11th ISCA Speech Synthesis Workshop (SSW 11)
26-28 August 2021, Budapest, Hungary

Accent Modeling of Low-Resourced Dialect in Pitch Accent Language Using Variational Autoencoder

Kazuya Yufune[1], *Tomoki Koriyama*[1], *Shinnosuke Takamichi*[1], *Hiroshi Saruwatari*[1]

[1]Graduate School of Information Science and Technology, The University of Tokyo, Japan.

kazuya_yufune@ipc.i.u-tokyo.ac.jp, t.koriyama@ieee.org

Abstract

Realizing text-to-speech (TTS) system of dialects is useful for personalizing TTS systems. However, TTS for many dialects of pitch accent languages is not realized because of low-resourced problem. Among many dialects of pitch accent languages, this paper focuses on Osaka dialect of Japanese, one of the most challenging pitch accent languages. For Japanese TTS system, accent labels are known to be necessary as input to synthesize natural speech. In rich-resourced dialect, human-resourced approaches and dictionary-based approaches are often used to annotate accent labels for training and inference, but such approaches are unfeasible and time-consuming for low-resourced dialects. In this paper, we propose accent extraction model that utilizes vector quantized variational autoencoder (VQ-VAE) to prepare accent information from speech, and accent prediction models that utilize decision tree and deep learning techniques to predict accent information from the input text. The models were examined with corpus of Osaka dialect, whose accent labels do not exist. The results showed that accent extraction model succeeded in extracting accent information of Osaka dialect from speech utterances as latent variable. It also showed that the accent of synthesized speech by accent prediction models were not better than baseline, but it had advantages such as interpretability.

Index Terms: pitch accent, speech synthesis, Japanese dialect, VQ-VAE, accent label, latent variable

1. Introduction

Text-to-speech (TTS) systems with dialects makes speech applications diverse. For example, personalizing TTS with the speaker's dialect can be an alternative form of voice output for patients who have progressive dysarthria and want to speak in their dialects [1]. For another example, dialect TTS systems could be adopted for local characters to speak in the local dialects.

For pitch accent languages such as Japanese, it is known that accent information of input texts has an important role for TTS to synthesize natural-sounding speech [2, 3]. For example, in Japanese, a change in pitch makes a difference between words. Changing the pitch of "chopsticks" (/ha'shi/) differentiates the meaning into "bridge" (/hashi'/) or "edge" (/hashi'/). Though these words have the same phonemes /hashi/, Japanese speakers distinguish them by the pitch accent. In Japanese TTS system, without inputting the accent information as accent labels, an acoustic model cannot capture the pitch fluctuations appropriately, resulting in unnatural (sometimes even wrong) synthetic speech. Hence, accent labels need to be correctly given from text in pre-processing for Japanese TTS systems. For TTS of the Tokyo dialect, accent labels are annotated typically by professional annotators or dictionary-based approaches such as OpenJTalk [4]. Since the Tokyo dialect is the standard dialect of Japanese, TTS of this dialect can utilize rich resources such as professional annotators and an accent dictionary.

However, TTS systems for many dialects of pitch accent languages have been suffering from low resource problems. Specifically, recorded speech data set is not sufficient for modeling of fundamental frequency (F0) curves of accents even if we use an end-to-end TTS framework [2]. Although it is true that accent labels improve the synthetic F0 curves, annotating pitch accent labels requires professional annotators familiar with both the target dialect and the pitch accent system. Moreover, the accents of dialects are rarely summarized as an accent dictionary. Therefore, we should investigate the dialect TTS system under the condition that accent labels are not sufficiently provided.

In this paper, we focus on the Osaka dialect, which is among the dialects of Japanese and significantly different from the Tokyo dialect. To overcome the low resource problems, we propose two frameworks: accent extraction models for accent modeling in training, and accent prediction models for accent modeling in inference. The accent extraction models are used to extract latent representations of accent from speech. As the accent extraction models, we use not only variational autoencoder (VAE) [5], which was successful in extracting sentence-level prosody representations [6], but also vector quantized VAE (VQ-VAE) [7] to express discrete characteristics of Japanese accent. Mora-level latent variable representation of accent using VAE and VQ-VAE enables an acoustic model to be trained without annotated accent labels. The accent prediction models are used to infer the latent variable representations. We examine the effectiveness of two accent prediction models using recurrent neural networks (RNNs) and decision trees. We also investigate the use of the accent dictionary of the Tokyo dialect as the input of the accent prediction models.

2. Japanese pitch accent of Tokyo and Osaka dialects

The label of the accent system of Japanese is defined as high or low for each mora, which fundamentally corresponds to a Japanese Hiragana/Katakana character [8]. In the Tokyo dialect, Japanese words have an accent nucleus position, where the label changes from high to low. In the case of two-mora nouns, the nucleus position is among "no-nucleus (0)," "1," or "2." An example of accent labels of two-mora nouns for the Tokyo dialect is shown in Figure 1. The last mora "wa" is a postpositional particle in Japanese. In this example,"ha-shi (edge)" has no accent nucleus, "ha-shi (chopsticks)" has nucleus position of "1," and "ha-shi (bridge)" has that of "2." Since these words have different accents, their corresponding accent labels are different. This accent information is in the accent dictionary for the Tokyo dialect.

10.21437/SSW.2021-33

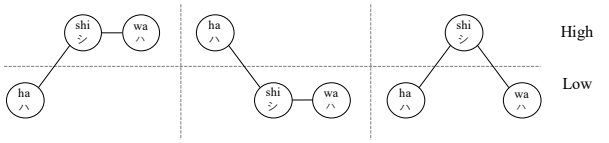


Figure 1: *Example of accent labels. left: "ha-shi (edge)"+"wa", center: "ha-shi (chopsticks)"+"wa", right: "ha-shi (ridge)"+"wa".*

Table 1: *Corresponding relationships of accent labels of two-mora nouns + postpositional particle "wa" between the Tokyo and Osaka dialects (H: High, L: Low)*

Tokyo dialect	Osaka dialect
L - H - H (no-nucleus)	H - H - H
H - L - L (nucleus position 1)	L - H - L
L - H - L (nucleus position 2)	H - L - L

The Osaka dialect is spoken in around Osaka prefecture[1]. When constructing a TTS system for this dialect, accent labels of the dialect are needed as the input, but the accent dictionary of the Osaka dialect does not exist. One of the available resources related to Japanese pitch accent is the accent dictionary of the Tokyo dialect. However, since the Osaka dialect has an accent system which is completely different from that of the Tokyo dialect, the accent dictionary of the Tokyo dialect is not suitable as it is for estimating the accent of the Osaka dialect. On the other hand, there are some corresponding relationships between the Tokyo and Osaka dialects [9]. For example, it is known that the accents of two-mora nouns of the dialects correspond to each other as Table 1 shows.

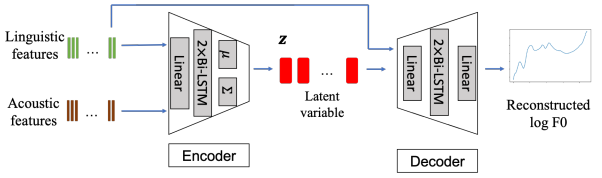
3. Related work

There have been many studies that focused on representation learning of prosody from acoustic features, including [10, 6, 11, 12]. Zhao et al. [11] proposed a model that reconstructs speech waveform with VQ-VAE [7] and down-sampled frame-level F0-related latent representation extracted from F0 curve. Hodari et al. [12] succeeded in improving prosody of synthesized speech by learning word-level prosody representations from referenced mel-spectrogram using VQ-VAE, and predicting them from the context in inference. Kenter et al. [6] proposed a hierarchical VAE [5] model that can synthesize a variety of prosodic features such as F0 by using sentence-level prosody embeddings. In this study, we examine VAE and VQ-VAE models for accent modeling of the Osaka dialect, with mora-level latent representation learning of pitch accent.

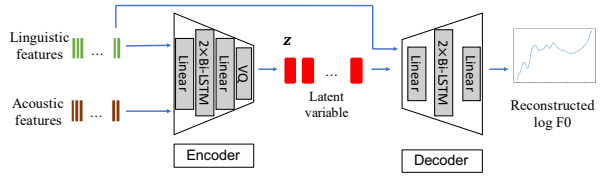
4. Latent-variable-based accent extraction models for Japanese using VAE models

We propose latent-variable-based accent extraction models for the expression of pitch accent of Osaka dialect. Specifically, we utilize VAE and VQ-VAE, which have been successful in prosody modeling described in the previous section. We adopted mora-level latent representation for accent modeling, as Japanese pitch accent of all dialects is defined for each mora as described in Sec. 1. We assume that we only have texts and speech utterances, and that accent labels are unavailable, which often happens, especially when targeting at low-resourced dialects.

[1]The second-largest metropolitan area in Japan.



(a) *VAE structure. Linear means linear layer, bi-LSTM means bi-directional long short term memory (LSTM) cells layer*



(b) *VQ-VAE structure. VQ means vector quantization layer which quantized the output of the previous linear layer.*

Figure 2: *Structure of accent extraction models*

4.1. Structure of accent extraction models

We propose accent extraction models that use VAE and VQ-VAE to extract the accent information from speech samples as latent variables. The model structures are shown in Figure 2. First, the encoder takes time-series frame-level linguistic $\boldsymbol{x}$ and acoustic features $\boldsymbol{y}$ as the input, and outputs latent variables $\boldsymbol{z}$ for each mora. In the second bi-LSTM layer of the encoder, the output of the last frame of each mora is propagated to the next layer, which results in transforming the frame-level features into mora-level features. The decoder takes frame-level linguistic features $\boldsymbol{x}$ and the mora-level latent variables $\boldsymbol{z}$ as the input, and predicts F0 curve for the speech $\hat{\boldsymbol{y}}_{F0}$. By providing linguistic features $\boldsymbol{x}$ that have no accent information of either the Tokyo nor Osaka dialect, we expect that the latent variables represent the accent information extracted from the acoustic features.

4.2. VAE model

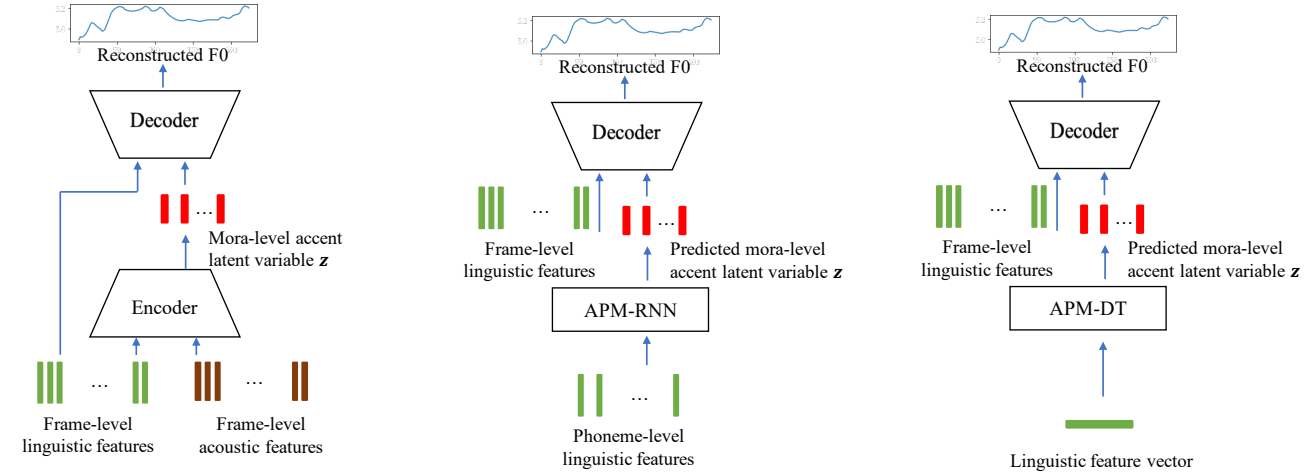
In this section, we propose an accent extraction model with VAE, which is often used in unsupervised learning of latent representations of speech [6]. Figure 3a shows the structure of the VAE model. The boxes of "$\boldsymbol{\mu}$" and "$\boldsymbol{\Sigma}$" in the figure mean linear layers that output mean vector $\hat{\boldsymbol{\mu}}$ and diagonal variance matrix $\hat{\boldsymbol{\Sigma}}$, respectively. The posterior distribution of latent variable $\boldsymbol{z}$ is defined as a Gaussian distribution with mean $\hat{\boldsymbol{\mu}}$ and variance $\hat{\boldsymbol{\Sigma}}$. Following [5], we define the loss function $\mathcal{L}_{\mathrm{VAE}}$ as follows:

$$\mathcal{L}_{\mathrm{VAE}} = \sum_{i=1}^{N_d} \left\{ \left\| \boldsymbol{y}_{\mathrm{F0}}^{i} - \hat{\boldsymbol{y}}_{\mathrm{F0}}^{i} \right\|^2 + D_{\mathrm{KL}}[\mathcal{N}(\hat{\boldsymbol{\mu}}^i, \hat{\boldsymbol{\Sigma}}^i) || \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})] \right\} \tag{1}$$

where N_d is the number of speech samples, $\mathcal{N}$ means Gaussian distribution and D_{KL} means the Kullback–Leibler divergence. $\boldsymbol{I}$ means an identity matrix.

4.3. VQ-VAE model

Since the Japanese accent information that people perceive is discrete as described in Sec. 2, we adopt VQ-VAE, which quantizes the latent space, to take advantage of this discrete characteristic of the Japanese pitch accent. Figure 3b shows the structure of the VQ-VAE model. The vector quantization layer quantizes the output of the previous linear layer. Following [7], we



(a) *Accent extraction model using VAE and VQ-VAE*

(b) *APM-RNN*

(c) *APM-DT*

Figure 3: *Summary of proposed models. (a) is accent extraction model and (b) and (c) are accent prediction models.*

define the loss function $\mathcal{L}_{\mathrm{VQ-VAE}}$ as follows:

$$\mathcal{L}_{\text{VQ-VAE}} = \sum_{i=1}^{N_d} \Big\{ \|\boldsymbol{y}^i_{\mathrm{F0}} - \hat{\boldsymbol{y}}^i_{\mathrm{F0}}\|^2 + \|sg(\boldsymbol{z}^i_{\mathrm{uq}}) - \boldsymbol{z}^i\|^2 + \beta \|sg(\boldsymbol{z}^i) - \boldsymbol{z}^i_{\mathrm{uq}}\|^2 \Big\} \quad (2)$$

where $\boldsymbol{z}_{\mathrm{uq}}$ are the values of $\boldsymbol{z}$ before quantization by the VQ layer, and function $sg(\cdot)$ stops the gradient. β was set to 1 in the experiment.

5. Accent prediction models

In this section, we propose two accent prediction models that predict accent latent variables from linguistic features, for synthesizing speech of the Osaka dialect with only text. The relationship between the proposed accent extraction models and the accent prediction models is shown in Figure 3. The accent prediction models predict accent latent variables, and the F0 curve is synthesized by inputting the predicted accent latent variables into the decoder. One uses RNNs and the other uses a decision tree [13]. For both of these two models, there are two candidates that take different input features. One takes only the linguistic features of text as the input. The other takes the linguistic features of the text and accent information of the Tokyo dialect. Using accent information of the Tokyo dialect as the input is possibly useful for the models to learn the corresponding relationships between the accents of Tokyo and Osaka dialects, hypothesizing that there generally exist the correspondences as described in Sec. 2. We examine the impact of accent information of the Tokyo dialect on predicting accents of the Osaka dialect by comparing the results of these two candidates.

5.1. Accent prediction model using RNNs (APM-RNN)

This model uses RNNs to predict the accent latent variables. Since this model adopts deep learning techniques, it can capture more complex features than the decision tree model. It takes phoneme-level linguistic features and accent information of the Tokyo dialect, and outputs the accent latent variables. The structure is almost the same as the decoder of the accent extraction models. The differences are that this model does not take the latent variable as the input, and that the outputs of this model are accent latent variables, not F0 curve.

5.2. Accent prediction model using decision tree (APM-DT)

This model uses a decision tree to predict the accent latent variables. For this model, we expect robustness, because the correspondences as shown in Table 1 are so simple that RNN models may be too expressive. Since the decision tree can output only a scalar value, we define a decision tree model for each mora index. As the input, this model takes linguistic feature vector, and accent latent variables of preceding moras to consider time series feature of accent. When predicting the accent latent variable of a four-mora word, four decision trees are used.

6. Experiments

6.1. Experimental conditions

We used a subset of the JSUT corpus, BASIC5000 [14], which consists of 5000 utterances of sentences spoken in the Tokyo dialect by a female speaker, and OSAKA3696, which consists of 3696 utterances of phrases spoken in the Osaka dialect by a male speaker. The phrases were composed of 258 verbs, 156 adjectives and 930 nouns and each phrase consisted of one content word and a positional particle. Since Japanese verbs and adjectives are conjugated depending on the postpositional particle or auxiliary verb, all conjugated forms were recorded for each verb and adjective with postpositional parts. Nouns were recorded with the postpositional particle “wa” because the accent of a noun affects the accent of a postpositional particle. We used 3000 utterances of BASIC5000, and 3126 utterances of OSAKA3696 for training, 285 of OSAKA3696 for validation, and 285 of OSAKA3696 for testing. The reason we also used the Tokyo dialect corpus was to make training stable.

Based on the context label of Japanese HTS [15], the linguistic feature vector for the accent extraction model was defined as a 444-dimensional one, which consisted of phoneme information, parts of speech, and one-hot speaker embedding. The linguistic feature vector for the APM-RNN was defined as a 442-dimensional one, which consisted of phoneme information, parts of speech. The accent information vector of the Tokyo dialects (Tokyo accent vector) was defined as a 91-dimensional one. For the APM-DT model, we used a phrase-level 159-dimensional vector including parts of speech as the linguistic feature vector. As the accent information vector of the Tokyo

Table 2: *RMSEs of reconstructed F0 [cent] using extracted accent latent variable*

model	F0 RMSE [cent]
VAE	216
VQ-VAE	**172**
NO-ALV	247

dialect, we used the same Tokyo accent vector as the APM-RNN.

The sampling rate of all speech signals was 48 kHz, and the frame shift length was set to 5 ms. The acoustic features were defined as the 0–59th mel-cepstral coefficients, continuous log F0, five-band aperiodicity, first and second derivatives of all these parameters, and a voiced/unvoiced flag. WORLD [16] was used for parameter extraction and waveform synthesis. As pre-processing of F0, trajectory smoothing [17] with a 10 Hz cutoff frequency was used. The number of classes of VQ-VAE latent space was set to 2, on the basis of the accent system of Japanese as described in Sec. 2. The basic structure of the DNN models (encoder, decoder, APM-RNN) consisted of a linear layer, 2 $\times$ bi-directional LSTM layer with 734 cells, and a linear layer. For the VAE encoder, the last linear layer was replaced with two linear layers of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as shown in Figure 2a. For the VQ-VAE encoder, VQ layer was added as the last layer as shown in Figure 2b. The maximum depth of decision tree was set to 11.

6.2. Evaluations of accent extraction models

6.2.1. Objective evaluations of accent extraction models

To evaluate the performance of the accent extraction models, we calculated the root mean squared errors (RMSEs) of the F0 curves reconstructed by the accent extraction models. The results are shown in Table 2. "NO-ALV" means a model that directly predicted F0 curve without accent latent variable (ALV), and had the same structure as the decoder. The F0 RMSEs of both the VAE and VQ-VAE models were smaller than NO-ALV, which did not use the accent latent variable. This implies that the proposed accent extraction models succeeded in extracting accent information as latent variables. Moreover, the RMSE of the VQ-VAE model was 172 cent, which was smaller than that of the VAE model. This implies that the discrete representation of the two classes was more suitable for representing high/low Japanese accent.

6.2.2. Subjective evaluations of accent extraction models

To confirm the effectiveness of the accent extraction models also in a subjective evaluation, we conducted an XAB test on the accent reproducibility. The evaluation was done by 30 listeners on our crowdsourcing system with speech samples vocoded with reconstructed F0 curve, original mel-cepstrum, and original band aperiodicity. The listeners were asked to answer which of two accents of synthetic speech samples was closer to the original one. Table 3 shows the results. As shown in the Table, the F0 curves created by the VAE and VQ-VAE models were significantly closer to the original speech than that of the NO-ALV. Moreover, the F0 curve created by the VQ-VAE model was significantly closer to the original than that of the VAE model. The effectiveness of the proposed accent extraction models and quantization were confirmed also in the subjective evaluation.

Table 3: *XAB test results of accent extraction models*

model A				p-value	model B
VAE	**0.591**	vs.	0.401	$< 10^{-5}$	NO-ALV
VQ-VAE	**0.700**	vs.	0.300	$< 10^{-5}$	NO-ALV
VAE	0.375	vs.	**0.625**	$< 10^{-5}$	VQ-VAE

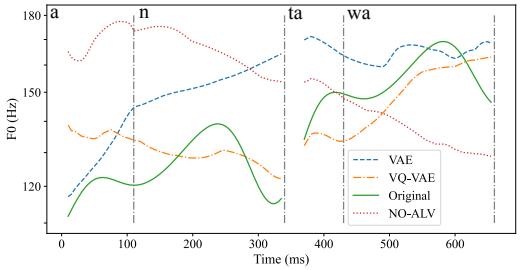


Figure 4: *F0 plot of "a-n-ta-wa" synthesized by accent extraction models. The vertical dotted lines mean the borders of mora.*

6.2.3. Synthesized F0 curves

The VQ-VAE model succeeded in extracting the accent than the VAE model. Here, we look into the the reconstructed F0 curves of the models. Figure 4 shows F0 curves for the phrase "a-n-ta-wa"(a noun "a-n-ta" and a postpositional particle "wa") synthesized by the accent extraction models. The accent of the original speech signal was low/low/low/high. Since only "wa" had high accent in the phrase, the original F0 curve had a higher value for "wa" than the others. The predicted F0 curve of the VQ-VAE model had the same tendency as the original one. However, the F0 curves of the other methods were different from the original one. The F0 curve of the VAE model had high values not only in "wa", but also in "n" and "ta". The F0 curve of the NO-ALV was a simple declination, which is far from the original one. The VQ-VAE model succeeded better in reconstructing the F0 curve of the original speech than the other models.

6.2.4. Examples of accent latent variable of VQ-VAE model

We confirmed that the VQ-VAE model succeeded in extracting accent information as latent variable better than the VAE model. In this section, we check how the extracted latent variables look like. Examples of the extracted latent variables extracted by the VQ-VAE model are shown in Figure 5. Compared with manual labels that we annotated to a part of the corpus, we found that, one of the classes (Class 1) of the latent space tended to correspond to high, and the other (Class 2) tended to correspond to low. Since the VQ-VAE had better results, we adopted the VQ-VAE model as our accent extraction model. The accent latent variables extracted by the VQ-VAE model were used as the teacher labels for the accent prediction models.

6.3. Results of accent prediction models

6.3.1. Objective evaluation of predicted F0

To measure the quality of the F0 curves predicted by the accent prediction models, we calculated the RMSEs of the F0 curves for three parts of speech (verbs, nouns, and adjectives) in OS-AKA3696. Table 4 shows the objective evaluation results of the predicted F0 curves. "W/" means that the input included accent labels of the Tokyo dialect, and "W/O" means that the input did not include them. The F0 RMSE of the APM-RNN W/ was the smallest (256 cent), while that of the APM-DT W/O was

Phrase 1: スレバ (su-re-ba)

Mora	ス (su)	レ (re)	バ (ba)
Annotated accent label	High	Low	Low
Class of extracted latent variable	1	2	2

Phrase 2: ゼンブハ (ze-n-bu-wa)

Mora	ゼ (ze)	ン (n)	ブ (bu)	ハ (wa)
Annotated accent label	Low	High	Low	Low
Class of extracted latent variable	2	1	2	2

Phrase 3: オイシイ (o-i-shi-i)

Mora	オ (o)	イ (i)	シ (shi)	イ (i)
Annotated accent label	Low	Low	High	Low
Class of extracted latent variable	2	1	1	2

Figure 5: *Example of accent latent variables extracted by VQ-VAE model*

Table 4: *RMSE of reconstructed F0 [cent] for each part of speech*

model	all	verb	noun	adjective
APM-DT W/	313	289	351	323
APM-DT W/O	321	287	368	322
APM-RNN W/	**256**	**239**	**334**	**215**
APM-RNN W/O	272	241	365	222

the largest (323 cent). The RMSEs of F0 of the APM-DT were much larger than those of the APM-RNN, which implies that APM-DT was not expressive enough to predict the accent latent variables. All models with the Tokyo accent labels had better prediction results compared with those without the Tokyo accent labels. As for the difference among parts of speech, the effect of adding accent labels of the Tokyo dialect was relatively small in verbs and adjectives compared with nouns. One of the causes may be that the accents of verbs and adjectives of the Osaka dialect have a few fundamental patterns. For example, accent labels of an n-mora adjective are fundamentally defined as high/.../high/low/low. This may make the accent of them easy to predict without the accent information of the Tokyo dialect.

6.3.2. Subjective evaluation of predicted F0

In addition to the objective evaluation, We conducted XAB tests on the accent reproducibility of the predicted F0 curves to check the prediction performance of the APMs. This subjective evaluation was done by two groups with speech samples vocoded with predicted F0 curves, original mel-cepstrum, and original 5 band aperiodicity. One was done by 30 listeners on our crowdsourcing system. The other was done by 30 listeners who speaks Osaka dialect. The listeners were asked which of two accents was similar to the original one, in the same way as Sec. 6.2.2. Table 5 shows the results of the evaluation by our crowdsourcing system, and Table 6 shows those by Osaka citizens. The results of both evaluations were similar. As both of the tables show, the APM-RNN W/ had significantly better performance than the APM-DT W/. There was no significant difference between the W/ models and the W/O models.

Since the degradation of RMSE in nouns by adding the Tokyo accent labels were larger than other parts of speech, we additionally conducted a subjective evaluation experiment, by limiting the test utterances to nouns. The experiment was done only on our crowdsourcing system, as the results of Osaka citizens and our crowdsourcing system were similar. The results are shown in Table 7. The APM-RNN W/ was significantly better at reproducing the accents of the Osaka dialect nouns than the APM-RNN W/O. It is estimated that adding accent labels of the Tokyo dialect was useful in predicting the accents of nouns of the Osaka dialect.

Table 5: *Subjective evaluation of predicted F0 by crowdsourcing system*

model A				p-value	model B
APM-DT W/	0.375	vs.	**0.625**	$< 10^{-5}$	APM-RNN W/
APM-DT W/	0.519	vs.	0.481	0.35	APM-DT W/O
APM-RNN W/	0.498	vs.	0.502	0.96	APM-RNN W/O

Table 6: *Subjective evaluation of predicted F0 by Osaka citizens*

model A				p-value	model B
APM-DT W/	0.334	vs.	**0.666**	$< 10^{-5}$	APM-RNN W/
APM-DT W/	0.533	vs.	0.467	0.12	APM-DT W/O
APM-RNN W/	0.511	vs.	0.489	0.64	APM-RNN W/O

6.3.3. Predicted F0 curves

The APM-RNN succeeded better in reproducing the accents of the Osaka dialect than the APM-DT. Here, we look into the predicted F0 curves of an adjective. Figure 6 shows the predicted F0 curves for a five-mora adjective “a-ri-ga-ta-i”, whose accent labels of the Osaka dialect are high/high/high/low/low. Since the last two moras of the term have low accent, the values of original F0 of them tend to be smaller than those of former three moras. The predicted F0 curve of the APM-RNN W/ had a similar tendency to the original one, which can be perceived as the same accent high/high/high/low/low. However, The F0 curve of the APM-DT W/ fell around third mora “ga”, which can be perceived as a wrong accent, high/high/low/low/low.

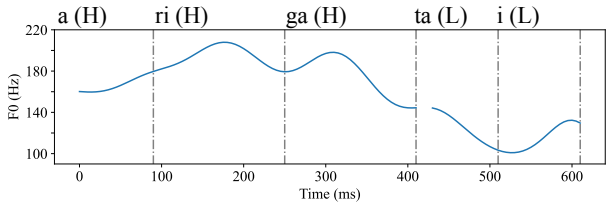
7. Conclusions

In this paper, we have proposed accent extraction models and accent prediction models for automatic accent modeling of the Osaka dialect. The result showed that the proposed accent extraction model succeeded in extracting accent information as latent variable using VQ-VAE. This model will make it possible to train an acoustic model that synthesizes natural F0 curve without annotated accent labels, which is one of the problems that TTS of Japanese non-Tokyo dialects is suffering from. For the accent prediction models, the result showed that the APM-RNN reproduced the accent of the Osaka dialect better than the APM-DT, and adding the accent labels of the Tokyo dialect is useful for predicting the accent of nouns of the Osaka dialect.

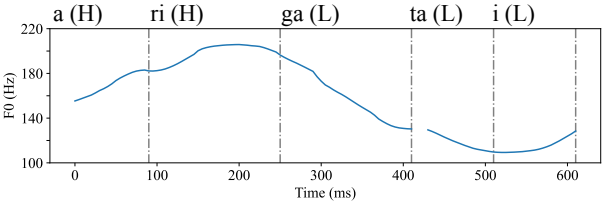
Combining the proposed accent extraction models and accent prediction models enables us to synthesize speech of texts without speech samples. Although the RMSEs of the proposed prediction models were still larger than that of the model without accent latent variables (NO-ALV), the proposed synthesis

Table 7: *Subjective evaluation of predicted F0 of nouns*

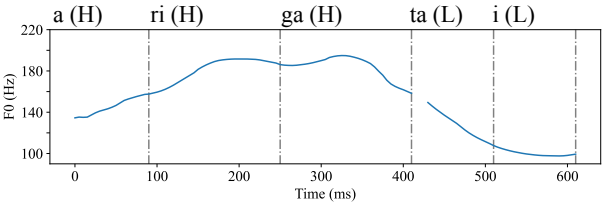
model A				p-value	model B
ALV-DT W/	0.526	vs.	0.474	0.35	ALV-DT W/O
ALV-RNN W/	**0.657**	vs.	0.343	$< 10^{-2}$	ALV-RNN W/O



(a) *Original*



(b) *APM-DT W/*



(c) *APM-RNN W/*

Figure 6: *Predicted F0 curve of an adjective, "a-ri-ga-ta-i". The horizontal and vertical axes mean Time and F0 respectively, and the dashed lines and the labels above mean the border of mora, the phoneme of mora and how people can perceive the accent of the mora.*

methods have some advantages such as:

- Interpretability:
  Looking into the input accent latent variables enables us to understand how the accents of speech utterances were synthesized.
- Controllability:
  Changing the input accent latent variables enables us to easily modify the accent of synthesized speech into more natural one.

Moreover, the proposed accent extraction models are possibly useful for an accent analysis of low resourced dialects, since they can easily visualize the accent information only with texts and speech utterances, even without professionals of the accent of the dialect.

Future work includes:

- Apply the proposed models to other dialects of pitch accent languages including Japanese
- Research model structures of the accent extraction models for better representation of the accent
- Incorporate modification systems or other input features into the proposed accent prediction models for better prediction
- Extend the proposed models to extract other features of speech signals such as emotion and dialog acts.

8. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 18K18100, 19K20292.

9. References

[1] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, and V. Strom, "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis," *Speech Communication*, vol. 52, no. 2, pp. 164–179, 2010.

[2] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6905–6909.

[3] T. Koriyama and T. Kobayashi, "Semi-supervised Prosody Modeling Using Deep Gaussian Process Latent Variable Model." in *INTERSPEECH*, 2019, pp. 4450–4454.

[4] "openjtalk," http://open-jtalk.sp.nitech.ac.jp/.

[5] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *International Conference on Learning Representations*, 2014.

[6] T. Kenter, V. Wan, C.-A. Chan, R. Clark, and J. Vit, "CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *International Conference on Machine Learning*, 2019, pp. 3331–3340.

[7] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete Representation Learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 6309–6318.

[8] S. Kawahara, "The phonology of Japanese accent," *The handbook of Japanese phonetics and phonology*, pp. 445–492, 2015.

[9] H. Kindaichi, "Akusento no bunpu to hensen," *Iwanami kouza nihongo*, vol. 11, pp. 129–180, 1977, in Japanese.

[10] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[11] Y. Zhao, H. Li, C.-I. Lai, J. Williams, E. Cooper, and J. Yamagishi, "Improved prosody from learned f0 codebook representations for vq-vae speech waveform reconstruction," *arXiv preprint arXiv:2005.07884*, 2020.

[12] Z. Hodari, A. Moinet, S. Karlapati, J. Lorenzo-Trueba, T. Merritt, A. Joly, A. Abbas, P. Karanasou, and T. Drugman, "Camp: a two-stage approach to modelling prosody in context," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6578–6582.

[13] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[14] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, 2020.

[15] "HTS," http://hts.sp.nitech.ac.jp/.

[16] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[17] S. Takamichi, K. Kobayashi, K. Tanaka, T. Toda, and S. Nakamura, "The NAIST text-to-speech system for the Blizzard Challenge 2015," in *Proc. Blizzard Challenge workshop*, 2015.