Which Resemblance is Useful to Predict Phrase Boundary Rise Labels for Japanese Expressive Text-to-speech Synthesis, Numerically-Expressed Stylistic or Distribution-based Semantic?

Hideharu Nakajima, Hideyuki Mizuno, Osamu Yoshioka, Satoshi Takahashi

NTT Media Intelligence Labs., NTT Corporation, Yokosuka, Kanagawa, JAPAN

Abstract

To establish Expressive Text-to-speech synthesis, current research studies both the processing of input text and the rendering of natural expressive speech. Focusing on the former as a front-end task in the production of synthetic speech, this paper investigates a novel feature for predicting phrase boundary tone labels which transcribe local fundamental frequency (F0) changes frequently appearing at phrase end positions in expressive speech. To this end, we examined a kind of distribution-based semantic features consisting of i) word surface strings, ii) their part-of-speech tags taken from a phrase and iii) the pause existence/non-existence at the final position of the phrase, which are different from conventional numerically-expressed stylistic features such as positions and lengths and distances of the phrase. Through experiments on Japanese expressive speech such as conversational speech and advertisement speech, we confirmed that the proposed features attain performance equal to or better than conventional features. These results suggest that the distribution-based semantic features might be useful to predict phrase boundary rise labels for conversational speech and might be useful equal to conventional numerically-expressed stylistic feature for advertisement speech.

Index Terms: Expressive text-to-speech synthesis, feature engineering, prosody, tone label, phrase boundary rise

1. Introduction

New application domains of text-to-speech (TTS) synthesis are being targeted to move beyond the conventional fields of news reading and providing information. These new domains include voice-assisted e-commerce, call center automation, and storytelling for entertainment [1]. Due to this expansion, human-like expressive speaking style is more strongly demanded instead of the neutral reading style used in news reading. In this context, prosodic features are known to play an important role in reproducing expressive speech and this paper focuses on textual features to predict expressive fundamental frequency (F0).

One of the important characteristics of expressive speech is the variability of F0 behavior at the ends of phrases. The canonical shape of reading style F0 is a rise followed by a decline to the right phrase edge in Tokyo Japanese, because most sentences are declarative when providing information. In expressive speech, some F0 samples rise again at the end of the phrase even though the sentence is not a question. Called "boundary pitch movement (BPM)" [2] or "phrase boundary tone", this pattern starts the "rise" near the end of the phrase and sometimes repeats "fall" and "rise", resulting in such various behavior as "rise", "rise and fall", and "rise and fall and rise". Phrase boundary tone conveys a pragmatic meaning beyond the meaning of words in the phrase, for example, the speaker's intentions as regards utterance continuation and a request for the listener's agreement; it also plays a role in dialogues ("dialogue function"). The following two sentences (ex.1 and 2) are taken from [2] and they contain identical segments in italic forms, but are spoken with different phrase boundary tone; ex.1 cued by "?" shows phrase boundary tone which rises to a high level, ex.2 cued by "!" shows phrase boundary tone which rises sharply to a mid-level. Moreover, there is another example; ex.3 is spoken with no apparent rise in phrase boundary tone.

(ex.1) *sô na no*? asking "Is that true?"
(ex.2) *sô na no*! insisting "It's true!"
(ex.3) *sô na no*. disappointed

The three sentences convey different pragmatic meanings, "asking", "insisting", and "disappointed" as shown to the right of each sentence. As the existence and kind of phrase boundary tone are not always specified by the above punctuation symbols (?, ! and .), phrase boundary tone itself or symbols representing phrase boundary tone should be predicted from the text to be synthesized.

The F0 contour including phrase boundary tone is symbolized by the tone labels of "Tone and Break Indices (ToBI)" [3]. Although there are several studies on the generation of F0 contour from ToBI labels (with HMM speech synthesis [4], with command and response model [5], and with regression models [6]), there have been relatively few studies on the prediction of symbol labels of F0 as ToBI from text. This is because, for Tokyo Japanese, text processing of TTS can predict and has predicted the tone labels or equivalent information except for the phrase boundary tone labels from the knowledge of canonical F0 form and the accent positions if present.

Therefore, in order to realize expressive text-to-speech synthesis, the remaining phrase boundary tone labels expressing BPM should be predicted from the text, even if any marks as in the above examples are not specified in the text. To the best of our knowledge, no study has examined phrase boundary tone label prediction for expressive speech. Conventional prediction studies targeted English radio news reading style speech [7]. They use numerical and categorical features to predict if F0 rises again or not at the end of each phrase (H% or L% in ToBI). As one of the categorical features, part-of-speech (POS) tags were used with other features as summarized in [7]. However, the word surface string has not been used. The importance of word surface string for Japanese has already been revealed in conventional analytical studies (e.g., [8]). That paper summarized the three tuple relations; *phrase end word surface string, dialogue function and phrase boundary tone*. It showed that different phrase end words have different phrase boundary tones, even though the words have the same POS tags. Thus, us-

ing only POS tags might not be sufficient to distinguish phrase boundary tones. Moreover, POS may not be able to realize the dialogue function either. This is because dialogue functions may be decided by the contents of the message and the contents are represented by words, not POS. Although the recognition of such pragmatic meaning as dialogue function (asking, insisting, utterance continuation, etc.) might be useful to our goal, no accurate recognizer has been established. Instead of a symbol to represent pragmatic meaning, recent computational linguistics studies use both i) distribution-based information such as word vector and ii) distributional/semantic similarity for processes such as word sense disambiguation and text classification[9, 10, 11].

This paper studies, as a first step, the prediction of phrase boundary rise/fall labels, by treating, as rise, all patterns starting with "rise" such as "**rise**", "**rise** and fall", and "**rise**, fall and rise" and by treating, as fall, those starting with "fall" at phrase ends. We also focus on prediction features such as distribution-based semantic features with the expectation that distribution-based semantic features can accurately distinguish phrase boundary tone and that the features capture some form of pragmatic meaning.

2. Distribution-based semantic feature and prediction

This research aims to clarify which is more useful in predicting phrase boundary tone labels, conventional numerically-expressed stylistic features or distribution-based semantic features.

2.1. Label prediction as a classification task

We consider label prediction to be the classification of phrase end boundary tone into two categories; i) rise again (as H% in ToBI), ii) no rise (as L%). This classification is executed at the end of each *accent phrase (or accentual phrase)* that corresponds to "minor phrase" [12]. This is because accent phrase occupies the lowest level among hierarchical prosodic structures [12], and because the numbers are small but a few phrase boundary tones occur at accent phrase end as 3.2 shows later.

We use Classification and Regression Tree (CART) as a machine-learning based classifier. CART is constructed from training data which consists of i) the correct labels showing rise or not and ii) the input features explained in the following section 2.2.

2.2. Distribution-based semantic features

As distribution-based semantic features, we use i) several *word surface strings* taken from the end to the head of an accent phrase, ii) their *POS tags*, and iii) *pause existence/non-existence at the end of the accent phrase*. These features yield a feature vector for prediction. The vector, especially components of word surface strings and their tags, can be expected to express the meaning of the phrase. Thus, when predicting phrase boundary rise/fall label for a new phrase (usually called "test data"), the classifier predicts the label based on the vector similarity between training data and test data.

It is not necessary for prediction that the input feature vector be completely the same as the vector in CART training data. Even if only some portion of the vector is similar to the training data, that portion does contribute to the final judgment. This is because CART first locates useful features at the node near the root of the tree. Thus, we expect that data sparseness, for example around rare proper nouns, does not degrade prediction accuracy since compensation is provided by the surrounding words in the distribution-based semantic feature.

To focus on the phrase boundary rise label prediction, we make domain-specific and speaker-specific classifiers. One reason of domain-specificity is that tone can be domain (situation) specific. The following example, taken from a conventional analysis study [13], explains that when the word "anata" is uttered with intention of "calling on" in *street domain (situation)*, the rising tone is used as in

$anata\nearrow$, $kasawo\ wasurete\ imasuyo$

(="you left your umbrella"),

but in *home domain (situation)* with the same "calling on" intention, the rising and falling tone is used as in

$anata\nearrow\searrow$, $okite\ kudasai$

(= "darling, isn't it time to get up?").

This example also implies that pragmatic meaning cannot be a deciding factor for phrase boundary rise prediction because both cases have the same intention of "calling-on". Also one reason of speaker-specificity is that though the prediction target is a symbol transcribing the form of F0, the phrase boundary tone itself is F0 and relates to the speaker's physical characteristics.

3. Expressive speech database

3.1. Domain and size

We used an expressive speech database [14] that includes natural prosodic speech gathered in the following two domains: i) telephone call center and ii) advertisements for mass media such as television. This paper refers to the call center data subset as **op**erator's speech "OP", and that of sales pitches for **ap**pealing products "AP". The sentences are real sentences used in both domains.

All speech is uttered in a communicative way as people talk to people. AP data includes instances of bright greetings, commodity explanations, cheerful sharing of feelings about the commodity, appealing to buy, and cheerful acknowledgement; OP data includes bright greetings, confirming call reasons, providing information, cheerful acknowledgements, serious-sounding apologies, and closing thanks. Shifting the perspective on emotion, cheerful and bright expressions in the AP and OP domain can be regarded as *joy*, serious-sounding expressions in the OP domain as *sad*. Thus these data can be regard as expressive speech.

The prediction targets in this research are the phrase boundary rise labels in BPMs. This database includes these labels at each rising position, for example,

OP) $\sim\ nan\ desu\ ga\nearrow$ (= "Speaking about $\sim$"), presenting discussion topic

AP) $Kantan\ de\ syo\nearrow$ (= "it's easy, isn't it?"), asking agreement

In addition to the expressive style speech, this database also includes neutral reading style speech. Three female professional narrators used both reading and expressive styles to utter all recorded sentences in each domain. The number of recorded sentences differed with the speaker, speaking style, and domain. To allow comparison of speakers and speaking styles, we used the utterances commonly recorded for all speakers and in both speaking styles for each domain. Table 1 shows statistical summaries of this expressive speech database. Numbers of accent phrases and total lengths of speech are average values among the three narrators, because these numbers differed with the speaker even in the same number of sentences. By comparing expressive F0 and reading F0 of the same sentence, different

Table 1: *Summary of expressive speech database.*

domain	OP	AP
# of sentences	104	152
# of accent phrases	1,061	1,550
# of speakers	3	3
Total length [min.]	14	20

Table 2: *Occurrence ratio [%] of phrase boundary rise labels.*

domain	OP		AP	
style	expressive	reading	expressive	reading
w/o pause	0.7	0.1	0.1	0.1
with pause	16.6	4.2	6.4	2.7
all	6.8	1.8	2.7	1.2

phrase boundary tone behavior was observed under the condition that both utterances (two styles) put accent boundaries and pauses at the same positions [15].

3.2. Occurrence ratios of phrase boundary rise labels

Table 2 shows the occurrence ratios [%] of phrase boundary rise labels in the database shown in Table 1. Ratios are listed instead of frequencies to allow comparison of different domains. "expressive" denotes expressive style and "reading" reading style. "all" row shows the occurrence ratios of phrase boundary rise labels among all accent phrases in each domain and style; "with pause" and "w/o pause" rows show occurrence ratios among the phrase which is or is not followed by a pause.

As the "all" row of Table 2 shows, phrase boundary rise labels occur in a few percent of all accent phrases. Occurrence ratios are higher in expressive style than in reading style. Moreover, occurrence ratios are doubled or tripled in the accent phrase with pause, roughly meaning that intonational phrases have high ratios. Hence, the existence or non-existence of pause might be one useful prediction feature. This coincides with the use of "depth of phrase break" in conventional prediction research [7].

We also used the Corpus of Spoken Japanese (CSJ)[16], a publicly available dataset, to compare the occurrence ratio of phrase boundary rise label in interview data in CSJ (18 to 21%) with the ratio in the conversational domain of OP in our database (16.6%). Since the ratios were almost the same, we believe that the proposal made in this paper has generality.

3.3. Classification potential of features

Here, we examine the classification potential by calculating the information gain (mutual information) offered by various features. Information gain (IG) is defined as the difference between entropy without condition (denoted as H(Y), where Y is class) and conditional entropy (denoted as H(Y|X), where X is classification feature and Y is class); IG=H(Y)-H(Y|X). Large information gain means high classification potential.

The feature X includes i) word surface string, POS, and the combination of word surface string and its POS, which yield distribution-based semantic features, and ii) conventional numerically-expressed stylistic features used in [7]. As the conventional features were originally designed for English, we replaced some of them with their Japanese equivalents. The total number of features examined was about 50. Distance, location, length were measured in number of words and mora. The "mora" is a unit often used in Japanese phonology. A Japanese morphological analyzer, JTAG, was used for word segmentation and part-of-speech tagging [17]. The output of the tagger has been corrected manually to focus this research on clarifying useful feature for phrase boundary rise label prediction.

Table 3: *Entropy and information gain (as classification potential).*

speaker id		#1	#2	#3
domain = **OP**				
Entropy		0.48	0.27	0.30
IG	POS	0.22	0.15	0.16
	word	**0.27**	**0.20**	**0.22**
	word_POS	**0.30**	**0.21**	**0.22**
	dist. to next *InP* (in *mora*)	0.15	0.05	0.06
	dist. to next *InP* (in word)	0.15	0.05	0.06
	dist. from last *AcP* (in word)	0.11	0.03	0.03
	dist. from last *prom.* (in mora)	0.05	0.03	0.03
	AcP length (in word)	0.08	0.03	0.03
	AcP end pause	0.14	0.04	0.05
	AcP end word length (in *mora*)	0.01	0.03	0.01
	InP location in utterance	0.00	0.01	0.00
domain = **AP**				
Entropy		0.17	0.20	0.18
IG	POS	0.12	0.17	0.13
	word	**0.14**	**0.18**	**0.15**
	word_POS	**0.15**	**0.19**	**0.15**
	dist. to next *InP* (in *mora*)	0.03	0.04	0.03
	dist. to next *InP* (in word)	0.03	0.04	0.03
	dist. from last *AcP* (in word)	0.03	0.05	0.04
	dist. from last *prom.* (in mora)	0.01	0.02	0.01
	AcP length (in word)	0.03	0.05	0.04
	AcP end pause	0.03	0.03	0.03
	AcP end word length (in *mora*)	0.01	0.01	0.01
	InP location in utterance	0.01	0.01	0.01

Table 3 shows the entropy and information gain (IG) after conditioning by each feature (as X) for each speaker. Because the number of phrase boundary rise labels differs with the speaker, even for the same number of sentences in the same domain, the entropy values without feature conditioning are also different. In Table 3, "word" means "word surface string", "POS" means "part-of-speech", "word_POS" means the combination of word surface string and its POS, "*dist.*" means "distance" of the accent phrase whose phrase boundary label is judged, "*prom.*" means "prominence", "*InP*" means "intonational phrase (major phrase [12]), " *AcP*" means "accent phrase", and "*InP* location in utterance" is one of three (first, middle, and last) locations in the utterance where the intonational phrase (which includes the accent phrase whose phrase boundary label is judged) belongs to. The *word* and *word_POS* features are newly introduced distribution-based semantic features and others are Japanese equivalents of those used in conventional research [7]. Though other conventional numerically-expressed stylistic features [7] were also examined, Table 3 only lists the features whose information gains are larger.

As shown in Table 3, conventional numerically-expressed stylistic features exhibited much lower information gain values than word and POS features. Among distribution-based semantic features such as "POS", "word" and "word_POS", features relating to "word" showed higher classification potential (i.e., information gain) than "POS".

4. Experiments

4.1. Experimental procedure and conditions

This section confirms the contribution of distribution-based semantic features to the prediction of phrase boundary rise labels.

Table 4: *Experimental setting.*

setting id	feature
c_1	distribution-based semantic features (word_POS)
c_2	distribution-based semantic features (word, POS)
c_3	conventional features

This experiment considered three settings based on the combination of features shown in Table 4.

Settings c_1 and c_2 are used to confirm the performance of our distribution-based semantic feature proposal. In summary, distribution-based semantic features includes followings;

1. the existence/non-existence of pause at the end of the accent phrase
2. N word surface strings and their POS tags taken from the end to the head of the accent phrase

Thus, Japanese accent type was not used as distribution-based semantic features. When the number of words is fewer than 'N' in the phrase, the shortfall is rectified by padding with "NULL" labels. Word surface string and its POS are combined as one feature in setting c_1, and separately used as two features in setting c_2.

Setting c_3 is used to confirm the performance of the conventional features and classifier [7]. Thus, we use "Wagon"[18] as CART for all three settings to compare features. As the conventional features were originally designed for English [7], we replaced some of them with their Japanese equivalents in the same way used for classification potential calculation in section 3. The conventional method [7] takes features from the next phrase where the phrase boundary rise/fall is predicted. All data was divided into 5 subsets and 5-fold cross validation was conducted. Domain and speaker dependent classifiers were constructed using 4 subsets as training data. All performance characteristics were calculated from test data not used in any training phase.

4.2. Evaluation measure

We use Cohen's Kappa value [19, 20] as the evaluation measure. Cohen's kappa value is the agreement ratio between correct answers and predicted answers. This measure is appropriate because it evaluates both rise and fall labels. As Table 2 shows, the case that F0 does not rise is much more frequent than the case that F0 rises. This measure can eliminate the bias of these amounts. This value ranges from -1 to 1 and a value near 1 means strong agreement. When the value is over 0.6, the two sequences can be regarded as exhibiting "agreement in practice".

4.3. Results and discussion

As the entropy-based criterion is commonly used for CART construction and tree pruning is also often used to avoid overtraining, we examined the use of both, however, the result was a severe drop in performance. Thus, the results were obtained by CART constructed with count-based criteria and without pruning.

Cohen's Kappa values are plotted in Fig 1. White bar charts were obtained by setting c_1, slash-hatched bar charts were obtained by setting c_2, gray bar charts were obtained by setting c_3. The nine bar charts on the left hand side are for the OP domain, and the remaining nine are for the AP domain. A bar chart is shown for every speaker from #1 to #3 for each domain. Vertical thin line bars indicate the 95% confidence intervals.

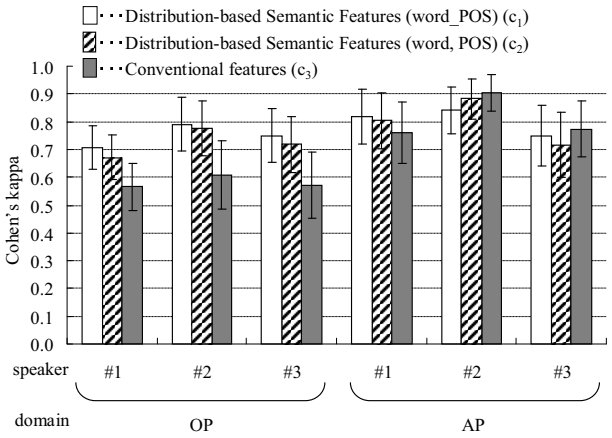

Figure 1: *Prediction performance measured with Cohen's kappa.*

In the OP domain, the distribution-based semantic feature proposal (c_1 and c_2), which introduces word surface strings, was better with 5% level statistical significance (white and slash-hatched bar charts), than the conventional feature without word surface strings (gray bar charts, c_3). These improvements coincide with the large information gain values shown in Table 3 obtained by introducing word surface strings as in distribution-based semantic features. Also, the proposal yielded kappa values above 0.6, that is, the use of distribution-based semantic features yielded "agreement in practice". However, as phrases of ex.1 to 3 in section 1, the meaning of the phrase, whose boundary tone label is judged, is sometimes decided by previous and next phrases. Future work includes the expansion of the phrase region to take into account this characteristic.

In the AP domain, there was no statistically significant difference between proposed and conventional features, high agreements were obtained by three settings. As shown in Table 3, entropy values themselves were low (about 0.2) before the introduction of any feature, the introduction of word surface strings did not give large information gain nor label prediction improvement from the performance obtained with POS.

5. Conclusions

This paper investigated the use of distribution-based semantic features for predicting phrase boundary rise labels. The use of distribution-based semantic features for such prediction is new and represents a departure from conventional numerically-expressed stylistic features. Experiments confirmed that our distribution-based semantic features can accurately predict phrase boundary rise. We measured the agreement ratio between correct answers and prediction results by Cohen's kappa. The use of distribution-based semantic features improved the kappa value in the conversational domain (OP in our database) with statistical significance, compared to the conventional numerically-expressed stylistic features. In the advertising domain (AP in our database), the use of distribution-based semantic features yielded higher kappa values but there was no statistical significant difference between the proposed and conventional features. This result suggests that the distribution-based semantic features might be useful in predicting phrase boundary rise labels for conversational speech as shown with OP data and might match conventional numerically-expressed stylistic features for advertisement speech.

6. References

[1] Tang, H., Zhou, X., Odisio, M., Hasegawa-Johnson, M., and Huang, T. S., "Two-stage prosody prediction for emotional text-to-speech synthesis", *Interspeech*, 2138–2141, 2008.

[2] Venditti, J. J., Maeda, K., and van Santen, J. P. H., "Modeling Japanese Boundary Pitch Movements for Speech Synthesis", *3rd ESCA/COCOSDA Workshop on Speech Synthesis*, 317–322, 1998.

[3] Maekawa, K., Kikuchi, H., Igarashi, Y., and Venditti, J., "X-JToBI: an extended J-ToBI for spontaneous speech", *ICSLP*, 1545–1548, 2002.

[4] Koriyama, Nose, T., Kobayashi, T., "On the Use of Extended Context for HMM-based Spontaneous Conversational Speech Synthesis", *Interspeech*, 2657–2660, 2011.

[5] 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 Jpn*, Vol.J81-D-II, No.6, 1058-1064, (in Japanese), 1998.

[6] Black, A. W. and Hunt, A. J., "Generating F_0 contours from ToBI labels using linear regression", *ICSLP*, 1385–1388, 1996.

[7] Ross, K. and Ostendorf, M., "Prediction of abstract prosodic labels for speech synthesis", *Computer Speech and Language*, vol.10, no.3, 155–185, 1996.

[8] Ishi, C. T., "The Functions of Phrase Final Tones in Japanese: Focus on Turn-Taking", *Journal of the Phonetic Society of Japan*, vol.10, no.3, 18–28, 2006.

[9] Wikipedia, "Distributional semantics". Online: http://en.wikipedia.org/wiki/Distributional_semantics, accessed on 22 Mar., 2013.

[10] Schütze, H., "Distributional semantics", 2011. Online: http://www.ims.uni-stuttgart.de/lehre/teaching/2011-SS/stats/distsem.pdf, accessed on 22 Mar., 2013.

[11] Manning, C. and Schuetze, H., *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

[12] Pierrehumbert, J. B., and Beckman, M.E., *Japanese Tone Structure*, Linguistic Inquiry Monograph Fifteen, The MIT Press, 263, 1988.

[13] Kubozono, H. and Tanaka, S., *Introduction to Japanese Pronunciation - Theory and Practice*, Kuroshio, (in Japanese), 1999.

[14] Nakajima, H., Miyazaki, N. Yoshida, A. Nakamura, T. and Mizuno, H., "Creation and Analysis of a Japanese Speaking Style parallel Database for Expressive Speech Synthesis", *Oriental COCOSDA*, paper-id=30, 2010. Online: http://desceco.org/O-COCOSDA2010/proceedings/paper_30.pdf, accessed on 22 Mar., 2013.

[15] Nakajima, H., and Sagisaka, Y, "F0 analysis for Japanese conversational speech synthesis", *8th Symposium of Natural Language Processing* (in IEEE Xplore), 137–142, 2009.

[16] Corpus of Spontaneous Japanese. Online: http://www.ninjal.ac.jp/csj/, accessed on 22 Mar., 2013.

[17] Fuchi, T., and Takagi, S, "Japanese morphological analyzer using word co-occurrence-JTAG-", *Coling-ACL*, 409–413, 1998.

[18] Edinburgh Speech Tools Library. Online: http://festvox.org/docs/speech_tools-1.2.0/book1.htm, accessed on 22 Mar., 2013.

[19] Cohen, J., "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, vol.XX, no.1, 37–46, 1960.

[20] Carletta, J., "Assessing Agreement on Classification Tasks: The Kappa Statistics", *Computational Linguistics*, vol.22, no.2, 249–254, 1996.