A method for predicting stressed words in English Jazz Chants

Ryo Nagata[1]*, Toshiaki Marueki*[1]*, Kotaro Funakoshi*[2]*, Tatsuya Kitamura*[1]*, Mikio Nakano*[2]

[1]Konan University, Japan
[2]Honda Research Institute Japan Co., Ltd., Japan
{rnagata,t-kitamu} @ konan-u.ac.jp., {funakoshi,nakano}@jp.honda-ri.com

Abstract

To acquire a second language, one must develop an ear and tongue for the correct stress and intonation patterns of that language. In English language teaching, there is an effective method called *Jazz Chants* for working on the sound system. In this paper, we propose a method for predicting stressed words, which play a crucial role in Jazz Chants. The proposed method is specially designed for stress prediction in Jazz chants. It exploits several sources of information including words, POSs, sentence types, and the constraint on the number of stressed words in a chant text. Experiments show that the proposed method achieves an F-measure of 0.936 and outperforms the other methods implemented for comparison. The proposed method is expected to be useful in supporting non-native teachers of English when they teach chants to students and create chant texts with stress marks from arbitrary texts.

Index Terms: language learning, stress prediction, teaching material generation, Jazz Chants, stress-timed rhythm

1. Introduction

To acquire a spoken language, one must develop an ear and tongue for the correct stress and intonation patterns of the spoken language. This is normally difficult for those who are acquiring a second language whose sound system is not similar to that of their first language. An example pair would be English and Japanese in which the sound systems are quite different.

In English language teaching, there is an effective method called *Jazz Chants*[1] for working on the sound system. "A chant is a rhythmic expression of natural language which links the rhythms of spoken American English to the rhythms of traditional American jazz — the rhythm, stress and intonation pattern of what children would hear from an educated native speaker in natural conversation [1]". In chants, each stressed word is pronounced (i) with an extra emphasis[2] (often with physical activities such as clapping or jumping) and (ii) with an equal time interval (i.e., isochronism). To support this, stressed words are sometimes (but not always) marked with the asterisk * or underlined in teaching materials for chants (Hereafter, teaching materials for chants will be referred to as *chant texts*). An example of a chant text is as follows [1]:

* * * *
Frank, Hank, walk to the bank.
* * * *
Jill, Phil, run up the hill.

[1]Jazz Chants® is a registered trademark of Oxford University Press. In this paper, Jazz Chants will be simply referred to as *chants*.

[2]In chants, each stressed word is somewhat exaggeratedly pronounced to acquire the rhythm, stress and intonation patterns.

Teachers and children read chant texts out loud, putting stress on the marked words.

Since chants require only sound and physical activities to teach, they are especially suitable for children who are not yet familiar with written language. In addition, Graham [1] shows that the use of chants has the following three advantages in language learning and teaching:

1. Acquiring stress and intonation patterns
2. Memorizing everyday phrases
3. Learning grammar and vocabulary

At the same time, the use of chants has a drawback for non-native speakers of English. It is crucial to recognize stressed words in chants. However, chant texts often do not mark stressed words because chants were originally designed for teachers who are native-speakers of English and who naturally recognize where to place the stresses. By contrast, non-native speakers of English, even teachers of English, have difficulties in recognizing stressed words in some cases. For instance, those who were not originally teachers of English but of other subjects are now in charge of English language teaching in primary schools in Japan. To reduce this difficulty, it is preferable that teaching materials for chants should explicitly mark stressed words for non-native teachers of English as well as for learners of English.

In order to predict stresses in chants, one could apply conventional pitch-accent prediction methods such as [2, 3]. However, although stresses in chants share similar properties with pitch accents, they seem not to be identical. Stresses in a chant text have special properties as will be described in Sect. 2. It is likely that one will have to modify the conventional pitch-accent prediction methods to achieve a good performance in stress prediction in chants. Nagata et al. [5] investigated how well a simple Hidden Markov Model (HMM) based method works on stress prediction in chants. They showed that the problem can be solved as a sequence labeling problem using HMMs where the input is a sequence of words or part-of-speech (POS) tags obtained from the chant text in question. At the same time, it was argued that, in stress prediction for chants, it is crucial to consider the properties of chants such as a constraint on the number of stressed words in a chant text.

Accordingly, we propose a stress prediction method specially designed for chants. This method exploits several sources of information including words, POSs, sentence types, and the constraint on the number of stressed words, which are relevant in stress prediction for chants. The proposed method is expected to be useful in supporting non-native teachers of English when they teach chants to students; it can provide them with the information about which word gets stressed in a given chant text. It should also be useful for them to create their own teaching materials, which teachers often do, from arbitrary texts. Note that

it is often the case that native speakers of English are not readily available in certain countries including Japan. In addition to supporting teachers, it can be applied to a chanting robot that interactively teaches the English rhythm based on chants as Nagata et al. [5] originally proposed. It is crucial for such chanting robots to real-timely recognize stressed words in the utterances.

The rest of this paper is structured as follows. Section 2 explores chants in more detail, which is necessary to discuss the proposed method. Section 3 describes the proposed method. Section 4 describes and discusses experiments conducted to evaluate the proposed method.

2. Looking into chants

There are some basic tendencies in which words get stressed in chants. Content words such as nouns and verbs tend to get stressed more often than function words such as determiners and prepositions. This implies that information on POSs is crucial for stress prediction. Also, information on words plays an important role since some of the words that fall into the same POS category get stressed and others do not. For example, while the words *you* and *it* fall into the same category *pronoun*, the former tends to get stressed more often than the latter. Therefore, information on both words and POSs needs to be considered in stress prediction.

One factor which is not as obvious as words and POSs is sentence types. In questions, interrogatives such as *where* and sometimes auxiliaries such as *does* get stressed as in *Where is my hat?*. Correlated with this is the relation between sentence types. The determination of stressed words in a sentence is sometimes influenced by the type of its previous sentence. For example, if the previous sentence is a *where*-question as in the above example, one of the prepositions in the next sentence is likely to get stressed (e.g., *It's on the table.*).

Another important factor is the constraint on the number of stresses in a chant text; it is constrained to be a multiple of eight. This may seem to be somewhat odd, but is explained as follows. Chants are normally performed with music that progresses regularly in 4/4 time (recall that chants are formally *Jazz Chants*) where each beat corresponds to each stressed word. Music is often based on two bars (i.e., motive), which consists of eight beats in 4/4 time, or their multiples (e.g., 16 beats in four bars, 24 beats in six bars, ...). Consequently, the number of stresses in a chant text is constrained to be a multiple of eight. It should be emphasized that null stressed words are sometimes inserted in a chant text to satisfy the constraint (e.g., "*Black, yellow, brown. NULL. Jack fell down. NULL*" [1] where *NULL* denotes a null stressed word). Null stressed words are not actually pronounced but can be expressed with physical activities such as a clap.

3. Proposed Method

The stress prediction task can be solved as a sequence labeling problem. The sequence of observed values is the sequence of words in a given chant text. The labels are binary and denote whether the word gets stressed or not. Take for example a sentence in the textbook for chants [1]:

* * * *
Frank, Hank, walk to the bank.

This can be alternatively expressed with a sequence of labels *S* and *N*:

Frank/S, Hank/S, walk/S to/N the/N bank/S.

where S and N denote stress and not-stress, respectively (hereafter, S and N will be used to denote stress and not-stress).

To solve the sequence labeling problem, we use conditional random fields (CRFs) [4], which have been shown to be effective in sequence labeling. One of the reasons why we use CRFs is that it can handle several sources of information. As discussed in Sect. 2, the determination of stressed words in chants relies on several factors including information on words, POSs, and sentence types. Also, as we will see bellow, CRFs have several favorable properties in stress prediction.

We use four types of features in stress prediction: (i) words, (ii) lemmas of words, (iii) POSs, and (iv) sentence types. For (i) to (iii), we set the window size to five: current word, two previous words, and two following words. In addition, we include bi-grams and tri-grams derived from them in the features: bi-grams consisting of the previous word and the current word, and the current word and the following word; tri-grams consisting of the previous word, the current word, and the following word. For (iv), we consider the combinations of the current word (or the lemma of the current word), the type of the sentence in which the current word appears, and the type of the previous sentence; the sentence types are declarative, *yes/no*-question, *what*-question, *where*-question, *when*-question, *who*-question, *why*-question, and *how*-question. These are the features we use in the proposed method. In this paper, we limit ourselves to first-order Markov model features to encode inter-label dependencies.

With CRFs and these features, we can make basic predictions. First, we break down the input chant text into feature vectors. Then, we put the feature vectors into CRFs to obtain predictions and the corresponding probabilities.

To satisfy the constraint on the number of stresses in a chant text, we can exploit the conditional probabilities predicted by CRFs. We search the N-best prediction results for the label sequences that satisfy the constraint. Among them, we can simply choose the one that maximizes the conditional probability as the prediction result. This is another advantage of using CRFs.

In addition to the constraint, we consider the distribution of the length between stress-intervals. Here, we define a stress-interval as an interval between a stressed word and the word before the next stressed word[3]. For example, there are three stress-intervals *Frank*, *Hank*, and *walk to the* in *Frank/S, Hank/S, walk/S to/N the/N bank/S*. Theoretically, one can put as many words as one wants in a stress-interval in English. Practically, however, too many words in a stress-interval (or too long stress-interval) make it difficult to pronounce the stress-interval properly. Accordingly, the length of stress-interval is expected to be distributed among certain lengths. In other words, there might be a prediction error in a too long stress-interval predicted by CRFs.

To consider the distribution of the length of stress-intervals, we have to solve two technical problems: (1) how to measure the length of stress-intervals and (2) how to combine the distribution with CRFs. In this paper, we measure the length of a stress-interval by the number of words in it[4]. To solve the second problem, we assume that the length follows the Gaussian distribution. Under this assumption, we can calculate the probability of the length of a stress-interval once we estimate the

[3] If the first word in a chant text is not stressed, then the stress-interval is between the first word and the word before the first stressed word. Similarly, if the last word is not stressed, then the stress-interval is between the last stressed word and the last word.

[4] We also used the number of syllables instead of the number of words. However, it did not make any difference in the prediction performance. Therefore, we selected the number of words as the length of stress-intervals, which is much easier to count.

mean and variance of the length. Then, we can combine it with CRFs by simply multiplying the probabilities (this is another reason why we use CRFs).

We calculate the probability of the length of stress-intervals as follows. We first estimate the mean and variance of the length. To do this, we predict stressed words in the chant text in question by using the CRFs mentioned above. At this point, we do not consider the constraint on the number of stressed words. Instead, we use the label sequence that maximizes the probability obtained by the CRFs. To formalize the calculation, we will denote the number of stress-intervals in the prediction result as M. Also we will denote the length of the m-th stress-interval as l_m. Then, we estimate the mean and variance by:

$$\mu = \frac{1}{M}\sum_{m=1}^{M} l_m \tag{1}$$

and

$$\sigma^2 = \frac{1}{M-1}\sum_{m=1}^{M}(l_m - \mu)^2, \tag{2}$$

respectively. Using Eq. (1) and Eq. (2), we can calculate the probability of the length l by

$$f(l) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(l-\mu)^2}{2\sigma^2}\}. \tag{3}$$

Since we have M stress-intervals in a chant text, we take the geometric mean of the probabilities, which is given by

$$F = \sqrt[M]{\prod_{m=1}^{M} f(l_m)}. \tag{4}$$

This value evaluates how good a prediction result is, solely relying on the length of stress-intervals.

To make the final prediction, we combine the geometric mean with the prediction results obtained by the CRFs. For the N-best results obtained by the CRFs that satisfy the constraint on the number of stressed words, we calculate

$$s = FP \tag{5}$$

where P denotes the conditional probability given by the CRFs. We choose the one that maximizes Eq. (5) as the final prediction. In other words, we make a prediction considering features around the word in question, the constraint on the number of stressed words, and the length of stress-intervals. If none satisfies the condition, we simply choose the label sequence that maximizes the probability obtained by the CRFs.

4. Experiments

For evaluation, we used 71 chant texts in the textbook[1][5], which are manually annotated with stresses. In the experiments, we assumed that null stressed words were given and we excluded them from the evaluation. In all, the 71 chant texts consisted of 2,396 tokens and 1531 stressed words.

To measure the performance, we used recall, precision, F-measure, and accuracy. Recall and precision were defined by

$$R = \frac{\text{Number of stressed words correctly predicted}}{\text{Number of stressed words}} \tag{6}$$

and

$$P = \frac{\text{Number of stressed words correctly predicted}}{\text{Number of words predicted to be stressed}}, \tag{7}$$

respectively. F-measure was defined by

$$F = \frac{2RP}{R+P}. \tag{8}$$

Accuracy was defined by

$$A = \frac{\text{Number of chant texts without prediction error}}{\text{Number of chant texts}}. \tag{9}$$

All measures were calculated by leave-one-out cross-validation [6] (one text was left out each time).

For comparison, we implemented five methods in addition to the proposed method. The first is a baseline where all tokens are tagged as S (Baseline). The second is the conventional method [5] based on the POS tri-gram HMMs (HMM)[6]. The third and fourth are based on CRFs, but use only word features and POS features, the same used in the proposed method, respectively (CRF word only and CRF POS only). The fifth is the proposed method without the constraint of the number of stressed words and the length distribution (CRF). The sixth is the proposed method without the length distribution (CRF constraint).

Table 1 shows the experimental results. Table 1 reveals that all CRF-based methods outperform the HMM-based method. Even "CRF word only" or "CRF POS only" perform better than the HMM-based method. This is because the CRF-based methods exploit information before and after the word in question including bi-gram and tri-gram features unlike the HMM-based method. The CRF-based method further improves when it combines POS features with word features as we expected (i.e., CRF). Basically, POSs are informative for determining which word to stress as Table 1 shows that "CRF POS only" performs better than "CRF word only". However, information on words is required in some cases. For instance, the word *I* tends to get stressed and the word *it* does not although both fall into the same POS category *PRP*. In other words, it is crucial to exploit both sources of information in stress prediction.

Table 1: Experimental results

Method	R	P	F	A
Baseline	**1.00**	0.639	0.780	0.281
HMM POS	0.914	0.853	0.883	0.423
CRF word only	0.915	0.893	0.904	0.451
CRF POS only	0.933	0.903	0.918	0.465
CRF	0.944	0.923	0.933	0.507
CRF constraint	0.944	0.925	0.934	**0.592**
Proposed	0.946	**0.926**	**0.936**	**0.592**

R: Recall, P: Precision, F: F-measure, A: Accuracy

The information on sentence types should be useful for stress prediction. However, with or without the sentence type features, the CRF-based methods performed similarly in most cases; "CRF" achieved an F-measure of 0.934 without the sentence type features. A possible reason for this is that the information on sentence types may not be efficiently coded in the

[5]We corrected some stress marks, which seemed to be typographical errors in three chant texts out of 71 by consulting a professional chants trainer and the accompanying CD.

[6]We chose the POS tri-gram HMMs because they perform better than the word tri-gram HMMs according to Nagata et al. [5].

feature vectors. Thus, we may need to explore a more efficient way of coding the information on sentence types.

At first sight, the constraint on the number of stressed words has no or very little effect on stress prediction, comparing "CRF constraint" with "CRF". However, it should be noted that "CRF constraint" improves in accuracy (0.507 to 0.592). As already explained, chant texts tend to satisfy the constraint on the number of stressed words and "CRF constraint" (and the proposed method) make predictions, trying to satisfy the constraint. This implies that it is expected to require no human intervention 60% of the time when the proposed method is applied to annotating stressed words in a given chant text. This is an advantage of the proposed method in supporting non-native teachers of English when they teach chants to students or create their own chant texts. Especially, the proposed method tended to make no prediction errors in chant texts that did not contain dialoges and/or special intentions (which will soon be explained below).

By contrast, the distribution of the length between stress-intervals seems to have little effect. It was not often the case that too short or too long stress-intervals occurred in the prediction results in the experiments. Consequently, the distribution of the length rarely seemed to contribute to the prediction. Another possible reason is that we assumed that the length followed a Gaussian distribution, which is a continuous distribution. However, the length is discrete because it is measured by the number of words (or syllables). Therefore, we might need to use other distributions.

So far, the discussion has shown that the proposed method is effective in stress prediction. However, there are still some false positives and negatives. False positives and negatives often occur when stressed words are determined by a special intention of the chant text. Take for example the sentences *What does he want? He wants one egg.* In the standard manner, it is stressed as follows: *What/S does/N he/N want/S? He/N wants/S one/N egg/S.* However, one could put stress on the word *one* instead of the word *he* to intend that he wants ONE egg. The proposed method does not really handle the intention of a given chant text. It is indeed difficult to understand and deal with such an intention, as in these cases, with using existing techniques. Considering this, it would be a better strategy to interactively determine which word gets stressed. Teachers can first apply the proposed method to obtain the standard stress prediction results, which do not contain special intentions. Then, they can modify the results according to their intentions. Although it would be difficult for machines to predict special intentions, it is relatively easy for teachers to articulate their intentions and specify which words contain the intentions when they create their own chants. For this, a software tool would be useful to modify stress prediction results. Alternatively, teachers can first mark stressed words that contain special intentions, and then a machine can be used to determine the rest. The proposed method should be useful to achieve it.

Correlated with this are dialogs such as question and answer (e.g., *Where's my hat? It is on the table*). Although the proposed method tries to deal with stress prediction in dialogs by using the sentence type features, it turned out that they did not work well in the experiments, as explained above.

Information on rhyme may also be useful in stress prediction. Rhyme is sometimes used to teach the stress and intonation pattern of English language. Take for example an excerpt from the chant text [1] *Mike/S likes/N to/N bike/S. Tim/S likes/N to/N swim/S.* Information on rhyme can be coded in the feature vector (e.g., rhyme=yes). It is expected to be useful when the word in question is unseen in the training data but is rhymed in the chant text.

In addition to reducing false positives and negatives, the proposed method needs to be improved in another area. In the experiments, we assumed that null stressed words were given. In a real application, however, one needs to generate null stressed words. This is a difficult task which comprises two problems: (i) how many null stressed words are needed and (ii) where null stressed words should be generated. A simple idea for solving the first problem is that if the prediction result does not satisfies the constraint on the number of stressed words, we can add some null stressed words to the prediction result and evaluate whether the probability improves or not. For the second problem, we should take into consideration that null stressed words are normally put between sentences. How to generate null stressed words will be one of our future works.

5. Conclusions

In this paper, we described a method for stress prediction for automatically predicting stressed words in chant texts. We proposed exploiting several sources of information which are relevant to stress prediction by using CRFs. We also proposed methods for satisfying the constraint on the number of stressed words in a chant text and for considering the distribution of the length of stress-intervals. The experiments showed that the proposed method achieved an F-measure of 0.936 and outperformed the other methods implemented for comparison. The proposed method is expected to be useful in supporting non-native teachers of English when they teach chants to students and create chant texts with stress marks from arbitrary texts.

In future work, we will explore methods for generating null stressed words. We will also explore how we can apply the proposed method to chanting robots that interactively teach English rhythm based on chants.

6. Acknowledgments

We would like to thank Miwako Oe, who is a professional Jazz chants trainer, for her useful comments on this work.

7. References

[1] Graham, C., Creating Chants and Songs, Oxford, 2006.

[2] Gregory, M.L. and Altun A., "Using Conditional Random Fields to Predict Pitch Accents in Conversational Speech," Proc. of 42nd Annual Meeting on Association for Computational Linguistics, pp.47–54, July 2004.

[3] Margolis, A. and Ostendorf, M., "Acoustic-based Pitch-accent Detection in Speech: Dependence on Word Identity and Insensitivity to Variations in Word Usage," Proc. of 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.4514–4516, Apr. 2009.

[4] Lafferty, J., Andrew, M., and Pereira, F., "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," Proc. of International Conference on Machine Learning, pp.282–289, June 2001.

[5] Nagata, R., Mizumoto, T., Funakoshi, K., and Nakano M., "Toward a chanting robot for interactively teaching English to children," Proc. of INTERSPEECH Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology, P2-13, Sep. 2010.

[6] Witten, I.H. and Frank, E., Data Mining — Practical Machine Learning Tools and Techniques with JAVA implementations, Morgan Kaufmann Publishers, 1999.