An Application of Bayes’ Theorem to Evaluate Synthetic Speech for Computer-Assisted Language Learning

MIN KANG, HARUMI KASHIWAGI*, YAHENG ZOU, KAZUHIRO OHTSUKI, MAKOTO KABURAGI**
Graduate School of Intercultural Studies, *School of Languages and Communication, **Institute for Promotion of Higher Education
Kobe University
Tsurukabuto 1-2-1, Nada, Kobe 657-8501
JAPAN
kang@kobe-u.ac.jp

Abstract: - The aim of this paper is to propose an approach that applies Bayes’ theorem to evaluate synthetic sounds. It is assumed that the approach can provide teachers a quantitative scale to determine whether a synthetic sound is suitable to use or not when they create materials for computer-assisted language learning. The raw data were collected in the evaluation experiments that were conducted to test student evaluation of synthetic speech in computer-assisted English vocabulary learning. Bayes’ theorem was adopted to estimate evaluation ratings of phonemes and scores of words.

Key-Words: - synthetic sound, computer-assisted language learning, evaluation, data analysis, Bayesian approach

1 Introduction
In the past few decades, there has been remarkable progress in speech synthesis. Speech outputs generated by synthesizers have been significantly intelligible although some still lack the naturalness. Text-to-speech (TTS) synthesizers have been used in a variety of fields such as services in help desks and voice response systems, applications for the visually impaired and so on [1, 2]. Recent years, speech synthesis can be broadly found in electronic dictionaries or online dictionaries. Its practice use has come to the stage that synthetic speech is adopted in foreign language learning.

On the other hand, computer assisted language learning (CALL) has become a popular way in language learning. In classroom learning, it is always a negative factor that students have few opportunities to expose themselves to the target language. The CALL is an effective learning style to improve the situation [3]. One can access web learning contents including texts, audios or videos outside classrooms so that he/she has more opportunity to expose himself/herself to the inputs that he/she has learned in classrooms. In the case, materials created by teachers play a key role in enhancing students’ language skills. However, preparing materials, especially audios or videos usually costs lots of efforts of teachers because they sometimes are required to have higher computer skills or difficultly find native-speakers to help record sounds.

A CALL system with a TTS synthesizer will definitely reduce the burden of teachers mentioned above. In fact, synthetic speech has been used in lots of CALL materials. The question is if synthetic sounds have the same effectiveness as natural sounds in enhancing listening skills of students. In our previous works [4, 5], we suggested that evaluating synthetic speech from the viewpoint of language learning should be one of solutions to answer the question. In this case, student opinion is a key factor to judge whether the speech is suitable to use as native-speaker’s in listener tests of quality assessment of a TTS system. We have investigated subjective evaluation of synthetic sounds by comparing with natural sounds for English words and sentences. The evaluation experiments were performed to measure how students assess the synthetic sounds and natural sounds. The results showed us that a large part of the synthetic sounds was assessed as same as natural sounds. We concluded that synthetic speech should be selectively utilized in language learning. In order to selectively utilize synthetic sounds, it is helpful for teachers if there are quantitative evaluation scores of synthetic sounds to help make the choice when they create listening materials. The scores may obtained by directly assess all sounds in the target language.
However, the way is not practical. Therefore, a method to predict the scores is necessary. The purpose of this paper is to propose an approach to estimate the evaluation of synthetic speech and provide quantitative evaluation scores for English words. Bayes’ theorem is applied to make the estimation of phoneme evaluation and calculate the evaluation scores of words according to the estimation. Two evaluation experiments are conducted to verify if the estimation possibly reflects the real evaluation of words.

2 Data for Estimation
Both of the data used for estimation and verification consist of values of 1 to 5 that mean ratings in a 5-point rating scale. The data were collected by five evaluation experiments in which three have been carried out in our previous works and present the raw data for the estimation in the paper. Another two evaluation experiments were conducted to compare the calculated scores with the measured values in order to verify the approach.

2.1 Criteria and Word-sets in Experiments
In quality assessment of a TTS system, “intelligibility/comprehensibility”, “naturalness”, “listening difficulty”, “acceptability” and so on are generally used as the criteria to ask native-speaker subjects’ opinions. For nonnative-speaker subjects, the criteria should be as simple as possible because of the problem of listening skill. We merely use one characteristic in the evaluation experiments to avoid confusing subjects. That is expressed as “聞き取りやすさ” in Japanese. In foreign language teaching in Japan, the expression is frequently used to ask students whether the voices they heard are easy to understand and acceptable. By the use of the criterion, subjects may easily give their opinions.

Table 1 shows the word-sets used in our previous works. According to the analysis for the data in Table 1, the sounds with the vowels such as /ei/, /u:/, /i/ were assessed to be easy to understand and acceptable while the sounds with fricative consonants were difficult to understand. In Table 2, we chose the words with the vowels of /ei/, /u:/ in the first word-setand mainly selected the words with fricative consonants in the second word sets. The sounds of the words were generated in normal speed by AT&T natural voices, a concatenative TTS system.

<table>
<thead>
<tr>
<th>No. of</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (15 words)</td>
<td>Vegetable, arrive, live, solve, fiftieth, theater, theory, Thursday, breathe month, familiar February photograph laugh wolf</td>
</tr>
<tr>
<td>II (12 words)</td>
<td>violet visitor active receive wave clothes death smooth tooth fortunate enough roof</td>
</tr>
<tr>
<td>III (9 words)</td>
<td>dish, look, today, keep, food, church, break, tower, go</td>
</tr>
</tbody>
</table>

2.2 Experiments and Subjects
In all experiments, the subjects were requested to listen to each sound and then select one choice in a multiple-choice question that was described as “Please give your evaluation of the sound by selecting one choice from the five ratings” The five ratings corresponded to the 5-point scale of +5 (Choice 1), +4 (Choice 2), +3 (Choice 3), +2 (Choice 4), and +1 (Choice 5). Choice 1 described that the speech was “very easy to understand and entirely acceptable,” and Choice 2 denoted that it was “easy to understand and mostly acceptable.” Choice 3 expressed that it was “neither easy nor difficult to understand and acceptable.” Choices 4 and 5 indicated that the speech was “difficult to understand and somewhat acceptable” and “don’t understand and unacceptable,” respectively. Meanwhile, the subjects were asked to write down the words and the translations what they heard.

All subjects are Japanese learners of English who are freshmen or second-year students in a variety of
majors. Empirically, there is little variance in listening skills of the students. There were 6 classes making a total student of 233 in Experiment I, II and III that give the raw data for estimation. Experiment IV and V were carried out in two classes, e.g. Class 1 and Class 2, with 41 and 42 students.

3 Application of Bayes’ Theorem

The key idea is if we can estimate the evaluation ratings of all phonemes, it may be easy to obtain an evaluation score of arbitrary word by the calculation of the mean of all expected values of the phonemes that constitute the word. An evaluation rating of a phoneme corresponds to the posterior probabilities of Choice 1, Choice 2, Choice 3, Choice 4 and Choice 5, given the phoneme.

3.1 Posterior Probability

Equation (1) expresses how to calculate the posterior probability that a phoneme Ph is assessed at the rating level i (i=1,2,…,5).

\[
P(i/\text{Ph}) = \frac{P(\text{Ph}/i)P(i)}{\sum_{m=1}^{n} P(\text{Ph}/m)P(m)}
\]  

(1)

where \(P(i)\) and \(P(m)\) represent the marginal probabilities that the rating i and the rating m are chosen, respectively. \(n\) denotes the largest number of the ratings and equals to 5 in this paper. \(P(\text{Ph}/i)\) means the conditional probability of the phoneme Ph, given the rating i.

3.2 Expected Values of Phoneme Evaluation

In the database for Experiment I, II and III, the evaluation ratings of words, evaluated words, words written by students, the translations written by students and other information such as subject number and name are stored. Here, we use the variable X to denote the evaluation ratings, the variable W to describe evaluated words. A subject gives one choice for each word so that the number of observed values \(N\) equals to the product that the number of subjects times the number of the evaluated words. Hence, The marginal probability \(P(i)\) (i=1,2,…,5) is computed by Equation (2).

\[
P(i) = \frac{n(\{x; x = i\})}{N}
\]  

(2)

where, \(n(\{x; x=i\})\) expresses the number of the rating i in the variable X. The conditional probability is calculated by Equation (3).

\[
P(\text{Ph}/i) = \frac{n(\{x; x = i\} \cap \{w; \text{Ph} \in w\})}{n(\{x; x = i\})}
\]  

(3)

where \(n(\{x; x=i\} \cap \{w; \text{Ph} \in w\})\) represents the number of the evaluated words that are assessed at the rating i and contain the phoneme.

Based on Equation (2) and (3), we can gain each posterior probability of a phoneme at each rating level by Equation (1), and then the expected value of the phoneme can be computed. Therefore, the evaluation of a word can be estimated to be the arithmetic mean of all expected values of the phonemes of that the word is composed if there is no difference in contribution to the evaluation of the word between phonemes. In the paper, we used the hypothesis and defined a quantitative evaluation score of a word as an arithmetic mean of all expected values of the phonemes that constitute the word.

Table 3. Expected values of the phonemes constituting the words in Experiment I, II and III.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Value</th>
<th>Consonant</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ï</td>
<td>2.00</td>
<td>j</td>
<td>2.28</td>
</tr>
<tr>
<td>ε**</td>
<td>2.37</td>
<td>p</td>
<td>2.30</td>
</tr>
<tr>
<td>ì*</td>
<td>2.39</td>
<td>g</td>
<td>2.48</td>
</tr>
<tr>
<td>θ*</td>
<td>2.44</td>
<td>θ</td>
<td>2.49</td>
</tr>
<tr>
<td>i*</td>
<td>2.46</td>
<td>f</td>
<td>2.52</td>
</tr>
<tr>
<td>ou</td>
<td>2.68</td>
<td>w</td>
<td>2.57</td>
</tr>
<tr>
<td>æ**</td>
<td>2.71</td>
<td>m</td>
<td>2.68</td>
</tr>
<tr>
<td>Λ*</td>
<td>2.74</td>
<td>r</td>
<td>2.72</td>
</tr>
<tr>
<td>ε*</td>
<td>2.77</td>
<td>b</td>
<td>2.80</td>
</tr>
<tr>
<td>θ</td>
<td>2.77</td>
<td>l</td>
<td>2.81</td>
</tr>
<tr>
<td>ι</td>
<td>2.80</td>
<td>n</td>
<td>2.83</td>
</tr>
<tr>
<td>u*</td>
<td>2.86</td>
<td>δ</td>
<td>2.93</td>
</tr>
<tr>
<td>Λ</td>
<td>2.92</td>
<td>t</td>
<td>2.96</td>
</tr>
<tr>
<td>θ*</td>
<td>3.03</td>
<td>v</td>
<td>3.11</td>
</tr>
<tr>
<td>au</td>
<td>3.13</td>
<td>k</td>
<td>3.12</td>
</tr>
<tr>
<td>ai*</td>
<td>3.18</td>
<td>s</td>
<td>3.13</td>
</tr>
<tr>
<td>u*</td>
<td>3.20</td>
<td>dʒ</td>
<td>3.15</td>
</tr>
<tr>
<td>æ*</td>
<td>3.25</td>
<td>jf</td>
<td>3.24</td>
</tr>
<tr>
<td>θ*</td>
<td>3.44</td>
<td>z</td>
<td>3.28</td>
</tr>
<tr>
<td>et*</td>
<td>3.48</td>
<td>d</td>
<td>3.43</td>
</tr>
<tr>
<td>et**</td>
<td>3.51</td>
<td>j</td>
<td>3.54</td>
</tr>
</tbody>
</table>

4 Estimation and Discussion
According to the section 3, we calculated the evaluation ratings of all phonemes that appeared in Experiment I, II and III. The data in Experiment IV and V were used as observed values in word evaluation to compare with evaluation scores computed by evaluation ratings of phonemes.

4.1 Evaluation Ratings of phonemes

Table 3 shows the evaluation ratings calculated by using of the data in Experiments I, II and III. Here, the symbol * and ** denote vowels carrying primary stress and secondary stress, respectively and the IPA are used to express the phoneme.

What do the expected values mean? We notice that /eɪ/ and /eɪ/ have comparably high expected values whereas /iː/ obtain the lowest one. As can be confirmed in Table 1, the values of /eɪ/ and /eɪ/ are mainly influenced by the data of “Thursday”, and “wave”, “break”, “today”, respectively. Because most of subjects judged the words as 3 or above, the value of /eɪ/ and /eɪ/ reflected the influence. On the other hand, the value of /iː/ was related to the words “theory”, “fiftieth” and “February” that were assessed as being “difficult to understand and somewhat acceptable” or “don’t understand and unacceptable” by lots of subjects. In consonants, the thing is the same. The /j/ appears merely in the words “February” and “familiar” that belonged to the word-set with fricative consonant /f/ and have lower evaluation values than the word-sets with other fricative consonants in our previous works. As analyzed above, the expected values do depend on the word evaluation ratings. In other words, the values, e.g. the evaluation ratings of phonemes, would properly reflect the phoneme evaluation if there is no biased in the raw data of word evaluation.

4.2 Comparing Observation to Expectation

We calculated the evaluation scores according to the values in Table 3 and the mean ratings based on the evaluation ratings for the words in Experiment IV and V. Figure 1 and 2 draw the results with the standard errors. The results were sorted by the evaluation scores.

The figures show that there is a good consistence in value for a part of words. According to the value of the vowel /iː/ in Table 3, we can infer that there may be a difficulty in understanding some words that include the vowel /iː/. As showed in Figure 1, the score (2.77) of “leave” was determined by the evaluation ratings of /l/, /iː/ and /v/ and was obviously influenced by the evaluation rating of /iː/. The agreement between the score and the mean ratings confirmed the inference. In fact, most of subjects wrongly recognized “leave” as “live”. On the other hand, there is a difference between the mean ratings and the score of “keep”. Subjects around 90% correctly understood the word.

In Figure 2, there is a large gap between the score and the mean ratings of “these”. The observed data in both classes are lower than the score that is equal to the arithmetic mean of /ð/ (2.93), /iː/ *(2.39) and /z/ (3.28). Clearly, the evaluation rating of /z/ plays a main role on causing the gap.

As a result, the estimation may properly reflect the real evaluation for some words, but many may not be correctly predicted. There are lots of factors causing
the result. The bias in sampling of the evaluated words may induce wrong evaluation ratings of phonemes. The evaluation of a word may not exactly depend on the arithmetic mean of the expected values of the phonemes that constitute the word. The weighted mean may be a better choice. In addition, syllable and prosody are indicated to be important factors that influence word recognition. Using the units in the estimation may improve the accuracy. Moreover, it is crucial aspect to introduce the evaluation data natural sounds into the estimation. The comparison between natural sounds and synthetic sounds may provide teachers more clear information to determine whether a synthetic sound should be use or not. The aspect has been included in our going-on researches.

Nevertheless, the result in the paper shows that the estimation approach possibly provide quantitative evaluation scores of words to help teachers choose synthetic sounds of words in language learning. However, more data is needed to definitely verify the efficacy of the approach and present more exact scores.

5 Conclusions

In the paper, we proposed an approach to estimate the evaluation of synthetic sounds and provide quantitative evaluation scores for English words. Bayes’ theorem was applied to calculate the evaluation ratings of phonemes and the evaluation scores of words were estimated according to the ratings. Two evaluation experiments were conducted to verify if the estimation possibly reflects the real evaluation of words. The verification denoted that the estimation possibly reflects the real evaluation of words but further studies are needed.

References: