Ann. Bull. RILP. No. 10, 191-198 (1976)

ACOUSTIC ANALYSIS AND SUBJECTIVE EVALUATION OF SUNG VOWELS

Mariko Tatsumi, Osamu Kunisaki, * and Hiroya Fujisaki

I. Introduction

The art of singing at high sonority is one of the important techniques to be acquired by operatic and concert singers.[1, 2] It not only provides the singer with a means of accomplishing sufficient loudness without excessive phonatory efforts, thereby helping to avoid possible impairment of the laryngeal mechanism which might be incurred by hours of strenuous singing, but also gives a "brilliance" to the voice which makes it stand out from the sounds of musical instruments. Although auditory evaluation of sonority is indispensable in voice training, it is subject to greater variability due to individual differences of referees, so that establishment of an objective measure of sonority is desired in order to serve as the basis for a subjective evaluation.

On the other hand, several studies have already been conducted on the acoustic characteristics of vowels sung by male singers, and these have revealed that, among other features, such vowels are commonly characterized by the presence of prominent spectral components around 3 kHz, where the human auditory sensitivity is the highest.[3-6] The spectral prominence often appears as one single peak in the case of a broad-band spectral analysis, and thus is generally referred to as the "singing formant." A detailed analysis based on the acoustic theory of speech production,[7] however, has revealed that such an enhancement of spectral components is actually realized by the adjacency of the third and the fourth, and possibly also of the fifth formant frequencies, caused mainly by lowering the larynx and widening the pharynx.[8, 9] The relationships between these acoustic characteristics and the auditory perception of the sonority of sung vowels, however, have never been analyzed.

The present investigation has been conducted to examine the possible correlation between the spectral characteristics of sung vowels and the auditory perception of their sonority through acoustic analysis and subjective evaluation. The basic approach and the steps of the present study are described schematically by Fig. 1. Vowels produced by a professional singer are analyzed acoustically on the one hand to determine with high accuracy the frequencies of the first five formants, while on the other hand they are evaluated subjectively with respect to their sonority. Results of the acoustic analysis and subjective evaluation are further analyzed to find the acoustic correlates of the perceptual sonority of sung vowels. In order to confirm the results obtained from the study of natural vowels, and to obtain more quantitative understanding of the roles of various acoustic factors in the perception of sonority, various sung vowels are then synthesized from a set of controlled parameters, and are evaluated by the same procedure as for the natural sung vowels. The steps in broken lines represent future

* Department of Electrical Engineering, Faculty of Engineering, University of Tokyo.

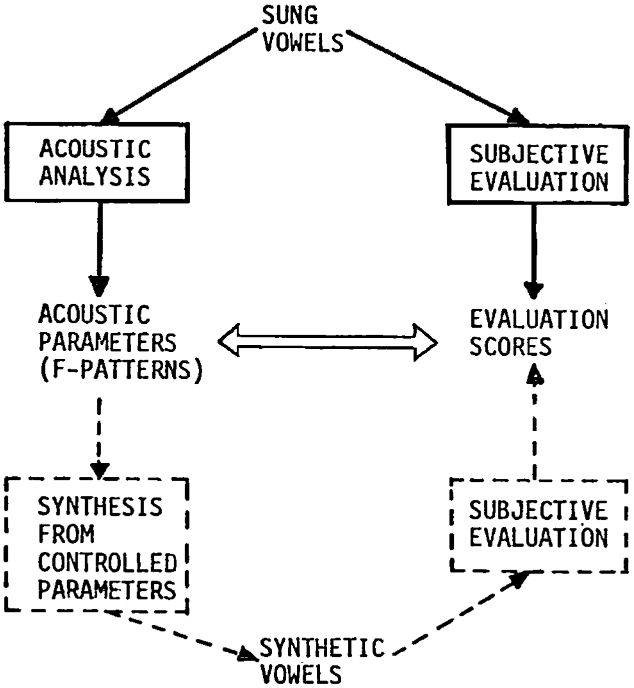


Fig. 1. Steps in the study of subjective and objective evaluation of vowels in singing.

stages of the research, so that only the results obtained from natural sung vowels will be described in this report.

II. The Singer and the Material

A 27-year-old high baritone singer with a relatively high sonority in singing provided the acoustic material. He has a pitch range from G_2 to G_4, and has been engaged in operatic performances for the past two years. Two kinds of sung vowels were recorded: one consisting of 13 samples each of the five Japanese vowels which the singer himself judged as being relatively high in sonority (to be denoted by Su-1); the other consisting of one sample each of the same vowels deliberately produced and judged by the singer as being low in sonority (denoted by Su-2). Both of these samples were sung approximately at B_3, corresponding to a fundamental frequency of 240 Hz. In addition to these sung vowels, two kinds of spoken vowels were also recorded for the sake of comparison with sung vowels: one consisting of five samples each of the five vowels sustained at the same pitch as the sung vowels (denoted by Sp-1); the other consisting of three samples each of the vowels sustained at approximately 150 Hz, which is the singer's normal speech pitch (denoted by Sp-2). The classification of materials are summarized in Table 1.

These vowel samples were sung or spoken in isolation and recorded in an anechoic chamber. The duration of sung and spoken vowels were

Table 1. Four types of sung and spoken vowels and number of samples analyzed for each of the five Japanese vowels.

Vowel Type	Pitch	No. of samples
Vowels sung with high sonority (Su-1)	B_3	13
Vowels sung with low sonority (Su-2)	B_3	1
Vowels spoken at the same pitch as above (Sp-1)	B_3	5
Vowels spoken at normal pitch (Sp-2)	D_3	3

approximately 2.5 sec and 1.5 sec , respectively. These samples were then read into a digital computer via an analog-to-digital converter at a sampling rate of 10 kHz and with an accuracy of 10 bits per sample, to be stored on a digital magnetic tape for further processing.

III. Analysis of Sung and Spoken Vowels

Although sound spectrograms were successfully used in the previous study [8] to disclose the detailed structure of the singing formant, they are not sufficient for the accurate specification of the spectral envelope and for the precise measurement of formant frequencies, to provide an acoustic correlate of sonority. Hence the present study utilized computer techniques both for the spectral analysis and for the extraction of formant frequencies.

Spectral analysis was performed on the portion of each of the vowel samples where the acoustic characteristics were found to be most stationary, using a hanning window of 102.4 msec width. The fundamental frequency was determined by means of the cepstrum technique, [10] and was then utilized to extract the spectral envelope represented by the amplitudes of all the harmonic components.

Extraction of formant frequencies was then performed by the method of Analysis-by-Synthesis, [11] i. e., by finding the optimum set of formant frequencies of a synthetic spectral envelope representing the closest approximation to the measured one. The model for the synthesis of the spectral envelope over the frequency range of 0 to 5 kHz is given by:

$$\begin{aligned} F(f) = 20 \log_{10} \left| \frac{S}{(S+\alpha)^2} \right| + \sum_{i=1}^{5} 20 \log_{10} \left| \frac{Si\ Si^*}{(S-Si)(S-Si^*)} \right| \\ + \left[0.43 \left(\frac{f}{F_1 n} \right)^2 + 0.00071 \left(\frac{f}{F_1 n} \right)^4 \right] + 20\beta \log_{10} \left| \frac{f}{1000} \right| \quad (dB), \end{aligned} \tag{1}$$

where

$$S = 2\pi jf, \qquad Si = 2\pi(jFi + \Delta Fi),$$

$$\Delta Fi = a(0.015Fi + 15), \qquad a = 1.0\sim 1.6,$$

$$\alpha = 200\pi, \qquad \beta = -1.0\sim 1.0.$$

The first term represents combined characteristics of the glottal source and the radiation; the second term indicates contributions of the first five formants to the vocal tract transfer function, the third term stands for the correction factor which approximates effects of the sixth and higher formants, [12] and the last term is an additional correction which may be varied to compensate for residual frequency characteristics of various origins. In this model, the five formant frequencies (F_1-F_5), the parameter in the higher-pole correction (F_1n), and the overall spectral slope (β) can be varied independently, to minimize the mean squared difference between spectral envelopes of the input speech and the model, evaluated at frequencies equal to integral multiples of the fundamental frequency (F_0) and expressed in dB^2. This method has been successfully applied to analysis and automatic recognition of sustained Japanese vowels. [13] Examples of power spectra as well as of spectral envelopes and parameters obtained by Analysis-by-Synthesis are shown in Fig. 2 for one sample each of the above-mentioned types of the vowel /e/. In comparison with spoken vowels (Sp-1 and Sp-2), sung vowels (Su-1 and Su-2) are found to be characterized by higher concentration of energy around 3 kHz. The concentration is apparently caused by the adjacency of three formant frequencies (F_3, F_4, and F_5) in the case of a sung vowel with higher sonority (Su-1), while it is caused by two formant frequencies (F_3 and F_4) in the case of lower sonority (Su-2). An appreciable downward shift of the second formant frequency (F_2) is also observed in Su-1. On the other hand, a change in the fundamental frequency from 156 Hz to 233 Hz is seen to have only minor influences on formant frequencies of spoken vowels, except possibly on the first formant frequency (F_1).

As a quantitative measure for the degree of concentration of the three higher formants, we may adopt the unbiased standard deviation of F_3, F_4, and F_5. Figure 3 shows an example of the spectrum of a sung vowel /e/ together with the mean M and the standard deviation σ of the three higher formant frequencies. A higher concentration of F_3, F_4, and F_5 is thus represented by a smaller value of their standard deviation around the mean. The mean values of σ are listed in Table 2 for the two types of each vowel, i.e., vowels sung at high sonority (Su-1) and those spoken at normal pitch (Sp-2); they clearly indicate the difference in concentration of F_3-F_4-F_5 between sung and spoken vowels.

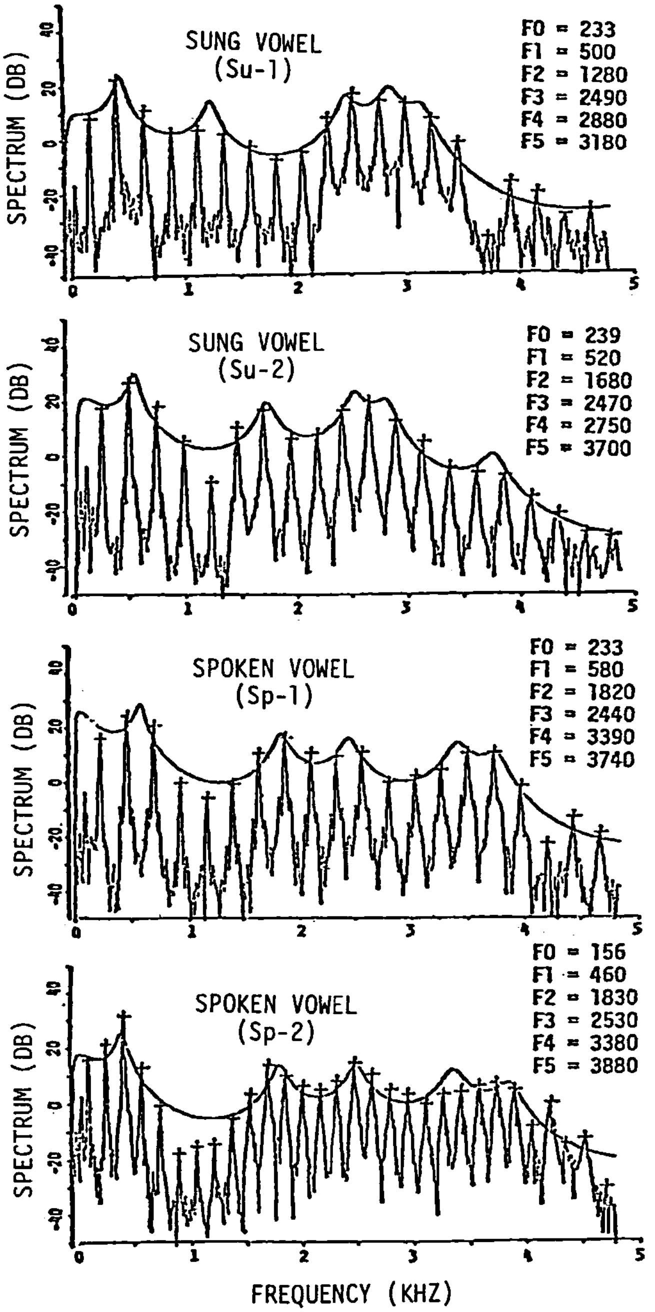
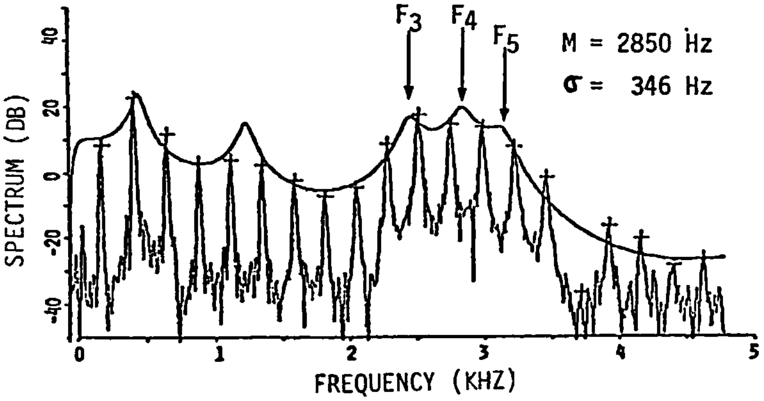


Fig. 2. Comparison of spectral envelopes and their parameters of four vowel types of /e/, obtained by Analysis-by-Synthesis.



$$\sigma = \sqrt{\sum_{i=3}^{5} (F_i - M)^2 / 2}, \quad \text{where} \quad M = \sum_{i=3}^{5} F_i/3 .$$

Fig. 3. Standard deviation (σ) of F_3, F_4, and F_5 as a measure for higher formant concentration.

Table 2. Mean standard deviation of F_3, F_4, and F_5 of sung vowels (Su-1) and spoken vowels (Sp-2).

	MEAN STANDARD DEVIATION OF F_3, F_4, AND F_5 (Hz)	
VOWEL	SUNG VOWELS (Su-1)	SPOKEN VOWELS (Sp-2)
/a/	376	654
/i/	247	473
/u/	273	727
/e/	413	619
/o/	333	539

IV. Subjective Evaluation of Sonority of Sung Vowels

Although evaluation of the quality of sung vowels may depend on a number of factors related both to static and dynamic characteristics of the glottal source and the vocal tract transfer function, results of the foregoing analysis indicate that the perceptual sonority of sustained sung vowels may be closely related to the concentration of higher formants around 3 kHz. In order to further examine the correlation between the proposed index of higher formant concentration (σ) and the perceived sonority of sung vowels, subjective evaluation tests were performed on the same samples of sung vowels /a/ and /e/ used in the acoustic analysis.

Eleven samples each were selected as stimuli from recorded materials of vowels /a/ and /e/ sung at higher sonority (Su-1). Separate test materials were prepared for each vowel by arranging the samples in a semi-random order at intervals of 6 sec. One set of test materials consisted of a total of 120 items containing 10 each of the 11 samples, preceded and followed by 5 dummy samples each. These materials were complied by a digital computer and were read out via a digital-to-analog converter with an accuracy of 8 bits, to be recorded on audio tapes for use in off-line experiments.

A 27-year-old female voice trainer, who has been in her profession for the past five years, served as the subject, making 100 judgments on each of the 11 stimuli. The evaluation was based on a five-point rating, giving the highest point "5" to those stimuli with the highest perceptual sonority. Figure 4 shows results of the subjective evaluation test of vowel /e/, expressed in terms of the mean evaluation score(E) and plotted against the standard deviation (σ) of the three higher formant frequencies for each of the 11 samples. The straight line in the figure indicates linear regression of the 11 data points. The results clearly indicate the existence of a strong negative correlation (-0.89) between E and σ, validating the assumption that the concentration of the three higher formants constitutes the main acoustic correlate of perceived sonority of sung vowels.

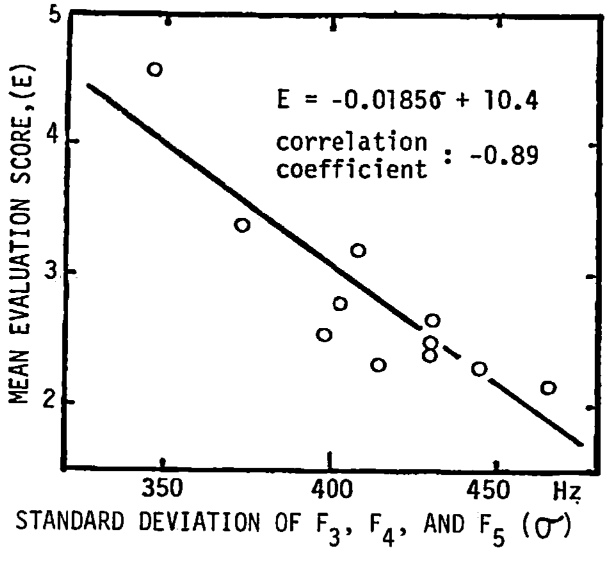


Fig. 4. Mean evaluation score versus standard deviation of F_3, F_4 and F_5 for each sample of sung vowel /e/.

Although the results confirmed the close relationship between subjective sonority and concentration of higher formants in sung vowels, the stimuli used in these experiments were natural utterances whose acoustic characteristics were not controlled, so that the results were not free from influences of factors other than concentration of higher formants. Furthermore, these stimuli were not necessarily suited for systematic investigation of effects of other factors that might possibly affect perceived sonority, such as mean value of higher formant frequencies, frequency of the second formant relative to those of higher formants, etc. Work is presently under way using synthetic vowels to obtain more precise estimates of effects of various factors to perceived sonority, as well as to find acoustic correlates of subjective characteristics other than sonority which may play important roles in the overall evaluation of the singing voice.

References

1. Vennard, W. (1967), Singing, the Mechanism and the Technique, Fischer Inc., New York.
2. Husler, F. and Y. R. Marling (1965), Singing, Faber and Faber Ltd., London.
3. Helmholz, H. L. v. (1862), Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik, Braunschweig.
4. Bartholomew, W. T. (1934), "A Physical Definition of 'Good Voice-Quality' in the Male Voice," J. Acoust. Soc. Am., 6, 25-33.
5. McGinnis, C. S., M. Elnick,and M. Kraichman (1951), "A Study of the Vowel Formants of Well-known Male Operatic Singers," J. Acoust. Soc. Am., 23, 440-446.
6. Fry, D. B. and L. Manén (1957), "Basis for the Acoustical Study of Singing," J. Acoust. Soc. Am., 29, 690-692.
7. Fant, G. (1960), Acoustic Theory of Speech Production, Mouton & Co., 's-Gravenhage.
8. Sundberg, J. (1970), "Formant Structure and Articulation of Spoken and Sung Vowels," Folia Phoniat., 22, 28-43.
9. Sundberg, J. (1974), "Articulatory Interpretation of the Singing Formant," J. Acoust. Soc. Am., 55, 838-844.
10. Noll, A. M. (1964) "Short-Time Spectrum and 'Cepstrum' Techniques for Vocal-Pitch Detection," J. Acoust. Soc. Am., 36, 296-302.
11. Bell, C. G., H. Fujisaki, et al. (1961), "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," J. Acoust. Soc. Am., 33, 1725-1736.
12. Fant, G. (1959), "Acoustic Analysis and Synthesis of Speech with Applications to Swedish," Ericsson Technics, No. 1.
13. Fujisaki, H., N. Nakamura,and K. Yoshimune (1970), "Analysis, Normalization, and Recognition of Sustained Japanese Vowels," J. Acoust. Soc. Japan, 26, 152-153.