TM-147

PROSODIC AIDS TO STRUCTURAL ANALYSIS OF CONVERSATIONAL SPEECH

AKIO KOMATSU, EIJI OOHIRA, AKIRA ICHIKAWA
Central Research Laboratory, Hitachi Ltd.
1-280 Higashi-Koigakubo, Kokubunji, Tokyo, 185 JAPAN

HIROHIDE ENDOH
Microelectronics Products Development Labs., Hitachi Ltd.
292 Yoshida, Totsuka, Yokohama, Kanagawa, 244 JAPAN

ABSTRACT

An algorithm for inferring the sentence structure of conversational speech using prosodic information is presented. The fundamental frequency (F0) contour shape, which is the major prosodic information parameter that best represents the structure of a spoken sentence, is analyzed in detail by means of F0 shape approximation with a sequence of linear lines. From such an analysis, boundaries between grammatical units can be detected. In addition, the relation between these grammatical units can be inferred by reffering to natural conversational speech styles. Structural hypothesis inference capabilities are investigated to determine the syntactic and semantic structure of a sentence through computer simulation experiments.

INTRODUCTION

The development of a system capable of understanding spoken language is a highly desirable objective because speaking is the most natural form of communication.

In oral communication, there are no spoken equivalents for commas or periods, and speakers are often almost indifferent to grammatical regulations. This means that, in oral communication, a wide range of sentence types must be dealt with, some of which are incorrect or ambiguous. Thus, to develop an effective oral communication system for computers, it is necessary to develop technology that will overcome this problem.

It is vital to understand the process by which human beings comprehend spoken language and the mechanism they use to transfer information by speech ([1], [2], [4], [5]). Human speech comprehension appears to mainly involve making use of prosodic information for syntactic structural analysis and phonetic information for semantic content analysis. Consequently, it is essential to utilize prosodic features to formulate preliminary syntactical structure hypotheses to attain computer understanding of conversational speech ([6]).

An algorithm for formulating structural hypotheses using prosodic information is proposed here. Additionally, several capabilities of the algorithm are evaluated by performance analyses of computer simulation experiments.

BASIC APPROACH

Prosodic information is composed of a fundamental frequency(F0), speech power, and rhythm (or speed). These factors relate to each other under constraints of grammar, expiration (breath), emphasis, emotion and so on. Among them, the F0 contour shape is of particular note. It is composed of phrase components and accent components[2]. Thus, the structure of a spoken sentence can be inferred by extracting phrase components from superimposed F0 contour shapes.

An algorithm for the automatic approximation of F0 shapes with a set of straight lines is essential to realize reliable F0 shape analyses. Based on such analyses, An algorithm that formulates structural hypotheses can be developed. The general functional configuration that is used to implement the algorithm is shown in Fig.1. In order to develop practical procedures, a model of conversation is defined and actual spoken sentences in the model are analyzed.

EXPERIMENTAL PROCEDURE FOR STRUCTURAL ANALYSES

Conversation Model

For the conversation model, the tasks of a PBX telephone operator are utilized. The operator understands what the user says. The speech in this conversation model is composed of three kinds of sentences (with abbreviations in the parentheses).

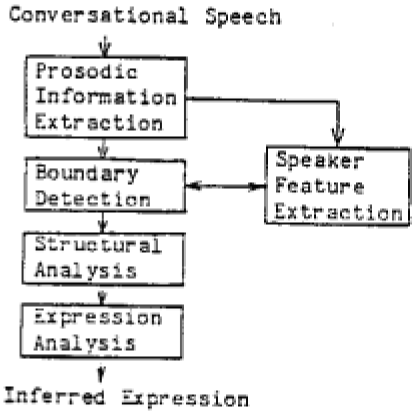


Fig.1 Basic functional block diagram of conversational speech analysis.

(1) Greeting sentence (<greet>) — A greeting is a typical opening sentence, but it is optional.

(2) Caller declaration sentence (<caller>) — He may declare his own name (<callernam>) and/or company name (<callerdep>). The caller declaration sentence is also optional.

(3) Callee declaration sentence (<callee>) — A declaration of the callee's name (<calleenam>) is indispensable in order for the operator to make the connection. A declaration of the callee's name may sometimes be accompanied by the department name (<calleedep>) or extension number (<calleetel>).

Prosodic Information Extraction

The incoming speech is processed as follows:

(1) A/D conversion(12 bit) sampled at 12kHz,

(2) Phonetic feature extraction : LPC analysis — 12-order autocorrelation coefficients and residual power with a 20 msec Hanning window in a 7.5 msec step,

(3) Prosodic feature extraction : Fundamental frequency calculation with a 40 msec rectangular window in a 7.5 msec step.

In order to obtain improvements in the accurracy of the F0 calculation, each F0 value is confirmed ng speech power and normalized residual power. Furthermore, by checking the F0 continuity, some adjustment of F0 value is performed.

Histogram of Fundamental Frequency

As the range of F0 varys depending on the speaker, an F0 histogram is necessary to decide the personalized threshold of the F0 baseline (Fmin). On this basis, the system detects the end of a sentence by checking the tail end of the F0 contour. The tail end, where F0 is close to Fmin, usually represents the end of the sentence. The threshold (Fe) for detecting the low F0 area should be decided from the distribution of the F0 histogram. However, at this stage, Fe is set experimentally at 10% above Fmin.

Boundary Detection

The boundaries between the grammatical units of incoming speech are detected first. Then, by analyzing the shape of the F0 contour, these grammatical units are classified into several tegories corresponding to the style of speech. For effective F0 contour analysis, the F0 contour shape is approximated by a sequence of straight lines.

(1) Boundary candidate detection

In general, pause in speech is a reliable index for obtaining candidates for boundaries between grammatical units. Experimentally, the length of the pause can be set at 300msec.

To confirm that a boundary candidate is actually the end of a sentence, the F0 value is compared with the Fe at the boundary. To more accurately confirm the end of a sentence, the amount the F0 value subsequently increases is examined (the threshold of increase is set at around 50Hz). This is because the start of the next sentence will show a rise in the F0.

(2) F0 contour approximation

For detailed analyses of F0 contour shapes, the F0 contour is approximated by a sequence of straight lines. This approximation is performed between the boundaries. The approximation algorithm procedure is as follows:

(a) Approximate with a straight line between the endpoints,

(b) Calculate the total error between the actual F0 value and the approximation line,

(c) If the total error is less than the threshold, e.g. 1.0 Hz/msec, go to (f). If not, continue on,

(d) Find the point where there is the largest difference between the F0 value and the approximation line,

(e) Make a new approximation line between the starting point and the point found in (d). Then, go back to (b),

(f) Reset the starting point to be at the point where the last approximation line obtained in (c) ended. If all of the area between the original endpoints has not yet been processed, go back to (a) to continue the approximation.

(3) Boundary classification

The boundaries between grammatical units can be classified into the following categories:

(a) Starting point of sentence : [s],

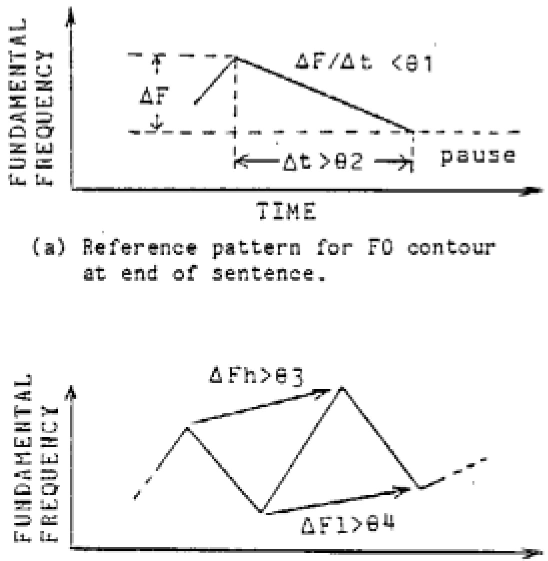
(b) Ending point of sentence : [e],

(c) Complex sentence boundary : [c].

A spoken sentences is not so grammatical that the definition of each categories is not rigorous. For example, the tail end of a long modifying phrase c be regarded as same as the complex sentence boundary.

The classification of boundaries is performed by comparing the F0 approximation line with the F0 contour reference pattern for [e] (for the end of a sentence) and [c] (for a complex sentence boundary), as shown in Fig.2(a) and Fig.2(b), respectively.

As each type of boundary has its own structural meanings, the grammatical unit between boundaries carrys some meaning corresponding to the style of speech. The possible speech styles are as follows:



(a) Reference pattern for F0 contour at end of sentence.

(b) Reference pattern for F0 contour at complex sentence boundary.

Fig.2 Reference patterns for F0 contours at syntactic boundarys of speech.

(a) Isolated unit : [SE] ([s]-[e]),
(b) Closing unit that follows another unit : [CE] ([c]-[e]),
(c) Starting unit that is followed by another unit : [SC] ([s]-[c]),
(d) Unit that follows another unit and also is followed by another unit : [CC] ([c]-[c]).

Structural Analysis

The syntax rules for written styles in PBX tasks can be represented as follows:

```
<sent>    --> <sent1>;<sent2>;<sent3>;<sent4>.
<sent1>   --> <greet>,<caller>,<callee>.
<sent2>   --> <greet>,<callee>.
<sent3>   --> <caller>,<callee>.
<sent4>   --> <callee>.
<caller>  --> <caller1>;<caller2>.
<caller1> --> <callerdep>,<callernam>.
<caller2> --> <callernam>.
<callee>  --> <callee1>;<callee2>;<callee3>.
<callee1> --> <calleedep>,<calleenam>.
<callee2> --> <calleetel>,<calleenam>.
<callee3> --> <calleenam>.
```

From these rules, 18 written styles (S1-S18) can generated. On the other hand, there is some iation between written style (like <greet>) and speech style (like [SE]) as follows:

[SE] includes <greet>,<callee1>,<callee2>,<callee3>.
[CE] includes <callee1>,<callee2>,<callee3>.
[SC] includes <calleedep>,<calleetel>,<caller1>,
<callerdep>,<callernam>.
[CC] includes <calleedep>,<calleetel>,<callernam>.

These come from the naturalness of speech. A <greet>, for example, spoken in [SC] style, which is supposed to be followed by another unit, is an intentionally artificial expression. Thus, it will not appear in natural conversational speech. Furthermore, there are several constraints on the sequences of speech styles. They are as follows:

[SE] can be followed by [SE] or [SC],
[CE] can be followed by [SE] or [SC],
[SC] can be followed by [CE] or [CC],
[CC] can be followed by [CE] or [CC], and,
speech can start with [SE] or [SC],
speech can end with [SE] or [CE].

From these constraints, the possible sequences of speech styles are as follows (the digit at the end the line represents the number of the corresponding written style in the list of S1-S18):

(t1) [SE]	3
(t2) [SE,SE]	3
(t3) [SC,CE]	8
(t4) [SE,SC,CE]	8
(t5) [SE,SC,CC,CE]	7
(t6) [SC,SE,CC,CE]	2
(t7) [SE,SC,CC,CC,CE] ...	2

Considering that a sequence of speech styles can be obtained by the boundary detection mentioned above, the hypotheses on the possible written styles can be limited to at most 8, down from 18 (S1-S18).

Expression Analysis

Based on the hypotheses obtained from the structural analyses, accurate expression analyses can be performed. A typical expression analysis, that of an idiomatic expression or fixed sentence form, can be accomplished by comparing the F0 contour with the template shape.

The expression inferred in this way can be confirmed by checking it with phonetic information. A typical Japanese greeting, e.g. OHAYOU GOZAIMASU, usually has an F0 contour shaped like a [He] character, i.e. a short rising F0 followed by a somewhat long tail for the F0 phrase component. At the tail end of the phrase, a fricative sound exists because the final vowel, /u/, is usually unvoiced.

For more detailed expression analyses, a word dictionary with prosodic information is necessary. Each word has its own F0 shape for the accent component. Consequently, boundary candidates of words can be obtained by inferring their accent components from a superimposed F0 and consulting with a prosodic dictionary. Regarding expression analyses, more research is required.

EXPERIMENTAL RESULTS

Several basic capabilities of the algorithm are tested by performance analyses of computer simulation experiments with several minutes of conversational speech by three speakers(KK, YK, and EO).

A typical process sequence for structural analysis of speech is shown in Fig.3. In this example, the sequence ([SE,SC,CE]) of the speech style is obtained by detecting the complex sentence boundary [c]. In addition, 8 possible written styles are inferred. In every possible case, the leading speech style ([SE]) is assigned to the greeting (<greet>).

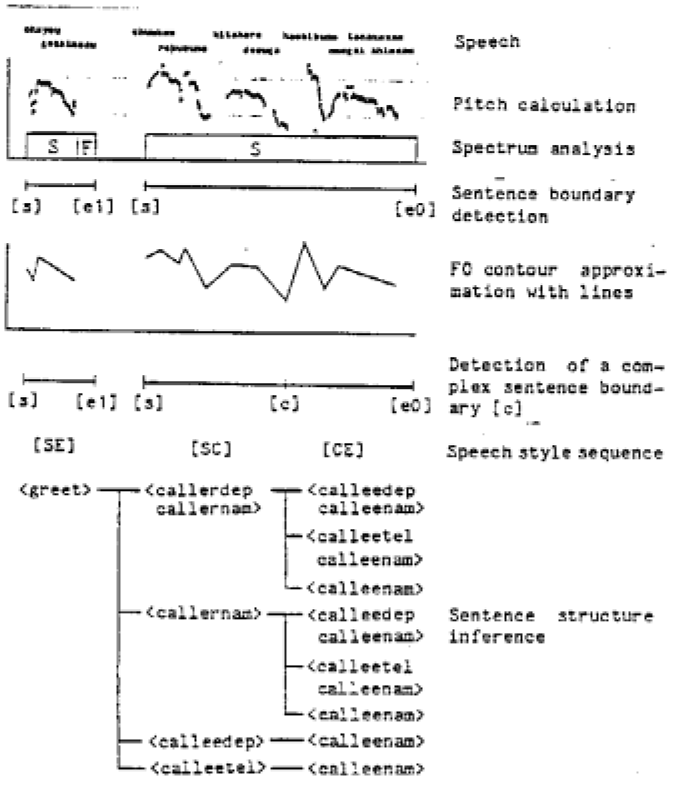


Fig.3 An illustrative example of experimental results from structural analysis.

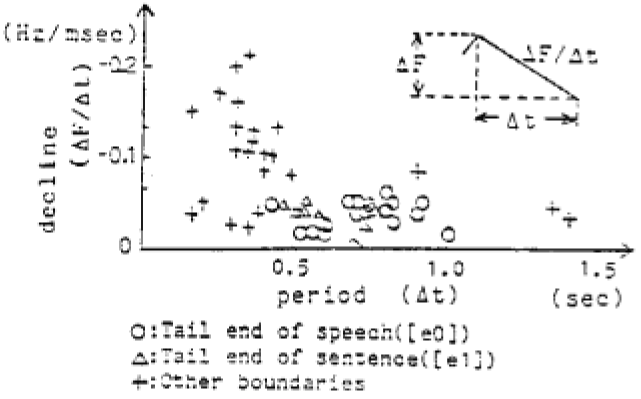


Fig.4 Experimental results on detection of sentence tail ends (speaker:KK).

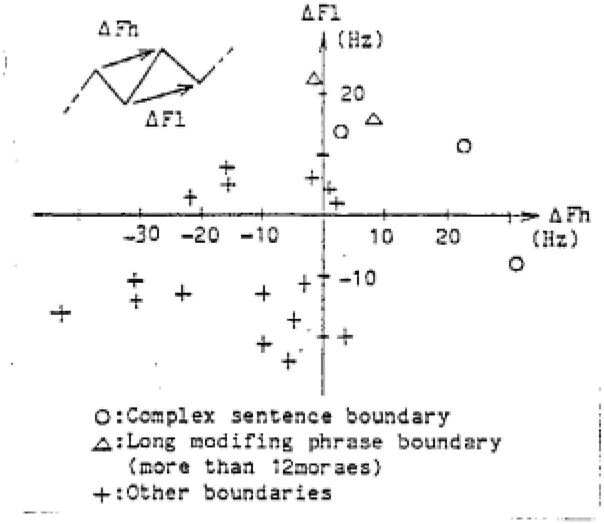


Fig.5 Experimental results on detection of complex sentence boundaries(speaker:KK).

In expression analysis, <greet> is confirmed by comparing an F0 contour shape with reference pattern and fricative sound at the tail end.

Regarding the detection of sentence tail end ([e]), the experimental data in Fig.4 shows the high possibility of accurate sentence tail end detection. The function of decline and the period of the F0 approximation line are used to distinguish sentence tail end from the others. In this experiment, three phrase tail ends are incorrectly detected as sentence tail ends.

Regarding the boundary detection of complex sentence or long modifying phrases, the experimental data of Fig.5 also shows the high probability of reliable separation. In the experiment, it is clear that the boundary candidates can be classified using the differences in F0 values for adjacent local minimum quantities and local maximum ones. However, in order to cover wider variations of expressions, a more complicated discrimination function may be required.

The capability of the structural hypothesis inference depends on the ability of the boundary detection, which is accurate as shown above. The speech style sequences of 16 speeches (out of 19 speeches) are inferred correctly (speaker:KK). Three mistakens come from three incorrect detections of sentence tails.

The same experiments are repeated on another speakers, and almost the same level of accuracy is obtained.

CONCLUSION

This paper has discussed prosodic aids to structural analysis of conversational speech. An algorithm for inferring the sentence structure of speech has been proposed. This algorithm is based on the F0 contour analysis using the linear approximation line. Several basic capabilities the algorithm are tested by performance analysis of computer simulation experiments. The experimental results show that the algorithm is applicable to structural analysis of conversational speech.

The research work reported in this paper was partially supported by ICOT (Insutitute 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] H. Fujisaki and K. Hirose, "Analysis of Voice Fundamental Frequency Contours for Declarative Sentence of Japanese," J. Acoust. Soc. Jpn. (E)5, 233-242 (1984).

[3] K. Hirose, H. Fujisaki and N. Takahashi, "Analysis and Synthesis of Fundamental Frequency Contours of Complex Sentences," Trans. of the Committee on Speech Research, Acoust. S Jpn. S82-40 (1982) (in Japanese).

[4] K.Hakoda and H. Sato, "The Investigation of Prosodic Rules in Connected Speech," Trans. of the Committee on Speech Research, Acoust. Soc. Jpn. S78-07 (1978) (in Japanese).

[5] K. Iwata, T. Ohono and K. Shirai, "Fundamental Frequency Control in Japanese Conversational Speech," Trans. of the Committee on Speech Research, Acoust. Soc. Jpn. S85-42 (1985) (in Japanese).

[6] A. Komatsu et al, "Phoneme Recognition in Continuous Speech," Proc. 1982 IEEE ICASSP, 883-886 (1982).