An Automatic Singing Impression Estimation Method Using Factor Analysis and Multiple Regression

Ai Kanato [†1] **Tomoyasu Nakano** [‡2] **Masataka Goto** [‡3] **Hideaki Kikuchi** [†4]
[†] Graduate School of Human Sciences, Waseda University, Japan
[‡] National Institute of Advanced Industrial Science and Technology (AIST), Japan
[1]53_like_pe-7-73[at]fuji.waseda.jp [2,3]{t.nakano, m.goto}[at]aist.go.jp [4]kikuchi[at]waseda.jp

ABSTRACT

This paper describes a method for estimating the impression of a singing voice via acoustic features. While much research has been conducted on singing impression, to date no method for determining appropriate words to represent the impressions created by a person's singing has been developed, primarily due to the lack of a comprehensive evaluation scale. We followed two steps: construction of such an impression scale, and development of models for estimating the impression score of each word. In the scale construction, two experiments were carried out. Firstly, 44 words were selected as relevant words based on subjective evaluation. Secondly, 12 words were selected as an impression scale, and three factors ("powerful", "cautious", and "cheerful") were extracted by factor analysis. To estimate impression scores, multiple regression models were constructed for each impression word with acoustic features. The models were tested by cross validation. The average R^2 value for the 12 words of the complete scale was 0.567, and the R^2 for the three factors were 0.863 (powerful), 0.381 (cautious), and 0.603 (cheerful).

1. INTRODUCTION

The purpose of this study is to develop an automatic estimation method for *singing impression words* via acoustic features of singing voice. We dealt with the words that describe singing impressions such as "cute" and "powerful", emotions such as "joyful" and "melancholy" [1], and singing skills such as "good" and "poor". Since most previous studies dealt with adjectives related to emotions only, a method for analytically determining which set of words is the most appropriate to describe an impression of a singing voice has yet to have been proposed. We have therefore investigated appropriate adjective words for a singing voice, and have constructed a model for estimating them from audio signals. Automatic singing impression estimation is useful for music information retrieval based on impression and facilitates sharing of singing impressions with many people. It also enables us to use common, intuitive words and meanings as an objective tool to evaluate the expression of singing emotion. Furthermore, if we could reveal the relationship between subjective evaluation and acoustic features, it is valuable in studying human perception of singing voices.

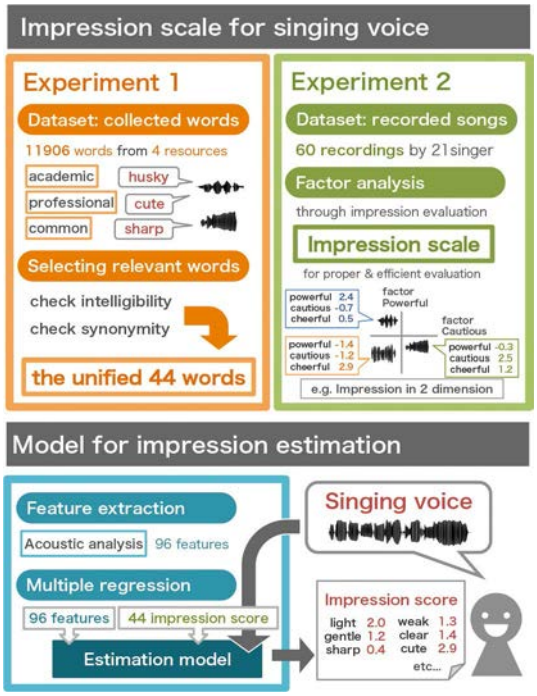

Figure 1. Proposed method to estimate impression.

Many studies have focused on relationships between emotions and acoustic features of singing voices [2]. For example, Kotlyar and Morozov explored emotional singing voices sung by 11 professional singers [3], and Scherer demonstrated that emotion estimation methods for speaking voices can be applied to singing voices [4]. The above studies have dealt with only specific domains (*i.e.*, emotion). None of them comprehensively investigated singing impression words or constructed impression scales in a bottom-up way. The lack of a comprehensive evaluation scale currently makes it difficult to study detailed differences of singing voices.

A popular approach of automatically estimating impressions from singing voices is to deal with singing skills. For example, Nakano *et al.* proposed an automatic singing skill evaluation method without score information of the

[1] Titze described that six primary emotions – fears, anger, joy, sadness, surprise, and disgust – are all commonly expressed vocally [1].

Copyright: ©2014 Ai Kanato et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

sung melody [5], and Tsi and Lee proposed an automatic singing evaluation system for Karaoke performances based on similarity computing between a user's singing voice and its original vocal which is estimated by using Karaoke track and the spectral subtraction method [6]. On the other hand, Daido worked on automatic estimation of singing enthusiasm [7]. However, emotion estimation is mostly investigated in speaking voices. For example, Luengo's research revealed that the spectral envelope features outperform prosody in an automatic emotion identification test [8], and Vlasenko showed that spectral formants are effective for estimating emotions [9].

Due to the lack of studies on scale construction of comprehensive impression of a singing voice, most previous studies tended to estimate the strength of a specific impression without selecting a proper set of impressions. This paper therefore presents a method of automatic impression estimation using factor analysis and multiple regression techniques. Our target genre is Japanese popular music sung by amateur female singers.

The remainder of this paper is structured as follows. In section 2, we explain the process of impression scale constructing. An impression scale was consequently constructed by factor analysis using the results of subjective evaluation. In section 3, we describe our estimation model constructed by multiple regression analysis. In section 4, we summarize the key points in this paper.

As an application example, we developed an automatic estimation system that outputs five words with high and low scores from 44 words. A website with song samples and estimation results obtained from the proposed method can be found at `http://shower.human.waseda.ac.jp/%7ekanato/icmc-smc2014/`)

2. IMPRESSION SCALE FOR SINGING VOICE

We define the "impression of a singing voice" as a subjective sense felt upon hearing the singing voice. We deal with impressions caused by vocal expressions, such as voice quality, vocalization, and pitch control. To focus on these elements, the song with a fixed melody, lyrics, key, and tempo was used for constructing our impression scale.

Singing skill is one particular example of the impression of a singing voice. Since the purpose of this study is to deal with impressions that are independent of singing skill, we do not deal with words that are directly used to evaluate singing skill, such as "skillful" and "good high tone voice."

The impression words should be easily understood and used by non-experts. We therefore avoid technical terms such as "vibrato" and "soprano" and inappropriate terms such as "inorganic" and "stateless," keeping only words with high intelligibility for singing voices. In addition, we reduce the number of synonyms such as "strong" and "powerful" to reduce the burden of evaluation by a human subject and make the *unified word set* that includes proper words to describe singing impression.

resource	token	type
1 Previous research [10, 11]	180	162
2 CD review	699	372
3 Twitter	10000	294
4 Video sharing site	1026	232
total number of words	11905	898

Table 1. Number of collected words.

2.1 Method

We carried out two experiments as follows. In experiment 1 (see 2.2), we collected various words and investigated their intelligibility and synonymity by subjective evaluation. In experiment 2 (2.3), we conducted subjective evaluation for singing voice, and factor analysis to reveal the factor in impression evaluation for singing voice.

In this section's investigations, all human subjects were Japanese college students, ages 20 to 24.

2.2 Experiment 1: Construct the unified word set for describing impressions of singing voices

We collected various words and conducted two experiments to select appropriate words to describe an impression of singing voice.

2.2.1 Dataset: The words describing impression of singing voice

To construct a reliable and valid scale, we collected various words that could be used to describe an impression of a singing voice. A total of 11,905 descriptive words were collected from four resources based on reasons for collecting important words in academic (1), professional (2), and common (3,4) usage. The number of collected words are described in Table 1.

Previous research Words were taken from two previous studies, which consisted of words from an affective value scale of music [10] and an impression scale in terms of the moods of classical music [11].

CD review We investigated 350 reviews of the Japanese popular music on the Web (RO69: `http://ro69.jp/`) within a two year period (June. 1st, 2010 – May. 31st, 2012) and extracted words that might represent the impression of a singing voice.

Twitter Words ending with "*– i*" and "*–na*" which are special features of Japanese adjectives, were automatically searched on Twitter (`http://twitter.com/`) for about one month (Aug 1st, 2012–Aug 28th, 2012) under various conditions (300 tweets each weekday, 500 tweets during the weekend).

Video sharing site We extracted relevant words from comments written by listeners on the most popular Japanese video sharing service NicoNico (`http://www.nicovideo.jp/`), searching for the term "utatte-mita" (rough translation: "Me Singing") and taking oldest (including comment of first impression) and latest 2 songs from each of the singers in the top 35 results (19 male and 16 female). 500 comments for each in the latest and earliest videos of each singer

Original Japanese	English Equivalent	*Original Japanese*	English Equivalent	*Original Japanese*	English Equivalent
amai	sweet	(*isshou kenmeina*)	(enthusiasm)	*seiryou no aru*	powerful
antei shiteiru	firm, stable	*josei tekina*	womanly	*sensaina*	sensitive
burikko mitaina	acting cute, lovely	*kakkoii*	cool	*shāpuna*	sharp
(*chuusei tekina*)	androgynous	*kanashii*	sad	*shoujo no youna*	like a girl
dansei tekina	manly	*karoyakana*	light	*shounen no youna*	like a boy
(*dosu ga kiiteiru*)	(threatening)	*kawaii*	cute	*shin no aru*	having core
genkina	active	*kikiyasui*	steady, comfortable	*shizukana*	silent, calm
hageshii	high-pitched	*kimochi yosasouna*	frank	*sukitootta*	clear
hana ni kaketa youna	nasal	*kokoro no komotta*	cordial	*tokuchou tekina*	characteristic
hasukīna	husky	*komotteiru*	veiled	*ureshisouna*	joyful
hibiki no aru	ringing, harmonic	*massuguna*	plain, straight	*utsukushii*	beautiful
(*huanteina*)	(unstable)	*mujakina*	ingenuous, innocent	(*youkina*)	(cheerful)
(*hurueteiru*)	(vibrating)	(*nobiyakana*)	(relaxed)	*yasashii*	gentle
ikioi ga aru	spirited	*ochitsuki no aru*	calm, careful	*yowai*	weak
iroke no aru	amorous, sexy	(*sawayakana*)	(fresh)		

Table 2. The unified word set for describing an impression of a singing voice (44 words) in experiment 1: parentheses indicate that a word was excluded from the factor analysis (2.3.3).

were then scanned manually for appropriate adjectives, resulting in 70 videos in total.

We selected 898 words from the 11,905 candidates by removing duplicates, and furthermore by removing inappropriate words such as proper nouns, resulting in 590 words which then went under an intelligibility check.

2.2.2 Intelligibility evaluation

We asked 20 human subjects to evaluate intelligibility, by classifying each word into one of two categories: "appropriate to describe the sung voice" and "not appropriate to describe the sung voice". Through this procedure, we reduced the number of words in our lexicon from 590 to 64 (examples of removed words include: "wet," "sturdy," and "flexible").

2.2.3 Synonymity evaluation

We then asked 10 human subjects to extract similar pairs in a round robin of 2016 pairs ($= 64 \times (64-1)/2$). From this, we found that 562 pairs of words were judged similar by more than 3 human subjects. These pairs were then used in a synonymity evaluation using a 7-point grade (1: not similar to 7: very similar) by 10 participants. As a result, pairs having high synonymity were unified (meaning one word was selected as most appropriate from a group of similar words; *e.g.*, "pure" and "pellucid" were discarded in favor of "clear") and the number of words was reduced from 64 to 44. We used these words as the unified word set for describing an impression of a singing voice.

2.2.4 Result

The unified word set of 44 words is shown in Table 2. The words from the evaluation results are naturally in Japanese (*e.g.*, "*amai*"), but Table 2 also shows an English translation (*e.g.*, "sweet"). Furthermore, we added three words ("likeability," "skillful", and "good match for melody/lyrics") to enable posterior investigations (see 2.3.2, 3) since these words are frequently observed in our word collecting methods (2.2.1) and important words for expressing a singing impression.

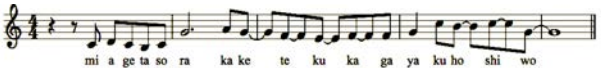

Figure 2. An original song used at the experiment 2.

2.3 Experiment 2: Construct an impression scale

We conducted subjective evaluation for singing voice, and factor analysis to reveal the factor in impression evaluation for singing voice.

2.3.1 Dataset: Recordings of the fixed song by various voices

To assess how well our chosen words correlated with sung audio, we asked 21 amateur singers (female university students aged 20–24 with various levels of experience) to sing a song in different styles. Rather than using an existing song, we composed an original piece to reduce possible bias in the evaluation (see 2.3.2). The singers were asked to sing in 7 styles, 1: in modal register, 2: head register, 3: expressively, 4: flatly (non-expressively), 5: sing as well as you can, 6: in a relaxed way and 7: in the style of a popular singer of their choice. The instructions to the singers were intentionally left somewhat ambiguous as the purpose was simply to collect different singing styles with a fixed lyrics, tempo and melody.

Based on the author's subjective evaluation, singing voices that showed no difference in impression were removed from the $147 (= 21 \times 7)$ recordings to reduce the participants' burden in the following experiments (2.3.2). In total, 60 recordings were selected for use in the experiment.

Note that throughout this paper, all singing samples were monaural recordings of solo vocal digitized at 16 bits / 44.1 kHz.

2.3.2 Impression evaluation for a singing voice

In this experiment, 19 participants were asked to rate, on a 7-point scale (1: not appropriate–7: very appropriate) the appropriateness of each of our unified word set (44 words), with three additional words/phrases: likeability, skillful,

	Word of scale		Factor of scale		
	Original Japanese	English Equivalent	*hakuryokusei* powerful	*teineisa* cautious	*akar* cheer
1	*ikioi ga aru*	spirited	**0.932**	0.044	0.0
2	*seiryou no aru*	powerful	**0.917**	0.188	-0.1
3	*yowai*	weak	**-0.898**	0.023	-0.0
4	*shizukana*	silent	**-0.752**	0.466	-0.1
5	*kikiyasui*	steady	0.146	**1.001**	0.2
6	*sukitootta*	clear	-0.127	**0.886**	0.2
7	*ochitsuki no aru*	calm	-0.286	**0.775**	-0.2
8	*hibiki no aru*	ringing	0.387	**0.756**	-0.1
9	*ureshisouna*	joyful	0.246	0.092	**0.9**
10	*karoyakana*	lightly	-0.037	0.358	**0.8**
11	*kawaii*	cute	-0.286	0.145	**0.8**
12	*mujakina*	ingenuous	-0.085	-0.359	**0.7**
Contribution ratio			0.292	0.292	0.262
Cronbach's α [12]			0.926	0.893	0.877

Table 3. Singing impression scale (12 words) in experiment 2: Each value in the "Factor of scale" columns indicates the factor loadings.

good match for melody/lyrics, for each of the 60 recordings. The subjects were free to listen to the melodies as many times as they liked. According to the result, we selected appropriate 36 (= 44 - 8) words for next factor analysis (after removing words with low correlation in human subjects for its commonality in evaluation, and high correlation in each of word for its synonymity).

2.3.3 Factor analysis with impression scores

Factor analysis was applied with 36 words using a maximum likelihood method and promax rotation, and the number of factors was determined by the scree test. The analysis was repeated with words incrementally removed until all factor loadings were less than 0.35.

2.3.4 Results

We select 12 words as an impression scale for a singing voice (the accumulated contribution ratio equaled 0.846) and named three factors to score words on: "powerful" (*hakuryokusei*) , "cautious" (*teineisa*), and "cheerful" (*akarusa*) (Table 3). These factors were expressed by summing the score of words that had high factor loadings.

Figure 3 shows each factor loading for the 12 words for each pair of the three factors.

The correlation of each factor is as follows: "powerful" and "cautious" is 0.189, "powerful" and "cheerful" is 0.229, "cautious" and " cheerful" is -0.132. It indicates that these three factors were almost independent.

For each of the three factors, Cronbach's α [12] (widely used as a measure of scale reliability) had a high value (above 0.85), indicating high internal consistency. The results from each gender were fairly consistent, although the order of each of the three factor's contribution differed, indicating that this scale is effective regardless of the listener's gender.

3. MODEL FOR IMPRESSION ESTIMATION

In this section, we explain how we determined effective acoustic features for estimating singing impression words through multiple regression analysis. The acoustic features are extracted without requiring information on the musical

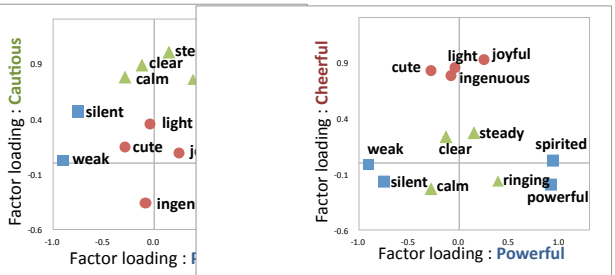

Figure 3. Factor loadings of 12 words (3 factors) in the impression scale are shown. Markers of each factor are colored (blue: "powerful," green: "cautious," red: "cheerful").

score or lyrics, so the model is applicable to a wide variety of songs.

3.1 Method

To construct an estimation model to automatically estimate singing voice impression, we carried out experiments as follows: First, we extracted 96 numerical features with acoustic analysis of singing voices. Secondly, multiple regression analysis was conducted with acoustic features and impression scores.

3.2 Acoustic features

Acoustic features are extracted from the F_0 (fundamental frequency), spectral envelope (the number of frequency bin is 2048) and aperiodic component estimated using STRAIGHT [13] once per millisecond.

Delta features in this analysis are calculated using the following equation:

$$R(\boldsymbol{y}) = \frac{\sum_{k=-K}^{K} k \cdot \boldsymbol{y}_k}{\sum_{k=-K}^{K} k^2} \quad (1)$$

where $\boldsymbol{y}$ is a feature vector for analysis, n is the length of $\boldsymbol{y}$, and $\boldsymbol{y}$ corresponds to the spectral envelope and F0 throughout this paper.

3.2.1 Spectral Envelope

The spectral envelope is important in defining the stationary voice quality of the singing voice and has been used in previous research [14]. We used both the linear and log spectral envelope, $S_{\mathrm{lin}}(f,t)$ and $S_{\log}(f,t)$ respectively, at time t and we extracted the following features (f is a frequency bin number).

Spectral centroid Spectral centroid is known as a *Timbral Texture Feature* [15]. This feature for each time $S_c(t)$ was extracted using equation (2) with $S_{\mathrm{lin}}(f,t)$ and $S_{\log}(f,t)$, and calculated mean and variance (4 features total), where N is the length of the frequency bin.

$$S_c(t) = \frac{\sum_{f=1}^{N}(f \cdot S_{\mathrm{lin|log}}(f,t))}{\sum_{f=1}^{N}(S_{\mathrm{lin|log}}(f,t))} \quad (2)$$

Spectral tilt Spectral tilt was extracted from equation (1) with $S_{\mathrm{lin}}(f,t)$ and $S_{\log}(f,t)$ substituted for $\boldsymbol{y}(k)$ for each frame t at various bandwidths (0-3, 0-6, 0-9, 0-22.05kHz), and calculated its mean and variance (16 features total).

Singer's formant Singer's formant is a feature corresponding to the ringing of singing voice [16]. We calculated this using the power ratio in the range 2-4 kHz compared to the power elsewhere, and calculated its mean and variance (4 features total).

Harmonic component The strength of spectral amplitude under the F_0 (H1) is a measure of voice breathiness [17]. We calculated the ratio of H1 to H2 (the amplitude of second harmonics) by extracting the amplitude nearest F_0 and $2 \times F_0$ and the ratio of the sum of power at odd harmonics to that at even harmonics of $S_{\mathrm{lin}}(f,t)$ and $S_{\log}$. The mean and variance of them were extracted (8 features total).

Spectral peaks related to formants The spectral envelope includes formants (the spectral peak) and has been shown to be related to the impression made by a singing voice [14]. We picked the peak of the spectral envelope as a feature related to the formant. First, the low level cepstrum (dimensions 1-18) was extracted from the spectral envelope by using inverse Fourier transform to deal with vocal tract characteristics. Next, we picked the first two formants by referring to the bandwidth for each formant ($F_1 < 900\mathrm{Hz} < F_2 < 3300\mathrm{Hz}$), to extract mean and variance of $F_1(t)$ and $F_2(t)$ (4 features total).

3.2.2 Aperiodic component

STRAIGHT [13] can estimate the ratio of aperiodic component to power of spectral envelope. The range of value is 0 – 1.0. The higher the value is, the higher the aperiodic component in voice is.

Aperiodic component We calculated the sum total in the aperiodic component $A(f,t)$ for each frame and calculated mean and variance (2 features total).

Aperiodic component tilt We calculated the aperiodic component tilt of different bandwidths (0-6, 0-22.05kHz) using equation (1) with $A(f,t)$ substituted for $\boldsymbol{y}(k)$ for each frame t, and calculated mean and variance for each bandwidth (4 features total).

3.2.3 Dynamic feature

The features described so far are related to the static voice quality of the singing voice. However, the impression made by a singing voice is clearly affected by dynamic changes in the spectral envelope.

Power fluctuation The power was extracted for each time t with the equation $P(t) = \sum_{f=1}^{N}(S_{\mathrm{lin}}(f,t))$: Δ P was extracted with equation (1), and mean and variance were calculated (2 features total).

Spectral change We extracted $\Delta S_{\mathrm{lin|log}}$(f,t) in the time base using equation (1) and calculated the sum total on the frequency axis as a feature under the condition $\Delta S_{\mathrm{lin|log}}(t) = \sum_{f=1}^{N} \Delta S_{\mathrm{lin|log}}(f,t)$, where N is the frequency bin corresponding to the target maximum frequency (3kHz and 22.05 kHz) in the spectrum. The mean and variance were calculated in each condition (K=1 : 4 features total, K=25 in two conditions that include boundary or not: 8 features total). In addition, 1024-order DCT (discrete cosine transform) coefficients $C_{\mathrm{lin|log}}$ were extracted and calculated in the same way (K=1:4 features total, K=25:8 features total).

Formant fluctuation We extracted fluctuation from $F_1(t)$ and $F_2(t)$ on a time basis, using Equation (1) and calculated mean and variance (K=25, 4 features total).

3.2.4 Fundamental frequency

In this paper, frequency values will be referred to by *cents*, which are log-scale frequency values. In the Western temperament, a semitone corresponds to 100 cents. The cent value f_{cent} of frequency f_{Hz} is given as

$$f_{\mathrm{cent}} = 1200 \log_2(\frac{f_{\mathrm{Hz}}}{f_c}) + 4800 \quad (3)$$

The middle C $f_c (= 440 \times 2^{\frac{3}{12}-1} = 261.62\mathrm{Hz})$ corresponds to 4800 cent. The fundamental frequency is indicated with $F_0(t)$, and t means the time base axis.

Pitch interval accuracy We extracted two kinds of feature referring to the *pitch interval accuracy* (see [5]). The pitch interval accuracy is judged by fitting $F_0(t)$ to a semitone (100 cent) -width grid (16 features total)

Vibrato Vibrato is an important feature in the sung voice. We extracted a feature, which refers to the rate, extent and *vibrato likeliness* (see [5]) of the audio. In this paper, we extracted the fluctuation as a vibrato for which the range of $F_0(t)$ extent is 30-150 cent and $F_0(t)$ intersects the average F_0 in the area (320 ms) more than five times. We extracted the fluctuation $f_d(t)$ of F_0 with the following equation and calculated the maximum, average, standard deviation, of each feature of vibrato from $f_d(t)$ (7 features total) , $F_0(t)$ (7 features total) and the ratio of the vibrato area to the whole area (1 feature).

$$f_d(t) = F_0(t) - f_l(t) \quad (4)$$

$f_l(t)$ means $F_0(t)$ filtered by a lowpass filter with a 5Hz cutoff frequency.

Pitch fluctuation Fluctuation in F_0 transitions (for example, *preparation* and *overshooting* [18]) are important to deal with regarding F_0. We extracted the fluctuation $D(t)$ with equation(1) substituting $F_0(t)$ for $\boldsymbol{y}(k)$ for each time point and calculated the mean and variance in the condition K=10, 25, 50 (6 features total).

For $D(t)$, we extracted the lower fluctuation area, and the ratio of this area to the whole area (1 feature) resulting in a more stable $F_0(t)$.

3.3 Multiple regression model to estimate singing impression

We next constructed a model for estimating singing impression using 60 recordings taken from our subjective evaluation (see 2.3.2). The model was constructed through multiple regression analysis with stepwise selection, owing to the large number of variables. The target impressions

(**A**) 12 words of impression scale

	word	closed	LOO	LOSO
powerful	spirited	0.757	0.726	0.727
	powerful	0.883	0.869	0.867
	weak	0.795	0.764	0.766
	silent	0.784	0.745	0.749
cautious	steady	0.335	0.278	0.238
	clear	0.549	0.483	0.447
	calm	0.442	0.391	0.363
	ringing	0.706	0.675	0.651
cheerful	joyful	0.359	0.285	0.299
	light	0.496	0.438	0.417
	cute	0.739	0.693	0.685
	ingenuous	0.675	0.626	0.599
	mean	0.627	0.581	0.567

(**B**) 3 factors of impression scale

word	closed	LOO	LOSO
Powerful	0.880	0.863	0.863
Cautious	0.481	0.416	0.381
Cheerful	0.676	0.628	0.603
mean	0.679	0.636	0.616

(**C**) Important words to evaluate singing

word	closed	LOO	LOSO
likeability	0.401	0.339	0.281
skillful	0.333	0.303	0.267
good match for melody/lyrics	0.346	0.266	0.210

(**D**) Highest 10 words in unified word set (Table 2)

word	closed	LOO	LOSO
powerful	0.883	0.869	0.867
high-pitched	0.858	0.820	0.807
weak	0.795	0.764	0.766
desperate	0.812	0.762	0.777
threatening	0.786	0.756	0.752
silent	0.784	0.745	0.749
like a girl	0.776	0.727	0.624
spirited	0.757	0.726	0.727
gentle	0.786	0.723	0.700
manly	0.768	0.721	0.683
mean of 44 words	0.614	0.555	0.541

Table 4. Multiple regression analysis $\hat{R^2}$ values: the higher $\hat{R^2}$ is, the more accurate the model. LOO means leave-one-out, and LOSO means leave-one-singer-out cross validation.

were 15 impressions (the 12 words and 3 factors shown in Table 3).

In addition, a model using the words considered important to a singing evaluation "likeability", "skillful", "good match for melody/lyrics" and a 32 (= 44 - 12) word model were also constructed to provide a basis for comparison with respect to impression construction.

3.3.1 Dataset preparation

This analysis was conducted with acoustic features and impression scores of 60 recordings. To prevent the increasing risk of multi-collinearity [19], which can degrade a model's stability in multiple regression analysis, we removed 17 acoustic features whose correlation coefficients with respect to each other exceeded 0.9.

After this preprocessing, we dealt with a standardized impression score as an independent variable and the standardized 79 acoustic features as the dependent variables. The impression score for each word was scored by subjective evaluation (2.3.2), and 3-factor scores were calculated using the sum of scores for related words having high factor loadings for a specific factor.

3.3.2 Model construction

We applied multiple regression analysis with step-wise selection of 79 acoustic features for each impression. To construct a reliable model, we used equation $V_i = 1/(1 - R_i^2)$ to calculate the variance inflation factor (VIF), which is a measure of the risk of multi-collinearity caused by variables being highly correlated with other variables, for each variable in the regression model. i is the number of variables in the model, and R_i^2 is calculated by multiple regression with variable i as a dependent variable and all other variables independent. If any of the variables had a large V_i, greater than 10.0, the feature was removed and V_i was again calculated. The average number of independent variables for each model was 6.61.

3.3.3 Evaluation of the model

To confirm the accuracy of the model, we calculated the adjusted R-square $\hat{R^2}$ (the coefficient of determination) calculated via sum of squares with residual error of observed variable and estimated value, divided by difference of observed variable and averaged (see equation 5) in the 60 recordings: The member of an equation $\hat{R^2}$, to prevent overestimation caused by large number of independent variables. N means the number of sample, and P means the number of variables in the model.

$$\hat{R^2} = 1 - \frac{\sum_{n=1}^{N} (y_n - \hat{y}_n)^2/(N-P-1)}{\sum_{n=1}^{N} (y_n - m)^2/(N-1)} \quad (5)$$

Furthermore, we used leave-one-out cross-validation (LOOCV) and K-fold cross-validation for each singer (*i.e.*, leave-one-singer-out cross validation (LOSOCV), K=21, where the average number of training samples was 57.14). For this test, the more closely the results are clustered, the higher the model's accuracy is.

3.3.4 Result of model construction

Table 4 shows $\hat{R^2}$ for the closed dataset and the cross-validation results for each model. In the unified word set, $\hat{R^2}$ of the LOSO cross-validation for "powerful" and "high-pitched" both exceeded 0.8; in addition, 9 of 44 words had values greater than 0.7, and 18 of 44 words had value greater than 0.6. This indicates that these models can estimate an impression score with high accuracy.

3.4 Discussion

First, we describe the relation between 3 factors and features, and secondly describe important features to estimate various impression.

3.4.1 3 factors and features

Table 5 shows all features related to the 3 factors' estimation and the PRC (Partial Regression Coefficient). The features of these are almost independent of each factor except "spectral tilt (0-6kHz)". Referring to the correlation of 3 factor (2.3.4), the correlation between each factor was low, meaning it is natural that the features related to these factors are almost independent. In addition, it is important that all factors be related to the static and dynamic features.

factor	PRC	feature
Powerful	0.44	Pitch fluctuation: M
	−0.32	Local spectral change (K=1): M
	−0.28	Sum of the aperiodic component: M
	−0.25	Length of high stability for F_0
	−0.22	Aperiodic component tilt (0-6kHz): SD
Cautious	0.43	maximum of vibrato likeliness
	−0.35	Spectral tilt (0-6kHz): M
	0.34	Power fluctuation: M
	0.28	Pitch interval accuracy
Cheerful	0.31	Spectral tilt (0-6kHz): M
	0.29	Spectral change (0-3kHz): SD
	0.13	Aperiodic component tilt (0-6kHz): M
	0.11	Local change of DCT (K=1): SD

Table 5. Features related to three factors (M is average and SD is standard deviation of 1 recording) and the PRC.

3.4.2 *Important features to estimate*

Table 6 shows impression words can be estimated from each feature. "+" in left columns means positive PRC, the higher the features are, the higher are the impression scores are, with "−" indicating an inverse relationship. Important features to estimate various impressions were described below.

Spectral tilt Spectral tilt was calculated with three kinds of bandwidth, and each provided a different impression. Table 6 shows which bandwidth was effective for estimating an impression with each model. The tilt in the range of 0-3kHz contributes to estimate "Powerful", like "high-pitched," "active," and "silent". From this we may gather that the gentler this line's tilt, the higher impression score. In addition, the range of 0-6kHz contribute to attributes such as "cheerful", like "like a girl," and "cute", the range of 0-9kHz contribute to "cordial," "sharp," "cool," and so on.

However these divided spectral tilts are not independent from each other because of the overlap of spectral envelope in calculation. Therefore, these features may need additional consideration.

Pitch interval accuracy in F_0 This feature is often used as an indicator used for skill evaluation [5]. In this research, it contributes to the impression described by words related to like cautiousness (*e.g.*, "steady" and "cordial") and beauty (*e.g.*, "beautiful" and "clear") (Table 6).

Vibrato Vibrato is a feature that is effective for evaluating singing skill.

"Vibrato likeliness" is the ratio of vibration in the 5-8 kHz range, and it is corresponded to the degree of similarity to sine wave. Results suggest that the maximum of vibrato likeliness is effective for estimating impressions associated with "beautiful", 'ringing", "firm", and so on (Table 6). In the model of "skillful", only this feature was accepted. Therefore, the likeliness of vibrato is more important than the length of vibrato in the estimation of "skillful." Furthermore, the maximum of vibrato extent contributes to "cordial", so singing with a wide vibrato extent contributes to the impression "cordial".

Spectral tilt	
0 - 3kHz	
+	active spirited
+	high-pitched
−	calm silent
0 - 6kHz	
+	like a girl cute active
−	relaxed powerful ringing
0 - 9kHz	
−	cordial sharp cool
−	frank fresh

Pitch interval accuracy	
Pitch interval accuracy F_0	
+	steady gentle sensitive
	beautiful firm relaxed
	clear cordial womanly

Vibrato	
vibrato likeliness	
+	steady clear firm
+	ringing beautiful
+	relaxed silent
+	(likeability) skillful
maximum of vibrato extent	
+	cordial
average of vibrato extent	
+	relaxed
ratio of vibrato area to all	
+	womanly sensitive
−	like a girl like a boy
−	ingenuous acting cute
−	active

Table 6. Important features to estimate various impressions. "+" in left columns means positive PRC, the higher the features are, the higher are the impression score, with "−" indicating an inverse relationship.

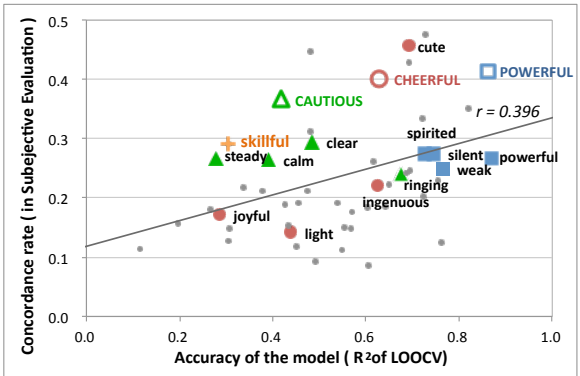

Figure 4. Model accuracy and concordance rate of subjective evaluation (2.3.2). The words in black are included in the 12-word impression scale, the words in color (except for "skillful") are factors, and gray plots indicate other words in the unified word set.

3.5 Consideration of gaps in model's accuracy and concordance of human subject

Declines in the accuracy of the models were mainly due to a lack of features and inconsistent scoring in the subjective evaluation. Figure 4 shows the relation between model accuracy and the concordance rate in the subjective evaluation (2.3.2) calculated using the sum of the correlative coefficient for a round robin; the correlation was low (r = 0.396). 12 words in impression scale and 3 factors (blue: "powerful," green: "cautious," red: "cheerful") are described in plot and others are plotted in gray color without text. Words with low concordance values (in the lower-left area of the figure) meant that the impression was interpreted in several different senses by the evaluators. Therefore, to improve the accuracy of these models, we have to be more aware of differences in the understanding of participants, and should try to provide words with more precise meanings. Conversely, words plotted in the higher x (concordance) and lower $\hat{R}^2$ (model accuracy) indicate that the models need more appropriate features for the estimation. For example, the concordance value of "skillful" is

higher than that of "silent", "powerful" or "weak", but the model accuracy is lower. We believe that the low accuracy of "skillful" was due to the distribution of each singer's skill (*e.g.*, pitch accuracy, vibrato and vocalization).

We studied the features *pitch interval accuracy* and *vibrato features*, inspired by previous research [5] but none of the singers in this study performed so poorly as to be considered an outlier that it makes *pitch interval accuracy* low. So it was not enough to evaluate "skillful" with *pitch interval accuracy features* and *vibrato features* in our stimulus. It means that even if *pitch interval accuracy* was similar in songs, there are difference that affect subjective evaluation. Therefore, we have to consider different features for this.

In addition, there were differences between the results of LOOCV and those of LOSOCV. There are several possible reasons for these differences. For example, the decrease of LOSOCV means each singer has accurate data in some recordings for model construction, therefore accuracy was decreased because of the lack of beneficial data for construct model. This indicates the impression of the low LOSOCV model is little different in same singer, in a word, the impression may depend on individuality of singer.

4. CONCLUSION AND FUTURE WORK

To determine which words are most appropriate to describe the impression made by a singing voice based on acoustic features, we have developed the following: First, an impression scale of 12 words that were constructed based on collecting words from existing studies, social media and the web, with factor analysis. Three factors were extracted for evaluating a singing voice: "powerful", "cautious", and "cheerful". Second, the estimation model for each impression was made by multiple regression analysis with acoustic features and impression score. These models were tested by LOOCV; the average coefficient of $\hat{R}^2$ for 12 words in impression scale with LOOCV was 0.581, and those for each factor were 0.863 (powerful), 0.416 (cautious), and 0.628 (cheerful).

Some words related to "powerful" can be estimated with high accuracy, but this is not the case for words related to "cautious", so this feature needs to be improved. In this paper, we extracted features without information of musical score and lyrics, and dealing with average and variance to reduce the effect from them. Therefore these model may be applicable with various songs, but to improve more robust model, we have to investigate more various songs, singer, features, and model construction.

Acknowledgments

We would like to thank Matt McVicar for his helpful comments and proof-reading. This work was supported in part by CREST, JST.

5. REFERENCES

[1] I. Titze, *Principles of voice production*. Prentice Hall, 1994.

[2] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 46, no. 12, pp. 227 – 256, 2003.

[3] G. M. Kotlyar and V. P. Morozov, "Acoustical correlates of the emotional content of vocalized speech," *Sov.Phys.Acoust.*, vol. 22, no. 3, pp. 208–211, 1976.

[4] K. R. Scherer, "Expression of emotion in voice and music," *Journal of Voice*, vol. 9, no. 3, pp. 235 – 248, 1995.

[5] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features." in *Proc.INTERSPEECH 2006*, 2006.

[6] W.-H. Tsai and H.-C. Lee, "Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Trans. on ASLP*, vol. 20, no. 4, pp. 1233–1243, 2012.

[7] R. Daido, S. Hahm, M. Ito, S. Makino, and A. Ito, "A system for evaluating singing enthusiasm for karaoke," in *Proc. of ISMIR 2011*, 2011, pp. 31–36.

[8] I. Luengo, E. Navas, and I. Hernaez, "Feature analysis and evaluation for automatic emotion identification in speech," *IEEE Trans. on M*, vol. 12, no. 6, pp. 490–501, 2010.

[9] B. Vlasenko, D. Prylipko, D. Philippou-Hbner, and A. Wendemuth, "Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions." in *Proc.INTERSPEECH 2011*, 2011, pp. 1577–1580.

[10] T. Taniguchi, "Construction of an affective value scale of music and examination of relations between the scale and a multiple mood scale (in Japanese)," *The Japanese Journal of Psychology*, vol. 65, no. 6, pp. 463–470, 1995.

[11] R. Hirae and T. Nishi, "Mood classification of music audio signals (in Japanese)," *The Journal of the acoustical society of Japan*, vol. 64, no. 10, pp. 607–615, 2008.

[12] J. M. Cortina, "What is coefficient alpha? an examination of theory and applications." *Journal of Applied Psychology*, vol. 78, no. 1, pp. 98 – 104, 1993.

[13] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[14] J. Sundberg, *The Science of the Singing Voice*. Northern Illinois University Press, 1987.

[15] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transs on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[16] K. Omori, A. Kacker, L. M. Carroll, W. D. Riley, and S. M. Blaugrund, "Singing power ratio: Quantitative evaluation of singing voice quality," *Journal of Voice*, vol. 10, no. 3, pp. 228 – 235, 1996.

[17] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[18] T. Saitou, M. Unoki, and M. Akagi, "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," *Speech Communication*, vol. 46, no. 34, pp. 405 – 417, 2005.

[19] D. E. Farrar and R. R. Glauber, "Multicollinearity in Regression Analysis: The Problem Revisited," *The Review of Economics and Statistics*, vol. 49, no. 1, pp. 92–107, 1967.