Emphasized Accent Phrase Prediction from Text for Advertisement Text-To-Speech Synthesis

Hideharu Nakajima, Hideyuki Mizuno, and Sumitaka Sakauchi
NTT Media Intelligence Labs., NTT Corporation,
1-1 Hikarino-oka, Yokosuka, Kanagawa, 239-0847 JAPAN

Abstract

Realizing expressive text-to-speech synthesis needs both text processing and the rendering of natural expressive speech. This paper focuses on the former as a front-end task in the production of synthetic speech, and investigates a novel method for predicting emphasized accent phrases from advertisement text information. For this purpose, we examine features that can be accurately extracted by text processing based on current Text-to-speech synthesis technologies. Among features, the word surface string of the main content and function words and the part-of-speech of main function words in an accent phrase are found to have higher potential on predicting whether the accent phrase should be emphasized or not through the calculation of mutual information between emphasis label and features of Japanese advertisement sentences. Experiments confirm that emphasized accent phrase prediction using support vector machine (SVM) offers encouraging accuracies for the system which requires emphasized accent phrase locations as context information to improve speech synthesis qualities.

1 Introduction

The introduction of corpus-based speech synthesis methods such as unit selection synthesis ((Hunt, et al., 1996) etc.) and Hidden Markov Model speech synthesis ((Zen, et al., 2009) etc.) makes expressive speech synthesis possible if an adequate speech database is prepared. However, the synthesized speech often fails to recreate emphasis or phrase boundary tone, even though both are key characteristics of expressive speech. The location markers of emphasis and phrase boundary tone have been confirmed useful in improving expressive speech synthesis qualities; they form part of the context information for speech synthesis (Meng, et al., 2012; Maeno, et al., 2014; Strom, et al., 2007; Yu, et al., 2010).

For establishing Text-To-Speech (TTS) synthesis for expressive speech, it is necessary to predict locations of emphasis and phrase boundary tone *from the input text*. The phrase boundary tone occurs at the phrase end, and existence/non-existence of the tone can be accurately classified, from the text to be synthesized, by using machine learning approaches (Nakajima, et al., 2013; Ross, et al., 1996). Thus, this paper focuses on the remaining target of emphasis positions. In this work, we use the word "emphasis (emphasized)" to denote portions that are perceptually more salient to the listeners in a sentence.

In human speech, emphasis can be regrouped at least into four functions based on analysis in conventional literatures as (Hovy et. al, 2013; Sridhar et. al, 2008) (bold portions show emphasized words and phrases).

1. expressing linguistic "focus": (e.g., " **Taro** did." (as an answer to "who did ...?"))
2. expressing "contrast": (e.g., "not A **but B**")
3. expressing "element of surprise": (e.g., "I heard he was sick, but he had **much energy**.")
4. disambiguating grammatical structure: clarifying parallel and dependency structure (e.g., to distinguish "{old men} and women" from "**old** {men and women}" in "old men and women")

28th Pacific Asia Conference on Language, Information and Computation pages 170–177

This paper focuses on items 1 to 3. For the purpose of establishing TTS for expressive speech, item 4, structural disambiguation, is hard to resolve when the text has ambiguities. On the other hand, it is not a problem when there is no ambiguity; the prosodic structure can be accurately fixed by following the clear structure.

Emphasis on location of focus, contrast, and element of surprise (items 1 to 3) are related to the novelty status of the information to be conveyed; status is normally obtained from the context. In the conversation domain, conversation history is the previous context. Consider, for example, the example of item 1. The query "who?" is answered by "Taro", which is new information to the questioner and is often focused on and emphasized in the responder's speech. In the story telling domain, the sentences before the current sentence form the context, and are the source for judging the novelty status of information in the current sentence.

In some domains, however, the previous context does not always exist, for example, as in sales pitches or advertisements in mass media services. Sales pitch sentences are composed by copywriters based on their belief of what consumers will find newsworthy and only the sentences are read aloud and broadcasted. The sentence does not include the background that copywriters considered before fixing the sales pitch. Thus, narrators, actors/actress, directors, or producers decode the sales pitch sentence to extract which portions should be emphasized when read aloud. This suggests that it is possible to predict emphasized portions from the words of the sentence being synthesized.

This paper focuses on emphasis in Japanese advertisement sentences and defines accent phrases as the prediction unit, while words have been used as the unit for predicting emphasis in the conversation domain (Hovy et. al, 2013). Exclamation marks are one of the characters indicating emphasis in written texts; they are often observed in advertisement sentences and must be a good cue for emphasis prediction. The expressive speech database, explained in Section 2, includes examples of Japanese emphasized words (in bold style) with exclamation marks ('_' denotes word delimiter and translations are indicated by parentheses):

ex.1) その_**前**_に! (**before** that!)
ex.2) **楽しめ**_る! (you can **enjoy**!)
ex.3) 110_種類_**以上**! (**more than** 110 types!)
ex.4) 水換え_**不要**! (**don't need** water exchange!)

The words immediately before exclamation marks are not always emphasized as in the Japanese word sequences of ex.1 and 2. However, the marks must have influence on emphasized words beyond their intermediate neighbors. As units longer than words might effectively include this long distance influence and accent phrases are one of the important units for Japanese speech synthesis and some studies on Japanese speech synthesis have adopted accent phrases as a unit of emphasis and confirmed improvements in speech wave generation (Maeno, et al., 2014), we adopt accent phrases as the prediction unit as well.

This paper proposes a method for predicting emphasized accent phrases from sales pitch sentences to establish expressive TTS. As far as we know, this is the first paper that proposes the emphasis prediction from Japanese sales pitch sentences and adopts accent phrases as the prediction unit. Section 2 describes the expressive speech database used in this paper. Section 3 analyzes the distributions of emphasized accent phrases in terms of linguistic expressions and their locations in both sentences and intonation phrases. Section 4 explains our method of predicting emphasized accent phrases and its experimental confirmation.

2 Expressive speech database

2.1 Target domain

This paper targets sales pitch texts for expressive speech synthesis. Given the increase of Internet-oriented advertisements, it is essential to establish technologies that can convert advertisement text to speech with emphasis in the appropriate positions to ensure that the advertisements reach the consumers.

As ambiguous and misleading messages are not suitable as advertisements, we can expect that sales pitch texts do not include ambiguities, and so we can focus research efforts on emphasis prediction. Sales pitch texts are written in Japanese and are Japanese sentences collected from advertisement pages on the Internet (Nakajima, et al., 2010). These include expressions that appear frequently in sales as "発売中 (now on sale)" and "〜円 (Yen)" and describe impressions and explanations of commercial products.

Table 1: Emphasis labels

	accent phrase base count
emphasized	853
not-emphasized	1,506
	word base count
emphasized	1,010
not-emphasized	4,727

2.2 Emphasis labels

Although human annotators can tag speech data with emphasis labels, research has showed little agreement between human annotators (Hovy et. al, 2013), and thus prediction targets cannot be fixed. As a practical solution, we asked one human subject to act as a recording director and decide emphasized accent phrases with the guideline that "labels are put at accent phrases that tend to be emphasized in commercial message conveyed through mass media."

The sales pitch database (Nakajima, et al., 2010) includes 248 utterances, which are divided into 363 texts (hereafter, sentences) by punctuation marks, and include 2,359 accent phrases as in Table 1. Emphasis labels were assigned to 853 accent phrases (36.2% of all accent phrases) as shown in Table 1. As 89% of the labels coincided with the labels set by at least one of the 3 annotators (based on listening to speech data), the labels extracted from the text are considered appropriate as emphasized labels. As reference, we also labeled emphasized words in the emphasized accent phrases as in Table 1.

2.3 Features for analysis

As this study focuses on features contributing to emphasis prediction, we added correct linguistic features as follows: *word boundaries, part-of-speech (POS), accent phrase (AP) boundaries, pause positions*. These features can be accurately extracted by text processing modules in conventional TTS. The number of POS and lemma (Fuchi, et al., 1998) were 62 and 1,571, respectively.

We also automatically extracted, from above features, *main content and function word in each accent phrase* by rules frequently used in Japanese dependency parsing studies ((Imamura, et al., 2007) etc.). We also used these features in defining the portion between pauses as "intonation phrase (IP)", and entered the following binary information:

- *whether the IP is at the sentence end or not*,
- *whether the AP is at the end of IP or not*,
- *existence/non-existence of exclamation marks, punctuation marks and pause at the end of the AP*.

By predetermined table look up, we also added

- *existence/non-existence of expressions on commercial products' information, evaluation, and prices in the AP*, and
- *existence/non-existence of sales-appeal words and qualifying words in the AP*.

Each word in the utterance including multiple sentences is examined if the word is mentioned in previous sentences in the utterance and

- the *existence/non-existence of words showing newness in the AP*

are added as another feature. Above features can be accurately assigned automatically because ambiguities are small. While semantic roles were used in (Hovy et. al, 2013), they are not used in our research, because automatic semantic role labeling is still immature and its accuracy remains insufficient and because our aim is to establish TTS and requires mature text processing.

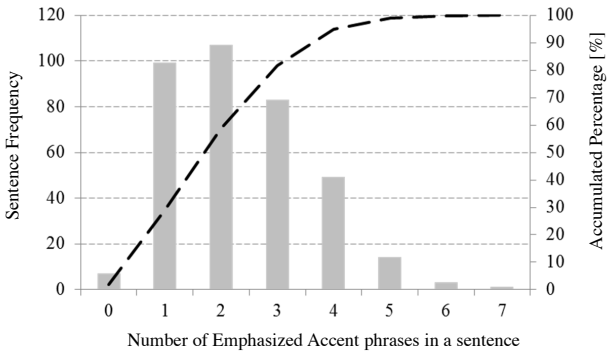

Figure 1: Sentence frequency associated with number of emphasized accent phrases in a sentence.

3 Emphasized accent phrase distributions

As shown in Fig.1, about 70 percent of the sentences in the database have more than 2 emphasized accent phrases. Unlike conversation (Hovy et. al, 2013), sales pitch speech synthesis requires the extraction of multiple emphasized accent phrases per sentence.

With a view to identify phrase location, emphasized accent phrase distribution is summarized in Table 2. Rows differ based on whether IP is emphasized (Emphasized IP (E-IP) or Not Emphasized

Table 2: Distribution of emphasized accent phrases (IP=Intonation Phrase, AP=Accent Phrase, NE=Not Emphasized, E=Emphasized, F=Final, NF=Not Final), bold phrases in samples are emphasized accent phrases in both Japanese and translations

Location			IP ratio (%)	E-AP ratio (%)	Samples
NE-IP			21.6	0	
E-IP			78.4	100	
	NF-IP	NF-AP		26.1	… すぐに / 仕上げて …
					(… **soon** / do it up …)
		F-AP		16.5	… コレステロールが / **高めの方** …
					(… cholesterol / **person indicating higher** …)
	F-IP	NF-AP		20.5	**効果的に** / コリを / **ほぐして** / くれます
					(**effectively**/stiffness/**flexed**/will be)
		F-AP		36.8	… 乾燥肌で / **泣かないで！**
					(… dry skin **/do not cry!**)

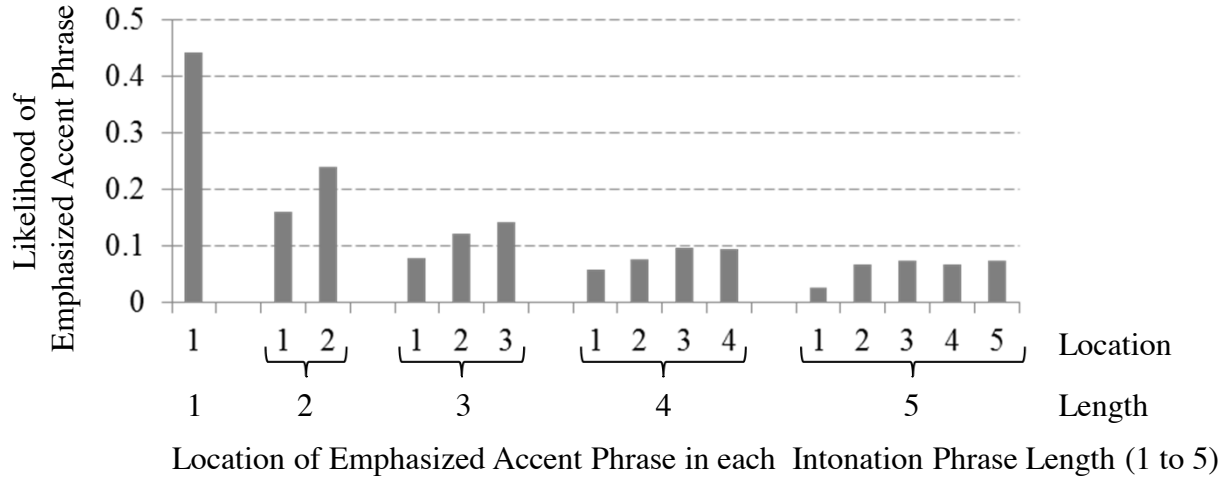

Figure 2: Likelihood of emphasized accent phrase by location in intonation phrase and its length.

IP (NE-IP)), whether IP exists at the end of sentence (Final IP (F-IP) or Not Final IP (NF-IP)), and whether AP exists at the end of IP (Final AP (F-AP) or Not Final AP (NF-AP)). Sample accent phrases are written in Japanese and divided by '/' and English translations for each accent phrase are written and divided by '/' in parentheses. The row of E-IP (Emphasized Intonation Phrase) shows that 78.4% of IPs have at least one emphasized AP.

The breakdown of E-IP lies in the four rows at the bottom of Table 2; the shares do not differ significantly (26.1, 16.5, 20.5 and 36.8 %). For detailed analysis, Fig.2 summarizes the likelihood of emphasized accent phrase by location in and length of intonation phrase whose lengths range from 1 to 5 (5 clusters correspond to length of intonation phrase). Upper number on the x axis denotes the location of emphasized accent phrase in each intonation phrase length. The larger the number is, the later in the intonation phrase does the emphasized accent phrase exist. Though later accent phrase locations showed higher likelihood of emphasized accent phrase, the likelihood values do not differ significantly. Thus, we decided to use *whether the IP is at the sentence end or not* and *whether the AP is at the end of IP or not* as location features in emphasized accent phrase distribution analysis.

We also measured the distance between two adjacent emphasized accent phrases; results are summarized in Fig. 3. 90% of emphasized accent phrases occurred within 0 to 4 accent phrases from the previous emphasized location. Thus, at most, the former

Table 3: Prediction potential

	Entropy $H(Y)$	0.94
1	Word surface string of the main content word in the AP	0.64
2	Word surface string of the main function word in the AP	0.15
3	Part-of-speech of the main function word in the AP	0.12
4	Whether the IP is at the sentence end or not	0.07
5	Existence/non-existence of exclamation marks at the end of the AP	0.07
6	Existence/non-existence of sales-appeal words in the AP	0.05
7	Existence/non-existence of expressions on commercial products' evaluation in the AP	0.05
8	Part-of-speech of the main content word in the AP	0.04
9	Whether the AP is at the end of IP or not	0.02
10	Existence/non-existence of pause at the end of the AP	0.02
11	Existence/non-existence of expressions on commercial products' information in the AP	0.01
12	Parallel structure	0.01
13	Existence/non-existence of punctuation marks at the end of the the AP	0.01
14	Existence/non-existence of expressions on commercial products' prices in the AP	0.01
15	Contrast structure	0.005
16	Existence/non-existence of words showing newness in the AP	0.001
17	Existence/non-existence of qualifying words in the AP	0.0006

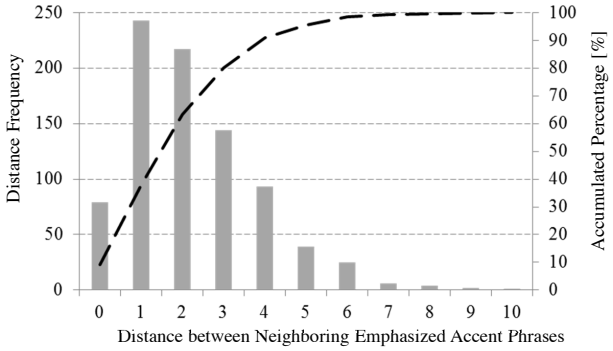

Figure 3: Distance between adjacent emphasized accent phrases.

4 and latter 4 accent phrases of the accent phrase might be a sufficient feature scope for emphasized accent phrase prediction.

To identify the promising features for emphasized accent phrase prediction, we also calculated the prediction potential of features (locations of accent phrases and linguistic expressions) based on the mutual information between those features and emphasis labels. Since the numbers of words and POS are large, we used the mutual information instead of the likelihood shown in Fig.2. When Y denotes emphasis label (emphasis or not), X each feature expression, $H(Y)$ entropy of Y, and $H(Y|X)$ is the conditional entropy of Y given X, then mutual information is calculated as $H(Y) - H(Y|X)$. The higher the mutual information value is, the greater is the contribution to emphasis prediction.

Table 3 lists prediction potentials in descending order with the first row showing entropy $H(Y)$. As the ratio of emphasized AP to not emphasized AP was almost 1 to 2, $H(Y)$ was 0.94 which is very high. Middle column in Table 3 lists the feature expressions mentioned so far and rightmost column shows mutual information values as prediction potential.

Word surface string of the main content word in the AP and *word surface string and part-of-speech of the main function word in the AP* showed higher mutual information (0.64, 0.15, 0.12, respectively) and are expected to contribute to emphasized accent phrase prediction. In the database, accent phrases accompanying exclamation marks at the end of the accent phrase are emphasized except for one sample, but too many accent phrases without the mark are emphasized, thus the mutual information was small (0.07). Though we also examined other binary features as "whether $\cdots$" and "existence/non-

Table 4: Range of parameters

Parameters	Range
dimension of polynomial kernel	1 to 4
cost of polynomial kernel	1 to 3
location index of features	-4 to 4
location index of past prediction results	-3 to -1

existence of $\cdots$" in Table 3 to confirm their contribution to prediction performance and the generality of features, their mutual information values were also small.

4 Emphasized accent phrase prediction

4.1 Prediction method

As more than 2 accent phrases are emphasized in an advertisement sentence as shown in Fig.1, we decided that the proposed method predicts multiple emphasized accent phrases in a sentence. As there are features that had few samples but whose probabilities are higher like exclamation marks, we consider emphasized accent phrase prediction as a classification problem between the existence/non-existence of emphasis. We used support vector machines (SVM) as classifiers and the features in Table 3 to establish and test the emphasized accent phrase prediction method.

4.2 Experimental conditions

The expressive speech database mentioned in section 2 were used for training and evaluating the SVM in 5-fold cross validation way. We used the polynomial kernel function of SVM and examined several parameter combinations of the kernel function (dimension and cost). Table 4 summarizes parameters and ranges. The dimension and cost are integers.

Others are indexes showing locations of accent phrases. 'i' denotes the location index of the accent phrase to be classified to emphasized or not, '-m' the location index of 'm' preceding accent phrase from i and 'n' the location index of 'n' following accent phrase from i. As we can use only past prediction results, maximum integer is '-1' for the location index of past prediction results.

For later description and discussion, F_{i-n}^{i+m} denotes the features between $(i-n)$ and $(i+m)$ locations, H_{i-n}^{i-h} the history of past prediction results between $(i-n)$ and $(i-h)$ locations, F_{i+1}^{i+m} a "future feature", F_{i-n}^{i-1} a "past feature," respectively.

Table 5: Accuracy definition ($\hat{E}$ and $\hat{N}$ are Emphasized and Not emphasized accent phrases as prediction results, E and N are Emphasized and Not emphasized accent phrases as answers,respectively, A, B, C, D are counts for each case, Accuracy is defined as $(A+D)/(A+B+C+D)\times 100$)

		Predicted results	
		$\hat{E}$	$\hat{N}$
Answers	E	A	B
	N	C	D

4.3 Evaluation measure

We used accuracy as the performance evaluation measure and evaluated the total accuracies of the proposed method using 5-fold cross validation. Accuracy is defined by the number of correctly predicted emphasis and not-emphasis ($A+D$ in Table 5) divided by the sum of the number of all 4 prediction results (in addition to the above 2 correct cases, the 2 other cases are that emphasis is erroneously classified as not-emphasis (B) and vice versa (C)): $Accuracy\ [\%] = (A+D)/(A+B+C+D)\times 100$.

4.4 Results

We examined 12 combinations of dimension (1 to 4) and cost (1 to 3) of the kernel function. Use of larger dimensions means combining more features. Better accuracies were obtained by larger dimensions than smaller dimensions. Cost values did not derive significant changes in accuracies for the same kernel dimension. Thus, we fixed dimension 4 and cost 1 and examined several scopes of features and history lengths of past prediction results.

Accuracy for test data varied from 74.1 to 77.4% under the feature scope changing from F_{i-4}^{i+4} to F_{i-1}^{i+1} and history changing from H_{i-4}^{i-1} to H_{i-1}^{i-1}. The smaller the feature scope and history length was, the better the accuracy was. As no use of future features F_{i+1}^{i+m} decreased accuracies slightly (0.2 to 0.6 points), future features somewhat contributes to prediction. No use of past prediction result H_{i-1}^{i-h} derived both slight increase (0.1 to 1.0) and decrease (0.2 to 0.3) of accuracies, but balance between recall

Table 6: Best prediction results at $F_{i-1}^{i+1} and H_{i-1}^{i-1}$ ($\hat{E}$, $\hat{N}$, E, N are the same as in Table 5)

		Predicted results		recall
		$\hat{E}$	$\hat{N}$	
Answers	E	548	305	64.2%
	N	228	1278	
precision		70.6%		
accuracy				77.4%

and precision of emphasized accent phrases became worse.

Based on these results and as we consider that both emphasized and not-emphasized cases should be correctly predicted, we chose using both future features and past prediction results. As a result, the best accuracy was 77.4% at F_{i-1}^{i+1} and H_{i-1}^{i-1} (-1 only), then recall and precision rates of emphasized accent phrase were 64.2% and 70.6%, respectively. Detailed prediction results were shown in Table 6.

As far as we know, there is no research for predicting emphasized accent phrases from Japanese advertisement text. As baseline calculations, if all the accent phrases are predicted emphasized ($\hat{E}$), accuracy is 36.2% and the recall and precision of emphasized accent phrases are 100% and 36.2%, respectively. On the other hand, if all the accent phrases are predicted non-emphasized ($\hat{N}$), accuracy is 63.8%, then both recall and precision of emphasized accent phrases are 0%. Thus, the proposed method offered 13.6 points higher accuracy than these above forced predictions.

Since Fig. 2 showed lowest likelihood of emphasized accent phrase at the top of each IP, we also examined another feature of *whether the AP is at the top of IP or not*. The feature showed smaller prediction potential 0.005 than the 9th feature in Table 3 (0.02) and did not offer prediction accuracy improvements.

5 Conclusion

This paper proposed a method for predicting which portions of an advertisement text should be emphasized; it uses only the text itself. The method uses accent phrases as the prediction unit and the features obtained by the text processing modules of current Text-to-speech synthesis systems. According to mutual information, features such as *word surface string of the main content and function word* and *part-of-speech of the main function word* offer higher prediction potential. Experiments showed the proposed method yielded encouraging accuracies for such an expressive TTS which uses emphasized accent phrase locations as a context information as (Maeno, et al., 2014). Accuracy improvement was left as a future work.

References

Takeshi Fuchi and Shin'ichiro Takagi. 1998. "Japanese morphological analyzer using word co-occurrence: JTAG" *Proceedings of Coling-ACL*, 409–413.

Dirk Hovy, Gopala Krishna Anumanchipalli, Alok Parlikar, Caroline Vaughn, Adam Lammert, Eduard Hovy, and Alan W. Black. 2013. "Analysis and Modeling of Focus in Context" *Proceedings of INTERSPEECH*, 402–406.

Andrew J. Hunt and Alan W. Black. 1996. "Unit selection in a concatenative speech synthesis system using a large speech database," *Proceedings of ICASSP*, 373–376.

Kenji Imamura, Gen'ichiro Kikui, and Norihito Yasuda. 2007. "Japanese dependency parsing using sequential labeling for semi-spoken language" *Proceedings of ACL*, 225–228.

Yu Maeno, Takashi Nose, Takao Kobayashi, Tomoki Koriyama, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, and Osamu Yoshioka. 2014. "Prosodic variation enhancement using unsupervised context labeling for HMM-based expressive speech synthesis," *Speech Communication*, 57: 144–154.

Fanbo Meng, Zhiyong Wu, Helen Meng, Jia Jia and Lianhong Cai. 2012. "Hierarchical English emphatic speech synthesis based on HMM with limited training data," *Proceedings of INTERSPEECH*, Mon.P2b.09.

Hideharu Nakajima, Noboru Miyazaki, Akihiko Yoshida, Takashi Nakamura, Hideyuki Mizuno. 2010. "Creation and Analysis of a Japanese Speaking Style Parallel Database for Expressive Speech Synthesis" *Proceedings of Oriental COCOSDA*, paper id 30.

Hideharu Nakajima, Hideyuki Mizuno, Osamu Yoshioka, and Satoshi Takahashi. 2013. "Which resemblance is useful to predict phrase boundary rise labels for Japanese expressive text-to-speech synthesis, numerically-expressed stylistic or distribution-based semantic?" *Proceedings of INTERSPEECH*, 1047–1051.

Ken Ross and Mari Ostendorf. 1996. "Prediction of abstract labels for speech synthesis" *Computer Speech & Language*, 10(3): 155–185.

Virek Kumar Rangarajan Sridhar, Ani Nenkova, Shrikanth Narayanan, Dan Jurafsky. 2008. "Detecting prominence in conversational speech: pitch accent, givenness and focus" *Proceedings of Speech Prosody*, 453–456.

Volker Strom, Ani Nenkova, Robert Clark, Yolanda Vazquez-Alvarez, Jason Brenier, Simon King, and Dan Jurafsky. 2007. "Modelling prominence and emphasis improves unit-selection synthesis," *Proceedings of INTERSPEECH*, 1282–1285.

Kai Yu, François Mairesse, and Steve Young. 2010. "Word-level emphasis modelling in HMM-based speech synthesis," *Proceedings of ICASSP*, 4238–4241.

Heiga Zen, Keiichi Tokuda and Alan W. Black. 2009. "Statistical parametric speech synthesis," *Speech Communication*, 51(11): 1039–1064.