2016 10th Workshop of Tohoku University, University of Electronic Science and Technology of China and Yeungnam University

# Detection of singing mistakes from singing voice

Isao Miyagawa
Department of Communications Engineering
Tohoku University
Sendai, Japan
miyagawa@spcom.ecei.tohoku.ac.jp

Yuya Chiba, Takashi Nose, Akinori Ito
Department of Communications Engineering
Tohoku University
Sendai, Japan
{yuya@spcom.ecei, tnose@m,
aito@spcom.ecei}.tohoku.ac.jp

*Abstract*— **We investigate a method of detecting wrong lyrics from the singing voice for karaoke. In the proposed method, we compare the input singing voice and the reference singing voice using Dynamic Time Warping, and then observe the frame-by-frame distance to find the error location. However, the absolute value of the distance is affected by the speaker individuality of the reference and input singing voice. Thus, we attempted to normalize the speaker individuality by linear transformation. The results of the experiment showed that we could detect the wring lyrics with high accuracy when the different part of the lyrics was long.**

***Keywords-singing voice; frame-by-frame distance***

## I. INTRODUCTION

In the recent years, the karaoke culture is popular in all generations. People go to karaoke with friends or alone to release stress. Besides the backing sounds, the scoring game is very popular as a way of enjoying karaoke.

When singing a song alone, it is difficult to notice wrong lyrics. If the karaoke machine could point out the singing mistakes, it could be another game, and also it contributes improving the singing skill of the singer.

There are several studies on the evaluation of the singing voice for karaoke, but most of them are to evaluate musical characteristics or voice quality, such as the accuracy of the pitch and length of the singing voice [1], singing skill including singing technique [2], singing enthusiasm [3], and so on. However, the lyrics is not treated in these studies at all, because it is very difficult to recognize the lyrics from the singing voice [4].

Two types of methods can be considered for detection of wrong lyrics. One is to match the input singing voice and the text information of lyrics. For example, we make the model of the right lyrics using the Hidden Markov Model and calculate probability of the input voice [5]. However, singer adaption is necessary to evaluate the singing voice with high accuracy, which means that the singer-independent detection is difficult. On the other hand, there are “guide vocal” function in some karaoke [6]. This is to play the vocal part of the song by the natural or synthetic voice with accompaniment. By using this function, we can exploit the reference singing voice with right lyrics and detect error without preparing the text of lyrics. The error detection system does not depend on the language if we use the guide vocal function. Therefore, we aim at the detection of singing mistake with high accuracy by matching the input singing voice and reference singing voice.

## II. PROPOSED METHOD

### A. Overview of the method

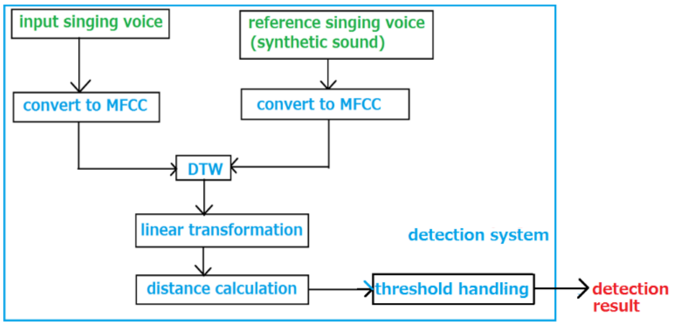

Figure 1. The proposed method

The overview of the proposed method is shown in Figure 1. The most basic method to measure the similarity of the feature vectors is to calculate the distance between vectors. The Euclidean distance between the feature vector of a frame in the input singing voice $\mathrm{y} = (y_1 \dots y_n)$ and the corresponding frame of the reference singing voice $\mathrm{r} = (r_1 \dots r_n)$ are determined by (1).

$$d(y,r) = \sqrt{\sum_{i=1}^{n} (y_i - r_i)^2} \tag{1}$$

We use this formula and determine that the part has a lyrics error when the distance is large. We use the MFCC (Mel Frequency Cepstral Coefficients) as the features, which is commonly used in the speech recognition. We calculate the frame-by-frame correspondence of the two sequences using the DTW(Dynamic Time Warping). The DTW is a nonlinear matching method of two sequences.

First, we convert the input and the reference singing voice into MFCC. Next, we compare the two MFCC sequences using the DTW based on the Euclidean distance. Finally, we detect the mistakes by thresholding the frame-by-frame distances.

### B. *Dynamic Time Warping*

The DTW is a nonlinear matching method of two sequences. We define the feature sequences of the input and the reference singing voice as $x_1 \dots x_n$ and $y_1 \dots y_n$, respectively. Then we calculate the cumulative distance $\mathrm{g}(i,j)$ of input voice frame (1~i) and reference voice frame (1~j) as follows.

$$\mathrm{d}(i,j) = \|x_i - y_i\|^2 \tag{2}$$

$$g(1{,}1) = d(1{,}1) \tag{3}$$

$$g(i,j) = d(i,j) + \min\begin{Bmatrix} d(i,j-1) + g(i-1,j-2) \\ g(i-1,j-1) \\ d(i-1,j) + g(i-2,j-1) \end{Bmatrix} \tag{4}$$

After calculating all the cumulative distances, the optimum correspondence is found by backtracing the minimum path of the cumulative distance.

## III. LINEAR TRANSFORMATION AND SMOOTHING

When we associate two sequences by the DTW and determine the distance, the distance of error location is larger when the singer of the input and reference singing voice is the same, but distances of the correct lyrics part is also large when the singers are different. Figure 2 and 3 show the square distances of the input and reference singing voice using DTW. Both of them are synthetic singing voice. Black color part indicates the wrong lyrics part. The vertical axis is the square distance, and the horizontal axis is the number of frame. Figure 2 shows the distance when the singers are the same, and Figure 3 shows that the gender of the singers is different and the key is 1 octave different. The frame-by-frame distance is large only in the error part in Figure 2, which is clearly different from the correct part. However, we can't discriminate the correct and error parts in Figure 3 because the frame-by-frame distance of the correct part is larger. This example suggests that we can't detect the error location using only the DTW. Therefore, we applied a method based on the linear transformation of the feature to adapt the feature vectors of a singer to another one's feature vectors [7]. Let the feature vector sequence of the input and reference singing voice as $(x_1 \dots x_n)$ and $(y_1 \dots y_n)$, respectively. Then we apply the linear transformation, $\mathrm{y} = \mathrm{A}x_i + e_i$, and we find the transformation matrix A that minimizes the sum of squares of error vector $e_i$.

$$Z = \sum_i \|e_i\|^2 = \sum_i \|y_i - Ax_i\|^2$$
$$= \sum_i (y_i - Ax_i)^T (y_i - Ax_i) \tag{5}$$

To minimize the Z

$$\frac{\partial}{\partial \mathrm{A}} \sum_i (y_i^T y_i - 2x_i^T A^T y_i + x_i^T A^T x_i) = 0 \tag{6}$$

To calculate this equation

$$\sum_i (-2y_i x_i^T + 2Ax_i x_i^T) = 0 \tag{7}$$

Thus,

$$C_{xx} = \sum_i x_i x_i^T \, , C_{yx} = \sum_i y_i x_i^T \tag{8}$$

$$A = C_{yx} C_{xx}^{-1} \tag{9}$$

This equation is assumed x and y are singing exactly the same lyrics, but a lyrics error is included in a real singing. At the time of linear transformation, it is inappropriate to use the error singing part because it becomes associated with different phonemes each other by force. Therefore, we exclude error-like part by using DTW to use only the collect part in the linear transformation. By doing so, we hold down that accociate an error part by force.

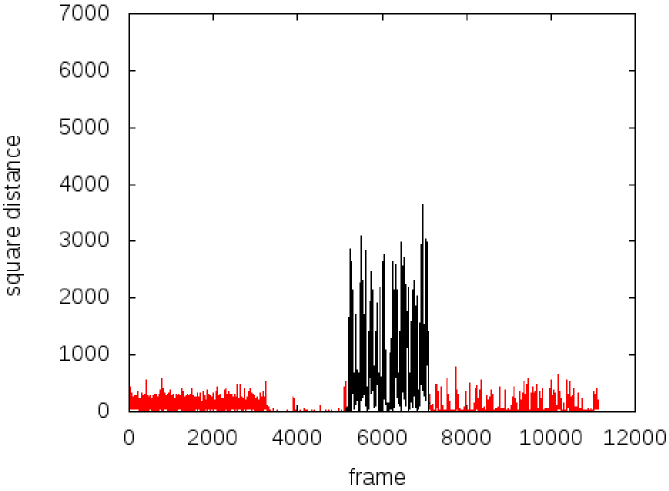

Figure 2. Square distance {Woman-woman(The same singer)}

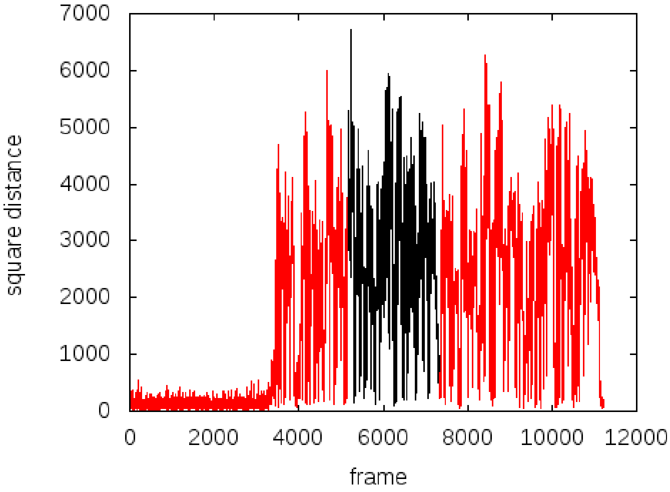

Figure 3. Square distance {Woman-man(1 octave below)}

Furthermore, we apply the linear transformation iteratively. In the first iteration, we roughly estimate the transformation matrix. Then we improve the estimation using correct lyrics part in the second iteration. By repeating DTW and linear transformation, we can reduce the distances of the correct parts while keeping the distance of the error location large.

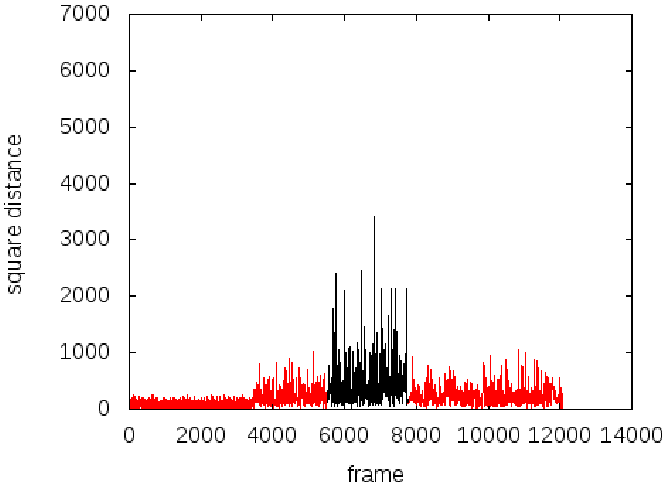

Figure 4. Square distance after linear transformation {Woman-man(1 octave below)}

An example of square distance after linear transformation are shown in Figure 4. This figure corresponds to Figure 3. The DTW and the linear transformation are iterated three times, and threshold is set to 3000. When comparing Figure 4 with Figure 3, the distances of the error part around 5500-7000 frame are larger than the other parts. From this example, it can be seen that the iteration of the DTW and linear transformation is effective. On the other hand, we can still see many peaks of the distance in the correct parts. Therefore, we introduce the smoothing of the distances by the moving average filter to reduce the influence of distance peaks of the correct parts.

## IV. EXPERIMENT

### A. *Detection experiment of the error lyrics*

We conducted an experiment to detect lyrics mistakes to investigate whether the proposed method could accurately detect the lyrics error. Singing voices of the11 male students were used as the input singing voice. We examined two songs, “Sekaini hitotsudakeno hana” and “Soramo toberu hazu”. 11 male singer sang the correct lyrics.

The reference singing voices are synthesized using Vocaloid. We prepared three reference singing voices: “Correct”, ”With different phrase”, ”With different words or syllables”. Thereafter “With different phrase” and “With different words or syllables” are defined as “large error” and “small error” voices. Here, we simulated the singing mistakes by changing lyrics of the reference singing voice. A reference voice with errors contains multiple lyrics mistake. Eight “large error” parts and nine “small error” parts are included in the reference singing voices, 17 error parts in total. There were six “large error” parts in the “Sekaini hitotsudakeno hana” and, There are two “large error” and nine “small error” parts in the “Soramo tobetu hazu”. As we examined the voices from the 11 persons, we have 187 errors to be detected.

The sound analysis condition are shown in Table 1. The DTW and linear transformation were iterated three times. After that, we smoothed the distance using the MA filter. The number of the smoothing frames was determined beforehand. The threshold is α times of the average of the all of the smoothed distances. If the distance of a frame is larger than the threshold, we detect it as the error part. We judged the correctness of the detection result by comparing the detection result with the original error part. Because the detected mistakes are given as frame sections, we regard the detected part as correct when the detected part and the labeled error part intersects each other. Otherwise, the detection is regarded as a false alarm. By changing the number of smoothing frames and α, we examined the change of F-measure.

TABLE I. SOUND ANALYSIS CONDITION

| | |
|---|---|
| Sampling frequency | 16kHz |
| Window length of MFCC calculation | 25ms |
| Frame shift | 10ms |
| Order of MFCC | 13(+power) |

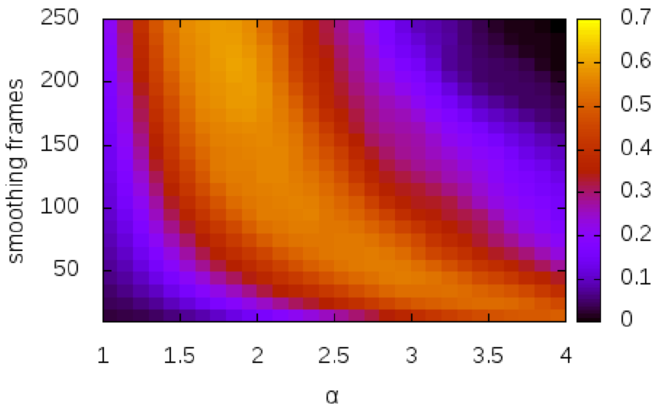

Figure 5. The result of F-measure

The F-measure with respect to the smoothing frame width and α is shown in Figure 5. The maximum value of F-measure is 0.61, which is not high enough. Therefore, we investigated the detection results of the large error and small error individually. The maximum F-measure of the large error is 0.90, whereas that of the small error is 0.23. From these results, the small errors are found to be difficult to detect.

### B. *The difference of detection by a person*

We investigated singer-by-singer difference of the detection performance. The results are shown in Figure 6. From Figure 6, we can see differences in the maximum values of F-measure, but the differences are not large.

### C. *The number of iterations of the linear transformation*

The DTW and linear transformation are iterated three times in the previous experiment but the optimum number iterations was not clear. Therefore we conducted an experiment to change the number of iterations of the linear transformation. The result are shown in Figure 7. From the figure, we found that the maximum value of F-measure does not change much with respect to the number of iteration. Therefore, only one linear transformations was considered to be enough. Note that the transformation in this experiment was conducted between the same genders; the optimum number of iterations for cross-gender transformation should be investigated.

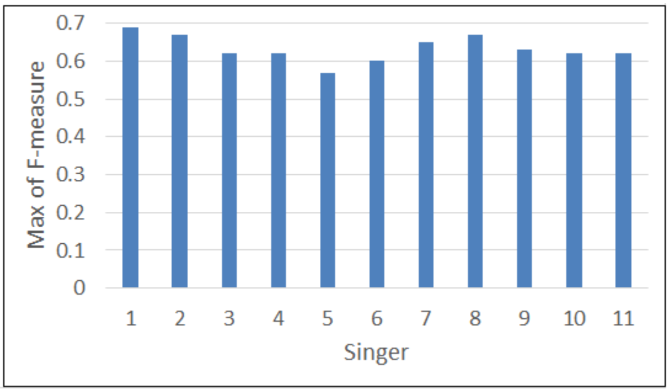

Figure 6. Max of F-measure for each singer

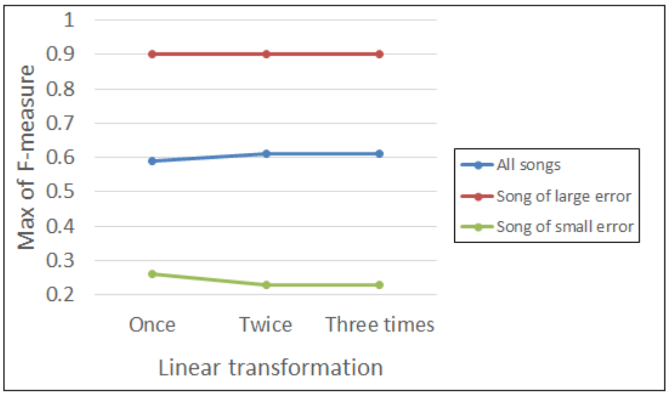

Figure 7. The number of times the linear transformation and the F-measure

### D. The phoneme and detection

We investigated whether the false detections are related to specific phonemes. The result is shown in Figure 8. The horizontal axis is the detected phoneme, and the vertical axis is the sum of the detection of that phoneme for all data of the 11 singers. From this result, the ratio of false detection was high when phoneme is /m/, /y/, and /z/. False detection ratio of /d/, /r/, /s/, and /t/ was also high. Among the vowels, ratio of false detection was high for /u/. The phoneme that was detected a lot as an error part was /n/, /r/, /t/. Furthermore, we found that the detection result was different singer by singer. In addition, the result may change by way of singing because the detection results of the phoneme are different singer by a singer.

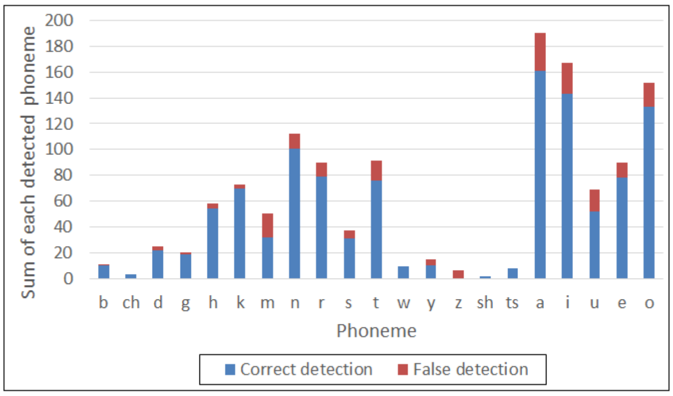

Figure 8. The number of correct detection and false detection

## V. CONCLUSION

In this research, we investigated a method of detecting wrong lyrics from the singing voice for karaoke. The result of the experiment showed that we could detect the wrong lyrics with high accuracy when the wrongly sang part was long. We investigated the singer-by-singer difference of the error detection. The result shows that the difference was small. Also, we found that one transformation was enough for singer normalization. We investigated the relation between the phoneme and error detection, but no distinct results were obtained.

## REFERENCES

[1] Takeuchi, H., Hoguro, M., Umezaki, T. “A KARAOKE System Singing Evaluation Method that More Closely Matches Human Evaluation”, The transactions of the Institute of Electrical Engineers of Japan. C Vol130, No.6, pp.1042-1053, 2010.

[2] Nakano, T., Goto, M., Hiraga, Y. “An Automatic Singing Skill Evaluation Method for Unknown Melodies”, Information Processing Society of Japan Vol48, No.1, pp.227-236, 2007.

[3] Daido, R., Ito, M., Makino, S., and Ito, A. “Automatic evaluation of singing enthusiasm for karaoke”, Compurter Speech&Language”, Vol28, pp.501-517, 2014.

[4] Mesaros, A. and Virtanen, T. “Automatic recognition of lyrics in singing”, EURASIP Journal of Audio, Speech and Music Processing, Vol2010, article No.4, 2014.

[5] Suzuki, M., Hosoya, T., Ito, A. and Makino, S. “Music Information Retrieval from a Singing Voice Using Lyrics and Melody Information”, EURASIP Journal on Advances in Signal Pricessing, Vol.2007.

[6] Panasonic: KARAOKE machine, JP-A-2001-42879,2001.

[7] Matsumoto, H. and, Inoue, H. ”A piece wise linear special mapping for supervised speaker adaptation”, Proc.ICASSP, Vol.1, pp.449-452, 1992