A SYSTEM FOR EVALUATING SINGING ENTHUSIASM FOR KARAOKE

Ryunosuke Daido*, Seong-Jun Hahm*, Masashi Ito†, Shozo Makino‡ and Akinori Ito*
Graduate School of Engineering, Tohoku University*
Tohoku Institute of Technology† Tohoku Bunka Gakuen University‡
{ryunosuke, branden65, aito}@spcom.ecei.tohoku.ac.jp*
itojin@tohtech.ac.jp† makino@ait.tbgu.ac.jp‡

ABSTRACT

Evaluation of singing skill is a popular function of karaoke machines. Here, we introduce a different aspect of evaluating the singing voice of an amateur singer: "enthusiasm". First, we investigated whether human listeners can evaluate enthusiasm consistently and whether the listener's perception matches the singer's enthusiasm. We then identified three acoustic features relevant to the perception of enthusiasm: A-weighted power, "fall-down", and vibrato extent. Finally, we developed a system for evaluating singing enthusiasm using these features, and obtained a correlation coefficient of 0.65 between the system output and human evaluation.

1. INTRODUCTION

Karaoke is a form of singing entertainment found worldwide, which enables anyone to sing like a professional. Karaoke machines not only provide backing music for singing, but also evaluate the singer's voice as another entertaining feature. Studies of analyzing the singing voice have been making progress. For example, Nakano et al. reported good results of a system for classifying "good" and "poor" singing based on SVM [2]. Mayor et al. proposed a categorization and segmentation system for singing voice expression using pre-defined rules and HMM [1]. In this paper, we describe our attempt to develop a new service for karaoke: a system for evaluating the singer's enthusiasm.

By "enthusiasm", we mean how eager the singer is to sing. The term "enthusiasm" for singing a song as used in this paper is a translation of the Japanese word *nessho*, which literally means "hot singing" and is often used for expressing the energy of a singer's performance. As karaoke is the entertainment for amateur singers, we believe that singing skill is not the only aspect worth evaluating because poor singers can never get a high score. However, even poor singers can sing enthusiastically, so we focused on this aspect. We consider that a system which evaluate singing enthusiasm would be an exciting service for amateur karaoke users.

Singing enthusiasm is similar to the emotion of music [3], especially the "arousal-calm" aspect. However, there are significant differences between enthusiasm and emotion. First, enthusiasm is not an expressed emotion. Karaoke is basically a form of self-entertainment, and most karaoke singers who sing enthusiastically are not trying to convey their enthusiasm to the audience but are just enjoying themselves. Also, enthusiasm is not an induced emotion, because a listener who listens to an enthusiastically-sung karaoke song does not necessarily become excited. In our opinion, enthusiasm is more like an attitude of singing, rather than an emotion.

As our study on objectively evaluating enthusiasm was a new attempt, there were several issues to investigate:

- Is a feeling of "enthusiasm" shared by many listeners?
- Is enthusiastic singing also perceived to be "enthusiastic" by listeners?
- What are the physical features related to enthusiasm?
- How can we build a system that evaluates enthusiasm automatically?

This paper is organized as follows. In Sections 2 and 3, we describe the procedures and results of analyzing a singing voice corpus and subjective evaluations, and show that humans can perceive enthusiasm appropriately. In Section 4, we describe our method for choosing acoustic features of a singing voice and discuss the efficiency of each feature. In Section 5, we describe an overview and evaluations of the system.

2. SINGING VOICE CORPUS

2.1 Selection of a Song

For this first study on singing enthusiasm using a simple and reliable scheme, we decided to use just one pop song for all the experiments. "Itoshi no Ellie" by the Southern All Stars (which was covered as "Ellie My Love" by Ray Charles) was finally selected as it satisfied the following conditions:

- Not too difficult for amateur singers to sing both "enthusiastically" and "normally" i.e., no extremely high, low or long notes.
- Well known by all the singers and human subjects of the subjective evaluation (Japanese, in our research).

All the recordings should be in the same key because differences of key may affect the subjective evaluations. Considering the vocal range of amateur singers, we chose to use C-Maj. transposed from the original key of D-Maj. As a result, the lowest note is E3 and the highest is G4 for male singers (it can be an octave higher for female singers). The tempo is 69-70 bpm.

2.2 Recording Procedures

Thirty-four singers participated in the recording, none of whom were professional singers. The sound accompaniment, which had been directly recorded using a karaoke machine beforehand, was played through headphones and the singers sang along to it. The singers sang into a microphone on a stand with a pop-filter attached to prevent handling noise and pop noise. The singers were instructed not to move much during the recording and stay almost a constant distance from the microphone. In order to obtain various voices with a wide range of enthusiasm and to label singers' intended enthusiasm to each voice, they were each asked to sing two times, once "enthusiastically" and once "normally". The singers themselves could choose in which style to sing first, and informed us before they sang.

The voices were recorded at 44.1-kHz/16-bit sampling in a soundproof chamber.

3. SUBJECTIVE EVALUATIONS

We conducted subjective evaluations for the following three purposes: (1) investigate whether humans can perceive singing enthusiasm using the same criteria, (2) investigate whether listeners can distinguish whether singers sang enthusiastically or not, and (3) investigate listeners' intuition about the enthusiasm, and obtain clues for choosing acoustic features for automatically evaluating singing enthusiasm.

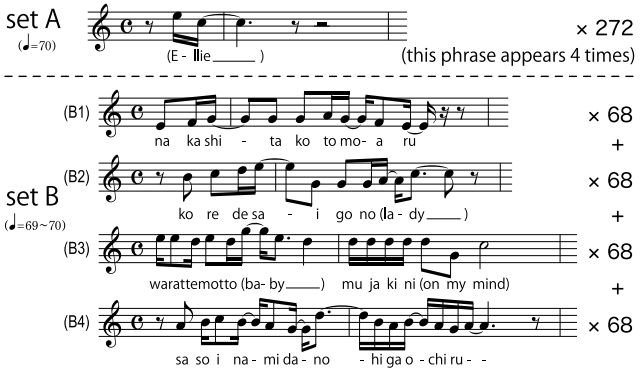


Figure 1. Stimuli for subjective evaluations (parenthesized words are English words)

Evaluation word	Value
enthusiastic	2
neither selected	1
not enthusiastic	0

Table 1. Evaluation words and the values for the subjective evaluations

3.1 Stimuli

For the subjective evaluations, we chose short stimuli (about 1.5 to 9 seconds) from the recordings to facilitate the decision-making. Figure 1 shows the prepared stimuli.

In this study, the absolute sound-pressure level (SPL) is of no interest because the SPL depends on not only the magnitude of a singer's voice but also the distance between the singer and the microphone. As our method should be applied to karaoke machines, it is difficult to measure the magnitude of the singer's voice precisely, so we decided to exclude the effect of absolute SPL, even though our preliminary experiment proved that absolute SPL is important for perception of enthusiasm. All the stimuli were normalized to the same power after passing through a high-pass filter (80 Hz cut-off) to reduce low-frequency noise.

As Figure 1 shows, two sets of stimuli were prepared. Set A was a collection of 272 stimuli of a phrase that appears four times in the song with the same melody and the same lyrics, and set B was a collection of four varieties of phrases, each of which was sung 68 times. (B1) is the beginning of this song, (B2) is from the early part, (B3) is from the middle part (the bridge or the climax) and (B4) is from the last part.

3.2 Evaluation Procedure

For each set of stimuli, 30 human subjects were asked to listen to the stimuli, and selected one of three evaluation words for each stimulus. Table 1 shows the evaluation words

and the associated values. Evaluations were conducted for each set of stimuli using the same procedure as follows:

1. The subjects listened to the stimuli through headphones in a soundproof chamber and the volume was fixed for all the subjects.
2. For training, the subjects evaluated 20 stimuli selected at random.
3. The subjects evaluated 100 stimuli for three times. The stimuli were selected so that each stimulus was evaluated by almost the same number of subjects. The stimuli used in the training phase were excluded.
4. After the evaluation, the subjects filled in a questionnaire about the vocal features they felt relevant to enthusiasm.

After the evaluation, one stimulus had 30 to 36 evaluation values given by 10 or 12 subjects. We took the average of all evaluation values, and the average was regarded as the result of the subjective evaluation for that stimulus.

3.3 Results

In order to investigate whether the subjects perceived singing enthusiasm consistently, we examined the correlation between the evaluation values given by a subject and the average of those given by all the other subjects.

Let $x_{si} \in \{0, 1, 2\}$ be an evaluation value for stimulus s given by the i-th subject. Let $\bar{x}_{si}$ be

$$\bar{x}_{si} = \frac{1}{N_s - 1} \sum_{j \neq i} x_{sj} \quad (1)$$

where N_s is the number of subjects who evaluated the stimulus s. Then calculate ρ_i, which is the correlation coefficient between x_{si} and $\bar{x}_{si}$ with respect to s. If ρ_i is high, it means that the i-th subject evaluated the stimuli in the same way as the other subjects. Note that we calculated ρ_i for sets A and B independently, which are represented by ρ_i^{A} and ρ_i^{B}.

Figure 2 is a histogram of ρ_i^{A} and ρ_i^{B}. This figure shows that the correlation coefficients are more than 0.7 for most of the subjects, so it is reasonable to suppose that the subjects perceived singing enthusiasm consistently. We can also observe that the correlation coefficients for set B are higher than those for the set A. This difference was caused by phrase-by-phrase differences in enthusiasm. Set A contained only one phrase, while set B had four phrases taken from different parts of the song. Different parts of the song had different enthusiasm; for example, phrase B1 (the first part) had smaller subjective evaluation values than phrase B3 (the hook line), which matches our intuition.

Next, we investigated the relationship between "intended enthusiasm" and "perceived enthusiasm." In this experiment, we asked singers to sing the song with two degrees

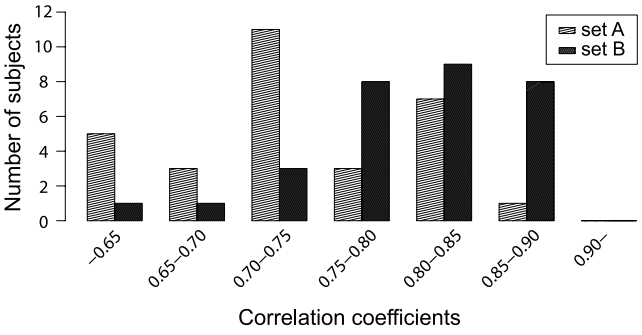
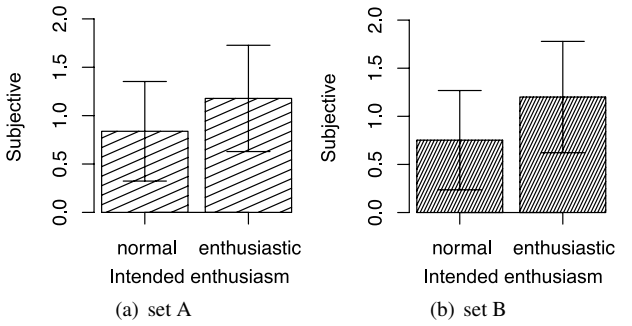


Figure 2. Correlation coefficients of the evaluations by the number of subjects



(a) set A (b) set B

Figure 3. Average of subjective evaluation for different singing styles (the error bars represent the standard deviation)

of enthusiasm: "enthusiastic" and "normal", to see whether this "intended enthusiasm" could actually be perceived by the subjects or not. To answer this question, we calculated the average of subjective evaluation values for the two "intended enthusiasm" sets. The results are shown in Figure 3. The paired Wilcoxon-signed rank test revealed significant differences (p<0.01) for both sets A and B, indicating that the subjects could distinguish the "intended enthusiasm" by listening to the voice.

Finally, we asked the subjects to describe the features of the singing voice that they felt were relevant to the perception of enthusiasm. Table 2 summarizes the features reported by the subjects. As the goal of this questionnaire was to identify acoustic features for automatically evaluating enthusiasm, we excluded opinions that were not related to the acoustic aspect of singing.

4. ACOUSTIC PARAMETERS

We examined several acoustic features for automatically evaluating enthusiasm based on the results of the questionnaire. The fundamental frequencies (F0) were extracted at 10-ms intervals using The Snack Sound Toolkit [4], and converted into log-scale (cent scale).

	loud voice
	strong attack
	sudden rise in loud voice
	loud voice on high notes
	articulated dynamics
	strong articulation of each note
	scooping up the pitch at the beginning
	pitched on key
	pitched higher than the correct note
	stable pitch
	voice with vibrato
enthusiastic	forceful voice
	shouting voice
	bright voice
	hoarse voice
	keeping forced voice until just before the release
	clearly pronounced lyrics
	articulated consonants
	strong breath sounds
	portamento
	some improvisation of rhythm
	some improvisation of melody
	getting into the rhythm
	soft voice
	monotonous voice
	pitched clearly off key
	pitched lower than the correct note
not	forceless voice
enthusiastic	dark voice
	muffled voice
	breathy voice
	released in short
	not getting into the rhythm

Table 2. Factors relevant to enthusiasm listed in the questionnaires

4.1 Examined Features

First, we focused on the loudness of the voice. Some subjects reported that they felt the "loud voice" was more enthusiastic, although all the stimuli were normalized to the same power. We guessed that this happened because the stimuli had different loudness levels. As the loudness depends not only on the power of the signal but also on its frequency, the "loud voice" might have larger loudness even though the physical power of all stimuli were equal. To investigate the relationship between loudness and enthusiasm, we calculate the A-weighted power of the stimuli, and examined a correlation between the A-weighted power and the enthusiasm. We used the A-weighted power instead of the loudness because it can be calculated more easily, and is widely used in acoustic measurements such as sound level meters. We designed an FIR filter which implements the A-weighting [6] shown in Figure 4, and calculated the power of the signals in dB after applying the filter.

Second, we focused on the change of power. There were several opinions on the change of sound power, such as "strong attack" or "strong articulation of each note." We examined the first derivatives of sound power (Δ power) of

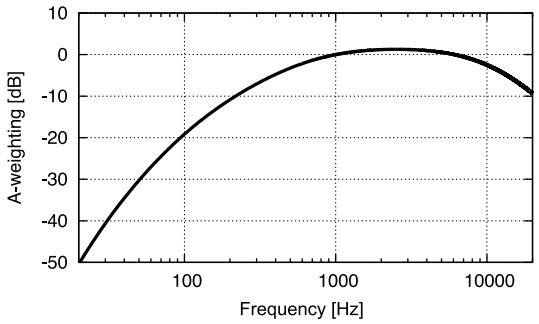


Figure 4. A-weighting curve

the voices as a physical feature expressing change of sound power, and took the maximum values for the feature. The Δpower was computed by

$$\Delta P(n) = \left\{ \sum_{k=-n_0}^{n_0} P(n+k)k \right\} / \sum_{k=-n_0}^{n_0} k^2 \qquad (2)$$

where $P(n)$ is the power at the n-th frame and n_0 is the number of side frames. The conditions were decided by the preliminary experiment: the frame size was 20 ms, the frame shift was 10 ms and the number of side frames was 4.

Third, we examined features related to F0 change at the beginning or end of a phrase. From the questionnaire, opinions concerning F0 change were observed such as "scoop-up" pitch at the beginning of phrases. Figure 5 shows an example of F0 with scoop-up and fall-down. Observing F0s of recorded voices, we found some of them were scooping up at the beginning, and some were falling-down at the end of phrases. The durations were within about 250 ms for both, and the frequency extent was under about 2000 cent for scoop-up, and under about 900 cent for fall-down. These features were described by Mayor et al. [1] as kinds of singing expressions, however no researches have revealed the relevance of the features to human perception of the singing voice.

As an acoustic feature that expresses these kinds of F0 change, we calculated the root mean square error (RMSE) value of F0 in regions of a constant duration, using Eq. (3):

$$E_{\mathrm{RMS}}(t_s, T) = \sqrt{\frac{1}{T} \sum_{t=t_s}^{t_s+T-1} \left(F_{\max}(t_s, T) - F0(t)\right)^2} \qquad (3)$$

$$F_{\max}(t_s, T) = \max_{0 \le t < T} F0(t_s + t) \qquad (4)$$

where $F0(t)$ is the fundamental frequency of the t-th frame, t_s is the beginning time of the calculation region, and T is the length of the region. The duration T was 200 ms. Here, a phrase is defined by a region not shorter than 500 ms with

continuous F0. We calculate two RMSE values corresponding to scoop-up and fall-down:

$$E_{\mathrm{up}} = E_{\mathrm{RMS}}(t_S, T) \tag{5}$$

$$E_{\mathrm{down}} = E_{\mathrm{RMS}}(t_E - T, T) \tag{6}$$

where t_S and t_E are the beginning and end of the phrase, respectively.

Finally, we examined vibrato-related features. Vibrato is one of the most basic features of the singing voice, and many studies have revealed its acoustic features. The results of the questionnaire suggested that vibrato is an important factor relevant to human perception of enthusiasm.

To detect vibrato, we computed "vibrato likeliness" proposed by Nakano et al. [2] Short-time Fourier transformation with a 32-point (320 ms) hanning window was applied to $\Delta F0(t)$ which is the first-order finite differential of $F0(t)$.

The amplitude spectrum $X(f,t)$ is expected to have a sharp peak range in the vibrato rate. Vibrato likeliness $P_v(t)$ is defined by Eq. (9) using the power $\Psi_v(t)$ and the sharpness $S_v(t)$.

$$\Psi_v(t) = \sum_{f=R_L}^{R_H} \hat{X}(f,t) \tag{7}$$

$$S_v(t) = \sum_{f=R_L}^{R_H} |\Delta_f \hat{X}(f,t)| \tag{8}$$

$$P_v(t) = \Psi_v(t) S_v(t) \tag{9}$$

where $\hat{X}(f,t)$ is $X(f,t)$ normalized over f, and $\Delta_f \hat{X}(f,t)$ is the first-order derivative of $\hat{X}(f,t)$ with respect to f. R_L and R_H are 5 and 8 Hz, respectively. Then we detect vibrato when $P_v(t)$ is higher than a threshold and $F0(t)$ crosses its regression line more than five times, as shown in Figure 6.

We derived three parameters of vibrato: (1) the rate V_r [Hz], (2) the extent V_e [cent], and (3) the ratio of time with vibrato in all the vocal regions V_t calculated as follows:

$$V_r = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2r_i} \tag{10}$$

$$V_e = \frac{1}{N} \sum_{i=1}^{N} e_i \tag{11}$$

$$V_t = \frac{1}{t_{F0}} \sum_{i=1}^{N} r_i \tag{12}$$

where N, r_i, and e_i are as shown in Figure 6 and t_{F0} is the total time of detected F0. However, if ($V_r < 5$ or $V_r > 8$) or ($V_e < 30$ or $V_e > 150$), the values were discarded because such values are likely to be caused by fine F0 fluctuation or analysis error. Note that the three vibrato parameters are 0 for voices when no vibrato is detected.

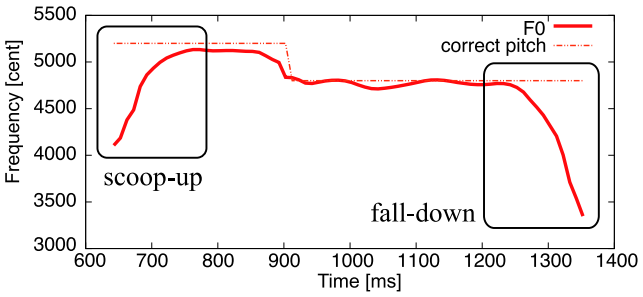


Figure 5. An example of scoop-up and fall-down

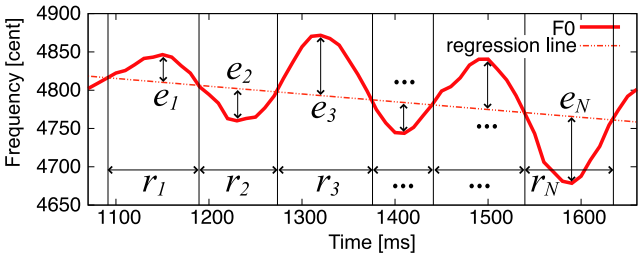


Figure 6. Calculation of vibrato-related feature

4.2 Results

As an evaluation of acoustic features, we calculated the correlation coefficient between individual features and the average human evaluation of enthusiasm. The results are shown in Table 3. From these results, we picked up three features that had relatively high correlations for both sets A and B: A-weighted power, E_{down}, and V_e.

The maximum Δpower and E_{up} gave only low correlation for set B. All of the three vibrato-related features gave relatively high correlation because the correlation between these three features are high (from 0.70 to 0.88), therefore we chose only one of these features.

The A-weighted power gave the best correlation among the examined features. From our observation, the A-weighted power seemed to be related to the quality of voice. The voice with high A-weighted power did not only sounded louder but also gave a clear and rich impression. The A-weight amplifies the frequency range around 3 kHz, which coincides with the frequency of the singing formant [5]. The A-weighted power and existence of the singing formant may be related, but the singing formant was not necessarily observed clearly in the voice even when the voice had high A-weighted power.

5. SINGING ENTHUSIASM EVALUATION SYSTEM

5.1 System Overview

Based on the observations described in the previous section, we constructed the Singing Enthusiasm Evaluation System

	Set A	Set B	B1	B2	B3	B4
A-weighted power	0.47	0.54	0.36	0.50	0.51	0.49
Max. Δpower	0.23	-0.22	0.05	0.13	-0.10	-0.09
E_{up}	0.20	0.07	-0.09	0.21	0.14	-0.12
E_{down}	0.35	0.36	0.13	0.38	0.29	0.50
Vibrato time V_t	0.37	0.30	0.42	0.25	0.30	0.36
Vibrato extent V_e	0.37	0.37	0.38	0.27	0.38	0.47
Vibrato rate V_r	0.37	0.37	0.39	0.29	0.38	0.47

Table 3. Correlation coefficients between acoustic parameters and subjective evaluations

(SEES), as outlined in Figure 7. The SEES consists of three subsystems: SEES front-end, core and back-end.

The SEES front-end consists of a high-pass filter for noise reduction, signal power normalizer, and F0 extractor. The SEES core is the main part of the system, and extracts the acoustic features: the A-weighted power, the RMSE for fall-down and the vibrato extent. The SEES back-end is the part where final evaluation values are computed by linear sum features. The multiplier coefficients correspond to the weights of the features and they must be determined beforehand. In our experiment, the coefficients were determined by a multiple linear regression analysis on set A using the subjective evaluation values as the response variables and feature values as the explanatory variables.

5.2 Evaluation of the System

Finally, we evaluate the system by comparing the system's output with the human evaluation values. Set A was used as a training set for determining the multiplier coefficient. We examined both sets A and B for testing the system, which corresponded with the closed test and open test, respectively.

The results are shown in Figure 8. The correlation coefficients between the system output and the human evaluation were 0.60 for set A (closed test), and 0.65 for set B (open test). We obtained good correlations not only for set A but also for set B, so we consider the system will produce stable evaluations for various melodies and lyrics.

6. CONCLUSIONS

In this paper we introduced "enthusiasm" as an aspect of evaluating the singing voice for karaoke, and obtained the following results by experiments.

First, subjective evaluations revealed that humans perceive singing enthusiasm almost consistently, and listeners can distinguish whether singers are singing enthusiastically or not only by listening to the voice.

Second, questionnaires revealed three effective acoustic features of voices: the A-weighted power, the RMSE for fall-down and the vibrato extent.

Finally, we developed a singing enthusiasm evaluation system using the three features and achieved correlation coefficients of more than 0.6 for unknown input.

As a future work, we need to evaluate our system using various inputs such as different songs that contain more variations of key, tempo, and genre.

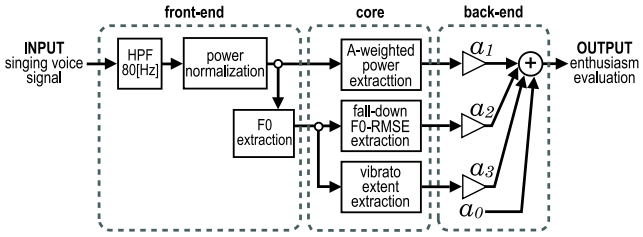
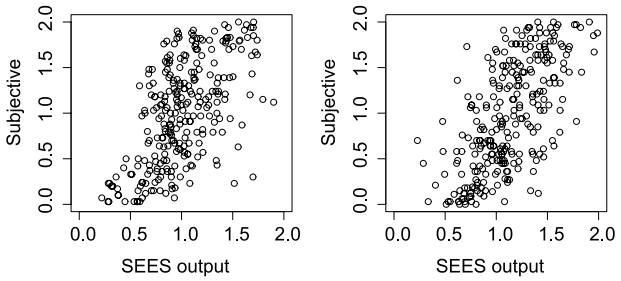


Figure 7. Overview of the SEES



(a) input: set A (closed) (b) input: set B (open)

Figure 8. Comparison of SEES output and subjective evaluations

7. REFERENCES

[1] O. Mayor, J. Bonada, A. Loscos: "The Singing Tutor: Expression Categorization and Segmentation of the Singing Voice," Proc. AES Convention, 2006.

[2] T. Nakano, M. Goto, Y. Hiraga: "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," Proc. Intespeech, pp. 1706–1709, 2006.

[3] K. R. Scherer, "Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them?," J. New Music Research, vol. 33, no. 3, pp. 239-251, 2004.

[4] K. Sjölander: "The Snack Sound Toolkit," http://www.speech.kth.se/snack/, 1997-2001.

[5] J. Sundberg: "Articulatory interpretation of the 'singing formant'," J. Acoust. Soc. Am., vol. 55, no. 4, pp. 838–844, 1974.

[6] Int. Electrotechnical Commission, "Electroacoustics – Sound level meters– Part 1: Specifications," IEC 61672-1, 2002.