TR-H-265

An Acoustical Study of Sound Production in Biphonic Singing, Xöömij.

Seiji ADACHI and
Masashi YAMADA (Osaka Univ. of Arts)

1999.2.8

ATR人間情報通信研究所
〒619-0288　京都府相楽郡精華町光台2-2　TEL: 0774-95-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
Telephone: +81-774-95-1011
Fax　　　: +81-774-95-1008

An acoustical study of sound production in biphonic singing, Xöömij

Seiji Adachi[†]
ATR Human Information Processing Res. Labs.
2-2 Hikaridai, Seika, Kyoto 619-02 Japan

Masashi Yamada
Dept. of Musicology, Osaka University of Arts
Higashiyama, Kanan, Osaka 585 Japan

[†] Present address: Faculty of Information Science and Technology, Aichi Prefectural University, Nagakute, Aichi, 480-1198 Japan. e-mail: `adachi@ist.aichi-pu.ac.jp`

A theory that the high melody pitch of biphonic singing, Xöömij, is produced by the pipe resonance of the rear cavity in the vocal tract is proposed. The front cavity resonance is not critical to the production of the melody pitch. This theory is derived from acoustic investigations on several three-dimensional shapes of a Xöömij singer's vocal tract measured by magnetic resonance imaging. Four different shapes of the vocal tract are examined, with which the melody pitches of F6, G6, A6 and C7 are sung along with the F3 drone of a specific pressed voice. The second formant frequency calculated from each tract shape is close to the melody pitch within an error of 36 cents. Sounds are synthesized by convolving a glottal source waveform provided by the Rosenberg model with transfer functions calculated from the vocal tract shapes. Two pitches are found to be successfully perceived when the synthesized sounds are listened to. In a frequency range below 2 kHz, their spectra have a strong resemblance to those of the sounds actually sung. The synthesized sounds, however, fail to replicate the harmonic clustering at 4-5 kHz observed in the actual sounds. This is speculated to originate from the glottal source specific to the 'pressed' timbre of the drone.

PACS numbers: 43.75.Rs, 43.70.Aj

INTRODUCTION

Biphonic singing, also known as throat-singing, is a vocal technique found in Central Asian cultures, by which one singer produces two voices with different pitches simultaneously. When listening to the performance, a high melody pitch can be perceived along with a low drone pitch. Xöömij, also transliterated variously as Khoomei, Xoomii, etc., is the name used in Mongolia and Tuva to describe this technique. In other regions, it may have other names. In this paper, we use Xöömij as the general term indicating all biphonic singing.

Up until now, two major theories have been proposed on the production of the two pitches by Xöömij singing: 1) The "double-source" theory[1], which asserts the existence of a second sound source such as a whistle-like mechanism formed by the narrowing of the false vocal folds in addition to the true vocal fold vibration, and 2) the "resonance" theory[2], which asserts that only a glottal sound source exists, but that a higher harmonic component is so emphasized by an extreme resonance of the vocal tract that it is segregated from the other components and heard as another pitch. The fact that the melody pitches producible by the singer are limited to the harmonic series of the drone supports the resonance theory. This has also been confirmed by a spectrum analysis of recorded Xöömij singing.[3] The resonance theory is thus currently considered to be consistent with the observations.

If the resonance theory is correct, the next question is: Which part of the vocal tract causes such an extreme resonance? Generally in biphonic singing, the singer divides his vocal tract into two cavities connected by a narrow opening using his tongue. Trân and Guillou[2] infer that both of the cavities contribute to the resonance producing the melody pitch. This is, however, just an inference from introspections of Trân as an amateur Xöömij singer and from their own spectrum analysis.

To answer the question posed above and to further test the resonance theory, the following procedures were performed: 1) measure the vocal tract shape while singing Xöömij, 2) examine the acoustical characteristics of the tract, and 3) synthesize sounds which can actually be perceived as two separate pitches by the listener. Following preliminary investigations[4]-[7], this research verifies the resonance theory. We further propose a theory stating that the resonance of the rear cavity, that is from the glottis to the narrowing of the tongue, produces the Xöömij melody pitch. The resonance of the front cavity, that is from the articulation by the tongue to the mouth exit, is not critical to the production of the melody pitch.

I. VOCAL TRACT MEASUREMENT BY MRI

A method of measuring three-dimensional (3-D) shapes of the vocal tract using magnetic resonance imaging (MRI) is rapidly becoming established.[8] This method has an advantage in that it is non-invasive and is capable of obtaining tomographic images in any direction. It also poses no known danger to the human subject being imaged. The main disadvantage of using this method is that the scanning time for acquiring an image is a few tens of seconds to several minutes. Fortunately, this does not become a serious problem for the purpose of measuring stationary vocal tract shapes while long tones such

Table 1

Figure 1

as in Xöömij are being sung.

In our measurement, one male subject (S.K.) who is able to sing various Xöömij tones with a high degree of stability was employed. He was instructed to produce four different Xöömij tones for measurement purposes. These tones have the same drone pitch of F3 and their melody pitches are F6, G6, A6 and C7. For comparison, measurements were also made for two ordinary monophonic tones of the vowel /ɑ/ sung by the same subject: One was phonated with a pressed voice as if producing the drone of Xöömij, and the other was phonated normally. No audio signal indicating the pitch was presented to the subject. All of the Xöömij and monophonic tones were sung on S.K.'s relative pitch, which was about 50 cents lower than the standard pitch of A4=440 Hz. Although intense scan noise was generated from the MRI equipment, the subject did not wear ear plugs. During the imaging process, we also recorded the singing tones. Note that these are not fit for the spectrum analysis, because of interference with the scan noise, but are sufficient for pitch detection. The detected sound frequencies of the melody and drone pitches of the Xöömij tones, and those of the monophonic tones are shown in Table 1.

The equipment used for the measurements was a Shimazu Magnetic Resonance Tomograph SMT-100GUX (static magnetic field density of 1.0 T) with an anterior neck coil. For each image production, 33 axial slices from the glottis to the palate were taken. The major scanning parameters were a repetition time of 1010 ms, an echo time of 15 ms and a slice thickness of 4 mm. There was no gap for imaging between slices. Each slice had a size of 258.1 mm × 258.1 mm and a pixel matrix of 256 × 256. It took 153 sec. for one image acquisition. During the image scanning, the subject was instructed to continue singing except for brief times (allowed for breathing). No pause in the scanning was taken while acquiring one image.

The 33 axial slices for each imaging sequence were interpolated by volume rendering software (VoxelView) to reconstruct the 3-D data. 2-D slices in any (not necessarily the axial) direction can be reproduced from this data. Figure 1 shows the mid-sagittal slices of the vocal tracts for the four Xöömij tones and the two monophonic tones. We can see in Figs. 1(a) to (d) that the singer narrows his vocal tract by using the tongue tip and the palate to produce the Xöömij tones. The tract is, therefore, divided into two cavities. We call the one from the glottis to the narrowing, the "rear cavity", and the other from the narrowing to the mouth exit, the "front cavity". Note that the tongue tip moves towards the rear as the melody pitch is raised from F6 to C7. No such narrowing is

Figure 2

Table 2

found in the vocal tract shapes for the monophonic tones, regardless of the voice timbre [Figs. 1(e) and (f)].

The superimposed lines on each vocal tract in Fig. 1 show the upper boundary, lower boundary, and the midline of the tract. Line segments with a constant interval of 5.162 mm are also drawn perpendicular to the midline. The boundaries were estimated by hand first, and the midline was then extracted by the same algorithm used in Ref. [9]. According to the positions and the angles of the line segments, cross-sectional slices were reproduced from the 3-D image data. Each slice was then analyzed to measure the area of the vocal tract at the position where that slice came from.

Figure 2 depicts the area functions, which indicate the cross-sectional areas as functions of the distance from the glottis along the midline, for all of the vocal tract shapes we measured for the Xöömij and monophonic tones. Table 2 lists numerical data for the area functions.

II. TRANSFER FUNCTIONS

The acoustic model employed in this paper is described in Appendix. This model assumes the transmission of a one-dimensional wave along the axis of the vocal tract. To construct the model, we modified a model for calculating the input impedance of brass instruments, developed by Caussé et al. [10]. The original model includes visco-thermal loss, i.e., loss due to the friction and thermal exchange between the air and the wall of the acoustic tube concerned, and loss due to the radiation from the exit. In our model, the same visco-thermal loss is considered. The radiation from the mouth is modeled by that from a piston with an infinite baffle. In addition to the losses above, our model includes the yielding wall effect[11] and loss due to the incomplete glottal closure.

Calculated volume velocity transfer functions for the Xöömij tones are depicted by solid lines in Figs. 3(a) to (d). The first and the second formant frequencies, and their bandwidths are listed in Table 3(a). Note that the second formant frequencies for the F6, G6, A6 and C7 vocal tract shapes are 1356, 1562, 1692, and 2066 Hz, respectively. These values are close to the sound frequencies of melody pitches of recorded F6, G6, A6 and C7 Xöömij tones within an error of 36 cents. This implies that the melody pitch is produced by the resonance corresponding to the second formant. Transfer functions for the monophonic tones are depicted in Figs. 3(e) and (f). Their formant structures,

Figure 3

Table 3

especially the relations between the first and the second formant frequencies, are those typically observed in the phonation of the vowel /ɑ/.

To examine how the rear and front cavities of vocal tract shapes for Xöömij tones contribute to the resonance, we calculated the transfer functions of tract shapes whose front cavity had been removed, i.e., having only the rear cavity and the narrowing by the tongue. Data regarded as the ends of narrowing are underlined in Table 2. Transfer functions without the front cavity are plotted by dashed lines in Figs. 3(a) to (d). We can see that the transfer functions without the front cavity have smaller magnitudes in the range of 2-3 kHz. This is because of the lack of the front cavity resonance, which should be in the same range and has a very large bandwidth. Small humps of the plots by the solid lines in Figs. 3(b) and (c), which are visible between the second and third formants, are evidence of this resonance.

In Table 3(b), the first and the second formant frequencies, and their bandwidths of the transfer functions without the front cavity are listed. We find that the frequencies are hardly changed by the removal of the front cavity. This suggests that the fundamental formant structure is determined by the rear cavity resonance, and that the effect of the front cavity resonance on the structure is small. The second formant producing the Xöömij melody pitch can, therefore, be directly associated with the rear cavity resonance. The front cavity may assist the melody pitch by enhancing the magnitude of the formant peak. This effect becomes apparent for the C7 tone, whose melody pitch is closer to the 2-3 kHz range than the other tones. It will, however, be found in the next section that the enhancement is not critical to the production of the melody pitch.

To further investigate the relation between formants and resonance modes of the rear cavity, we utilized a method by which 'craftsmen' often adjust the resonance frequencies of wind instruments, that is, we changed the original tract shape locally to see how the formant frequencies would be perturbed by the change. By reducing the cross-sectional area at several places in the vocal tract, the first and the second formant frequencies were calculated. The main observations were: The first formant frequency is not perturbed very much except for a reduction near the narrowing portion of the tract, whereas the second formant frequency is considerably decreased by reductions in the vicinity of data 12 to 16. The observations suggest that the Helmholtz resonance mode, which has no pressure node in the cavity, causes the first formant, and that the pipe resonance mode, which has one pressure node, causes the second formant. Consequently, we conclude that

Figure 4

the pipe resonance mode of the rear cavity produces the Xöömij melody pitch.

III. SOUND SYNTHESIS

Let us investigate whether tones having two pitches can actually be reproduced by the transfer functions for the Xöömij tones obtained in the previous section. To this end, we synthesized tones using the acoustic tube model.

In this synthesis method, a glottal flow waveform is convolved with the transmission impulse response of the vocal tract, which is the inverse Fourier transformation of the pressure-to-velocity transfer function. Here we assume that the Rosenberg model[12], capable of providing the glottal flow for various qualities of normal speech, can also be applied to Xöömij singing. The glottal volume flow $U_g(t)$ for one oscillation period T is modeled by

$$U_g(t) = \begin{cases} \frac{1}{2}\left[1 - \cos(\pi t/T_p)\right] & (0 \leq t \leq T_p) \\ \cos(\pi (t - T_p)/2T_n) & (T_p \leq t \leq T_s) \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

where T_p is the time for increasing the flow, and T_n is the time for decreasing the flow. The duty period T_s is the sum of T_p and T_n.

Figure 4 (a) depicts an example ($T_s/T = 0.4$ and $T_p/T_n = 3.0$) of the modeled waveforms. To mimic the characteristic 'pressed' timbre of the drone of Xöömij, we adjusted the parameter ratios as $T_s/T = 0.2$ and $T_p/T_n = 3.0$. As a result, $U_g(t)$ contains rich harmonics as shown in Fig. 4 (b). The oscillation period T was set to the inverse of the fundamental frequency of the drone pitch.

Four synthesized Xöömij tones of 4.0 sec. duration, and of 48 kHz sampling, were produced from the F6, G6, A6 and C7 transfer functions. The spectra of these tones are presented in Figs. 5(a) to (d). The dashed lines drawn in the same figure are spectrum envelopes estimated by linear prediction (LP). To do the estimation, the tones of the 48 kHz sampling were down sampled to 12 kHz, and the number of LP coefficients was set to 20.

The fundamental frequencies and the formant data estimated from the synthesized tones are listed in Table 4(a). We find that the estimated formant frequencies are different from the formant frequencies directly derived from the transfer functions listed in Table 3. The estimation errors of the first formant frequencies (F_1) of F6, G6, A6 and C7 tones are -108.7, -60.7, 120.9 and 62.3 cents, respectively. Those of the second formant frequencies (F_2) of F6, G6, A6 and C7 tones are -5.1, -37.0, 2.0 and -27.9 cents, respectively. By a simple test of listening to these synthesized tones, we could confirm

Figure 5

Table 4

that the same melody pitches heard when listening to the actual Xöömij tones are able to successfully be perceived in addition to the F3 drone[13].

Using the same parameters of $T_s/T = 0.2$ and $T_p/T_n = 3.0$, a pressed monophonic tone was synthesized from the transfer function shown in Fig. 3(e). A normal monophonic tone was also synthesized from the transfer function shown in Fig. 3(f) with $T_s/T = 0.4$ and $T_p/T_n = 3.0$. The spectra of these synthesized monophonic tones (solid lines), and the envelopes (dashed lines) are depicted in Figs. 5(e) and (f). A simple listening test confirmed that these tones are able to be heard as phonations of the vowel of /ɑ/, but that their timbres differ slightly[13].

We also synthesized tones from transfer functions calculated without the front cavity. Their spectra (solid lines) and the spectrum envelopes (dashed lines) are depicted in Fig. 6. The fundamental frequencies and the formant data are listed in Table 4(b).

The magnitudes of the second formants generally decrease due to the lack of the front cavity resonance. A simple listening test, however, showed that the melody pitches of all the tones (F6, G6, A6 and C7) are able to still be heard in addition to the drone. Consequently, we conclude that the melody pitch of the Xöömij tone is produced by the rear cavity resonance. The front cavity resonance may enhance the sensation of the melody pitch somehow, but this effect is not critical to the production of the pitch. A remaining issue is to quantitatively analyze to what extent the front cavity contributes to the sensation. In this case, a psychoacoustic experiment will probably be needed. This is, however, out of the scope of this paper.

The synthesized tones can be compared with tones actually sung by the same subject, S.K. The spectra (solid lines) and the estimated envelopes (dashed lines) of the Xöömij and monophonic tones are shown in Fig. 7. These were recorded separately from the MRI measurement, since the intense scan noise from the magnet interfered with the recording acceptable for spectrum analysis. The recording was done in an anechoic room, with a

Figure 6

Figure 7

Table 5

B&K 4003 microphone connected to a SONY DTC-A8 DAT recorder. The sampling rate was 48 kHz. During the recording, S.K. was in a supine position like he was in the MRI experiment.

In the frequency range up to around 2 kHz, where the first and the second formants appear, the spectra of the synthesized and recorded Xöömij tones for each pitch have a strong resemblance. In particular, the harmonic components having the largest magnitudes in the second formant of the synthesized and recorded Xöömij tones are the same: the 8th, 9th, 10th and 12th harmonics for the F6, G6, A6 and C7 Xöömij tones, respectively. It is these harmonic components that are segregated from the others and that can each be perceived as the melody pitch.

More minute observations can be made by comparing the formant data estimated from the synthesized tones listed in Table 4(a) with the data estimated from the recording tones listed in Table 5. The F_1 differences between the synthesized and recorded tones of F6, G6, A6 and C7 are 4.1, 10.6, -13.0 and -54.0 cents, respectively. The F_2 differences for the F6, G6, A6 and C7 tones are -21.9, -36.6, -9.2 and -46.6 cents, respectively. These are acceptable values, because they are comparable with the estimation errors between the formant frequencies derived from the transfer functions and those estimated from the synthesized tones. On the other hand, the bandwidths estimated from the synthesized and recorded tones are roughly of the same order. In particular, a common tendency is observed for the second formant bandwidth (BW_2) to become larger as the melody pitch is increased. These observations provide general support for the parameters of the acoustic model we set for the calculations.

In the higher frequency range above 2 kHz, however, the spectra do not have a strong resemblance. The recorded Xöömij tones have less energy in the 2-4 kHz frequency range, and, in contrast, show a clustering of the harmonics in the range of 4-5 kHz. For the synthesized Xöömij tones, no such harmonic structures are found. Their spectra show smooth reductions of the harmonic levels due to the character of the glottal waveform, except for some modulation due to the formant structure. The spectrum difference in this frequency range is one of the reasons why the timbre of synthesized Xöömij tones is artificial and different from that of recorded tones.

The energy suppression at 2-4 kHz may be due to zeros in the transfer function. These are generally produced by cross-modes of propagation in the side branches of the vocal tract. The possible branches are the piriform fossa[15], which is a pair of bilateral cavities in the hypopharynx, and the large volume under the tongue surface in the front cavity. Because we omitted the piriform fossa, and assumed no cross-mode propagation in the front cavity, this effect did not appear in the spectra of the synthesized tones.

Let us find the cause of the harmonic clustering in the range of 4-5 kHz. A similar clustering can be found in the recorded pressed monophonic tone [Fig. 6(e)]. On the other hand, no such clustering can be observed in the recorded normal monophonic tone [Fig. 6(f)]. It is therefore probable that the energy concentration at 4-5 kHz found in the Xöömij and pressed monophonic tones is due to the glottal flow waveform characterizing the 'pressed' timbre, rather than due to the resonance characteristics of the vocal tract.

The above comparison between the pressed and normal tones leads us to speculate as to the reason why the reproduction of the sound spectra of the Xöömij tones failed in the high frequency range. This is probably because the glottal waveform for Xöömij singing has spectral peaks in a specific frequency range, which can not be provided by the Rosenberg glottal source model. We believe that future studies should contain refinements of the source modeling to obtain a better agreement of the spectrum in the high frequency range. To estimate the actual glottal waveform for Xöömij singing, an inverse filtering method may be useful, such as the two-pass method[14] used to clarify the relation between the voice quality and the vocal fold vibratory pattern. A dynamical model of the vocal fold vibration, such as the two-mass model[16], may also be helpful in seeing how the glottal waveform is affected by the acoustic response of the vocal tract for Xöömij singing.

IV. CONCLUSION

To examine the sound production of biphonic singing, Xöömij, we measured 3-D shapes of the vocal tract, while our subject sang, using the MRI method. Transfer functions were calculated from the measured vocal tract shapes. It was found that the second formant frequency of each tract shape was close to the fundamental frequency of the melody pitch within an error of 36 cents. Tones were also reproduced from the transfer functions by synthesis based on the acoustic tube model. Two pitches could successfully be perceived when the synthesized sounds were listened to. In conclusion, we proposed a theory stating that the high melody pitch is produced by the pipe resonance of the rear cavity in the vocal tract.

This research clarifies the nature of the production mechanism of Xöömij, and this finding is verified through the synthesis of sounds actually having two pitches. It should be noted, however, that the synthesis was not at such a level that the timbre of the sound could be completely reproduced, and therefore, future studies should focus on the modeling of the glottal source.

The psycho-acoustical aspect of why a harmonic component emphasized by an extreme resonance of the vocal tract is segregated by the auditory system is another issue. Reference [17] addresses this problem.

ACKNOWLEDGMENTS

The authors are grateful to Yasuhiro Shimada and Ichiro Fujimoto of Takanohara Chuo Hospital for their helpful collaboration in the MRI measurements, to Shinobu Masaki for his useful suggestions on the image processing, to Jianwu Dang for his valuable discussions on the calculated transfer functions, and to Shinji Maeda of CNRS for letting the authors realize the importance of the glottal impedance.

APPENDIX: ACOUSTIC MODEL

Wave propagation in the vocal tract

We use the same values for the air constants as used in Ref.[10]. Some of them are dependent on the air temperature T. In this paper, we set $T = 25$ degrees Celsius for the vocal tract. The values of the constants are:

Sound of speed, $c = 331.45(1 + 0.0018T)$ m/s.

Air density, $\rho = 1.2929(1 - 0.0037T)$ kg/m^3.

Viscosity coefficient, $\mu = 1.708 \times 10^{-5}(1 + 0.0029T)$ kg/(s $\cdot$ m).

Thermal conductivity, $d = 5.77 \times 10^{-3}(1 + 0.0033T)$ cal/(m $\cdot$ s $\cdot ^\circ$ C).

Specific heat at constant pressure, $C_p = 240$ cal/kg $\cdot ^\circ$ C.

Specific heat ratio, $\gamma = C_p/C_v = 1.402$.

The yielding wall of the vocal tract can be modeled as a damped spring-mass system.[11] Three parameters characterizing the mechanical property of the vocal tract wall are:

Lowest angular resonance frequency of the tract when closed at both ends, $\omega_0 = 406/\pi$ rad/s, which is inversely proportional to the square root of the wall mass per unit length.

Ratio of the wall resistance to the mass, $a = 130\pi$ rad/s.

Squared angular frequency of the mechanical resonance, $b = (30\pi)^2$ (rad/s)^2.

The following symbols are also defined:

Angular frequency, ω rad/s.

Pipe radius, r.

Ratio of the pipe radius to the viscous boundary layer thickness, $r_v = r\sqrt{\omega\rho/\mu}$.

Ratio of the pipe radius to the thermal boundary layer thickness, $r_t = r\sqrt{\omega\rho C_p/\lambda}$.

The vocal tract can be considered as a succession of finite elements, each of which is a truncated cone. In each element, pressure $p(x)$ and particle velocity $u(x)$ satisfy the following one-dimensional wave equations:

$$\frac{\partial p}{\partial x}(x) = -Z_v\,u(x), \tag{A1}$$

$$\frac{\partial u}{\partial x}(x) = -(Y_t + Y_w)\,p(x) - \frac{2}{x}u(x), \tag{A2}$$

where x is the distance along the axis from the apex of the cone, Z_v is the series impedance per unit length including the viscous effect, Y_t is the shunt admittance per unit length including the thermal exchange, and Y_w is the shunt admittance per unit length due to the yielding wall. These are assumed to be constant over the length of the truncated cone, and modeled by

$$\begin{aligned} Z_v &= i\omega\rho\left(1+\frac{2}{r_v}(1-i)-\frac{3i}{r_v^2}\right), && \text{(A3)} \\ Y_t &= \frac{i\omega}{\rho c^2}\left[1+(\gamma-1)\left(\frac{\sqrt{2}}{r_t}(1-i)+\frac{i}{r_t^2}\right)\right], && \text{(A4)} \\ Y_w &= \frac{i\omega}{\rho c^2}\frac{\omega_0^2}{b+i\omega a-\omega^2}, && \text{(A5)} \end{aligned}$$

respectively. To calculate r_v and r_t in Eqs. (A3) and (A4), the pipe radius r is regarded as the arithmetic mean of the radii of the two ends. The propagation constant $\mathit{\Gamma}$ and the wave impedance ζ become

$$\mathit{\Gamma} = \sqrt{Z_v(Y_t+Y_w)}, \tag{A6}$$

$$\zeta = \sqrt{\frac{Y_t+Y_w}{Z_v}}, \tag{A7}$$

respectively. Integrating Eqs. (A1) and (A2) from one end at x_1 to the other at x_2, we have the following transmission matrix equation:

$$\begin{bmatrix} p_2 \\ u_2 \end{bmatrix} = L^{-1}(x_2)\, M\, L(x_1) \begin{bmatrix} p_1 \\ u_1 \end{bmatrix}, \tag{A8}$$

where

$$\begin{aligned} L(x) &= \begin{bmatrix} x & 0 \\ -\dfrac{1}{Z_v} & x \end{bmatrix}, && \text{(A9)} \\ M &= \begin{bmatrix} \cosh \mathit{\Gamma}(x_1-x_2) & \zeta \sinh \mathit{\Gamma}(x_1-x_2) \\ \dfrac{1}{\zeta}\sinh \mathit{\Gamma}(x_1-x_2) & \cosh \mathit{\Gamma}(x_1-x_2) \end{bmatrix}. && \text{(A10)} \end{aligned}$$

Multiplying the transmission matrices of all elements, we have an equation relating pressure p_{in} and particle velocity u_{in} at the input end of the vocal tract with pressure p_{out} and particle velocity u_{out} at the output end as follows:

$$\begin{bmatrix} p_{\text{in}} \\ u_{\text{in}} \end{bmatrix} = \widehat{M} \begin{bmatrix} p_{\text{out}} \\ u_{\text{out}} \end{bmatrix}, \tag{A11}$$

where

$$\widehat{M} = \prod_{\text{all elements}} L^{-1} M L \equiv \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}. \tag{A12}$$

The input and output volume velocities are defined by

$$U_{\text{in}} = A_{\text{in}} u_{\text{in}}, \quad U_{\text{out}} = A_{\text{out}} u_{\text{out}}, \tag{A13}$$

respectively, where A_{in} and A_{out} are the areas of the input and output ends, respectively.

Radiation load

The radiation at the output end, i.e., the mouth, can be modeled by the radiation from a round piston head surrounded by an infinite baffle. In the low frequency range up to a few kHz, where the wavelength is sufficiently larger than the length characterizing the size of the mouth, the radiation impedance Z_r is approximated by a lumped acoustic resistance R_r and an inductance L_r[18] as follows:

$$Z_r \equiv \frac{p_{\text{out}}}{U_{\text{out}}} = \frac{\zeta}{A_{\text{out}}} \frac{i\omega R_r L_r}{R_r + i\omega L_r}, \tag{A14}$$

where

$$R_r = \frac{128}{9\pi^2}, \quad L_r = \frac{8\sqrt{A_{\text{out}}}}{3\pi^{3/2} c}. \tag{A15}$$

By substituting Eqs. (A13) and (A14) into Eq. (A11), the volume velocity transfer function, pressure-to-velocity transfer function and input impedance of the vocal tract become

$$T_U \equiv \frac{U_{\text{out}}}{U_{\text{in}}} = \frac{A_{\text{out}}}{A_{\text{in}}} \frac{1}{m_{21} Z_r A_{\text{out}} + m_{22}}, \tag{A16}$$

$$T_{p/U} \equiv \frac{p_{\text{out}}}{U_{\text{in}}} = \frac{A_{\text{out}}}{A_{\text{in}}} \frac{Z_r}{m_{21} Z_r A_{\text{out}} + m_{22}}, \tag{A17}$$

$$Z_{\text{in}} \equiv \frac{p_{\text{in}}}{U_{\text{in}}} = \frac{1}{A_{\text{in}}} \frac{m_{11} Z_r A_{\text{out}} + m_{12}}{m_{21} Z_r A_{\text{out}} + m_{22}}, \tag{A18}$$

respectively.

Glottal impedance

In voice synthesis based on the acoustic tube model, a glottal volume velocity U_g is provided by a source model, e.g., the Rosenberg model. In a physical sense, U_g is a flow generated by a constant subglottal pressure p_0 when the supraglottal pressure is zero. Bernoulli's law yields

$$U_g = \sqrt{\frac{2p_0}{\rho}} A_g, \tag{A19}$$

where A_g is the glottal area made by the vocal folds. Note here that U_g is different from volume velocity U_{in} at the entrance of the vocal tract, because of the presence of p_{in}. To find the difference, we again use Bernoulli's law to obtain

$$U_{\text{in}} = \sqrt{\frac{2(p_0 - p_{\text{in}})}{\rho}} A_g. \tag{A20}$$

By considering small-amplitude oscillation, and by noting that U_{in}, U_g and A_g have both the time-averaged component and time-varying component, whereas p_{in} has only the time-varying component, we can linearize Eq. (A20) as follows:

$$U_{\text{in}} = U_g - \frac{p_{\text{in}}}{Z_g}, \tag{A21}$$

where Z_g is the glottal impedance defined by

$$Z_g = \frac{\sqrt{2p_0\rho}}{\bar{A}_g}, \tag{A22}$$

and where $\bar{A}_g$ is the average area of A_g. The second term on the r.h.s. of Eq. (A21) implies that a reverse flow toward upstream is generated, which is proportional to p_{in}. The ratio of p_{in} to the flow is equal to Z_g.

To estimate the value of Z_g, we assumed the typical p_0 and $\bar{A}_g$ to be

$$p_0 = 10\,\text{cmH}_2\text{O}, \quad \bar{A}_g = 0.5 \times 10^{-2} A_{\text{in}}, \tag{A23}$$

respectively.

From Eqs. (A16), (A18) and (A21), the volume velocity transfer function of the entire system including the vocal tract and the glottis becomes

$$\widehat{T}_U \equiv \frac{U_{\text{out}}}{U_g} = \frac{T_U}{1 + Z_{\text{in}}/Z_g}. \tag{A24}$$

Similarly, the pressure-to-velocity transfer function of the entire system becomes

$$\widehat{T}_{p/U} \equiv \frac{p_{\text{out}}}{U_g} = \frac{T_{p/U}}{1 + Z_{\text{in}}/Z_g}. \tag{A25}$$

In this paper, we calculated $\widehat{T}_U$ to discuss the acoustic properties of the vocal tract, and used the time-domain representation of $\widehat{T}_{p/U}$ for synthesis.

References

[1] B. Chernov and V. Maslov, "Larynx — double-sound generator," *Proc. 11th Int. Conf. of Phonetic Science,* Tallinn, Estonia 40–43 (1987).

[2] Q.H. Trân and D. Guillou, "Original research and acoustical analysis in connection with the Xöömij style of biphonic singing," In *Musical voices of Asia* (Heibonsha, Tokyo, 1980) pp. 162–173.

[3] T. Muraoka, K. Wagatsuma and M. Horiuchi, "Acoustic analysis of the Mongolian singing Xöömij," *Proc. Fall Meet. Acoust. Soc. Jpn.,* 385–386 (1983) (in Japanese).

[4] S. Adachi, S. Kinoshita, H. Tamagawa and M. Yamada, "MRI measurement of the vocal-tract shape while singing Xöömij and the synthesis based on the acoustic tube model," *Tech. Rep. Musical Acoustics* **MA 96-10** 9–16 (1996) (in Japanese).

[5] S. Adachi, S. Kinoshita, T. Komoike, H. Tamagawa and M. Yamada, "Study on sound production in Xöömij — Part 1: MRI measurement of the vocal-tract shape and the synthesis based on the acoustic tube model," *Proc. Spring Meet. Acoust. Soc. Jpn.,* 645–646 (1996) (in Japanese).

[6] T. Komoike, S. Kinoshita, M. Yamada, S. Adachi and I. Nakayama, "Study on sound production in Xöömij — Part 2: Perceptual experiment with synthesized sound," *Proc. Spring Meet. Acoust. Soc. Jpn.,* 647–648 (1996) (in Japanese).

[7] S. Adachi and M. Yamada, "An acoustical study of sound production in biphonic singing, Xöömij," *Proc. 1997 Japan-China Joint Meeting on Musical Acoustics,* 21–26 (1997).

[8] B.H. Story, I.R. Titze and E.A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am. **100** 537–554 (1996).

[9] J. Dang, K. Honda and H. Suzuki, "Morphological and acoustical analysis of the nasal and the paranasal cavities," J. Acoust. Soc. Am. **96** 2088–2100 (1994).

[10] R. Caussé, J. Kergomard and X. Lurton, "Input impedance of brass musical instruments — comparison between experiments and numerical models," J. Acoust. Soc. Am. **75** 241–254 (1984).

[11] M.M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," IEEE Trans. Acoust., Speech, Signal Processing **ASSP-35** 955–967 (1987).

[12] A.E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am. **49** 583–590 (1971).

[13] The synthesized tones can be heard on the World Wide Web at http://www.hip.atr.co.jp/~adachi/Xoomij/Sound/.

[14] D.G. Childers and C.K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," J. Acoust. Soc. Am. **90** 2394–2410 (1991).

[15] J. Dang and K. Honda, "Acoustic characteristics of the piriform fossa in models and humans," J. Acoust. Soc. Am. **101** 456–465 (1997).

[16] K. Ishizaka and J.L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," Bell Syst. Tech. J. **51**, 1233–1268 (1972).

[17] M. Yamada, "Stream segregation in Mongolian traditional singing, Xöömij," *Proc. Int. Sym. Musical Acoustics,* Dourdan, 540–545 (1995).

[18] J.L. Flanagan, *Speech Analysis, Synthesis and Perception,* 2nd Ed., (Springer-Verlag, New York, 1972) Chap. 3, pp. 36-38.

Figure Captions

Fig. 1. Mid-sagittal slices of the vocal tract during the singing of various Xöömij [(a) to (d)] and monophonic [(e) and (f)] tones. Each slice is reproduced from a 3-D image obtained by MRI measurement. All of the Xöömij tones have the same drone pitch of F3. Their melody pitches are (a) F6, (b) G6, (c) A6 and (d) C7, respectively. The ordinary monophonic tones of the vowel /ɑ/, whose pitch is F3, were sung with a pressed voice in (e), and normally in (f).

Fig. 2. Area functions of vocal tract shapes for Xöömij [(a) to (d)] and monophonic [(e) and (f)] tones.

Fig. 3. Transfer functions of vocal tract shapes for Xöömij [(a) to (d)] and monophonic [(e) and (f)] tones. The solid lines denote those of entire vocal tract shapes. The dashed lines in (a) to (d) represent those of vocal tract shapes without the front cavity, i.e., having only the rear cavity and the narrowing by the tongue.

Fig. 4. A glottal flow waveform (a) and a sound spectrum (b) assumed by the Rosenberg source model. The parameters are $T_s/T = 0.4$ and $T_p/T_n = 3.0$ in (a), and $T_s/T = 0.2$ and $T_p/T_n = 3.0$ in (b).

Fig. 5. Sound spectra of Xöömij [(a) to (d)] and monophonic [(e) and (f)] tones synthesized by the acoustic tube model. The 8th harmonic in (a), the 9th in (b), the 10th in (c) and the 12th in (d) are enhanced by the second formant resonances. These can each be heard as the melody pitch separated from the others composing a complex tone of the drone pitch.

Fig. 6. Sound spectra of Xöömij tones synthesized from vocal tract shapes without the front cavity. The same harmonic component in each of (a) to (d) as in Figure 5 are enhanced by the second formant resonances. Although the magnitudes of the components are reduced due to the lack of the front cavity resonance, each of these is still heard as the melody pitch.

Fig. 7. Sound spectra of Xöömij [(a) to (d)] and monophonic [(e) and (f)] tones actually produced by the subject singer. The largest harmonic components in the second formant are the 8th, 9th, 10th and 12th in (a), (b), (c) and (d), respectively. Clustering of the harmonics at 4-5 kHz is found for the Xöömij [(a) to (d)] tones and the monophonic pressed [(e)] tone. No such clustering is found for the monophonic normal [(f)] tone.

Table 1: Measured frequencies (Hz) of the drone and the melody pitches of Xöömij tones, and of monophonic singing voices.

	Xöömij tones				Monophonic tones	
	F6	G6	A6	C7	Pressed	Normal
Drone	168.6	170.4	169.2	169.8	166.0	165.0
Melody	1349	1530	1698	2042	—	—

Table 2: Equal interval (5.162 mm) area functions of vocal tract shapes for the Xöömij and monophonic tones. The data is listed in mm^2, and numbered from the input end to the mouth exit.

Data	F6	G6	A6	C7	Press	Norm
1	93	108	103	134	94	146
2	133	129	124	172	72	123
3	59	90	60	135	51	92
4	84	94	102	83	45	110
5	135	110	120	90	106	133
6	156	157	480	116	380	135
7	481	512	319	388	371	383
8	397	392	234	344	262	135
9	351	310	249	281	138	165
10	345	401	260	593	143	100
11	289	378	309	714	153	234
12	342	320	474	743	162	184
13	368	306	505	828	134	209
14	310	378	527	715	125	207
15	343	395	449	794	141	139
16	142	355	457	685	123	225
17	232	404	520	702	224	356
18	332	493	662	631	359	404
19	540	626	661	541	553	335
20	569	578	676	517	706	436
21	595	646	759	395	1127	414
22	684	639	612	256	1381	870
23	727	606	411	119	1386	780
24	598	429	211	50	1699	1005
25	418	196	49	30	1542	894
26	243	97	16	35	1289	1124
27	154	41	11	32	1133	1234
28	53	20	79	57	980	843
29	19	27	724	368	975	635
30	27	130	1059	875	859	628
31	80	1126	1149	1056	889	617
32	1030	1263	1165	931	856	498
33	1246	1160	887	1028	912	490
34	1001	744		966		
35	1143	767				

Table 3: First and second formant frequencies (F_1, F_2), and their bandwidths (BW_1, BW_2) of transfer functions calculated for (a) Xöömij and monophonic vocal tract shapes, plus (b) Xöömij tract shapes without the front cavity. These are listed in Hz.

(a)	Xöömij				Monophonic singing	
	F6	G6	A6	C7	Pressed	Normal
F_1	358.3	353.7	324.8	333.0	738.3	714.5
BW_1	42.8	44.0	48.0	45.3	57.4	63.4
F_2	1356	1562	1692	2066	1151	1271
BW_2	48.0	64.8	78.4	233.4	78.6	103.3

(b)	Xöömij, No front cavity			
	F6	G6	A6	C7
F_1	355.0	353.7	323.3	333.1
BW_1	42.4	43.8	47.6	44.8
F_2	1355	1563	1690	2052
BW_2	47.8	63.0	77.3	182.4

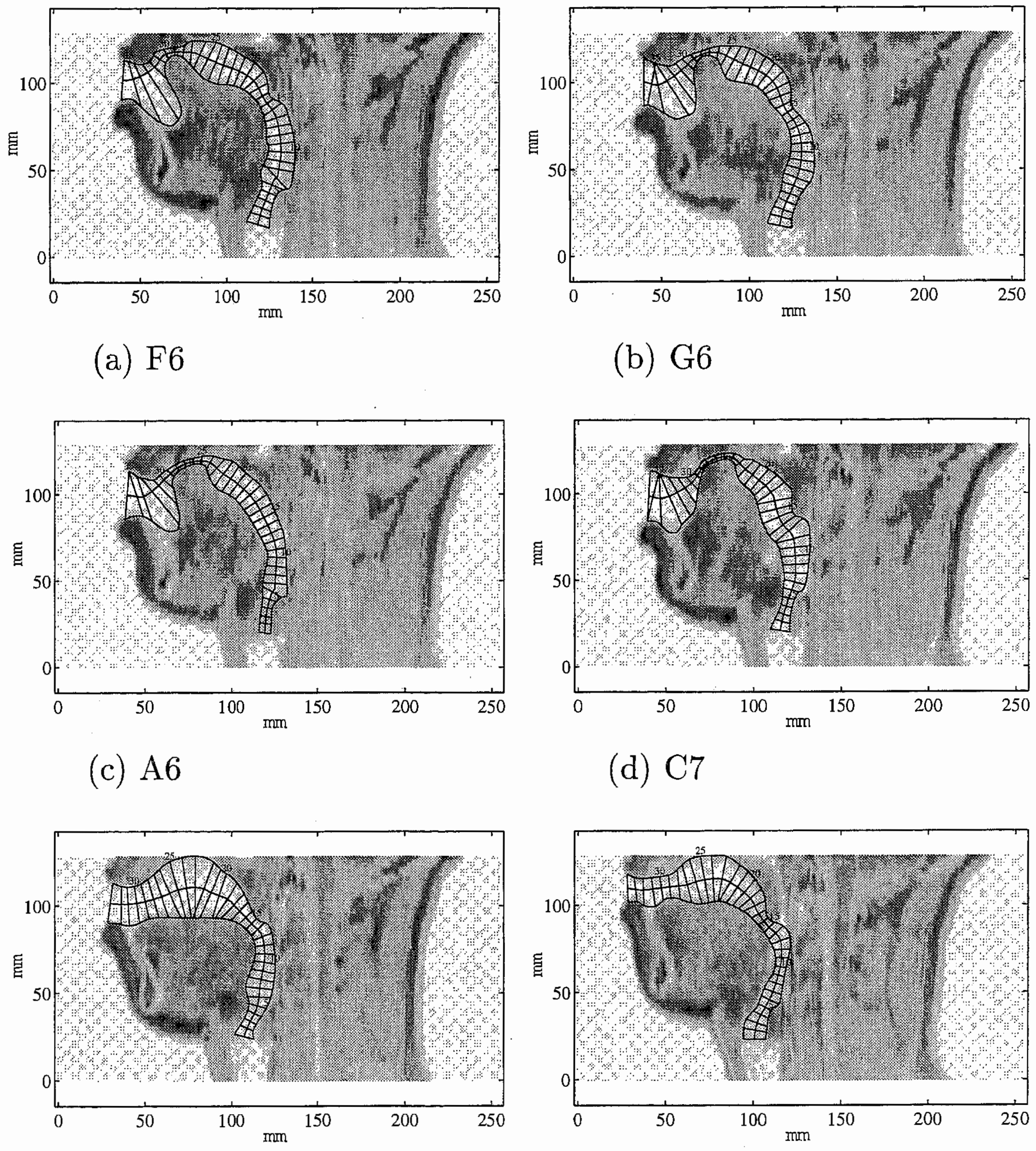
Table 4: Fundamental frequencies (F_0), first and second formant frequencies (F_1, F_2), and their bandwidths (BW_1, BW_2) of (a) synthesized Xöömij and monophonic tones, plus (b) Xöömij tones synthesized without the front cavity. These are listed in Hz.

(a)	Xöömij tones				Monophonic tones	
	F6	G6	A6	C7	Pressed	Normal
F_0	168.6	170.4	169.2	169.8	166.0	165.0
F_1	336.5	341.5	348.3	345.2	778.1	696.9
BW_1	25.8	42.7	39.7	35.5	253.1	257.7
F_2	1352	1529	1694	2033	1159	1301
BW_2	25.1	93.8	88.7	135.3	41.2	93.5

(b)	Xöömij tones, No front cavity			
	F6	G6	A6	C7
F_0	168.6	170.4	169.2	169.8
F_1	336.6	346.8	336.9	345.1
BW_1	24.9	41.9	40.7	37.3
F_2	1354	1529	1702	2029
BW_2	26.5	137.0	74.9	169.4

Table 5: F_0, F_1, F_2, BW_1 and BW_2 of recorded Xöömij and monophonic tones. Listed in Hz.

	Xöömij tones				Monophonic tones	
	F6	G6	A6	C7	Pressed	Normal
F_0	167.8	167.4	168.7	167.8	164.7	166.7
F_1	337.3	343.6	345.7	334.6	652.0	645.6
BW_1	54.5	27.0	57.7	48.9	90.4	106.2
F_2	1335	1497	1685	1979	1200	1158
BW_2	30.2	54.0	53.8	179.9	194.2	114.3



(a) F6

(b) G6

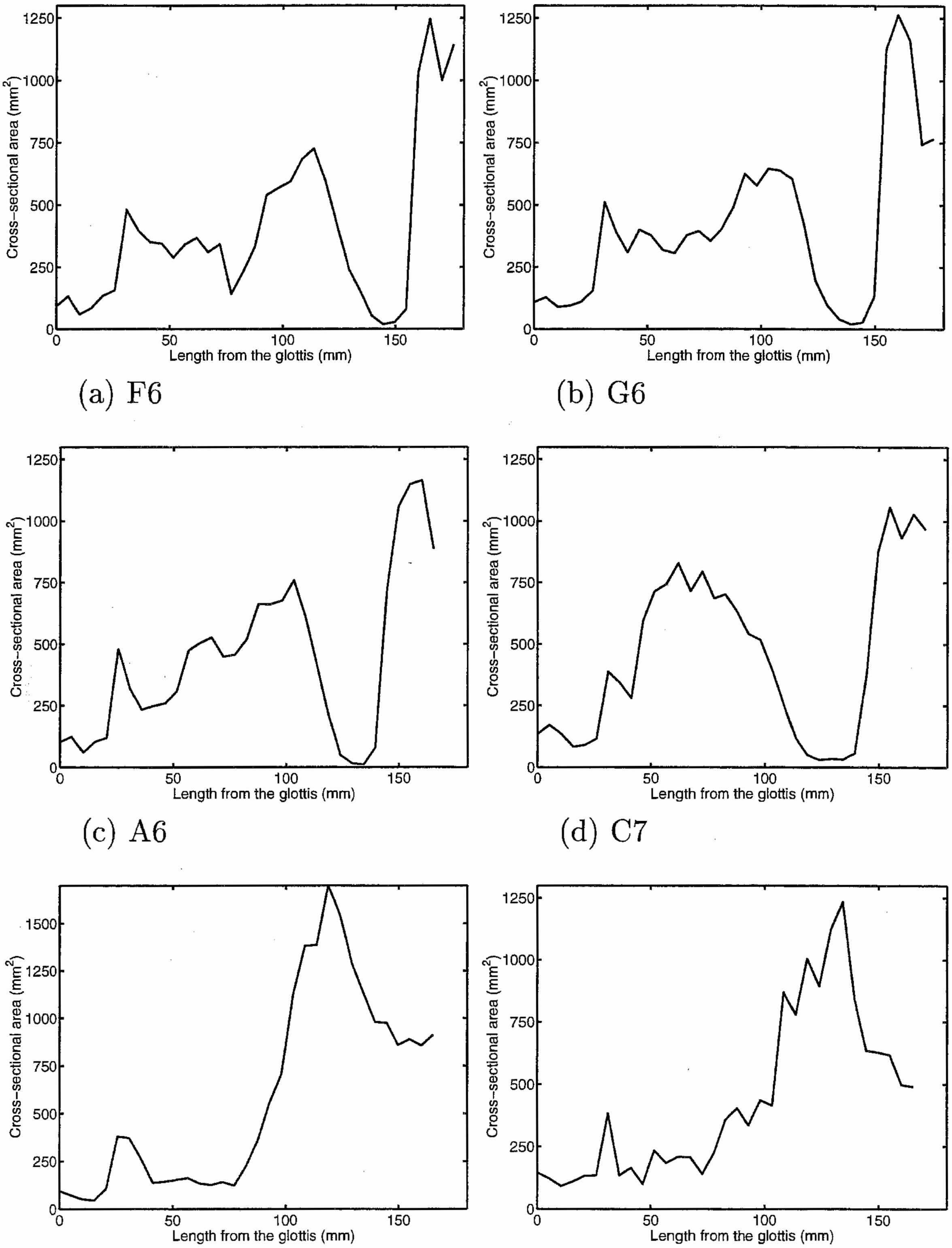
(c) A6

(d) C7

(e) ‘Pressed’ singing /ɑ/

(f) Normal singing /ɑ/

Figure 1:



(a) F6

(b) G6

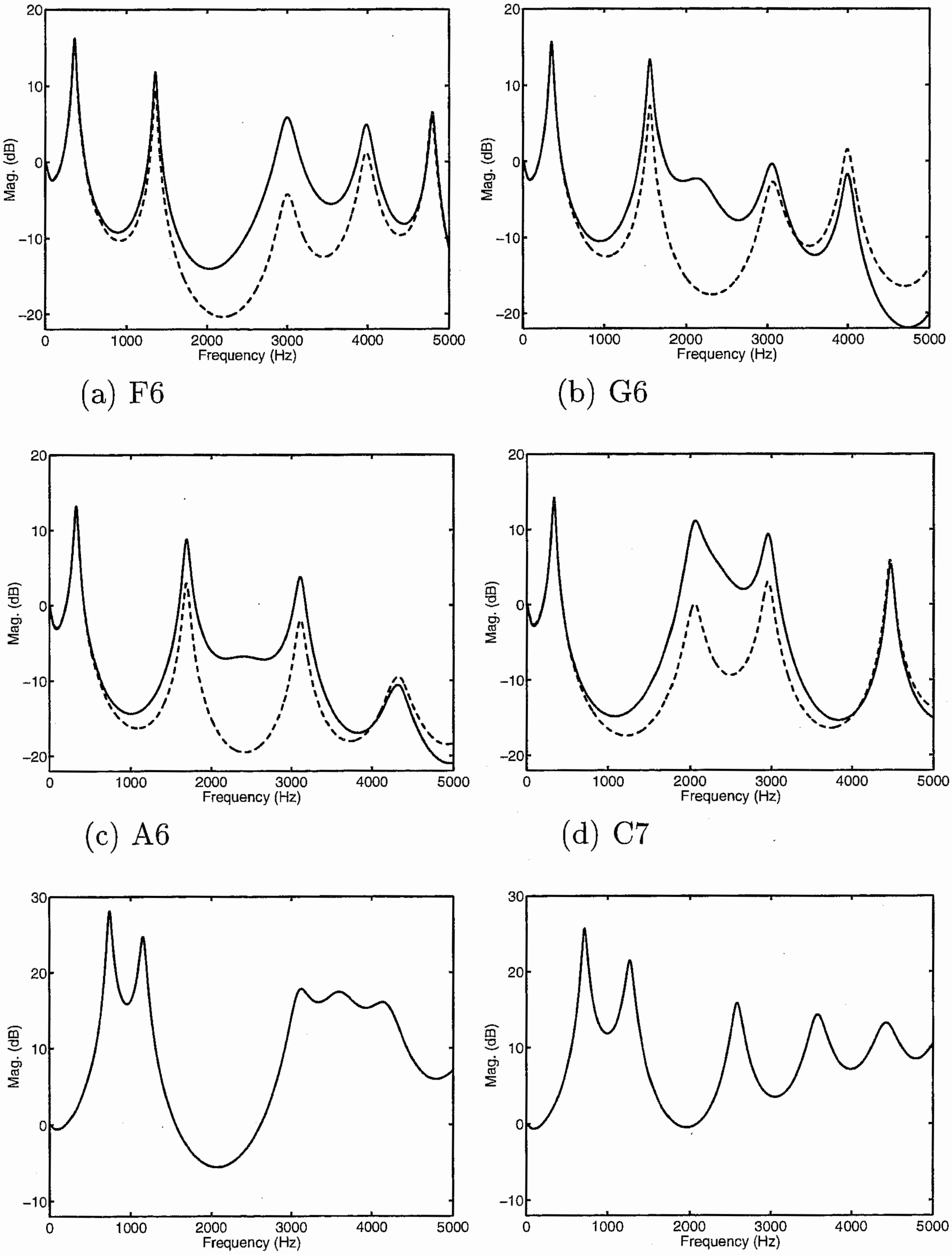
(c) A6

(d) C7

(e) 'Pressed' tone /ɑ/

(f) Normal tone /ɑ/

Figure 2:



(a) F6

(b) G6

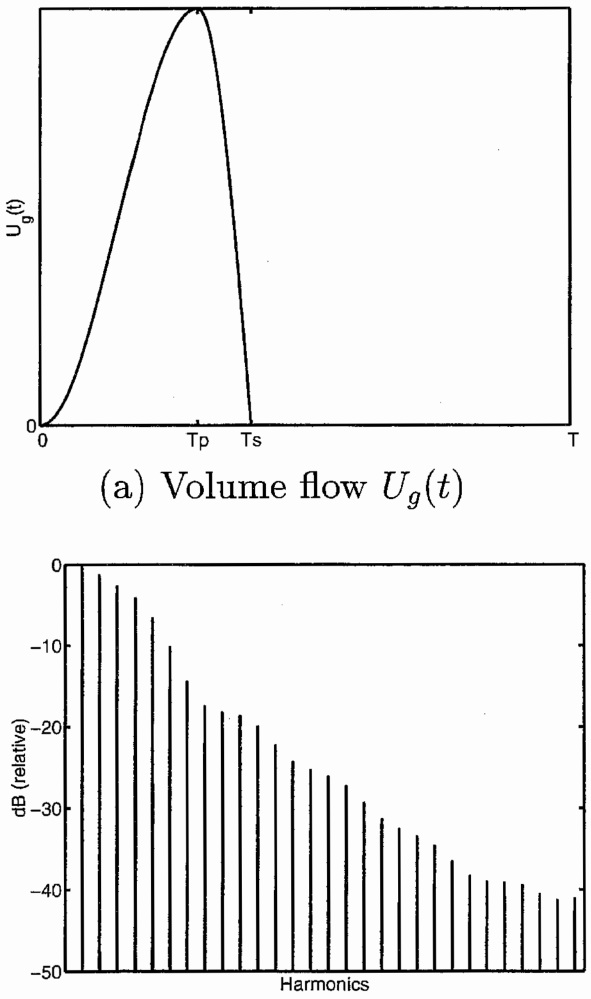
(c) A6

(d) C7

(e) 'Pressed' tone /ɑ/

(f) Normal tone /ɑ/

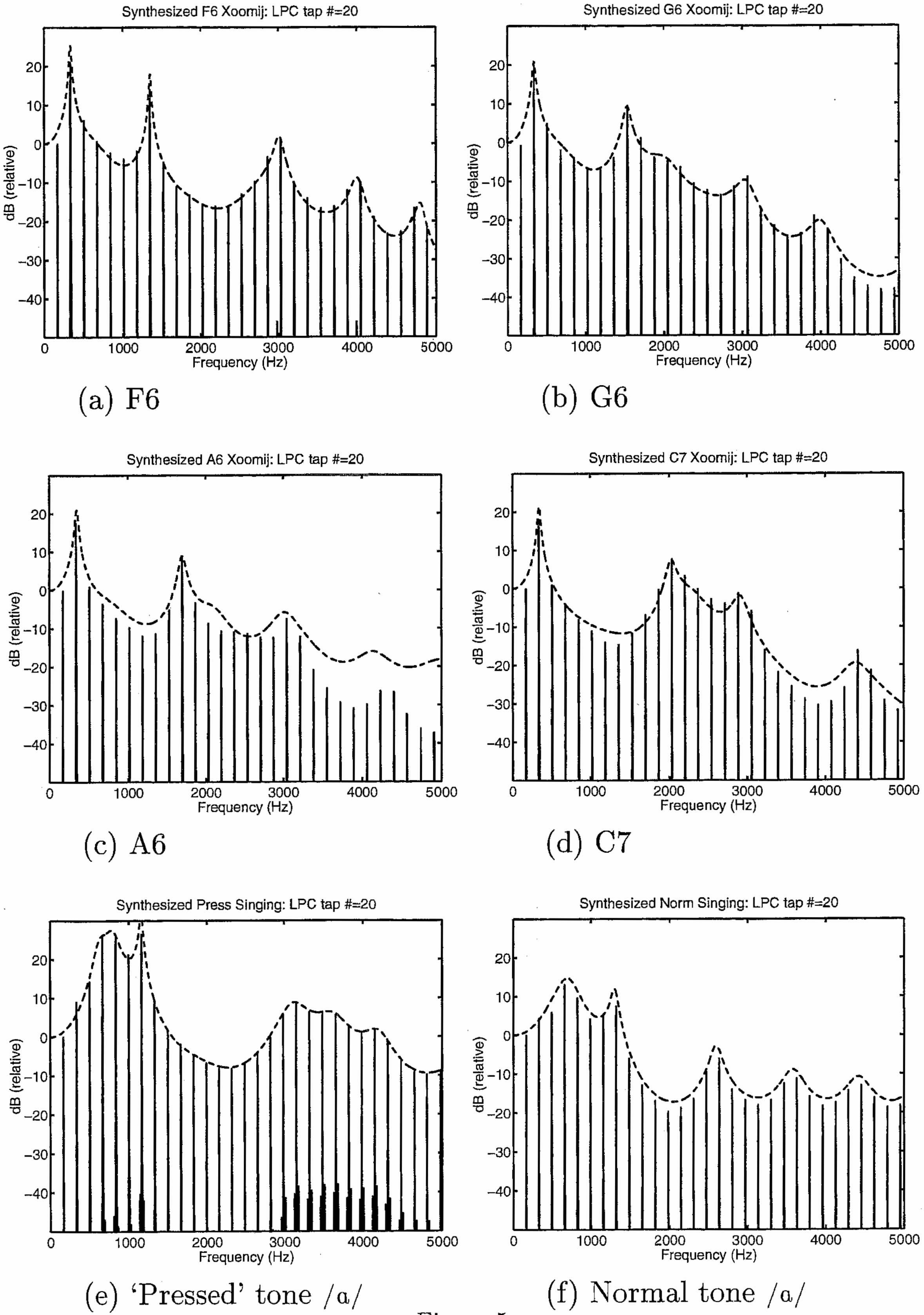
Figure 3:



(a) Volume flow $U_g(t)$

(b) Spectrum

Figure 4:



(a) F6

(b) G6

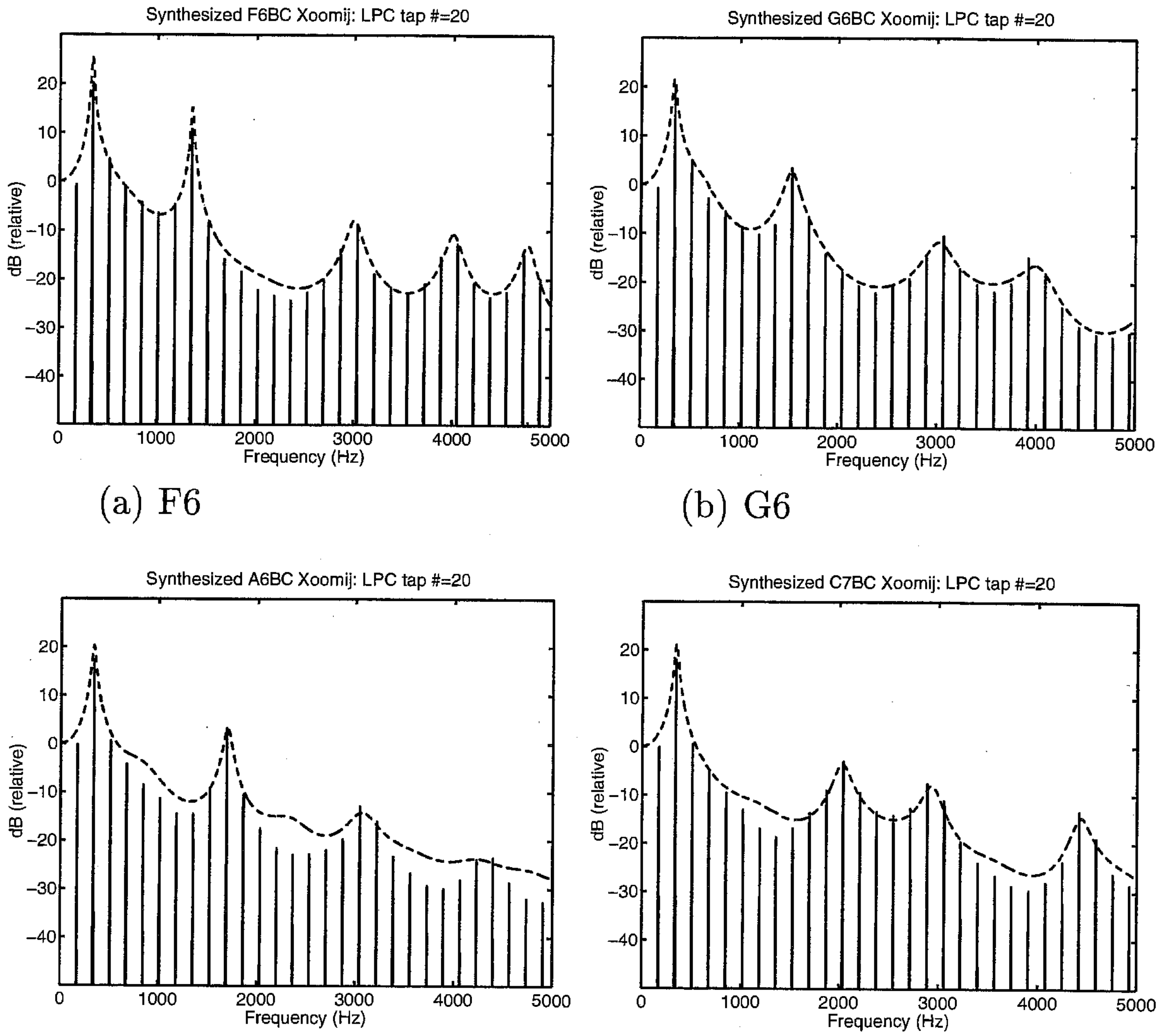
(c) A6

(d) C7

(e) 'Pressed' tone /ɑ/

(f) Normal tone /ɑ/

Figure 5:



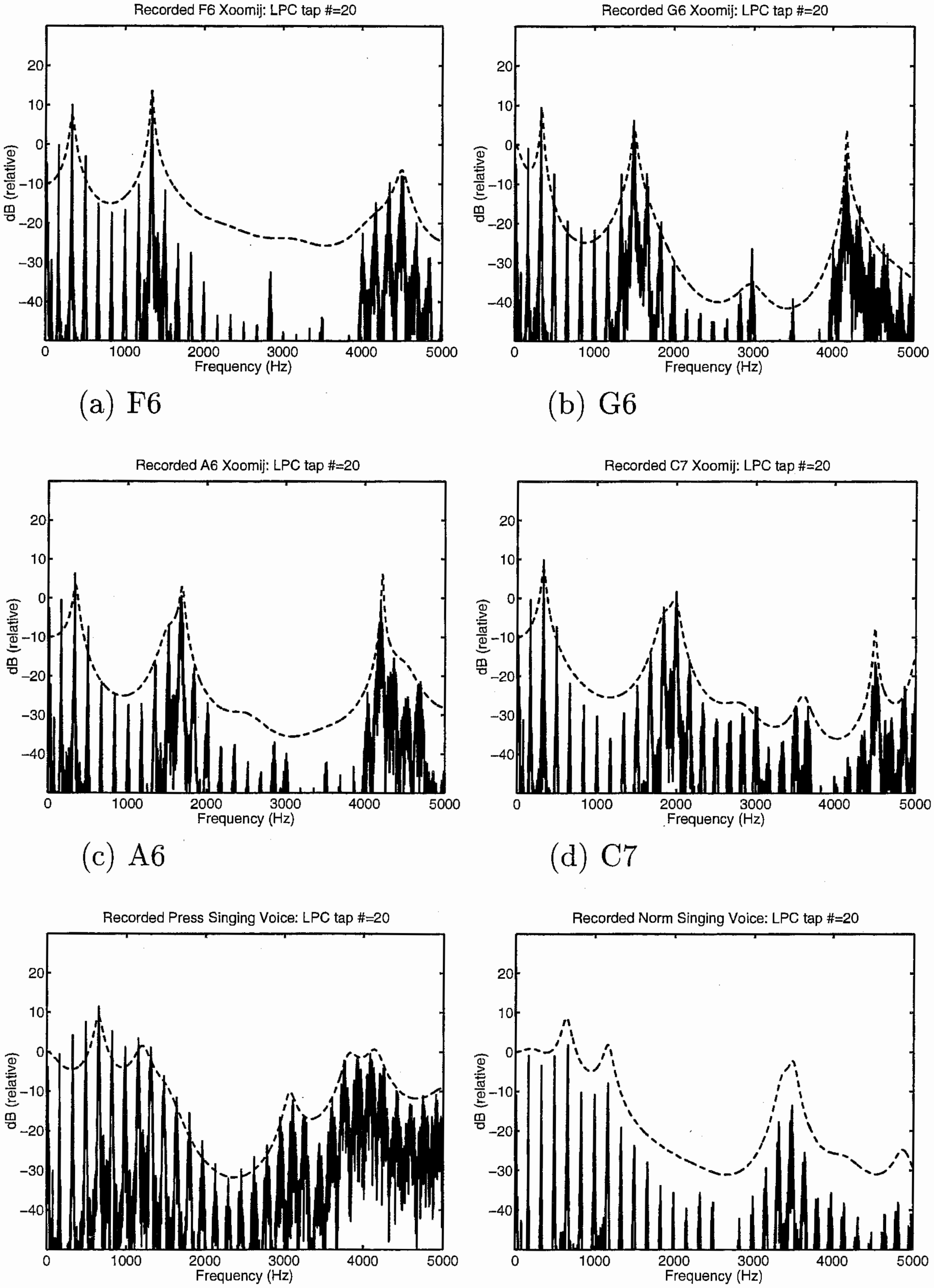
(a) F6

(b) G6

(c) A6

(d) C7

Figure 6:



Recorded F6 Xoomij: LPC tap #=20

(a) F6

Recorded G6 Xoomij: LPC tap #=20

(b) G6

Recorded A6 Xoomij: LPC tap #=20

(c) A6

Recorded C7 Xoomij: LPC tap #=20

(d) C7

Recorded Press Singing Voice: LPC tap #=20

(e) 'Pressed' tone /ɑ/

Recorded Norm Singing Voice: LPC tap #=20

(f) Normal tone /ɑ/

Figure 7: