2013 International Workshop on Nonlinear Circuits,
Communications and Signal Processing
NCSP'13, Israd of Hawaii, Hawaii, USA. March 4-7, 2013

A singing voices synthesis system to characterize vocal registers using ARX-LF model

Hiroki Motoda and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan
Phone/FAX:+81-761-51-1391/+81-761-51-1149
Email: {H.Motoda, akagi}@jaist.ac.jp

Abstract

This paper proposes a singing voices synthesis system to synthesize singing voices having characteristics of vocal registers, such as vocal fly, modal and falsetto. Human can sing songs naturally in wide range of frequency by training how to use vocal fold vibrations to represent vocal registers. However, even state-of-the-art singing voices synthesis systems cannot produce vocal registers appropriately. Naturalness of the synthesized singing voices using these systems is reduced in low and high frequency ranges. One of the methods for improving naturalness is adding characteristics of glottal sources for each vocal register. In this paper, the ARX-LF model that can formulate glottal sources for each vocal register by simulating human voice production mechanisms was applied. A model for controlling ARX-LF parameters corresponding to characteristics of glottal sources was constructed, and acoustic features corresponding to naturalness of singing voice were added. Singing voice data of each vocal register were analyzed by the ARX-LF model, and ARX-LF parameter values corresponding to glottal source of each vocal register were obtained. The control model was constructed using the results of the analysis. Singing voices were synthesized by the control model, and quality of the synthesized voices was evaluated. As the results, almost the same impressions were obtained from the synthesized singing voices as those from actual singing voices in each vocal register. Results revealed effectiveness of the proposed system for synthesizing singing voices to characterize vocal registers.

1. Introduction

Vocal register is a particular series of tones in the human voice that are produced by one particular vibratory pattern of the vocal folds and therefore possess a common quality. Human can sing songs naturally in wide range of frequency by training how to use vocal fold vibrations to represent vocal registers. Many reserchers investigated relationships between vocal registers and characteristics of glottal sources [1][2][3].

However, even state-of-the-art singing voices synthesis systems cannot produce vocal registers appropriately. Naturalness of the synthesized singing voices using these systems is reduced in low and high tone ranges. Characteristics of vocal registers in VOCALOID [4] depends on materials of stored singing data. Although SingBySpeaking [5] can synthesize natural singing voice in mid tone range using STRAIGHT [6], controlling of vocal quality corresponding to glottal sources is difficult in STRAIGHT. Thus, SingBySpeaking cannot produce vocal registers in high and low tone ranges appropriately. One of the methods for improving naturalness is adding characteristics of glottal sources for each vocal register. This paper considers a model that can represent glottal sources with characteristics of vocal registers for an advanced singing voice synthesis system. This paper proposes a singing voice synthesis system with the ARX-LF model [7] that can describe glottal sources for each vocal register by simulating human voice production mechanisms. A model for controlling ARX-LF parameters corresponding to characteristics of glottal sources was constructed. Singing voices were synthesized using the control model, and quality of the synthesized voices were evaluated.

2. Overview of the proposed system

2.1 ARX-LF model

Production mechanisms of human voices is simulated by an ARX-LF model as equation 1.

$$s(n) + \sum_{i=1}^{p} a_i(n)s(n-i) = b_0(n)u(n) + \varepsilon(n) \qquad (1)$$

$s(n)$ and $u(n)$ denote an observed speech signal and glottal source signal at time n, where $u(n)$ is approximated by LF glottal source model [8]. $a_i(n)$ and $b_0(n)$ are time-varying coefficients and $e(n)$ is residual. By applying the Z-transform onto the equation (assuming time invariance), one gets the following equation,

$$S(z) = \frac{1}{A(z)} \cdot U(z) + \frac{1}{A(z)} \cdot E(z) \qquad (2)$$

where $U(z), E(z)$ and $S(z)$ are Z-transform of glottal source, residual and voice. The shape of $u(n)$ is represented by fundamental periodno T_0, and four waveshape parameters T_p, T_e, T_a, and E_e as shown Figure 1. E_e is calculated by b_0. In order to simplify controlling the shape of $u(n)$, three parameters O_q, α_m, and Q_a are chosen as ARX-LF

parameters. O_q ($= T_e/T_0$) correnponds to the open quotient, α_m ($= T_p/T_e$) to the asymmetry coefficient, and Q_a ($= T_a/(1-O_q)T_0$) to the return phase quotient.

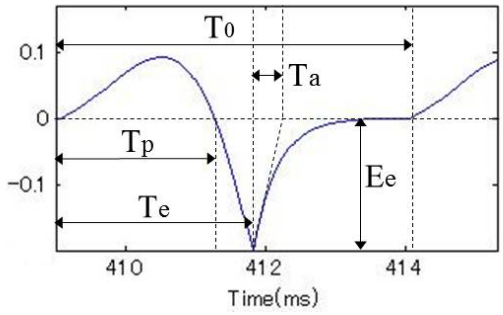

Figure 1: Glottal source signal approximated by LF model

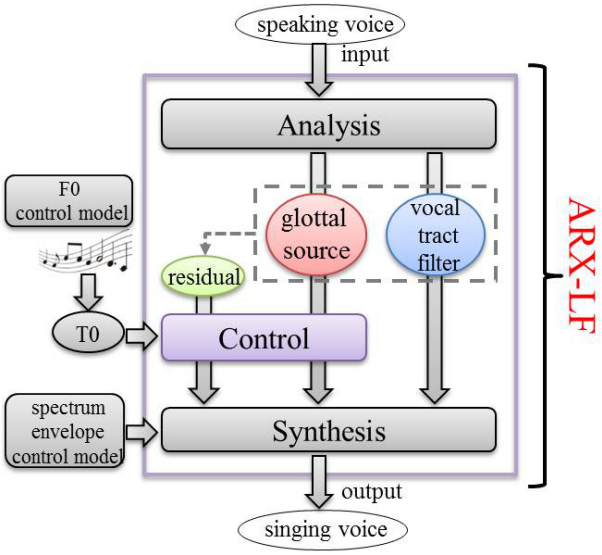

Figure 2: Block diagram of the proposed system

2.2 Procedures for synthesizing singing voices

Block diagram of the proposed system is shown in Figure 2. There are following five steps to synthesize singing voice.

1. Fundamental period(T0) is produced by a music score. T0 is obtained by the reciprocal of fundamental frequency(F0) generated by F0 model [5].

2. Glottal sources, vocal tract filters and residual are obtained in each period of speaking voices by the ARX-LF analysis.

3. Parameter values of the glottal sources are modified in accordance with T0 obtained in step1 by using a control model of the ARX-LF parameters.

4. Residual is modified in accordance with T0 obtained in step1.

5. Singing voice is synthesized using Equation 2 with vocal tract filter obtained in step2, modified glottal sourceus obtained in step3, and modified residual obtained in step4. Spectral envelope control model [5] is applied for improving naturalness of the synthesized singing voice.

Although literatures revealed relationships between vocal registers and vocal tract filter, for example [3], this paper focuses on charactaristics of glottal sources only. Vocal tract filter is not modified.

3. Model for controlling ARX-LF parameters

In this section, a model for controlling ARX-LF parameter values was proposed to represent characteristics of glottal sources for each vocal register. ARX-LF parameter values of each register were obtained by ARX-LF analysis, and the control model was constructed using the results of the analysis.

3.1 Analysis condition

Singing voice data were selected from the database "Nihongowo uta uta utau". Vowel /a/ was chosen from voiced sounds. In this study, three vocal registers, vocal fry, modal and falsetto were targeted. In order to obtain ARX-LF parameter values corresponding to glottal source of each vocal register, typical singing voices belonging to each register were selected. In this study, two parameters V_b and F_b are set to determine the boundaries of vocal registers voluntarily. V_b denotes the boundary between vocal fry and modal, and F_b denotes the boundary between modal and falsetto. V_b and F_b are set to 90 (Hz), 400 (Hz), respectively. Five samples of vocal fry and ten samples for each of modal and falsetto were chosen as the typical ones. Sampling frequency was 12 kHz.

3.2 Analysis results

Relationships between ARX-LF parameter values and characteristics of glottal sources for each vocal register were considered. Mean values of ARX-LF parameter values for each vocal register were shown in Table 1. It is known that the open quotient is larger in falsetto and smaller in vocal fry and typical O_q values were obtained; α_m is small in falsetto, because partial vibration of vocal fold in falsetto leads this result; and Q_a is large in falsetto, because turbulent flow generated by incomplete closure of vocal fold in falsetto produces this result. These results revealed that characteristics of glottal source for each vocal register can be represented by ARX-LF parameters.

Table 1: Mean parameter values of ARX-LF analysis of each vocal register

	O_q	α_m	Q_a
vocal fry	0.226	0.826	0.015
modal	0.434	0.824	0.025
falsetto	0.824	0.773	0.116

3.3 Model for controlling each ARX-LF parameter

Models for controlling ARX-LF parameters are proposed to synthesize singing voices having characteristics of vocal registers. Each ARX-LF parameter is modified in accordance with F0 of music score $F0_{-syn}$. In order to preserve speaker individuality, F0 of speaking data $F0_{-ori}$ and each ARX-LF parameter obtained by ARX-LF analysis O_{q-ori}, α_{m-ori} and Q_{a-ori} are used. ARX-LF parameters for each vocal register are interpolated in a linear manner. Scattered maps of parameter values are shown in Figures 3 to 5.

• **Control model for O_q**

O_q is widely varied among vocal registers as shown in Figure 3. In addition, there is a sligft tilt for each vocal register. In order to represent tilt, regression line of each vocal register were calculated by a least-squares method, as shown in Table 2. An O_q control model is constructed based on the regression line. A modified parameter value O_{q-syn} is calculated by the following equations.

$$O_{q-syn} = O_{q-ori} + y_{oq}(F0_{-syn}) - y_{oq}(F0_{-ori}) \quad (3)$$

$$y_{oq}(x) = a_{oq} \cdot log_2 x - b_{oq} \quad (4)$$

where a_{oq}, b_{oq} are decided by x, V_b and F_b value. If $x < V_b$, values of vocal fry is chosen, in Table 2. If $V_b \leq x < F_b$, values of modal is chosen. If $F_b \leq x$, values of falsetto is chosen.

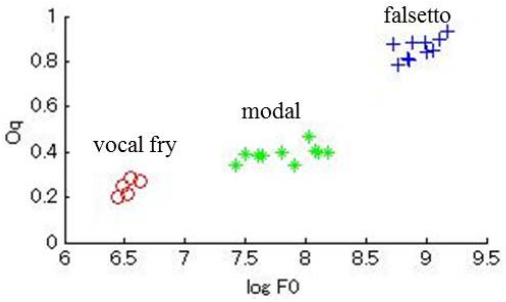

Figure 3: Distribution map of O_q of each data

Table 2: Tilt and intercept of O_q of each vocal register

	tilt a_{oq}	intercept b_{oq}
vocal fry	0.119	-0.529
modal	0.050	0.047
falsetto	0.207	-1.001

• **Control model for α_m**

In order to represent partial vibration of vocal fold in falsetto, the following control model for α_m is constructed. α_m is modified if $F0_{-syn}$ is high. α_m is widely varied in falsetto as shown in Figure 4. Pamameter α_r is set up to decide rate of change voluntarily. A modified parameter value α_{m-ori} is calculated by the following equations.

$$\alpha_{m-syn} = \begin{cases} \alpha_{m-ori} \cdot \alpha_r & (F_b \leq F0_{-syn}) \\ \alpha_{m-ori} & (F0_{-syn} < F_b) \end{cases} \quad (5)$$

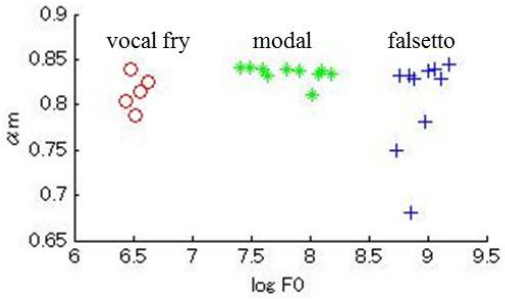

Figure 4: Distribution map of α_m of each data

• **Control model for Q_a**

In order to represent turbulent flows generated by incomplete closure of the vocal fold, a control model for Q_a is constructed. Q_a is large in falsetto as shown by Figure 5. In addition, there is a slight tilt in each of vocal register. As the same as O_q, the regression line of each vocal register was calculated, as shown in Table 3. A modified parameter value Q_{a-syn} is calculated by the following equations.

$$Q_{a-syn} = Q_{a-ori} + y_{qa}(F0_{-syn}) - y_{qa}(F0_{-ori}) \quad (6)$$

$$y_{qa}(x) = a_{qa} \cdot log_2 x - b_{qa} \quad (7)$$

where a_{qa}, b_{qa} are decided by x, V_b and F_b value. If $x < V_b$, values of vocal fry is chosen, in Table 3. If $V_b \leq x < F_b$, values of modal is chosen. If $F_b \leq x$, values of falsetto is chosen.

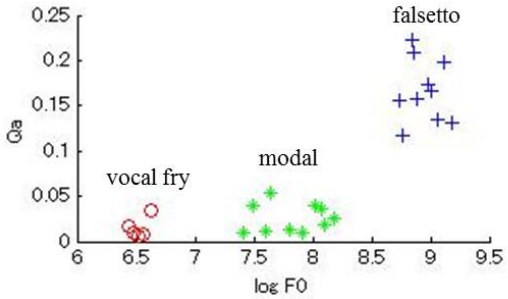

Figure 5: Distribution map of Q_a of each data

Table 3: Tilt and intercept of Q_a of each vocal register

	tilt a_{qa}	intercept b_{qa}
vocal fry	0.038	-0.232
modal	0.009	-0.004
falsetto	0.015	0.035

4. Evaluation of synthesized singing voice

In this section, quality of the synthesized singing voices in wide frequency ranges was evaluated. In order to evaluate reproducibility of falsetto and vocal fry, Experiment I and Experiment II were carried out.

Experiment I: synthesized singing voices in modal and falsetto registers were compaired.

Experiment II: synthesized singing voices in modal and vocal fry registers were compaired.

4.1 Synthesiszed singing voice

Six singing voices were synthesized for each experiment. In Experiment I, singing voices in modal register as Mf1, Mf2, Mf3 and those in falsetto register as F1, F2, F3 were synthesized. F0s were set to 247 Hz(B3), 277Hz(C4 ♯), 294 Hz(D4), 330 Hz(E4), 370 Hz(F4 ♯), 440 Hz(A4), respectively. In Experiment II, singing voices in modal as Mv1, Mv2, Mv3 and those in vocal fry as V1, V2, V3 were synthesized. F0s were set to 130 Hz(C3), 123Hz(B2), 110 Hz(A2), 87 Hz(F2), 82 Hz(E2), 73 Hz(D2), respectively, and α_r = 0.9, V_b = 100, and F_b = 310.

4.2 Experimental condition

The procedure for the Experiments was the Scheffe's Paired Comparison. Seven male listeners were participated in the experiments. The number of trials was 30 (=6 $\times$ 5) in each experiment. Degrees of "Breathy" that is typical impression of falsetto was evaluated in Experiment I. Degrees of "Rough" that is typical impression of vocal fry was evaluated in Experiment II. In order to simplify the decision task, "Breathy" voices and "Rough" voices were learned using actual singing voices in advance.

4.3 Results and Discussions

Population parameter σ was estimated from results of each Experiment using Ura Variation. σ of Experiment I is shown in Table 4. If σ is large positive value, singing voice was perceived as "Breathier". σ is widely varied between modal and falsetto. This result ensured that characteristics of falsetto can be represented by the proposed system. In addition, σ is larger if tone ranges are higher. This indicates that adding characteristics of glottal sources for each vocal register is important in wide tone ranges.

In Experiment II, similar results are obtained. σ of Experiment II is shown in Table 5. If σ is larger positive value, the singing voice was perceived as "Rougher". This result suggested that characteristics of vocal fry can be represented by the proposed system. As the results, reproducibility of vocal registers by the proposed system were ensured.

Table 4: Results of Experiment 1

	Mf1	Mf2	Mf3	F1	F2	F3
σ	-1.35	-1.29	-1.01	0.88	1.09	1.67

Table 5: Results of Experiment 2

	Mv1	Mv2	Mv3	V1	V2	V3
σ	-1.51	-1.41	-1.05	0.76	1.36	1.86

5. Conclusions

This paper proposed a singing voices synthesis system with the ARX-LF model to synthesize singing voices having characteristics of vocal registers. A model for controlling ARX-LF parameters was constructed to represent charactaristics of glottal sources for each vocal register. Quality of the synthesized singing voices in wide frequency ranges was evaluated. Results revealed that characteristics of vocal registers can be represented appropriately by the proposed system.

In this study, boundaries of vocal registers were not analyzed, and vocal tract filter was not modified. In the future work, we investigate transitions of ARX-LF parameters boundaries of vocal registers, and construct a model for controlling vocal tract filter.

References

[1] Roubeau, B., Henrich, N., Castellengo, M,. "Laryngeal vibratory mechanisms: The notion of vocal register revisited," Journal of Voice, 23(4), 425-438, 2009.

[2] Titze, I.R., "Principles of Voice Production," Allyn & Bacon, 1994. References.

[3] Tokuda, I., Zemke, M. kob, M., Herzel, H., "Biomechanical Modeling of Register Transitions and the Role of Vocal Tract Resonators," Journal of Acoustic Society of America 127(3), 1528-1536, 2010.

[4] Kenmochi, H., Ohshita, H., "VOCALOID — Commercial singing synthesizer based on sampleconcatenation," INTERSPEECH, 4011-4010, 2007.

[5] Saitou, T., Goto, M., Unoki, M., Akagi, M., "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," WASPAA, 215-218, 2007.

[6] Kawahara, H., "STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," Acoustic Science and Technology, 27(6), 349-353, 2006.

[7] Vincent, D., Rosec, O., Chonavel, T., "Estimation of LF glottal source parameters based on arx model," INTERSPEECH, 333-336, 2005.

[8] Fant, G., Liljencrants, J., Lin, Q., " A four-parameter model of glottal flow, "STL-QPSR, 85(2), 1-13, 1985.