社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－ＨＩ－122 (18)
2007－SLP－ 65 (18)
2007／2／10

音声合成のための自動アクセントラベリング

立花 隆輝† 長野 徹† 倉田 岳人† 西村 雅史† 馬場口 登‡
† 日本アイ・ビー・エム東京基礎研究所 〒242-8502 神奈川県大和市下鶴間 1623-14
‡ 大阪大学大学院工学研究科 〒565-0871 大阪府吹田市山田丘 2-1
{ryuki,tohru3,gakuto,nisimura}@jp.ibm.com babaguchi@comm.eng.osaka-u.ac.jp

あらまし

人間の音声のみから、全モジュールについて統計的学習を行って自動的にテキスト音声合成システムが構築できれば、現在よりはるかに多様な声質の合成音声を日常の様々な場面で利用できるようになるだろう。そのためには音声に対してアクセントラベルの付与を自動的に行う必要がある。しかしアクセントの違いによる音響的特徴量の変化は微小であるため高精度の推定は従来困難であった。本論文では日本語のアクセント句境界推定とアクセント型推定の精度を改善するため、音響的モデルと言語的モデル、話者依存モデルと非依存モデルを組み合わせた利用を提案する。実験では、各モデルを独立して利用した場合と比較して、組み合わせた場合の精度が優れていることが確認できた。

キーワード テキスト音声合成 アクセント認識 アクセント句境界

Automatic Accent Labeling for a Text-to-Speech System

Ryuki TACHIBANA†, Tohru NAGANO†, Gakuto KURATA†,
Masafumi NISHIMURA†, Noboru BABAGUCHI‡
† Tokyo Research Lab., IBM Japan, 1623-14 Shimotsuruma Yamato Kanagawa 242-8502
‡ Graduate School of Engineering, Osaka University, 2-1 Yamadaoka Suita Osaka 565-0871
{ryuki,tohru3,gakuto,nisimura}@jp.ibm.com babaguchi@comm.eng.osaka-u.ac.jp

Abstract

If we could automatically build a text-to-speech (TTS) synthesis system by stochastically training every modules of the system only from the speech of a human, we would be able to use various synthetic voices in greater diversity of day-to-day situations. Automatic determination of the prosodic labels for the speech is necessary for this purpose. However, the subtleness of physical features makes accurate labeling difficult. In this paper, we propose a method that can accurately determine prosodic labels using both the acoustic and linguistic models, and using speaker-dependent and speaker-independent models. Our experiments on Japanese accent determination show the effectiveness of the combination of the models.

Key Words Prosody recognition, mora accent, prosodic phrase boundary, text-to-speech synthesis

1 Introduction

The recent improvements on text-to-speech (TTS) synthesis systems made it possible to reproduce the acoustic characteristics of human narrators by using stochastic training [1]. Many modules of the systems including prosody models and phonetic segment databases have become trainable based on the speech corpora.

However, there are two modules that cannot be automatically built in the conventional systems: the text processing module and the speech corpus itself.

For the text processing module, though it is still common to use rule-based approaches, it is costly to maintain numerous rules. As for the speech corpus, a large speech corpus with accurate prosodic labels is necessary for training of prosody models However, manual labeling of accurate prosodic labels is expensive and time-consuming.

Totally Trainable TTS (T^4S) System

To tackle these problems, we are working on a totally trainable TTS system every components of which including the text processing module can be automatically built from the speech corpus [2]. The system configuration is illustrated in Fig 1.

The build process of the system first obtains the alignments of the speech segments by using an automatic speech recognition technology. The obtained speech segments are stored in the speech segment database. The prosodic labels are then automatically estimated by analyzing the speech segments. The prosodic labels are used for training of the language models of the text processing module [3] as well as training of the prosody models. The trained modules are used in the runtime of the speech synthesis.

Among the modules of the whole system, we propose a novel approach for the automatic prosody labeling module in this paper. Since errors in the estimated prosodic labels result in poorness of the trained modules and unnatural synthetic sound, the accuracy of the estimation is important for the whole system.

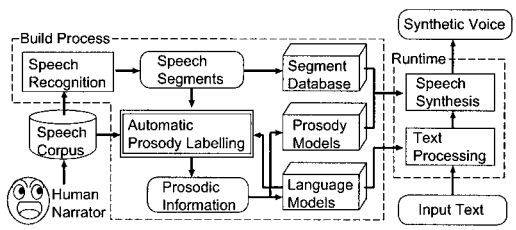

Figure 1: System configuration of the T^4S system. The double-lined square is the module described in this paper.

2 Automatic Prosody Labelling

Recent work in this field has gained significant insights into the challenges of automatic prosody labelling. Wightman et al. [4] showed consistent prosodic labels were determined by using a decision tree and a Markov chain model. Hirose et al. [5, 6] showed HMM-based prosodic pattern modeling can be used for prosodic phrase detection. The work by Chen et al. [7] showed that use of a linguistic model with an acoustic model can improve the accuracy. Though the accuracy of automatic labeling has approached the agreement rate of human labelers, the accuracy of accents is still below 90%. There is clearly room to improve this.

For further accuracy improvement, we propose explicit use of the prosodic structure of the language as constraints for the prosodic labels. Though the prosodic structure is handled by Ma et al. [8], this old approach simultaneously determines the multiple layers of the structure. In contrast, we split the labeling problem into multiple layers and address each of the layers from the top down. We use different acoustic models and linguistic models for different layers (Table 1). In addition, for maximum leverage over the linguistic constraints, the proposed method uses a speaker-independent model only for the linguistic model of the accent determination. This speaker-independent model is devoted to describing knowledge about the possible correct accentuations in the language, and this is not dependent on the individual speakers. This combination of models makes the best use of the prosodic structure of the language in a proactive manner and enables accurate accent determination without requiring large speaker-dependent training corpora (which are costly and time consuming to supply).

2.1 The Prosodic Structure of Japanese

In this paper, we assume the prosodic structure of Japanese as illustrated in Fig 2. That is, a sentence utterance consists of intonational phrases

Table 1: Four models

	Acoustic model	Language model
Prosodic phrase boundary detection	Speaker-dependent GMM	Speaker-dependent decision tree
Accent determination	Speaker-dependent decision tree	Speaker-independent n-gram

(IPs), which are separated by periods of silence. An IP consists of prosodic phrases (PPs). A PP is a group of words that are uttered in a prosodic combination. We assume that the accent type of an N-mora PP can be either one of the accent type 0 to the accent type $(N-1)$. The phonemes in a word are grouped into morae. A mora consists of one or zero consonants and a vowel, and is a phonetic unit similar to a syllable.

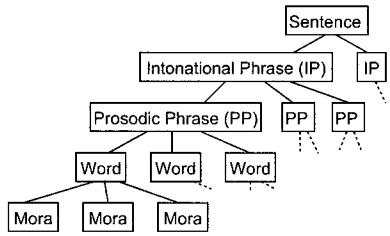

Figure 2: Assumed prosodic structure of Japanese

3 Method

The proposed method hierarchically determines the mora accents of the input utterance using the following steps: (1) Obtain word boundaries and the part-of-speech (POS) of each word by analyzing the sentence with a morphological analyzer [3], (2) Align the phonemes by using an ASR-based phoneme alignment tool, (3) Separate the utterance into IPs at the pauses, (4) Separate each of the IPs into PPs (Section 3.1), and (5) Determine the accent type for each of the PPs (Section 3.2).

3.1 Prosodic Phrase Boundary Detection

The objective in this layer is to determine the PP boundaries among all of the word boundaries in the given IP. The IP boundaries are excluded from the candidate list for the PP boundaries. We let the word sequence of the IP $\mathbf{W} = (w_1 w_2..w_l)$, where w_i denotes the i-th word of the IP and l is the number of words in the IP. $\mathbf{B} = (b_1 b_2..b_{l-1})$ is a sequence of the locations of the PP boundaries, where $b_i = 1$ denotes the presence of a PP boundary just after w_i. The other possible value of b_i is 0, used if there is no PP boundary at that location. $\mathbf{V} = (v_1 v_2..v_{l-1})$ is a sequence of the acoustic feature vectors observed at the word boundaries. The objective of can be restated as a search for the $\mathbf{B}$ that maximizes the conditional probability for given $\mathbf{W}$ and $\mathbf{V}$.

$$\mathbf{B}_{max} = \underset{\mathbf{B}}{\mathbf{argmax}}\, P(\mathbf{B}|\mathbf{W},\mathbf{V}) \quad (1)$$

$$= \underset{\mathbf{B}}{\mathbf{argmax}}\, \frac{P(\mathbf{V}|\mathbf{W},\mathbf{B})P(\mathbf{B}|\mathbf{W})}{P(\mathbf{V}|\mathbf{W})} \quad (2)$$

$$= \underset{\mathbf{B}}{\mathbf{argmax}}\, P(\mathbf{V}|\mathbf{W},\mathbf{B})P(\mathbf{B}|\mathbf{W}), \quad (3)$$

where '**argmax**' is the operator that returns the value of the argument that maximizes the following term. $P(\mathbf{V}|\mathbf{W})$ can be ignored when we only want to find $\mathbf{B}_{max}$. $P(\mathbf{V}|\mathbf{W},\mathbf{B})$ can be obtained by using the acoustic boundary model, while $P(\mathbf{B}|\mathbf{W})$ is a linguistic probability calculated by using the linguistic boundary model. Speaker-dependent models are used for these models because PP formation is dependent on the speaker's style. Since the presence of a PP boundary at a word boundary has an effect on the neighboring word boundaries, we search for $\mathbf{B}_{max}$ by using the Viterbi algorithm.

3.1.1 Acoustic Boundary Model

We ignore $\mathbf{W}$ in $P(\mathbf{V}|\mathbf{B},\mathbf{W})$ and approximate its value by a purely acoustic probability $P(\mathbf{V}|\mathbf{B})$ focusing on the presence of the boundary. In addition, we ignore the effect of the neighborhood as follows.

$$P(\mathbf{V}|\mathbf{B}) \simeq \prod_{i=1}^{l-1} P(v_i|b_i). \quad (4)$$

$P(v_i|b_i)$ is calculated by using multivariate Gaussian Mixture Models (GMM) trained by using the training corpus.

The feature vector, v_i, is a three dimensional vector whose components are (1) the change of the logarithmic fundamental frequency (F0) in the preceding mora, (2) the logarithmic F0 gradient in the following mora (g_2 in Fig. 3), and (3) the change of the logarithmic F0 gradient at the point ($g_2 - g_1$ in the figure). As shown in the figure, a minimum point near the word boundary and maximum points in the neighboring morae are searched for when calculating these features.

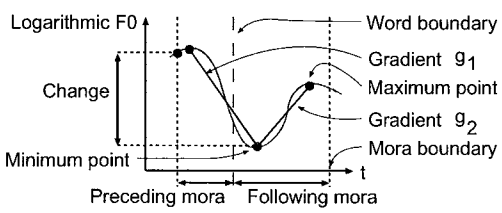

Figure 3: Features for acoustic boundary model

3.1.2 Linguistic Boundary Model

The linguistic probability is calculated by using the following equations.

$$P(\mathbf{B}|\mathbf{W}) = P(b_1, .., b_{l-1}|\mathbf{W}) \quad (5)$$

$$= P(b_1|\mathbf{W}) \prod_{i=2}^{l-1} P(b_i|b_1, .., b_{i-1}, \mathbf{W}) \quad (6)$$

$$\simeq P(b_1|w_1, w_2) \prod_{i=2}^{l-1} P(b_i|b_{i-1}, w_i, w_{i+1}) \quad (7)$$

The conditional probability $P(b_i|b_{i-1}, w_i, w_{i+1})$ is calculated from a decision tree trained by using the learning corpus. The reason for the approximation in the equation is that we found that b_{i-1} and the information on the nearest words (w_i and w_{i+1}) were the only important factors. The part-of-speech is used for the information of a word.

3.2 Accent Determination

The objective in this layer is to determine the accent sequence, $\mathbf{A} = (a_1 a_2 .. a_m)$, for the given PP, where a_i has a value of `H` or `L`. Since the scope of this section is limited to the PP, we use $\mathbf{W} = (w_1 w_2 .. w_n)$ as the word sequence of the PP and the $\mathbf{V} = (v_1 v_2 .. v_m)$ as the sequence of the acoustic feature vectors. The value of n is the number of words in the PP, and $m = \sum_{i=1}^{n} m_i$ is the total number of morae in the PP, where m_i denotes the number of morae in the word w_i. The objective can be restated as

$$\mathbf{A}_{max} = \underset{\mathbf{A}}{\mathbf{argmax}}\, P(\mathbf{A}|\mathbf{W}, \mathbf{V}) \quad (8)$$

$$= \underset{\mathbf{A}}{\mathbf{argmax}}\, P(\mathbf{V}|\mathbf{W}, \mathbf{A}) P(\mathbf{A}|\mathbf{W}), \quad (9)$$

where $P(\mathbf{V}|\mathbf{W})$ is ignored again. $P(\mathbf{V}|\mathbf{W}, \mathbf{A})$ and $P(\mathbf{A}|\mathbf{W})$ are calculated by using the acoustic model and the linguistic model of this layer, respectively. We can obtain $\mathbf{A}_{max}$ by simply comparing $P(\mathbf{V}|\mathbf{W}, \mathbf{A})P(\mathbf{A}|\mathbf{W})$ for all of the cases of $\mathbf{A}$, since the number of possible sequences for $\mathbf{A}$ is only m.

While a speaker-dependent model is necessary for the acoustic model for handling the F0 contour, we use a speaker-independent model for the linguistic model. This is because the variety of words is huge and it is impossible to learn the linguistic probabilities of accent types for words only from a small speaker-dependent corpus. Another reason is that there are indeed speaker-independent linguistic constraints on "correct" accent sequences.

3.2.1 Acoustic Accent Model

The components of the feature vector v_i are (1) the normalized logarithmic F0 at the beginning of the current mora, (2) the normalized logarithmic F0 change in the current mora, and (3) the logarithmic F0 gradient in the current mora (Fig. 4). We approximate $P(\mathbf{V}|\mathbf{W}, \mathbf{A})$ using the multiplication of $P(v_i|\mathbf{W}, \mathbf{A})$ calculated by using a decision tree trained with the training corpus. A multivariate GMM is trained for each of the leaves of the tree.

$$P(\mathbf{V}|\mathbf{W}, \mathbf{A}) \simeq \prod_{i=1}^{m} P(v_i|\mathbf{W}, \mathbf{A}) \quad (10)$$

$$\simeq \prod_{i=1}^{m} P(v_i|a_{i-1}, a_i, m, i, (m-i)) \quad (11)$$

For $\mathbf{W}$ and $\mathbf{A}$, the necessary input variables for the tree are the accents of the previous mora (a_{i-1}) and the current mora (a_i), the number of morae in the PP (m), and the distance to the PP beginning (i), and the distance to the PP end ($m - i$).

The normalized logarithmic F0 ($\tilde{F0}$) is the logarithmic F0 normalized to fall in the range of [0,1] according to

$$\tilde{F0} = \frac{F0 - F0_{min}}{F0_{max} - F0_{min}}, \quad (12)$$

where $F0_{min}$ and $F0_{max}$ are the minimum and maximum logarithmic F0s in the PP, respectively.

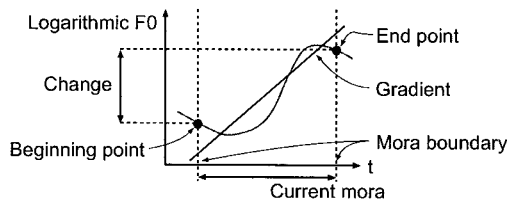

Figure 4: Features for acoustic accent model

3.2.2 Linguistic Accent Model

The linguistic prosody $P(\mathbf{A}|\mathbf{W})$ is obtained by using stochastic accent estimator [3] with the following normalization:

$$P(\mathbf{A}|\mathbf{W}) = \frac{P'(\mathbf{A}, \mathbf{W})}{\sum_{\mathbf{A}} P'(\mathbf{A}, \mathbf{W})}, \quad (13)$$

where the summation of the denominator is used for the m possible accent sequences.

$$P'(\mathbf{A}|\mathbf{W}) = \prod_{i=1}^{n} P(\mathbf{A}_i|\mathbf{A}_1, .., \mathbf{A}_{i-1}, \mathbf{W}_1, .., \mathbf{W}_i), \tag{14}$$

where $\mathbf{A}_i$ is the accent sequence of the i-th word, and is a part of $\mathbf{A}$. In other words, if the indices of the first and last mora of the word $\mathbf{W}_i$ are j and k, respectively, $\mathbf{A}_i$ is $(a_j, .., a_k)$. The conditional probability $P(\mathbf{A}_i|\mathbf{A}_1, .., \mathbf{A}_{i-1}, \mathbf{W}_1, .., \mathbf{W}_i)$ is calculated using the stochastic accent estimator.

The stochastic accent estimator is trained by using a large speaker-independent corpus.

4 Experiments and results

We conducted experiments to evaluate the performance of the proposed method. In the experiments, we compared different combinations of the components of the method.

4.1 Corpus

The speech corpus we used in the experiments is a reading of excerpts of the ATR phonetically balanced text corpus and consists of 503 sentences [9]. We used 100 sentences for training the speaker-dependent models and for testing. Each sentence in the corpus was segmented into words and each word was manually annotated with its POS, its phoneme sequence, its accent sequence, and its PP boundaries.

The speech data was recorded by an adult female using a laryngograph and a microphone. The F0 contours were obtained by smoothing the pitch mark periods obtained from the laryngograph. The statistics of the test corpus are shown in Table 2.

Table 2: Statistics of the test corpus

# of sentences	100
# of intonational phrases	377
# of prosodic phrases	686
# of words	1,813
# of morae	3,342
# of morae with an H accent	1,729 (51.7%)

4.2 Compared methods

We compared the following three combinations for PP boundary detection. The combinations are also shown in Table 3.

BA, BL and **BAL** The PP boundaries are detected by using either one or both of the acoustic and linguistic boundary models. **BAL** is our new proposed method for PP boundary detection.

We compared the following nine combinations for accent determination.

BN-TA, BN-TL and **BN-TAL** The mora accents were determined by using either one or both of the acoustic and linguistic accent models with a single-layer approach. That is, PP boundary detection was not done. The accent sequences of the PPs were determined from all of the possible combinations.

BC-TA, BC-TL and **BC-TAL** For given correct PP boundaries, the mora accents were determined by using either one or both of the acoustic and linguistic accent models.

BAL-TA, BAL-TL and **BAL-TAL** Based on PP boundaries detected by **BAL**, the mora accents were determined by using either one or both of the acoustic and linguistic accent models. **BAL-TAL** is our new proposed method.

Table 3: The experimental results of the compared combinations. N, C, A, and L stand for "not-used", "correct", "acoustic", and "linguistic", respectively. While the performance numbers for PP boundary detection is the F measure, that for accent determination are mora accuracy (%).

	PP boundary				Accent		Accuracy
	N	C	A	L	A	L	
BA			√				0.657
BL				√			0.781
BAL			√	√			0.862
BN-TA	√				√		62.4
BN-TL	√					√	85.5
BN-TAL	√				√	√	69.7
BC-TA		√			√		84.5
BC-TL		√				√	89.7
BC-TAL		√			√	√	94.6
BAL-TA			√	√	√		84.1
BAL-TL			√	√		√	87.8
BAL-TAL			√	√	√	√	92.7

4.3 Results

The results of the PP boundary detection are shown in the upper part of Table 3. The accuracy is shown in the F measure. Note that the IP boundaries were ignored in the calculations of these values. For **BAL**, the accuracy of detecting PP boundaries was 93.7%.

It can be seen that using both the acoustic and linguistic models (**BAL**) produced the best results. The poor precision of the acoustic model (**BA**) is an intrinsic problem of this model. This is because the acoustic features observed at a non-PP-boundary word boundary next to a real PP boundary are sometimes very similar to those observed at the real PP boundary, especially when the word sandwiched between these boundaries is very short. For example, postpositionals such as "wa", "ga", and "o" have only one mora.

The results of accent determination are shown in the lower part of Table 3. The mora accuracy is expressed as percentages.

Again, the combination of the acoustic and linguistic models showed the best performance (**BAL-TAL** > **BAL-TL** and **BAL-TA**, and **BC-TAL** > **BC-TL** and **BC-TA**). In addition, we can see the effectiveness of the proposed layered approach by comparing the result of the proposed method (**BAL-TAL**) and that of the non-layered approach (**BN-TAL**). Though the errors in PP boundary detection resulted in a 1.9% decrease for accent determination accuracy (**BAL-TAL** < **BC-TAL**), the combined result is still over 90% and is better than the other approaches, the text-only processing (**BN-TL**) and the non-layered approach (**BN-TAL**).

5 Conclusion

In this paper, we proposed an automatic accent labeling method that makes the best use of the prosodic structure of the language by combining the acoustic and linguistic models, and the speaker-dependent and speaker-independent models. The method showed 92.7% mora accuracy using trained speaker-dependent models with a speaker-dependent training corpus containing only 100 sentences.

The combinations of the acoustic models and the linguistic models resulted in the best performance compared to the other models. Though the errors of the prosodic phrase boundary detection affected the accuracy of the accent determination in the subsequent stage, we confirmed that the separation of the problem into multiple layers was effective even with this amount of errors.

References

[1] E. Eide, A. Aaron, R. Bakis, P. Cohen, R. Donovan, W. Hamza, T. Mathes, M. Picheny, M. Polkosky, M. Smith, and M. Viswanathan, "Recent improvements to the IBM trainable speech synthesis system," in *Proc. of ICASSP*, 2003, pp. I–708–I–711.

[2] M. Nishimura, R. Tachibana, T. Nagano, and G. Kurata, "A study on totally trainable TTS system," in *Fall Meeting of Acoustic Society of Japan*, 2006, pp. 237–238, (in Japanese).

[3] T. Nagano, S. Mori, and M. Nishimura, "A stochastic approach to phoneme and accent estimation," in *Proc. of INTERSPEECH*, September 2005, pp. 3293–3296.

[4] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 469–481, Oct 1994.

[5] K. Hirose and K. Iwano, "Accent type recognition and syntactic boundary detection of japanese using statistical modeling of moraic transitions of fundamental frequency contours," in *Proc. of ICASSP*, May 1998, pp. I–25–I–28.

[6] K. Hirose and K. Iwano, "Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition," in *Proc. of ICASSP*, June 2000, pp. III–1763–III–1766.

[7] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," in *Proc. of ICASSP*, May 2004, pp. I–509–I–512.

[8] X. Ma, W. Zhang, Q. Shi, W. Zhu, and L. Shen, "Automatic prosody labeling using both text and acoustic information," in *Proc. of ICASSP*, April 2003, pp. I–516–I–519.

[9] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, "Speech database user's manual," Tech. Rep., ATR TR-I-0166, 1990.