A Talking Robot and Its Autonomous Learning of Speech Articulation for Producing Expressive Speech

Hideyuki Sawada

Kagawa University, Takamatsu-city, Kagawa, Japan
sawada@eng.kagawa-u.ac.jp

Abstract. The author is developing a talking robot by reconstructing a human vocal system mechanically based on the physical model of human vocal organs. The robotic system consists of motor-controlled vocal organs such as vocal cords, a vocal tract and a nasal cavity to generate a natural voice imitating a human vocalization. By applying the technique of the mechanical construction and its adaptive control, the robot is able to autonomously reproduce a human-like vocal articulation using its vocal organs. In vocalization, the vibration of vocal cords generates a source sound, and then the sound wave is led to a vocal tract, which works as a resonance filter to determine the spectrum envelope. For the autonomous acquisition of the robot's vocalization skills, an adaptive learning using an auditory feed-back control is introduced. In this manuscript, a human-like expressive speech production by the talking robot is presented. The construction of the talking robot and the autonomous acquisition of the vocal articulation are firstly introduced, and then the acquired control methods for producing human-like speech with various expressions will be described.

1 Introduction

Auditory and speech functions play an important role in the human communication, since vocal sounds instantly and directly transmit information using our vocal and auditory organs. In vocal communication, we present verbal information using words and phrases, and also emotional expressions in voices are simultaneously transmitted to listeners, so that smooth and flexible communication would be established among speakers and listeners. Various vocal sounds are generated by the complex movements of speaker's vocal organs, and this mechanism contributes to generate different voices that include speech expressions and speaker's individuality. By realizing the mechanisms of human speech and auditory systems using mechanical systems and computers, a new innovative robotic system will be introduced, which are utilized in the flexible speech communication with humans.

Various techniques have been reported in the research of human speech productions. For example, in the production of human voices, algorithmic syntheses historically have taken the place of analogue circuit syntheses, and became widely used techniques [1]-[4]. Sampling methods and physical model based syntheses are typical techniques these days, which are expected to provide realistic vocal sounds. In

© Springer International Publishing Switzerland 2015

P. Sinčák et al. (eds.), *Emergent Trends in Robotics and Intelligent Systems*,
Advances in Intelligent Systems and Computing 316, DOI: 10.1007/978-3-319-10783-7_10

addition to these algorithmic synthesis techniques, a mechanical approach using a phonetic or vocal model imitating the human vocal system would be valuable and notable solutions.

Vocal sounds are generated by the relevant operations of the vocal organs such as a lung, trachea, vocal cords, vocal tract, tongue and muscles. The airflow from the lung causes a vocal cord vibration to generate a source sound, and the glottal wave is led to the vocal tract, which works as a sound filter as to form the spectrum envelope of a particular voice. The voice is at the same time transmitted to the auditory system so that the vocal system is controlled for the stable vocalization. Different vocal sounds are generated by the complex movements of vocal organs under the feedback control mechanisms using an auditory system.

The articulation of vocal organs for the appropriate vocalization is acquired by repeating trials and errors of hearing and vocalizing vocal sounds in the age of infants. If any part of the vocal organs or the auditory system is injured or disabled in the age of infants, we might be involved in the impediment in the vocalization. Even in an adult age, we can learn new vocalization skills such as the mimicry of speech of other people and the utterance of foreign words that include different sounds from the mother language.

Mechanical constructions of a human vocal system to realize human-like speech have been reported so far [5]-[7]. In most of the researches, however, the mechanical reproductions of the human vocal system were mainly directed by referring to X-ray images and FEM analysis, and the adaptive acquisition of control methods for natural vocalization have not been considered so far. In fact, since the behaviors of vocal organs have not been sufficiently investigated due to the nonlinear factors of fluid dynamics yet to be overcome, the control of mechanical system has often the difficulties to be established.

The authors are developing a talking robot by reproducing a human vocal system mechanically [8],[9]. An adaptive learning using an auditory feedback control for the acquisition of vocalizing skill is introduced. The fundamental frequency and the spectrum envelope determine the principal characteristics of a sound. The former is the characteristics of a source sound generated by a vibrating object, and the latter is operated by the work of the resonance effects. In vocalization, the vibration of vocal cords generates a source sound, and then the sound wave is led to a vocal tract, which works as a resonance filter to determine the spectrum envelope.

The robot consists of motor-controlled vocal organs such as vocal cords, a vocal tract and a nasal cavity to generate a natural voice imitating a human vocalization. By introducing the auditory feedback learning with an adaptive control algorithm of pitch and phoneme, the robot is able to autonomously acquire the control skill of its mechanical system to vocalize stable vocal sounds imitating human speech. After the learning, the relations between articulatory motions and their produced sounds are established in the robot brain, which means that by listening to a certain vocal sound, the robot estimates its articulatory motion to autonomously generate vocal sounds.

In this study, we try to realize a talking robot that mimics the human-like expressive speech to establish a speech communication with a human. The speech expression is important for the smooth speech communication to transmit emotions to human listeners, and the robotic speech with human-like expressions would realize flexible vocal communication. In the first part of the paper, the structure and the adaptive control method of mechanical vocal cords and vocal tract are briefly described, and then the control method to reproduce human-like speech with various expressions is presented. In this study, the robot recognizes speech expressions given by a human speaker, and reproduces the expression in the robotic speech by articulating its mechanical vocal systems.

2 Construction of a Talking Robot

The talking robot mainly consists of an air pump, artificial vocal cords, a resonance tube, a nasal cavity, and a microphone connected to a sound analyzer, which respectively correspond to a lung, vocal cords, a vocal tract, a nasal cavity and an auditory system of a human. The construction and the overview of the talking robot are shown in Figure 1.

An air from the pump is led to the vocal cords via an airflow control valve, which works for the control of the voice volume. The resonance tube as a vocal tract is attached to the vocal cords for the manipulation of resonance characteristics. The nasal cavity is connected to the resonance tube with a rotary valve settled between them for the control of nasal sounds. The sound analyzer plays a role of the auditory system. It realizes the pitch extraction and the analysis of resonance characteristics of generated sounds in real time, which are necessary for the auditory feedback learning and control. The system controller manages the whole system by listening to the vocalized sounds and calculating motor control commands, based on the auditory feedback control mechanism employing neural network learning. The relation between the voice characteristics and mo-tor control commands are stored in the system controller, which are referred to in the generation of speech articulatory motion.

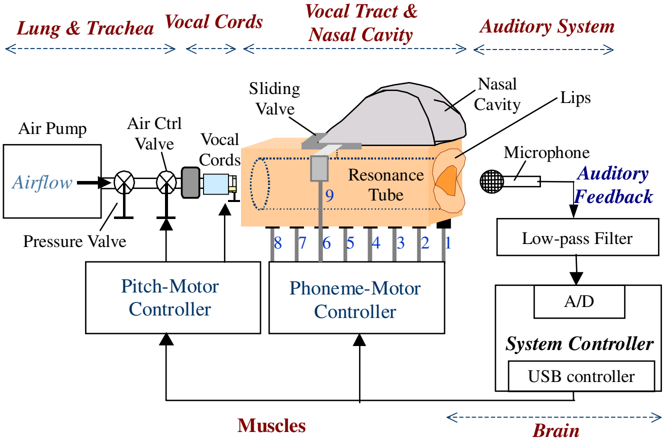


Fig. 1. System configuration of the talking robot

The characteristics of a glottal wave which determines the pitch and the volume of human voice are governed by the complex behavior of the vocal cords. It is due to the oscillatory mechanism of human organs consisting of the mucous membrane and muscles excited by the airflow from the lung. We employed an artificial vocal cord used by people who had to remove their vocal cords because of a glottal disease. The vibration of a rubber with the width of 5 mm attached over a plastic body makes vocal sound source. We measured the relationship between the tensile force and the fundamental frequency of a vocal sound gener-ated by the artificial vocal cord, and found that the fundamental frequency varied from 110 Hz to 350 Hz by the manipulations of a force applying to the rubber. The relation between the produced frequency and the applied force was not stable but tended to change with the repetition of experiments due to the fluid dynamics. The artificial vocal cord is, however, considered to be suitable for our system not only because of its simple structure, but also its frequency characteristics to be easily controlled by the tension of the rubber and the amount of airflow. The instability of the vibration would be compensated by employing an auditory-feedback control as humans do in daily speech and singing songs.

3 Learning of Vocal Articulations

In this study, we introduce an adaptive learning algorithm of the robotic voice articulations for achieving a talking and singing performance of the robot. The algorithm consists of two phases. First in the learning phase, the system acquires two maps in which the relations between the motor control values and the char-acteristics of generated voices are described. One is a motor-pitch map, which associates motor control values with fundamental frequencies. It is acquired by comparing the pitches of generated sounds with the desired pitches included in speaking phrases. The other is a motor-phoneme map, which associates the con-trol values of vocal tract motors with the phonetic characteristics of words to be generated. Then in the performance phase, the robot gives a speaking perfor-mance by referring to the obtained maps while pitches and phonemes of produced voices are adaptively maintained by hearing its own output voices.

A three-dimensional Self-Organizing Map (3D-SOM) is employed to autonomously associate vocal tract shapes with generated vocal sounds. The associated relations will enable the robot to estimate the articulation of the vocal tract for generating particular vocal sounds even if they are unknown vocal sounds, owing to the inference ability of the SOM.

The 3D-SOM has three-dimensional mapping space, and the characteristics could be located three-dimensionally, so the probability of miss location could be decreased. In this study, we employ two 3D-SOMs, one for constructing the topo-logical relation among the control commands and the other for establishing the relations of the phonetic characteristics [8-9]. After the learning of two 3D-SOMs, two 3D-SOMs will be associated with each other, based on the topological relations of motor control commands with phonetic characteristics. We call this algorithm a dual-SOM. The structure of the dual-SOM is shown in Figure 2, which consists of two self-organizing maps that arrange the mapping cells three dimensionally. One is a 3D-Motor_SOM that describes the topological relations of various shapes of the vocal tracts, in which close shapes are arranged in close locations with each other, and the other is

3D-Phonetic_SOM, which learns the relations among phonetic characteristics of generated voices. The talking robot generates various voices by changing its own vocal tract shapes. Generated voices and vocal tract shapes have the physical correspondence, since different voices are produced by the resonance phenomenon of the articulated vocal tract. This means that similar phonetic characteristics are generated by similar vocal tract shapes.

By adaptively associating the 3D-Phonetic_SOM with the corresponding 3D-Motor_SOM, we could expect that the talking robot autonomously learns the vocalization by articulating its vocal tract. After the learning of the relationship between the 3D-Phonetic_SOM and the 3D-Motor_SOM, we inputted human voices from microphone to confirm whether the robot could speak by mimicking human vocalization. Figure 3 shows the spectra of vocalized Japanese vowels /a/ and /i/. The first and second formants, which presented the characteristics of different vowel sounds, were formed as to approximate the human vowels, and the sounds were well distinguishable by human listeners. The robot also successfully acquired the other Japanese vowels /u/, /e/ and /o/ and nasal sounds, and the first and second formants were formed as to appear in vowels vocalized by a human. Figure 4, on the other hand, presents the obtained vocal tract shapes for /a/ and /i/ vocalization given by the robot. We verified that all the shapes given by the robot reproduced the actual human vocal tract shapes by the comparison with MR images.

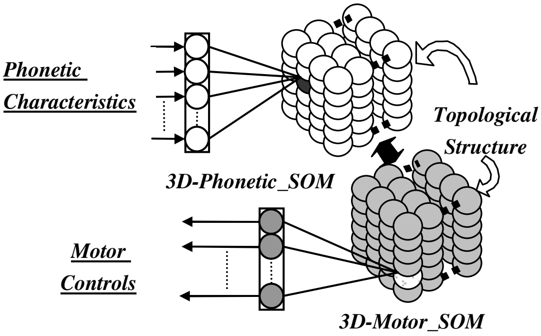
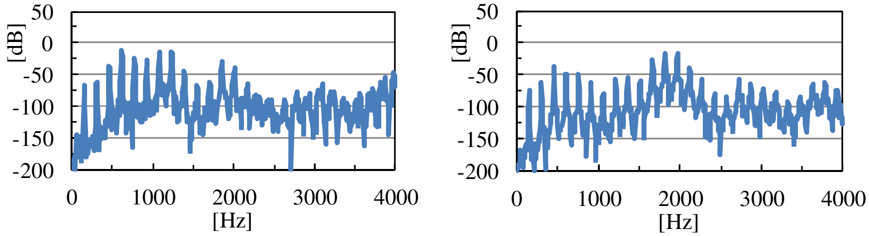
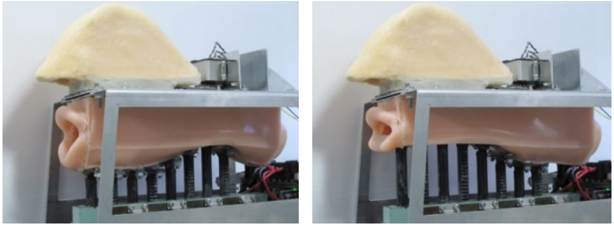


Fig. 2. Structure of the dual-SOM



(a) Robot's /a/ vowel (b) Robot's /i/ vowel

Fig. 3. Spectra of generated vowels



(a) /a/ vocalization (b) /i/ vocalization

Fig. 4. Two different vocal tract shapes given by the talking robot

4 Mimicry of Human Expressive Speech

A human generates speech by controlling their own vocal organs not only for just uttering words, but for changing the speech expressions such as the volume, speech speed and the intonation. The speech expression is important in the speech communication to transmit emotions to listeners. These expressions generated by the articulation of vocal organs also present the individuality of a speaker, and a listener instantly recognizes who is talking to, when he just receives a phone call and hears the first voice. This means that a voice transmits the individuality of the speaker, and the speech expression is an important factor for the smooth vocal communication.

To realize the human-like speech expressions employing the talking robot, we firstly studied human speech by paying attention to physical factors, which are the voice volume, pitch and speech speed.

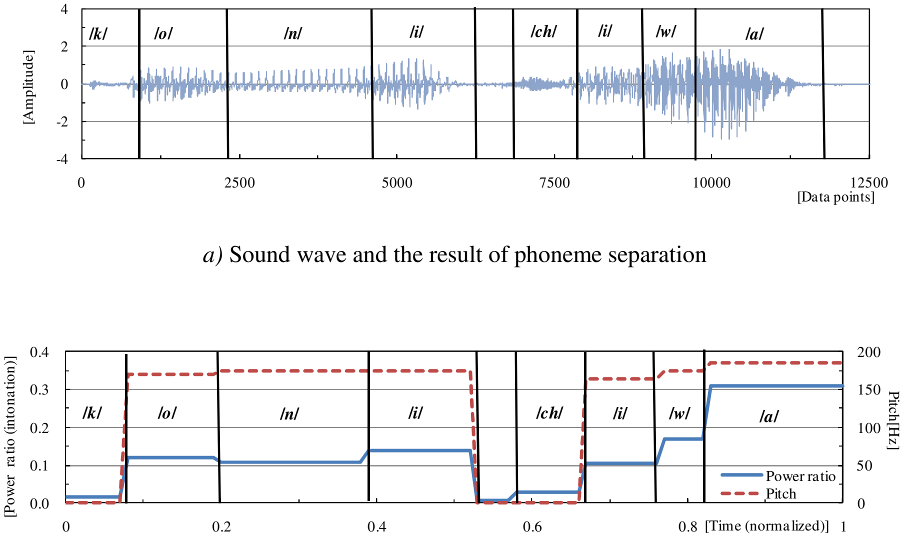
4.1 Human Speech

The three physical factors of speech expressions are extracted from human speech. We recorded a human speech voice in a PC, and divided the speech signal into phonemes to find the boundaries of each vowel and consonants, then calculated the volume, pitch and speech speed for each phoneme. A template matching method was employed to find the boundaries of each phoneme. The templates were obtained by calculating the sound parameters of human vocaliza-tion. Tenth-order partial autocorrelation (PARCOR) coefficients were employed as sound parameters in this study, and five Japanese vowels and nasal sound /n/ were selected as templates for the matching of each phoneme.

The template matching was executed to find the phoneme boundaries in human speech. For the matching, the window frame of 64 [msec] was settled, and PARCOR coefficients were calculated. The Euclidean distances between templates and calculated sound parameters were obtained, and the one which marked the smallest value among the templates were selected as the matching result.

The vowels and consonants are distinguished by referring to the average power of each window frame. If the calculated power is less than the threshold, the phoneme is defined as a consonant. On the other hand, if the power is greater than the threshold, the phoneme is defined as a vowel or a nasal sound, and the template matching was executed to recognize the vowel name. Figure 5 shows the analysis results of a speech /kon-nichi-wa/, a Japanese greeting, given by a human. The sampling frequency was set to 8 [kHz] in this study. The abscissa shows the sampling data points, and the ordinate shows the amplitude. As a result, the speech sequence was well partitioned into phonemes by the introduced method using the template matching. The speech speed and the change of the volume are corresponded with the time duration of each phoneme and the amplitude, respectively. In addition, we found that when the power becomes greater, the pitch also becomes higher. This result means that there is a relation between the pitch and the power of a vocal signal.

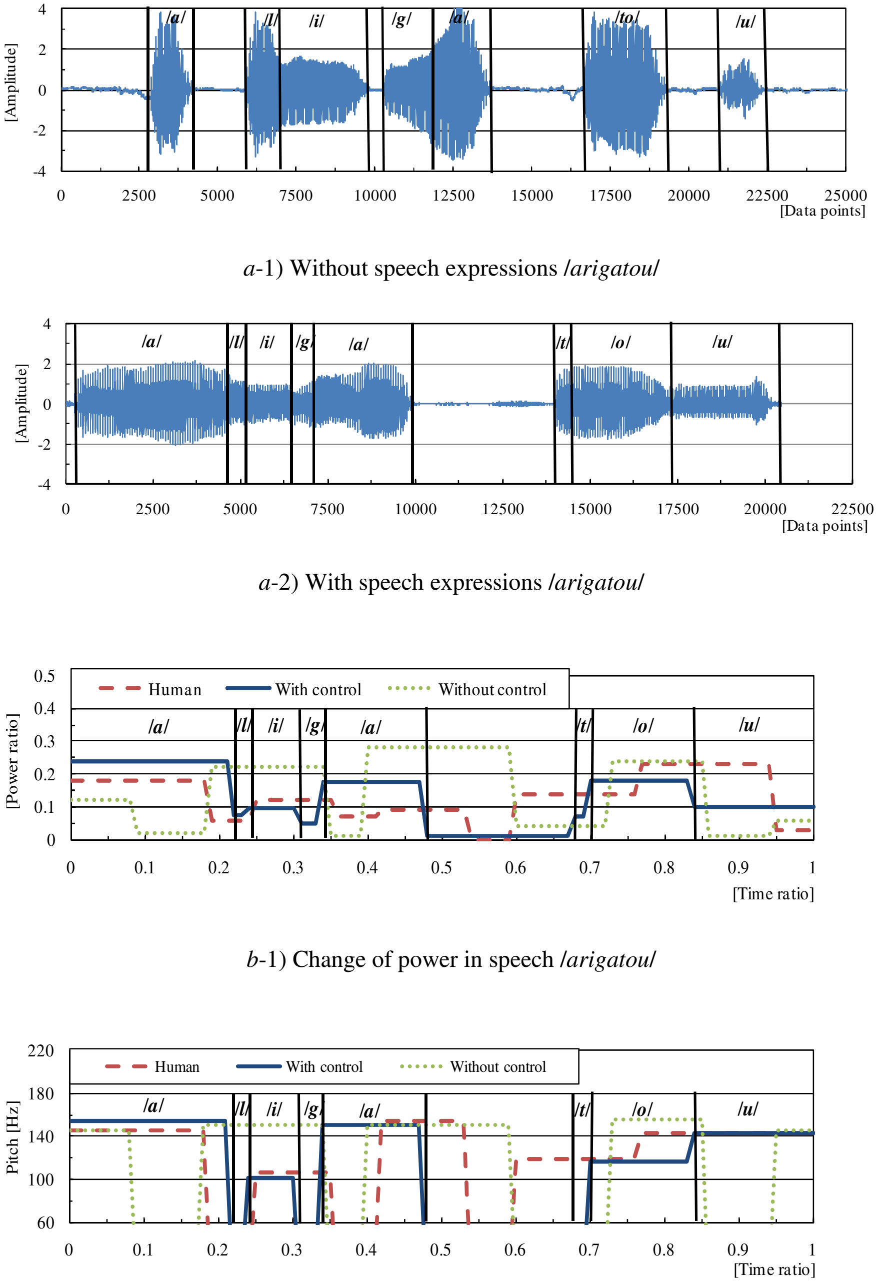
We examined the vocalizations of a human and the talking robot to validate the relation between the power and the pitch. The randomly uttered voices with changing the volume and pitch given by one subject, together with vocal sounds given by the talking robot with the random operations of vocal articulations were recorded. From the results, both human and the talking robot indicated the similar characteristics, in which the power increases its value in accordance with the higher pitch. With this result, we understood that as the opening and closing values of the airflow control motor were associated with human speech expressions, the talking robot could reproduce the human-like speech expressions.



a) Sound wave and the result of phoneme separation

b) Temporal change of power and pitch

Fig. 5. Analysis results of human speech /kon-nichi-wa/



a-1) Without speech expressions /*arigatou*/

a-2) With speech expressions /*arigatou*/

b-1) Change of power in speech /*arigatou*/

b-2) Change of pitch in speech /*arigatou*/

Fig. 6. Comparison of speech expressions

4.2 Reproduction of Human-Like Expressive Speech

The time duration, pitch and power would be controlled to reproduce the human-like speech expressions. The time duration can be managed by the temporal planning of opening and closing the motor for the airflow control, and the power and pitch will be manipulated by the opening and closing values of the airflow motor. The extraction of stable pitch from consonants is not easy due to their non- periodicity, so we pay attention to the signal power. If the appropriate relation between the motor angle and the power is established, the estimation of the motor angle from the human speech would be possible. An experiment for the estimation was conducted, and we calculated the powers of phonemes obtained by the talking robot. A regression formula was obtained by the method of least squares as presented below;

$$P = 0.02\,\theta + 1.10, \quad 0 < \theta < 110. \tag{1}$$

In the formula, θ indicates the motor angle and P indicates the power. If the angle becomes 110 degrees or greater, the generated sound becomes unnatural, and we set the maximum angle with 110 degrees.

The motor angle is calculated by referring to the power of phonemes, and the time durations of each phoneme are obtained by measuring the speech speed of a human speech. For human-like speech expressions, the calculated power and time durations are reflected to the talking robot. Figure 6 shows the sound signal of a speech /arigatou/ reproduced by the talking robot. An initial robotic speech without speech expressions is presented in the figure a-1), and we find that all the vowel and consonant vocalizations are made by the equal time durations. On the other hand, the figure a-2) shows the expressive robotic speech, in which the extracted volume and the time scale are reflected. Comparing with a human speech, the sound wave gets similar with each other by the proposed control method. The comparison of the temporal intonation and pitch changes among the three vocalizations are shown in the figures b) and c) respectively, and we find that both characteristics of robotic speech with expressive control get similar to human speech. Furthermore, by adequately controlling the timing of opening and closing the motors, the talking robot has successfully generated the human-like speech expressions.

For the assessment of the expressive speech vocalized by the robot, listening experiments were conducted and the speech was evaluated by questionnaires. Ten able-bodied subjects listened to 3 Japanese speeches with and without expressions given by the robot, which were /kon-nichi-wa/, /arigatou/ and /sayo-nara/, and evaluated them from the viewpoint of the naturalness and their preference. All the subjects preferred the expressive speech, and reported that the clarity of expressive speech is much higher than the one without expressions.

5 Conclusions

This paper introduced a talking robot constructed mechanically with human-like vocal chords and a vocal tract. By employing the adaptive learning and control-ling of the mechanical model based on the auditory feedback, the robot success-fully acquired the vocal articulation as a human baby did when he grew up, and autonomously generated vocal sounds whose pitches and phonemes were uniquely specified.

The human-like expressive speech production was also realized and evaluated in the study. The human speech was firstly analyzed, and physical factors for the human-like speech were extracted. The control method of the airflow motor which determines the pitch and volume of a speech was acquired to be given to the robotic speech, and the talking robot successfully reproduced the human-like expressive speech. The mimicry of human speech was evaluated by the listening experiment. All the subjects preferred the expressive speech in comparison with speech without expressions, and reported that the clarity of expressive speech is much higher than the one without expressions.

For the future work, we will study emotional factors from various human speeches, and try to let the talking robot to understand and mimic the expressive speech for realizing smooth speech communication.

Acknowledgements. This work was partially conducted by the doctoral study project of the late Dr. Mitsuki Kitani. He passed away one year later from his receipt of the doctoral degree. The author would like to dedicate this paper to Dr. Kitani for his great contribution to the research outcome.

References

1. Flanagan, J.L.: Speech Analysis Synthesis and Perception. Springer (1972)
2. Rodet, X., Benett, G.: Synthesis of the Singing Voice. In: Current Directions in Computer Music Research. PIT Press (1989)
3. Hirose, K.: Current Trends and Future Prospects of Speech Synthesis. Journal of the Acoustical Society of Japan, 39–45 (1992)
4. Smith III, J.O.: Viewpoints on the History of Digital Synthesis. In: International Computer Music Conference, pp. 1–10 (1991)
5. Fukui, K., Shintaku, E., Honda, M., Takanishi, A.: Mechanical Vocal Cord for Anthropomorphic Talking Robot Based on Human Biomechanical Structure. The Japan Society of Mechanical Engineers 73(734), 112–118 (2007)
6. Sawada, K., Osuka, K., Ono, T.: For the Realization of Mechanical Speech Synthesizer - Proposal of a model of tongue for articulation. Robotic Society of Japan 17(7), 1001–1008 (1999)
7. Miura, K., Asada, M., Yoshikawa, Y.: Unconscious Anchoring in Maternal Imitation that Helps Finding the Correspondence of Caregiver's Vowel Categories. Advanced Robotics 21(13), 1583–1600 (2007)
8. Kitani, M., Hara, T., Hanada, H., Sawada, H.: A Talking Robot and Its Singing Performance by the Mimicry of Human Vocalization. In: Human-Computer Systems Interaction: Backgrounds and Applications Part 2. Advances in Intelligent and Soft Computing, vol. 99, pp. 57–73 (2012) ISSN 1867-5662
9. Kitani, M., Hara, T., Sawada, H.: Voice articulatory training with a talking robot for the auditory impaired. International Journal on Disability and Human Development 10(1), 63–67 (2011)