論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	An Automatic Singing Voice Evaluation Method for Voice Training Systems
著者(和文)	ﾊﾟｻｰﾄｳﾞｨｯﾀﾔｰｶｰﾝ ﾊﾟｻｰﾄ, 岩野 公司, 古井 貞熙
Authors(English)	Prasertvithyakarn Prasert, Koji Iwano, Sadaoki Furui
出典(和文)	日本音響学会2008年春季講演論文集, Vol. , No. 2-5-12, pp. 911-912
Citation(English)	, Vol. , No. 2-5-12, pp. 911-912
発行日 / Pub. date	2008, 3

An Automatic Singing Voice Evaluation Method for Voice Training Systems*

◎Prasertvithyakarn Prasert, Koji Iwano, and Sadaoki Furui (Tokyo Institute of Technology)

1 Introduction

In response to the increasing requirements for voice training systems, automatic singing voice evaluation methods need to be developed.

However, most of the research dedicated to information retrieval from human's voice focuses on speaking, not singing voice, and most of the research on singing voice focuses on fundamental properties such as pitch or rhythm and not the quality of the voice itself. Our research focuses on the evaluation of the spectral features of singing voice to measure its quality.

This paper proposes a GMM (Gaussian Mixture Model)-based singing voice evaluation method, in which 2 classes of the voice quality, good and bad singing voices, are separately modeled, and evaluates its performance.

2 Data Specification

The first step of the research is to define the quality of singing voice. We decided to use a standard vocalization in the classical music theory to define "good" voice from "bad" one.

Since there are not many singing voice databases available, and available ones do not have the label of the quality of voice, we decided to create the database by ourselves.

The database consists of voices of 10 veteran male singers from Tokyo Institute of Technology mixed-voice chorus Chor Kleines (Gold medal from All-Japan chorus competition for 10 years in a row). All singers sang with 2 styles of voice. For the first style of voice, we told them to sing "with a good classical voice for choir". We labeled this voice data as "good voice". For the other style of voice, we told them to imitate "the voice of the non-experience singer". We labeled this voice data as "bad voice". From now on we will use the term "good voice" and "bad voice" as defined here.

For both good and bad voices, each singer sang a phrase "DA-ME-NI-PO-TU" for 5 times, each time with the same melody but with different keys. The reason we chose this phrase is that, it is usually used in normal voice training and that the singers were familiar with it.

The record of the singing voice was performed in the recording room to ensure no noise effects. The voice was sampled at 16 kHz.

There was no guarantee that the voice sang by the singers actually had the same quality as they meant to; there were possibilities that some singers sang "bad" voice when they meant to sing "good" one, or even other singers might still sing a "good" one when they meant to imitate the voice of the non-experienced singers. Therefore, all of the voice files were evaluated by 40 volunteers consisting of 28 male and 12 female experienced-singers from the same choir. The evaluators judged whether each voice was good or bad (1: very bad – 5: very good).

Figure 1 shows the averaged scores for each singer's good and bad voices. It is found that human can easily recognize bad voice from good ones. However, in the case of one singer among 10, labeled as singer 6 in Figure 1, most of the evaluators evaluated this singers "bad" voice as a rather "good" one. We decided to exclude this singer's data from the further experiments.

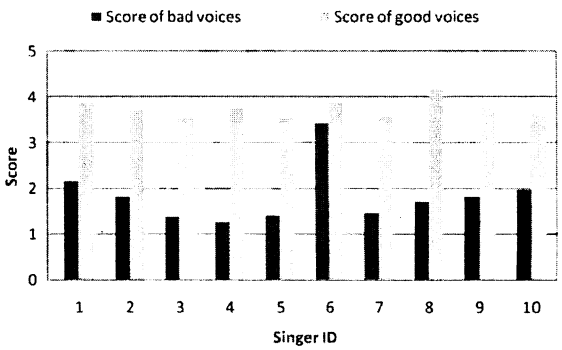

Figure 1: Evaluation of singer's good and bad voice by human

3 Voice Evaluation System

The system's final objective is to evaluate, in other words, classify, input singing voice data into either good or bad ones. First, the input singing voice phrase is automatically segmented into vowel periods. With the trained good and bad vowel models, a score for classification is computed.

The score of an observed vowel v's feature vector $\boldsymbol{o}_v$ denoted by $q(\boldsymbol{o}_v)$ is calculated by

$$q(\mathbf{o}_v) = \left\{\log p(\mathbf{o}_v|M_{gv}) - \log p(\mathbf{o}_v|M_{bv})\right\}/l_v \quad (1)$$

where $p(\boldsymbol{o}_v|M_{gv})$ and $p(\boldsymbol{o}_v|M_{bv})$ are likelihoods of the observed vowel feature vectors in the phrase for good and bad singing voice models, respectively. l_v refers to the length (number of frames) of a vowel v.

The score of the whole singing phrase $q(\boldsymbol{o})$ is simply calculated by

$$q(\mathbf{o}) = \sum_v l_v q(\mathbf{o}_v) \Big/ \sum_v l_v \quad (2)$$

To classify the voice into good and bad classes from a score, we set a threshold value θ, and define a discriminant function as

$$z = q(\mathbf{o}) - \theta \quad (3)$$

If z is positive, the voice is classified as a good one, and vice versa.

* ボイストレーニングシステムのための歌声発声の自動評価,
パサートウィットヤーカーン パサート, 岩野公司, 古井貞熙 (東工大)

3.1 Model Construction

In the step of model construction, we manually segmented each phrase into phonemes. With the GMM method, we trained both good and bad singing voice models for each vowel: /a, e, i, o and u/. Please note that consonant models were not trained.

In this research, we investigated 2 types of feature vectors, MFCC (Mel-Frequency Cepstrum Coefficient) with 11 dimensions and the FBANK vectors consisting of log mel-filterbank coefficients with 12 dimensions.

3.2 Dimension Weighting

Although there are 11 or 12 components in the feature vector, not all of them are relevant to voice quality. Thus, to enhance the vocal quality evaluation, we decided to take into account the relevance of each component in the evaluation.

To achieve this, feature vectors are modeled by a multi-stream method [1,2] as

$$\log p(\mathbf{o}|M) = \sum_s \lambda^s \log(\mathbf{o}^s|M) \qquad (4)$$

where o^s refers to an s-th stream (dimension) of a feature vector o, λ_s refers to a weighting coefficient for s-th stream, and M refers to a model.

With this, our score of a vowel $q(o_v)$ is denoted as

$$q(\mathbf{o}_v) = \sum_s \lambda_v^s q(\mathbf{o}_v^s) \qquad (5)$$

where λ_v^s refers to the weighting coefficient of an individual stream s for a feature vector of a vowel v, and the score $q(o_v^s)$ can be denoted by

$$q(\mathbf{o}_v^s) = \left\{\log p(\mathbf{o}_v^s|M_{gv}) - \log p(\mathbf{o}_v^s|M_{bv})\right\}/l_v \qquad (6)$$

where $p(o_v^s|M_{gv})$ and $p(o_v^s|M_{bv})$ are likelihoods of an individual stream s of the observed vowel phoneme belonging to a phrase of a good singing voice and a bad one.

The weight λ_v^s of each dimension is calculated by the LDA (Linear Discriminant Analysis)-based method proposed in [2].

3.3 Automatic Segmentation of Vowels

As a pre-processing, we implemented an automatic vowel segmentation module. Since we do not know whether the input voice is a good or bad one, the models for segmentation were trained with both good and bad voices for each vowel.

Since we used only vowel models, consonant frames were segmented into vowel periods as a result of segmentation.

4 Experimental Results

With the method of user cross validation, experiments were performed for various mixture numbers of GMM, namely from 1 to 32. However, no significant difference was observed. The threshold θ for the discriminant function (3) was set by finding a point where false positive and false negative values of all the testing data are the same.

Figure 2 shows successful classification rates of voice quality in the case of GMM with a mixture of 4, with or without using weights calculated by LDA. It can be seen that by applying appropriate weighting to each dimension, the successful classification rate rises by approximately 10 percent. The classification rates reached 93.3% for FBANK and 97.8% for MFCC. Note that the classification rate of MFCC surpasses that of FBANK.

Figure 3 shows the weighting coefficient of each FBANK component calculated by applying LDA and averaged over 5 vowels. It can be seen that 2nd and 8th dimension have relatively large effects in voice quality classification. The 8th dimension (related to the energy band of 2–3.3 kHz) matches a so-called singers' formant referring to the special formant formed by a clustering of the third, fourth and fifth vowel formants [3]. In classical literatures on singer's voice, it was reported to be observed at the energy band between 2.4 – 3.6 kHz [4].

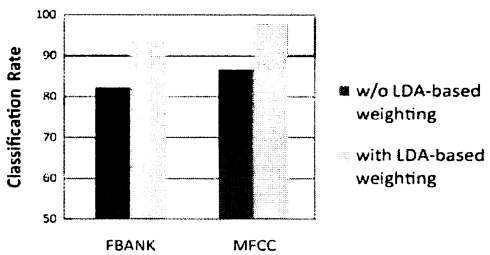

Figure 2: Classification rate

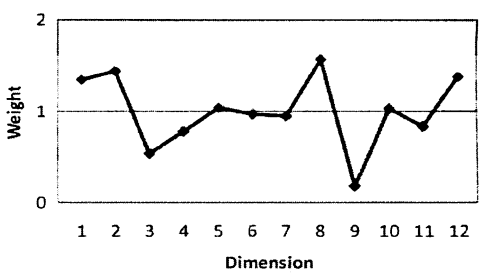

Figure 3: Weighting coefficient for each FBANK component averaged over vowels

5 Conclusions and Future Work

This paper has proposed a singing voice quality evaluation method using multi-stream GMM and dimension weighting based on LDA. Using the proposed method, classification rates of 93.3% in the case of feature vector FBANK and 97.8% in the case of MFCC with Gaussian mixture 4 were obtained.

Since quality of singing voice is subjective, its research is very difficult to perform without a large amount of data to cover them. The area is still new and there are countless problems waiting to be challenged.

6 Acknowledgement

This work is supported in part by the 21st Century COE Program Framework for Systematization and Application of Large-scale Knowledge Resources.

References

[1] T. Asami et al., Proc. Interspeech, pp.2185-2188 (2005-9).

[2] K. Iwano et al., Proc. Interspeech, pp.2534-2537 (2006-9).

[3] T. J. Millhouse and F. Clermont, Proc. Australian Int. Conf. on Speech Science & Tech., pp. 253-258 (2006-12).

[4] J. Sundberg, Journal of Voice, vol.15, iss.2, pp.176-186 (2001-6).