A LINGUISTIC AND PROSODIC DATABASE
FOR DATA-DRIVEN JAPANESE TTS SYNTHESIS

* †*Atsuhiro Sakurai*, †*Takashi Natsume, and* †*Keikichi Hirose*

*Tsukuba R&D Center, Texas Instruments
7 Miyukigaoka, Tsukuba, Ibaraki, 305-0841, Japan
†Dept. of Information and Communication Engineering, Univ. of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

ABSTRACT

We propose a method to generate a database that contains a parametric representation of F0 contours associated with linguistic and acoustic information, to be used by data-driven Japanese text-to-speech (TTS) systems.

The configuration of the database includes recorded speech, F0 contours and their parametric labels, phonetic transcription with durations, and other linguistic information such as orthographic transcription, part-of-speech (POS) tags, and accent types. All information that is not available by dictionary lookup is obtained automatically. In this paper, we propose a method to automatically obtain parametric labels that describe F0 contours based on a superpositional model. Preliminary tests on a small data set show that the method can find the parametric representation of F0 contours with acceptable accuracy, and that accuracy can be improved by introducing additional linguistic information.

1. INTRODUCTION

Generating prosodic features from text is one of the most difficult problems faced by current TTS systems. Most TTS systems generate prosodic features by rules, based on the linguistic information produced by a morpho-syntactic analyzer. However, making such rules is an extremely complex and human-dependent task. This has fostered research on data-driven TTS systems, which try to automatically create rules from large databases containing prosodic and linguistic information.

However, this approach invariably stumbles on the problem of creating a reliable and sufficiently large database that correctly associates linguistic information and prosodic features. Labeling guidelines have been proposed with the objective of providing the framework for creating databases for data-driven TTS systems like ToBI and J-ToBI [1][2], with promising perspectives. However, some problems can be pointed out:

- J-ToBI labeled databases are still rare, especially those containing lexical information such as POS tags and accent types;
- In J-ToBI, prosodic features are not represented in a parametric format that allows straightforward conversion to physical quantities. In view of that, some research effort has been dedicated to extracting parametric representations of F0 contours from J-ToBI labeled databases [3], with partial success.
- Detection of prosodic events (tone and break index tiers) is highly human dependent, in spite of the attempts to achieve automatic labeling.

In view of the problems mentioned above, we propose:

- To use public domain linguistic databases as the solution for the lack of linguistic information such as POS tags. Such linguistic databases are becoming increasingly available as important resources for natural language processing applications;
- To include an efficient parametric representation of F0 contours in the prosodic database using a superpositional model (F0 model for now on) [4], which can be easily converted to physical F0 contours;
- To propose a method for automatic extraction of F0 model parameters based on F0 contours, phonetic transcription and POS tags.

In the next sections, we present the overview of a linguistic and prosodic database that can be constructed with the present method. We describe how each kind of information is obtained, and then focus on the problem of automatic generation of F0 model parameters.

2. THE LINGUISTIC AND PROSODIC DATABASE

The outline of our method for automatic generation of a linguistic and prosodic database is sketched in **Figure 1**. The database is designed to include the following information:

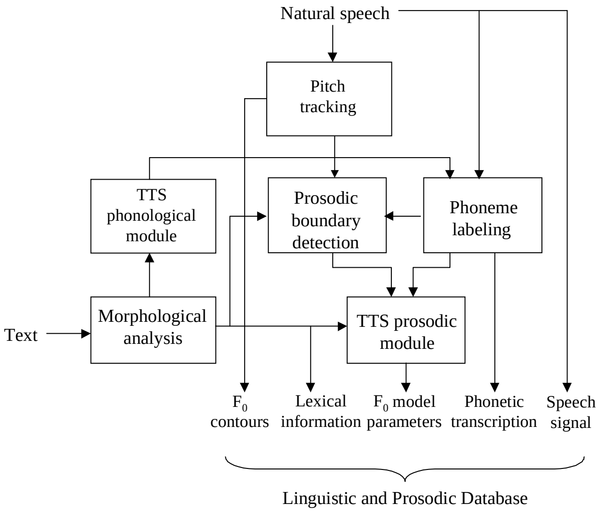

Figure 1: A system for automatic generation of a speech database containing linguistic information and prosodic labeling

- Speech signal: read from a text database, such as the Japanese Mainichi newspaper.
- F0 contours: automatically extracted [5] and automatically smoothed.
- Parametric representation of F0 contours using a superpositional model for F0 contour synthesis [4]: automatically detected (see next sections) and manually corrected.
- Phonetic transcription with durations: automatically obtained using HMM-based forced alignment [7]. The input to the HMM recognizer is driven by the phonological module of a TTS system, which performs grapheme-to-phoneme conversion based on lexical information.
- Lexical information: word-segmented Japanese texts, POS tags, and accent types. This information can be obtained automatically using a morphological analyzer [9] and an accent-type dictionary. In addition, public domain databases containing hand-corrected data can be used.

Among the items mentioned above, generating F0 model parameters is usually the most human-dependent task, since no method has been reported so far that gives a complete solution to the problem of automatically finding F0 model parameters given an F0 contour. Deriving these parameters from text is even more difficult - it represents the essential problem that TTS systems are supposed to solve. In [6], an interesting data-driven method is proposed for deriving F0 generation rules using a parametric representation of F0 contours, and a method is proposed for automatic calculation of F0 model parameters. However, the initial parameter values for the prosodic database must be assigned by hand. Here, we try to obtain these parameters automatically, using acoustic and linguistic information. The task consists of automatically detecting prosodic boundaries, and then obtaining a parametric representation of F0 contours using the F0 model. The prosodic module of a TTS system is used to generate this parametric representation. The method is described in the next sections.

3. AUTOMATIC GENERATION OF F0 MODEL PARAMETERS

3.1. Parametric Representation of F0 Contours

The F0 model is a superpositional model that expresses logarithmic F0 contours as the superposition of two types of components, generated by the response of a linear system to two different types of commands. One is an impulse-like command that triggers a relatively long rise-fall pattern of the F0 contour, which we denominate prosodic phrase. The other is a stepwise command that triggers a shorter rise-fall pattern of the F0 contour, which we call prosodic word. These commands are respectively called phrase and accent commands.

The problem of automatically deriving F0 model parameters from the F0 contour has been studied in the past with the objective of providing prosodic clues for syntactic parsing in speech recognition, with promising results. However, the performance obtained by such systems did not encourage practical implementation of a pre-processing module prior to speech recognition [10][8]. This is mainly due to the fact that the relation between syntactic structures and prosodic features is not well understood, and prosodic structures cannot be tightly associated with syntactic structure.

In this paper, we address the problem of generating F0 model parameters with the objective of constructing a database to be used by data-driven TTS systems. As a consequence, we are allowed to use other linguistic information (phonetic transcription and lexical information), in addition to acoustic features.

The problem of finding F0 model parameters is divided into two parts: detecting phrase and accent parameters. In this system, phrase and accent parameters are found separately using F0 contours, and the data are corrected by referring to the linguistic information mentioned above. After creating a complete model that includes phrase and accent parameters, the parameters are optimized using an analysis-by-synthesis scheme. In the subsections below, we describe the two processes involved in the determination of phrase and accent parameters: prosodic boundary detection and analysis-by-synthesis (AbS) optimization.

3.2. Prosodic Boundary Detection

The first step to determine phrase and accent parameters is to detect prosodic boundaries. In this paper, we classify prosodic boundaries into phrase and accent boundaries, which are defined here as boundaries between prosodic phrases and prosodic words. Sentence boundaries are not considered here, since they can be immediately detected from the pauses that precede them.

Phrase boundaries are detected using the method reported in [8], based on the decomposition of F0 contours into high-frequency accent components and low-frequency phrase components, using a low-pass filter that views F0 contours as signals in the time domain. Prosodic events corresponding to phrase commands are first detected, and then the delay with respect to segmental boundaries is compensated using a linear function that takes into account the derivative peak of the logarithmic F0 contour. For the present system, the cut-off frequency has been slightly increased (14 Hz) in order to reduce the number of deletions. The softer cut-off frequency could cause a higher insertion rate, but most insertions can be eliminated by checking phrase boundaries against segmental phoneme labels and POS tags.

After detecting the approximate position of phrase boundaries, they are approximated to the closest “bunsetsu” (Japanese syntactic unit containing a content word plus one or more particles) boundaries. Bunsetsu boundaries of the utterance are detected using segmental phoneme labels and POS tags.

Accent boundaries are also detected automatically using the method described in [11]. The positions of accent onset and reset are also corrected based on segmental phoneme labels and POS tags, with the following restrictions:

- Accent onsets are allowed to occur before the first mora (accent type 1) or between the first and second mora of the bunsetsu (other accent types).
- As a general rule, accent resets are allowed to occur at any mora boundary. However, in cases where the position of the accent nucleus can be promptly determined using simple rules, we also tried to place the accent reset command after the accent nucleus when the system fails to detect the position of the accent reset (see next section).

After phrase and accent boundaries are found, the prosodic module of a TTS system is used to assign initial amplitude and timing values for phrase and accent commands, using rules for TTS synthesis. These values, however, are greatly affected by the analysis-by-synthesis scheme that is carried out in order to adjust the parameters to the extracted F0 contour.

3.3. Analysis-by-Synthesis

Analysis-by-synthesis using the detected phrase and accent boundary positions is carried out in order to adjust the parameters to the extracted F0 contours. For the AbS process, phrase timing commands are allowed to oscillate ±80 ms around the initial values, and accent timing commands are given a variation range of ±40 ms (no more than the duration of one mora). Phrase magnitudes and accent amplitudes are allowed to vary from 0.0 to 0.60.

The initial values for timing commands are fixed based on the detected positions of phrase and accent boundaries and on offsets used by the TTS prosodic module: the offset represents the time lag from the command time to the actual segmental boundary. The TTS prosodic module uses offset values of 210 ms for phrase commands corresponding to prosodic sentence boundaries, 80 ms for other phrase commands, and 70 ms for accent commands.

In the AbS process, the parameters are adjusted by minimizing the root-mean-square error between the extracted F0 contour and the F0 contour generated using the parameters (in log scale). Since a limited variation range is given to the parameters in the AbS process, the final value of this difference (AbS error) is highly dependent on the initial configuration and initial values of the parameters. Thus, it is possible to use the AbS error to compare different configurations of F0 model parameters. One such example is shown in the next section.

4. EXPERIMENTAL RESULTS

We first carried out a preliminary test on 25 read-style speech samples extracted from the ATR continuous speech database, in order to evaluate phrase boundary detection. The system detected 32 out of 50 phrase boundaries corresponding to prosodic phrases, with 15 insertions. This score is acceptable, if we consider that phrase boundary assignment sometimes involves subjective decisions, and that mismatches do not necessarily represent failures, but in some cases denote that more than one configuration of phrase boundaries is possible.

Accent parameters represent, though, a tougher problem due to the larger number of parameters and the softer syntactic restriction of their occurrence. **Figure 2** shows an example of automatic detection of phrase and accent parameters when different levels of linguistic information are introduced. The sentence 'sono yookooni tsuite chotto otazune shitaindesukeredomo, yoroshii desuka?' ("I'd like to ask some questions about some important points, may I?") has been automatically analyzed using the method described in the previous section. In (a), accent command resets are allowed to occur at any mora boundary. In (b), however, lexical information related to accent type is introduced, and accent resets are allowed to occur only at the accent nucleus. The regions of the utterances between the vertical bars include the segment "tain", which contains an accent reset in (b), but not in (a). The figure shows that (b) represents a better parametric representation of the F0 contour. We also calculated the AbS errors: $9.6x10^{-2}$ for case (a), and $5.9x10^{-2}$ for case (b). This example shows how linguistic information can be used as a complement to acoustic information for automatic prosodic labeling.

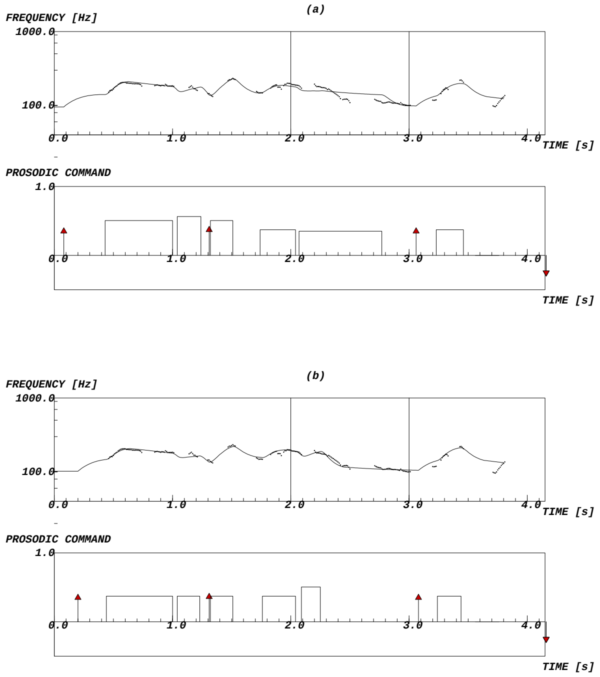

Figure 2: Extracted F0 contour (discontinuous line) and automatically generated F0 contour (continuous line), and their respective parametric representations for two cases: (a) without considering accent type, and (b) considering accent type.

5. CONCLUSION

A method to automatically generate a speech database that contains linguistic and prosodic information has been proposed and evaluated. The method includes automatic prosodic labeling of F0 contours using parameters based on a superpositional model for F0 synthesis. Preliminary tests show the validity of our method, even though some improvement is needed in the automatic detection of prosodic boundaries. For now on, we plan to study ways to further introduce linguistic information as a complement to the information extracted from prosodic features.

6. REFERENCES

1. Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., "TOBI: A standard for labeling English prosody," *Proc. ICSLP92*, pp.867-870, 1992.
2. Campbell, N. and Venditti, J., "J-ToBI: An intonation labelling system for Japanese," *Reports of Spring Meeting*, Acoust. Soc. Jpn., pp.317-138 (1995-9).
3. Hirai, T. and Higuchi, N., "Automatic extraction of the Fujisaki Model parameters using the labels of Japanese tone and break indices (J-ToBI) system," *Trans. IEICE*, Vol.J81-D-II, No.6, pp.1058-1064 (1998-6). (in Japanese)
4. Fujisaki, H. and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Jpn* (E), Vol.5, No.4, pp.233-242 (1984-10).
5. Hirose, K. Fujisaki, H. and Seto, S., "A scheme for pitch extraction of speech using autocorrelation function with frame length proportional to the time lag," *Proc. IACSSP'92*, 1, pp.149-152 (1992).
6. Hirai, T., Iwahashi, N., Higuchi, N., and Sagisaka, Y., "Automatic extraction of F0 control rules using statistical analysis," in *Advances in Speech Synthesis*, Springer, pp.333-346 (1996).
7. Young, S. et. Al., "The HTK Book, version 2.1", Cambridge University, 1996.
8. Sakurai, A. and Hirose, K., "Detection of phrase boundaries by filtering the fundamental frequency contour," *Proc. ICSLP'96*, vol.2, pp.817-820 (1996-9).
9. Matsumoto, Y., Kitauchi, A., et. al., "Japanese morphological analysis system ChaSen Manual," Nara Institute of Science and Technology, 1997.
10. Nakai, M., Singer, H., Sagisaka, Y., and Shimodaira, H., "Automatic prosodic segmentation by F0 clustering using superpositional modeling," *Proc. IEEE ICASSP'95*, Vol.1, pp.624-627 (1995-5).
11. Fujisaki, H., Hirose, K., and Seto, S., "A study on automatic extraction of characteristic parameters of fundamental frequency contours," *Proc. Fall Meeting of Acoust. Soc. Jpn.*, 2-6-18, pp.255-256 (1992-3). (in Japanese)