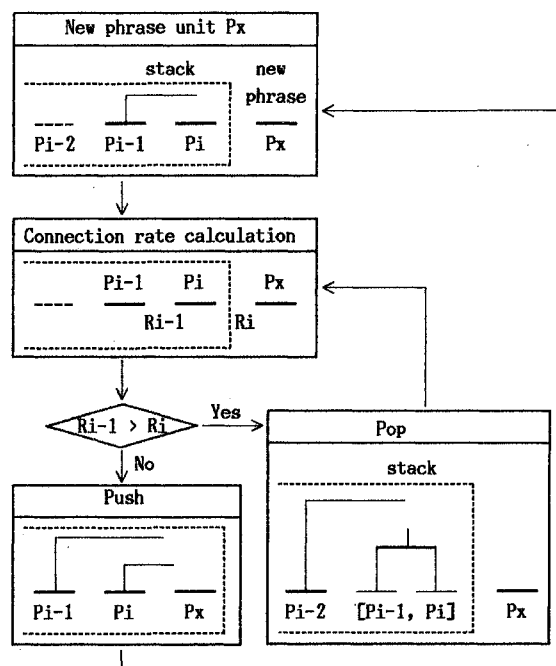


Fig. 4 Steps to generate parsing tree.

Furthermore, computer simulation experiments are performed for English speech spoken by a native male speaker, and a suitable parsing tree is obtained (Fig. 5d):

[[[The feasibility and],[validity],[of our algorithm]],
[[are confirmed by],[computer simulation],[experiments]]].

CONCLUSION

This paper discussed prosodical sentence structure inference of natural conversational speech. We proposed an algorithm for inferring the structure of a spoken sentence. This algorithm generates a parsing tree that represents a semantical relationship between phrases. We confirmed the feasibility and validity of our algorithm by computer simulation experiments using actual spoken sentences. We also confirmed that our algorithm is speaker independent and task independent. Furthermore, we confirmed that the algorithm is applicable to another language; i.e., English.

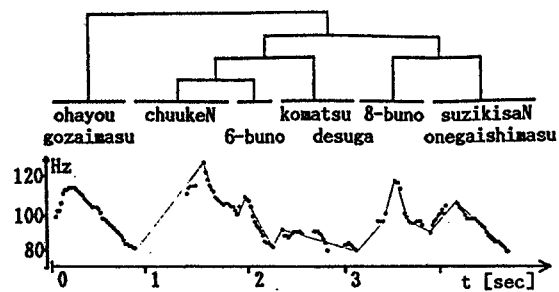
ACKNOWLEDGMENTS

The authors wish to thank Yoshinori Kitahara, for his helpful comments. They further thank Zenji Tsutsumi, Dr. Yoshito Tsunoda and Dr. Masakazu Ejiri for guidance and encouragement in the research.

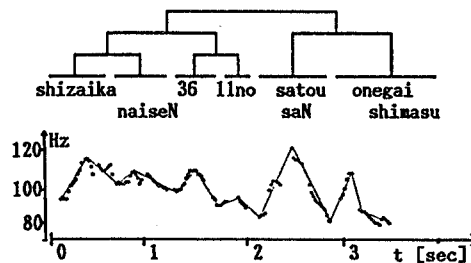
The research work reported on in this paper was partially supported by ICOT (Institute for New Generation Computer Technology).

REFERENCES

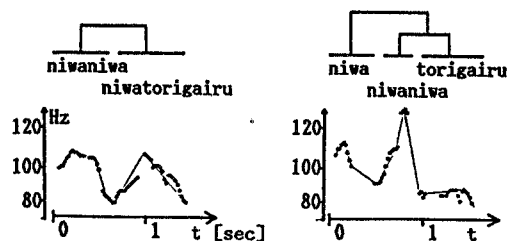
- [1] W. A. Lea, "A Prosodically Guided Speech Understanding Strategy," IEEE Trans. Vol. ASSP-23, No.1, 1975.
- [2] A. Waibel, "Prosody and Speech Recognition," PhD thesis, Computer Science Department, Carnegie Mellon University, 1986.
- [3] A. Komatsu, E. Oohira, A. Ichikawa, "Prosodic Aids to Structural Analysis of Conversational Speech," Proc. 1986 IEEE ICASSP, 42.15.1-42.15.4, 1986.
- [4] H. Fitch, "Relative Timing Measures of Acoustic Segments Aid Automatic Word Recognition," Proc. 1982 IEEE ICASSP, 1247-1250, 1982.
- [5] L. R. Rabiner, "On the Application of Energy Contours to the Recognition of Connected Word Sequences," AT&T Bell Lab. Tech. Journal, 63(9), 1981-1985, Nov., 1984.
- [6] H. Fujisaki and K. Hirose, "Analysis of Voice Fundamental Frequency Contour for Declarative Sentence of Japanese," Acoust. Soc. Jpn., (E)5, 233-242, 1984.
- [7] A. Komatsu, E. Oohira, A. Ichikawa, "Conversational Speech Understanding Based on Sentence Structure Inference Using Prosodics, and Word Spotting," Trans. IEICE Japan, Vol. J71-D, No.7, pp.1218-1228, 1988 (in Japanese).



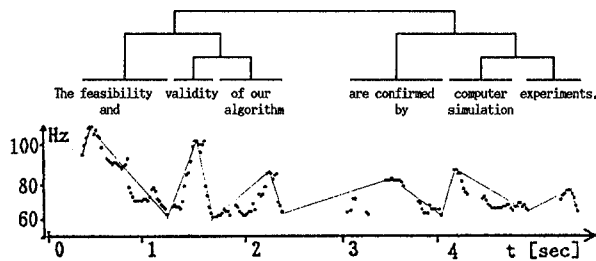
(a) Sample speech which has the same sentence structure as learned speech.



(b) Sample speech in PBX task.



(c) Sample speech of ambiguous sentence structure.



(d) English speech spoken by a native male speaker.

Fig. 5 Examples of parsing tree of speeches by prosodical sentence structure inference([7]).