Use of Linguistic Information of Accent Phrases for Automatic Extraction of F_0 Contour Generation Process Model Parameters

Keikichi HIROSE **Yusuke FURUYAMA**
Hirose-Minematsu Lab., School of Frontier Sciences, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033 JAPAN
hirose@gavo.t.u-tokyo.ac.jp furuyama@gavo.t.u-tokyo.ac.jp

Shuichi NARUSAWA **Nobuaki MINEMATSU**
Hirose-Minematsu Lab., School of Inf. Science and Tech., University of Tokyo
narusawa@gavo.t.u-tokyo.ac.jp mine@gavo.t.u-tokyo.ac.jp

Hiroya FUJISAKI
Prof. Emeritus, University of Tokyo
fujisaki@alum.mit.edu

Abstract

A method was developed to utilize linguistic information (lexical accent types and syntactic boundaries) to improve the performance of the automatic extraction of the F_0 contour generation process model commands. The extraction scheme is first to smooth the observed F_0 contour by a piecewise 3rd order polynomial function and to locate accent command positions by taking the derivative of the function. The phrase commands are estimated from the residual F_0 contour, which is obtained by subtracting the extracted accent components from the observed F_0 contour. If the results of automatic extraction differ from those estimated from the linguistic information, they are modified according to the several rules. The results showed that some errors could be corrected by the use of linguistic information; especially when the initial word of an accent phrase is type 0 accent. As a whole, the correct extraction rate (recall rate) was increased from 79.8 % to 82.3 % for phrase commands and from 81.6 % to 85.9 % for accent commands.

1 Introduction

Recent technology on speech information processing largely relies on the speech corpora. Several rather large speech corpora have already been developed. However, most of them only include phonemic labels with no prosodic labels. The major reason for this situation is that the corpora are mostly arranged for speech recognition study, where prosodic features are not so much utilized yet. However, control of prosodic features is an important issue in speech synthesis already and use of prosodic features will become mandatory for the future development of speech recognition.

The major problem on developing prosodic speech corpus (speech corpus with prosodic labelling) is that we have no good system for prosody annotation for Japanese. The well-known Tone and Break Indices (ToBI) system is a good candidate, and J-ToBI has already been developed with its extended version X-JToBI [Kikuchi and Maekawa. (2002)]. Several speech corpora have already been developed with the ToBI labelling, but the ToBI system has an unavoidable defect that it is not based on the quantitative definition of prosodic features. It requires "intuition" of the human labellers, and, therefore, the labelling results may fluctuate between labellers. Also, labelling by human is time consuming work. To cope with this situation, there were several attempts for the automatic ToBI labelling, though the results were still far from satisfaction and needed manual correction.

The generation process model for fundamental frequency contour (henceforth F_0 model) will offer us a prosody labelling system with quantitative definition. The model assumes two types of commands, phrase and accent commands, as model inputs, and these commands have been proven to have a good correspondence with linguistic and para-/non-linguistic information of speech [Fujiaski and Hirose (1984)]. For instance, we can easily define the prosodic boundary depth between accent phrases using the absence/existence and magnitude of phrase command between the phrases. The major problem with

useong the F_0 model for prosodic labelling is that we have no good method for automatically extracting the model commands from the observed F_0 contours.

From this point of view, several research works were conducted on the automatic extraction of model commands. Recently, by smoothing the observed F_0 contours using 3rd order polynomial functions and searching accent command location from their derivatives, we succeeded extracting model commands with a rather high accuracy [Narusawa, Minematsu, Hirose, and Fujiaski. (2002)]. However, there were still a number of cases where the extraction was done incorrectly. The erroneous results can be corrected by utilizing the linguistic information of the utterances, which is not taken into account in the current method. We have already reported a method using linguistic information, but it only looked at the timing relationship between accent command timings and corresponding segmental boundaries [Sakurai and Hirose (1999)]. This was because we had no good system available for obtaining the accent types of accent phrases. Recently, we have developed such a system under the project on "Development of Fundamental Software of Anthropomorphic Agents," and using the system, succeeded to include the linguistic information in the scheme of automatic extraction of model commands. In the current paper, the developed method is explained with some experimental results.

2 F_0 contour model

The F_0 model is a command-response model that describes F_0 contours in logarithmic scale as the superposition of phrase and accent components. The phrase component is generated by a second-order, critically-damped linear filter in response to an impulse called phrase command;

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & t \geq 0 \\ 0, & t \geq 0 \end{cases} \tag{1}$$

and the accent component is generated by another second-order, critically-damped linear filter in response to a step function called accent command:

$$G_a(t) = \begin{cases} \min\left[1-(1+\beta t)\exp(-\beta t), \gamma\right], & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{2}$$

where α and β are the time constants for the phrase and accent control mechanisms, respectively. Since these parameters are tightly related to the mechanical system of larynx, they are considered to be similar for all the utterances. Based on the former F_0 contour analysis results, they were fixed at 3.0 s^{-1} and 20.0 s^{-1}, respectively. The ceiling parameter γ was also fixed at 0.9.

An F_0 contour is given by the following equation:

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} \left\{ G_a(t - T_{1j}) - G_a(t - T_{2j}) \right\} \tag{3}$$

In the equation, F_b is the bias level, i is the number of phrase commands, j is the number of accent commands, A_{pi} is the magnitude of the ith phrase command, A_{aj} is the amplitude of the jth accent command, T_{0i} is the time of the ith phrase command, T_{1j} is the onset time of the jth accent command, and T_{2j} is the reset time of the jth accent command. A schematic view of the model is given in Fig. 1.

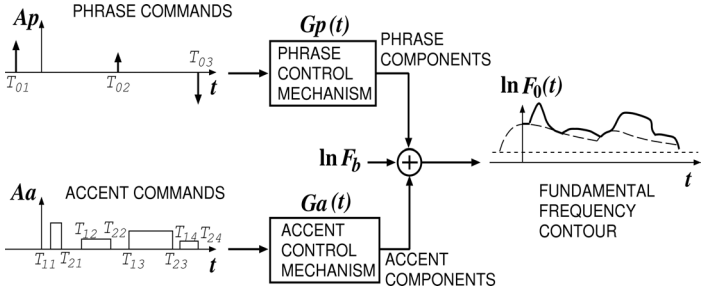

Fig. 1. Generation process model for sentence F_0 contours.

3 Original method (not using linguistic information)

The F_0 contour extracted from speech waveform using a pitch extraction algorithm may include pitch extraction errors, sharp F_0 movements due to articulation of speech sounds, and voiceless parts with no F_0 values. These portions may cause errors in the F_0 model parameter extraction. Therefore, the pre-processing of the method includes operations for these portions: correction of gross errors, removal of microprosody, and interpolation of voiceless consonants. In the current experiments, the pitch extraction was done using an algorithm developed by the authors [Hirose, Fujisaki, and Seto (1992)].

The gross errors are defined as sudden and large F_0 shifts due to pitch extraction errors and irregularity in vocal folds vibration, often observable at the end of utterances. First, frames with the power lower than a threshold were assumed to be voiceless, even if they were judged voiced and had pitch values by the pitch extraction algorithm. Then, the isolated F_0 clusters whose F_0 values differ from those of other parts, than at the threshold, were assumed as gross errors and substituted by the interpolated values.

The consonant articulation influences the vocal folds vibration and causes sharp F_0 movements near some voiceless consonants, especially plosives. These so-called microprosodic F_0 movements are not included in the F_0 model and, therefore, should be removed. Since the mcroprosody for voiceless plosives appears as the sharp down-drift in F_0 contour into the succeeding vowels, it can be removed by a rather simple logic.

A voiceless portion without F_0 was interpolated by the 3rd order polynomial, whose coefficients were decided from F_0 values immediately before and after the portion. If a portion without F_0 had duration larger than $1/\alpha$ (0.33.s), it was assumed as a pause and no interpolation was conducted.

After these processes, the F_0 contours were smoothed out by a piecewise 3rd order polynomial function. The advantage of using 3rd order polynomial over using other curves is that the derivative can be given straightforwardly by mathematical calculation.

The accent commands are obtained from the 1st order derivative of the smoothed F_0 contour. Basically an accent command appears in the derivative as a set of peak and valley. However, when an accent component is immediately followed by another component with similar or larger amplitude, the valley will not appear clearly. An algorithm of accent command detection was developed taking these features into account. Since the timings of derivative peak and valley correspond to the middle of the F_0 upward and downward movements, respectively, and delay from the command onset and reset times by $1/\beta$, the command location can be decided from the derivative peak and valley points easily. The command amplitude is estimated from the peak and valley values.

The estimation of phrase commands were conducted for the F_0 contour residual, which was obtained by subtracting accent components generated from the extracted accent commands from the smoothed contour. The onset of the first phrase component is estimated from the timing of the first peak of the residual by shifting it backward by $1/\alpha$. The magnitude of the component was estimated through a successive approximation process. Then the estimated phrase component was subtracted from the residual F_0 contour to obtain a new residual F_0 contour, which was used to extract the second phrase command in a similar way. The process starts from the utterance initial and ends when it reaches the utterance final.

The commands extracted by the above process are used as the initial parameter values for the analysis-by-synthesis process to obtain the optimised command values.

4 Extraction of linguistic information from the text

The text analysis was conducted using Japanese parsers *Chasen* [http://chasen.aist-nara.ac.jp/] and KNP [http://www-nagao.kuee.kyoto-u.ac.jp/projects/nl-resource/]. *Chasen* divides the input sentence into morphemes with their readings, part-of-speeches, accent types, accent attributes, and so on. *KNP* detects *bunsetsu* boundaries using information obtained from another parser *Juman* (*Juman* is used, because *KNP* cannot handle the outputs from *Chasen* directly.). It also tells which *bunsetsu* directly modifies another *bunsetsu*, and thus gives us *bunsetsu* boundary depth information. The *bunsetsu* is defined as a basic unit of Japanese grammar and pronunciation consisting of a content word (or content words) followed or not followed by a function word (or function words), and its accent type is somewhat

different from those of consisting words. While the accent phrase coincides with *bunsetsu* in many cases, it is defined in a different way: a unit with one accent component (sometimes, accompanied by secondary accent components). Two or more *bunsetsu* can be an accent phrase and vice versa.

The large vocabulary speech recognition system *Julius* [http://winnie.kuis.kyoto-u.ac.jp/pub/julius/] was utilized to obtain phoneme boundary locations through the forced alignment process.

The accent type of each accent phrase was decided applying the accent sandhi rules to consisting morphemes. The rules are those originally developed by Sagisaka and Sato (1983) and modified though perceptual experiment by the authors [Minematsu, Kita, Hirose (2003)]. They include rules for a content word and a function word, two content words, an affix and a content word, and so on. The accent type of an accent phrase can be decided by referring to the accent types and attributes of the morphemes. Figure 2 summarizes these processes.

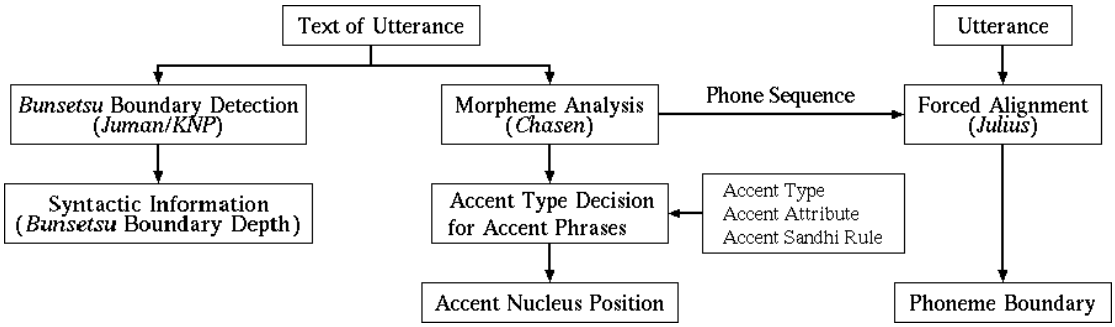

Fig. 2. Processes of extracting linguistic information of the text.

5 Relation of F_0 model commands and linguistic information

The speech corpus used for the experiment is the 15-minute recordings of a male announcer's speech from a radio program "From My Bookshelf." The relation between *bunsetsu* boundary depths and phrase commands was analysed for the corpus. The F_0 commands are those manually extracted, which are also used later as correct commands to evaluate the method. Table 1 summarizes the result according to the existence/absence of preceding pauses and boundary depth codes. They are obtained by *KNP* as shown in Fig. 3, which shows an example of parsing for the sentence "arayuru geNjitsuo subete jibuNno hoHe nejimagetanoda ([He] twisted all the reality to his side.)."

Table 1. Relation between manually-extracted phrase commands and linguistic information. The magnitudes and distances are shown as averages of extracted phrase commands. There are many cases without phrase commands at boundaries, especially when Boundary Depth Code is 1 and without pauses.

Pause at boundary	Boundary depth code	Magnitude	Distance between corresponding *bunsetsu* beginning (ms)
	Sentence beginning (F)	0.616	-251
No	1	0.293	-106
	2	0.303	-81
	3	0.363	-39
	4 or larger	0.389	-76
Yes	1	0.425	-200
	2	0.434	-170
	3	0.522	-192
	4 or larger	0.582	-193

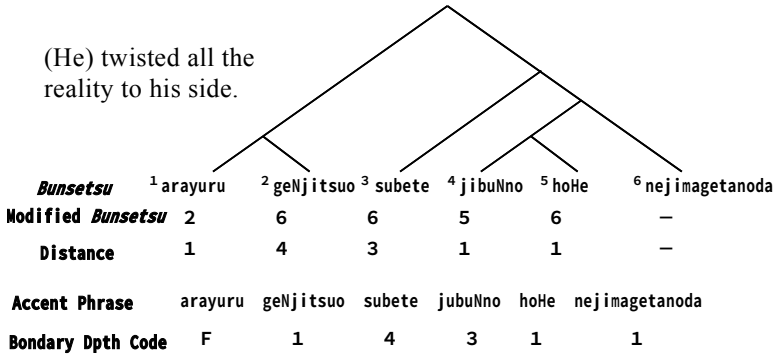

Fig. 3. Result of syntactic analysis by *KNP* and codes to represent *bunsetsu* boundary depth.

Table 2. Correspondence between manually-extracted accent commands and those estimated from the linguistic information.

		Type 1	Type 2	Type 0	Others
Number of commands		151	109	206	188
Onset	Number of corresponded commands	121	68	168	168
	Average distance from corresponding mora initial (ms)	-15.8	-78.3	-75.8	-74.5
Reset	Number of corresponded commands	124	68	130	153
	Average distance from corresponding mora end (ms)	55.8	19.1	-30.2	6.7

Also analysis on accent commands was conducted on the same corpus. Table 2 shows relation of the correct (manually-extracted) accent commands and those estimated from the linguistic information.

6 Modification of F_0 model commands using linguistic information (Developed method)

Based on the results in Section 5 and other knowledge on F_0 contour obtained in the former experiments, the following rules were developed for the modification of the extracted F_0 model commands:

1) When the extracted accent command onset (or reset) position coincide with that generated from linguistic information, and when the extracted accent command reset (or onset) position does not coincide, the reset (or onset) position is changed to the generated position. The generated position is obtainable from Table 2 by referring to the accent type.

2) When no accent command is extracted where the linguistic information requires an accent command, add a command with amplitude "average value minus a standard deviation."

3) When an accent command is extracted where the linguistic information requires no accent command, and when its amplitude is below a threshold, it is deleted.

4) When no phrase command is extracted at a *bunsetsu* boundary with a short pause or with the Boundary Depth Code equal to or larger than 2, a new phrase command is added. The magnitude of the phrase command is that listed in Table 1. When adding a new phrase command, the amplitude of the first accent command after the boundary is halved.

5) When a large and long-period accent command is extracted without phrase command immediately after a *bunsetsu* boundary, a phrase command is added. The magnitude of the phrase command is that listed in Table 1. When adding a new phrase command, the amplitude of the first accent command after the boundary is halved.

6) When a phrase command is extracted around the middle of a *bunsetsu*, it is deleted.

Using the commands modified by the above rules as initial parameter values for analysis-by-synthesis process, the final command values are obtained.

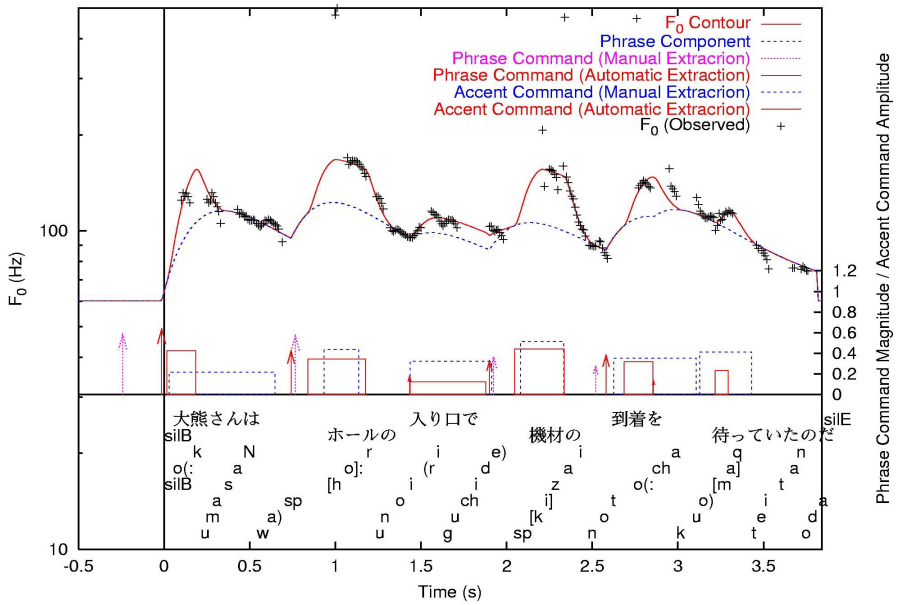

Fig. 4. F_0 command extraction by the original method for "oHkumasaNwa hoHruno iriguchide kizaino toHchakuo matteitanoda." (English translation of the sentence: Mr. Ohkuma had been waiting for the arrival of the machine parts at the hall entrance.)

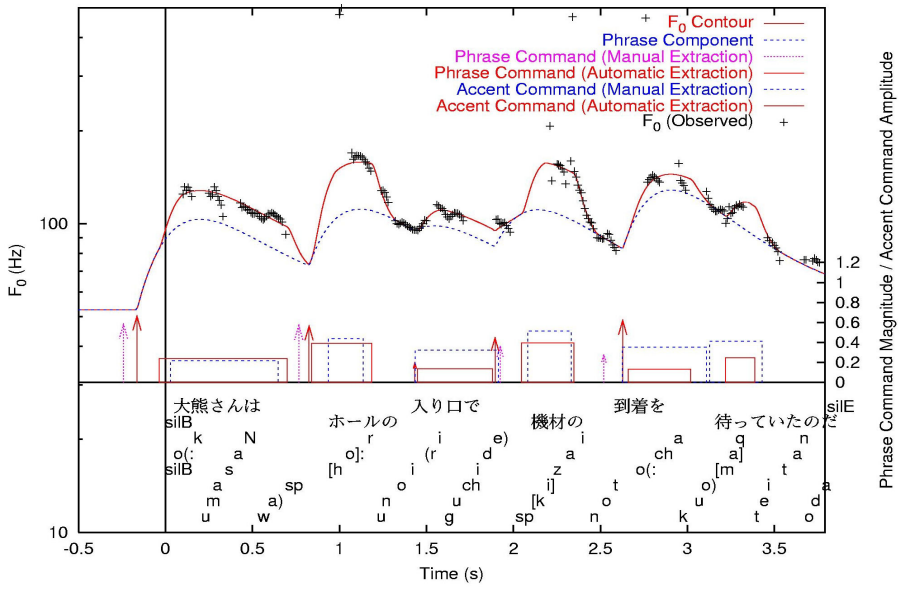

Fig. 5. F_0 command extraction by the developed method.

Figures 4 and 5 show the results of the automatic F_0 model command extraction by the original and developed methods, respectively. It is clearly shown that the extraction error at the sentence initial by the original method is corrected by the developed method using accent type information. In the case of Japanese, when a command of type 0 accent is located at the beginning part of an accent phrase, the

original method often cannot extract it correctly. Another example of this type of error correction by the developed method is observable at the accent phrase "toHchakuo matteitanoda," where an error in phrase command extraction is also corrected.

As the quantitative measures for command extraction performance, we adopted recall rate R, precision rate P, and F measure F, which are respectively defined as follows:

$$R = \frac{C}{C+D}, \qquad P = \frac{C}{C+I}, \qquad F = \frac{P \cdot R}{(P+R)/2}$$

where C, D, I respectively mean number of commands correctly extracted, number of deletion errors, and number of insertion errors. In this evaluation, the magnitude/amplitude of the command is not taken into account, because it can be modified through the analysis-by-synthesis process. Table 3 summarizes the result. It clearly shows improvements in the performance for the developed method.

Table 3. Comparison of F_0 model command extraction by the original and developed methods. The developed method utilizes linguistic information.

	Phrase command (396 commands)		Accent command (624 commands)	
	Original method	Developed method	Original method	Developed method
Number of correct commands	316	326	509	536
Number of deletion errors	80	70	115	88
Number of insertion errors	76	56	99	91
Recall rate R (%)	79.8	82.3	81.6	85.9
Precision rate P (%)	80.6	85.3	83.7	85.5
F measure F (%)	80.2	83.8	82.6	85.7

7 Discussion

As an index for the goodness of the extracted F_0 model commands, the mean square error between the F_0 contour extracted from the speech and that generated by the F_0 model using the extracted commands was calculated for the original and developed methods. It is interesting that the mean square errors are almost the same for both methods. This result implies that there is a certain limit in finding out the correct model commands only by the matching of F_0 contours; the linguistic information of the sentence is indispensable.

Since the developed method uses accent type information, which is specific for the Japanese language, it will not work for other languages as it is. By including such information specific for a language instead of accent type information, the method can work for the language. As already mentioned, there are many cases where the wrong extraction of type 0 accent command by the original method was corrected by the developed method. Since type 0 accent appears as a rather flat F_0 contour, its command extraction is difficult only from the F_0 contour derivatives. This feature is specific for Japanese, but there are similar cases for other languages, where accent command extraction becomes difficult because of flat F_0 contours.

8 Conclusion

Information on the F_0 model commands obtainable from linguistic information was successfully introduced to our method of automatic extraction of F_0 model commands. We are now applying the method to other corpora uttered by several speakers. Further study is also planned to introduce a statistical method, which becomes possible when the speech data with F_0 model commands increase.

The work is partly supported by Grant in Aid for Scientific Research of Priority Areas (#746).

References

Kikuchi, H. and K. Maekawa. 2002. Accuracy of prosodic labeling of spontaneous speech by X-JtoBI. *Record of Acoustical Society of Japan Fall meeting*: 259-260. (in Japanese)

Fujiaski, H. and K. Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of Acoustical Society of Japan (E)*, 5(4): 233-242.

Narusawa, S., N. Minematsu, K. Hirose, and H. Fujiaski. 2002. A method for automatic extraction of model parameters from fundamental frequency contours of speech. *Proc. IEEE ICASSP*, Orlando: 509-512.

Sakurai A. and K. Hirose 1999. Designing a parameter-based prosodic speech database. *Proc. Oriental COCOSDA Workshop*, Taipei: 5-8.

Hirose, K., H. Fujisaki, and S. Seto 1992. A scheme for pitch extraction of speech using autocorrelation function with frame length proportional to the time lag. *Proc. IEEE ICASSP*, San Francisco, 1: 149-152.

Nara Institute of Science and Technology, Morphological Analyzer ChaSen, http://chasen.aist-nara.ac.jp/.

Kyoto University, Japanese Syntactic Analysis System KNP http://www-nagao.kuee.kyoto-u.ac.jp/projects/nl-resource/.

Kyoto University, Large vocabulary continuous speech recognition decoder Julius, http://winnie.kuis.kyoto-u.ac.jp/pub/julius/.

Sagisaka, Y. and H. Sato 1983. Accentuation rules for Japanese word concatenation," *Trans. Institute of Electronics and Communication Engineers*, 66D(7): 849-856. (in Japanese)

Minematsu, N., R. Kita, K. Hirose 2003. Automatic estimation of accentual attribute values of words for accent sandhi rules of Japanese text-to-speech conversion. *Institute of Electronics, Information and Communication Engineers, Trans. Information and Systems*, E86-D(1): 550-557.