Human Speech Production Mechanisms

Masaaki Honda†

Abstract

We use language every day without devoting much thought to the process, but articulatory movement—the movement of the lips, tongue, and other organs—is among the subtlest and most adept of any actions performed by human beings. Here, I discuss the mechanisms of speech production, introduce functions for producing sound by controlling the movement of vocal organs, and consider models for achieving speech functions that will enable a computer to mimic a voice that it hears.

1. Introduction

Speech is a natural form of communication for human beings, and computers with the ability to understand speech and speak with a human voice are expected to contribute to the development of more natural man-machine interfaces. Computers with this kind of ability are gradually becoming a reality, through the evolution of speech synthesis and speech recognition technologies. However, in order to give them functions that are even closer to those of human beings, we must learn more about the mechanisms by which speech is produced and perceived, and develop speech information processing technologies that

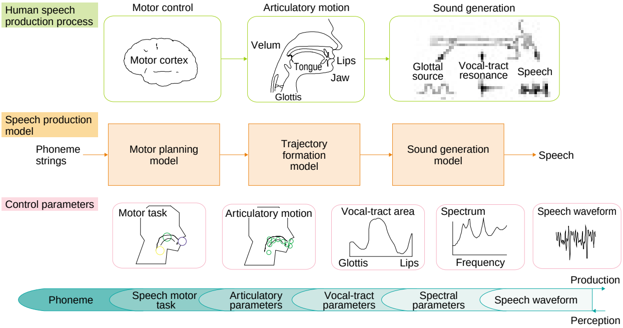


Fig. 1. Speech production process and the model.

† NTT Communication Science Laboratories
Atsugi-shi, 243-0198 Japan
E-mail: gomi@idea.brl.ntt.co.jp

make use of these functions. We use speech every day almost unconsciously, but an understanding of the mechanisms on which it is based will help to clarify how the brain processes information and will also lead to the development of more human-like speech devices through the imitation of these functions by computers.

Figure 1 shows the process of human speech and a computer model corresponding to it [1]. The mechanism of speech is composed of four processes: language processing, in which the content of an utterance is converted into phonemic symbols in the brain's language center; generation of motor commands to the vocal organs in the brain's motor center; articulatory movement for the production of speech by the vocal organs based on these motor commands; and the emission of air sent from the lungs in the form of speech. Here, I provide a brief explanation of these mechanisms and introduce a model for achieving these functions on a computer.

2. Mechanisms of speech production and speech synthesis

First, in order to simplify this discussion, I will explain how speech is created mechanically—that is, the mechanisms of speech synthesis—and then discuss how speech synthesis is related to the mechanisms involved in producing human speech. The first attempts to create speech using a machine date back to long before the birth of computers, to a speaking machine created by Wolfgang von Kempelen in 1771. It consisted of a bellows and a sound tube that simulated the vocal tract (the shape of the mouth from the vocal cords to the lips and nose), along with the reed of a musical instrument installed at one end of the sound tube. The flow of air sent from the bellows created a sound source by causing the reed to vibrate, and voices with a wide variety of tones were created by changing the shape of the resonance tube, thus modifying the resonance qualities of the sound tube. This speaking machine, which simulated the mechanisms of speech production using this type of sound source and vocal tract sound tube, became the model for speech synthesis devices that followed. The Vocoder (voice coder), developed by Homer Dudley in the 1930s, took a different approach. Rather than using the shape of the vocal tract itself, it recreated the sound resonance characteristics of the sound waves within the vocal tract by using electrical circuits as resonance filters. Many of the speech synthesis systems used widely throughout the world today utilize this principle through digital signal processing on computers or LSIs [2]. This same vocoder principle is also utilized in the digital speech coding technologies used in modern mobile phones.

The vocoder comprises a sound source generator and a voice tract resonance filter. When the sound source is a voiced sound, a cyclic pulse string is produced; when the sound source is unvoiced, a noise signal is produced. The pitch of the voice is controlled by the cycle of the cyclic pulse string. In the case of the vocal tract resonance filter, differences in tone resulting from phonemes such as the Japanese syllables "a" or "ka" are controlled by modifying resonance characteristics (voice spectrum characteristics) [2].

The principle of speech synthesis using a vocoder is to simulate the process of human speech production in the sense that the process of producing speech is expressed by the production of a sound source and resonance of sound. In human speech production as well, in the case of voiced sounds, the air flow sent from the lungs passes through an opening in the vocal cords, whose vibration causes that air flow to be converted into cyclic puffs of air that then become sound. In the case of unvoiced sounds like the syllable "sa," the air flow then passes through a narrow space formed by the tongue inside the mouth, a turbulent flow of air is produced, and this is emitted as a noise-like sound. These sounds sources themselves do not have any noticeable tone, but changing the position of the mouth creates sound wave resonance that differs depending on the shape of the vocal tract sound tube, and this results in the creation of various tones. When we speak, we move our jaws, tongue, and other parts of our mouth; in fact, this changes the shape of the vocal tract by changing the position of the mouth, and this in turn enables us to control sound resonance characteristics.

A speech production model that more directly simulates the physical process of human speech production comprises lungs, vocal cords, and the vocal tract. The vocal cords are expressed as a simple vibration model, and the pitch of the speech changes according to adjustments in the tension of the vocal cords. When the vocal cords close, their vibration results in voiced sounds; when they open, this vibration stops, and unvoiced sounds result. The vocal tract model is created as a non-uniform sound tube with differing cross-sectional areas, and the transmission of sound waves inside the sound tube is expressed by a digital filter. Vocal tract resonance characteristics are controlled according to the cross-sectional area (vocal tract area function) for various parts of the vocal tract up to the lips.

3. Mechanisms of articulatory movement

The speech production process has many levels, from the movement of vocal organs to the production of sounds. There are four hierarchical levels in speech production: the speech sound level, vocal tract shape level, vocal organ configuration level, and muscle contraction level. There is a "one-to-many" relationship among levels starting from the sound level and moving toward the muscle contraction level. For example, as is evident in the case of ventriloquism, sounds that seem very similar can be created using different vocal tract shapes. Similar vocal tract shapes can be created using differing vocal organ configurations; for example, the degree of mouth opening is determined by the relative position of both lips and the jaw. Furthermore, each individual vocal organ involves two or more competing muscles, and the vocal organ configuration is determined by their relative degree of contraction. This means that when speech with a given tone is created, the "one-to-many" relationship that exists among levels cannot be determined uniquely because there is always an excessive degree of freedom on the lower level.

This raises the questions of "What are the goals of articulatory movement?" and "How do humans determine specific human articulatory movements in the context of speech systems where these excessive degrees of freedom exist?" There is no doubt that the goal of articulatory movement is to transmit mainly language information through speech. It is also natural to assume that in speech systems where a "one-to-many" hierarchical relationship exists, the goal will be the highest level. In other words, we can assume that the goal in the context of articulatory movement is not to assign the position of the individual vocal organs (*e.g.*, the jaw, lips, or tongue), but rather to form the vocal tract that is most suitable for emitting a given sound and to achieve the acoustic phenomenon for speech. This type of approach can be seen within actual human articulatory movement.

Figure 2 shows the action of the tongue when the shape of the palate is changed using an artificial palate in the case of an utterance of the Japanese syllable "sha" [2]. Even when the shape of the upper jaw changes, the shape of the tongue responds instantaneously, changing to achieve the normal utterance of the syllable "sha." The results of these experiments suggest that the goal of articulatory movement is not only the absolute position of the various individual vocal organs, but also the shape of the vocal tract created by their relative positions.

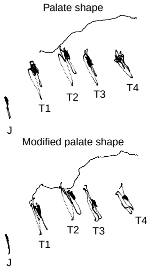


Fig. 2. Trajectories of jaw (J) and tongue (T1, T2, T3, T4) movements in production of /sha/.

Another feature of articulatory movement is that this is not a simple movement, but continuous movement for the purpose of uttering continuous sounds. For example, the movement of the mouth when uttering the sound "api" is not a simple connection of the individual mouth positions (articulations) for the phonemes /a/, /p/, and /i/. The articulation of the consonant /p/ is affected by the articulation of the vowel /i/ that follows it, and when it is uttering /p/, the tongue is already in position for the vowel /i/. The term "coarticulation" is used to refer to the phenomenon in which the mouth position for individual phonemes in the utterance of continuous sounds incorporates the effects of the mouth position for phonemes uttered immediately before and after. This is a very significant feature of articulatory movement.

4. Speech production models

Now, let me discuss the mechanisms of speech production models. As shown in Fig. 3, in a speech production model, a phoneme symbol (equivalent to a pronunciation key) for a spoken word is specified, a motor goal (task) for the articulatory movement is generated, and a motor trajectory for the vocal organ is calculated for the motor task assigned. The vocal tract shape is determined from the position of the vocal organs, and speech is produced by controlling the speech production model using the vocal tract

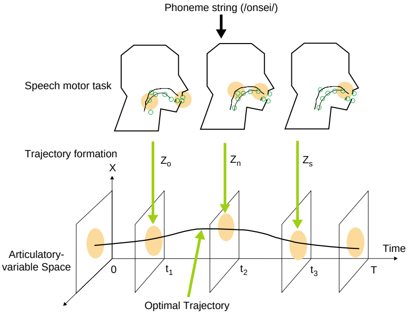


Fig. 3. Trajectory formation model of articulatory movements.

area [4].

Motor tasks are expressed by vocal tract variables that represent the local shape of the vocal tract. The vocal tract variables are: opening of the lower jaw, opening and protrusion of the lips, size and position of the narrowed segment of the vocal tract as formed by the tongue, and opening of the velum palatinum, which adjusts the connection to the nasal cavity in the vocal tract. The method of specifying the motor task differs depending on the phoneme. For example, in the case of /p/, only the opening of the lips and the velum palatinum are specified; other vocal tract variables are not. Vocal tract variables are expressed as a function of articulation variables that represent the configuration of individual vocal organs. For example, the vocal tract variable for the opening of the lips is a function of the jaw position and the height of both lips, and it is determined by the relative position of the jaw and lips. Each vocal tract variable is thus functions of multiple articulation variables, and there is a "one-to-many" relationship between them. For this reason, articulation variables that represent the configuration of vocal organs cannot be uniformly determined simply by specifying vocal tract variables. Furthermore, motor tasks are specified discretely in time as action targets, so there are an infinite number of motor trajectories for articulatory movement that fulfills the motor tasks. Therefore, in order to decide the articulatory movement, it is necessary to define a trajectory actually selected in the context of human articulatory movement from among an infinite number of motor trajectories that satisfy the motor task specified by the vocal tract variables. As illustrated in Fig. 3, when the motor task is represented as partial spaces within an articulation variable space, this issue of trajectory production is formulated as a problem of deciding the motor trajectory that passes through each partial space. In general, this is seen as a problem of indeterminacy, and the definition of a solution requires the introduction of new constraints.

Here, the cost function is assumed to be the volume defined as the weighted sum of [(the time derivative of motive force on the speech dynamic system)^2 + (motor speed)^2], and the motor trajectory is decided based on the constraint of minimizing this cost function. Qualitatively speaking, this motor index corresponds to the selection of the shortest possible trajectory while avoiding the use of unnecessary motive force energy for articulatory movement.

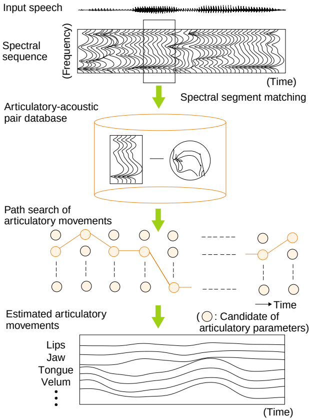
5. Speech mimicking model

A human can imitate the sound of an utterance he has heard, without necessarily being able to speak that particular language. This means that the person can estimate which mouth movements are necessary to produce a sound similar to the one he has heard.

For a computer to achieve this type of "Hearing-Mimicking Speech" function, it must estimate the mouth movement from the relevant speech. Generally, however, there is a one-to-many relationship between the acoustic quality of the speech and the shape of the sound tube corresponding to the vocal tract, so the position of the mouth cannot be uniquely determined from the utterance. In order to resolve this problem of indeterminacy, we must take into consideration the fact that the geometric shape of the sound tube is subject to constraints in terms of the positions that can be taken by the vocal organs and the fact that the changes in the sound tube over time are subject to constraints related to the movement of the vocal organs.

Figure 4 illustrates the process flow when estimating the movement of the vocal organs at the time of an utterance from the relevant speech [5]. The frequency of the input speech is analyzed, and the time line of the spectrum is derived. The articulatory movement and sounds are simultaneously observed in advance using an electromagnetic sensor system, and articulation/sound data—consisting of pairs composed of movement patterns for speech fragments about 100 ms long and temporal patterns for corresponding speech spectra—is stored in a database. The spectrum timeline for the input speech is matched with spectrum segment patterns in the database using spectrum segments of about 100 ms as the reference unit, and several patterns are selected from the database in order of their inter-pattern distance, from smallest to largest. Next, a determination is made regarding the movement pattern series with the highest degree of temporal continuity from among the selected movement pattern candidates, based on the constraint of temporal continuity as seen in articulatory movement.

In this method, a constraint related to the position and movement of the mouth that could actually take place is incorporated into a database created using observed articulatory motor data and speech data. In this way, it is possible to estimate the movement of the mouth from the speech with an average accuracy of about 1.9 mm. We have also confirmed that it is


Input speech
Spectral sequence
(Frequency)
(Time)
Spectral segment matching
Articulatory-acoustic pair database
Path search of articulatory movements

Time
(○: Candidate of articulatory parameters)
Estimated articulatory movements
Lips
Jaw
Tongue
Velum
⋮
(Time)

Fig. 4. Estimation of articulatory movements from speech acoustics.

possible to recreate the speech based on the mouth movement estimated from the original speech using a speech production model and to synthesize speech with tones that are extremely similar to the original speech. The model discussed here is a very close simulation of human "hearing-mimicking speech" functions and indicates that the functions demonstrated by human beings can be achieved even on computers.

6. Conclusion

I have provided a brief explanation of the mechanisms of speech production from the perspectives of speech-sound generation and articulatory movement and illustrated how these functions could be achieved on a computer. Currently, we are trying to develop a speaking robot by constructing a biomechanical vocal organ model that further incorporates the biological structures of human vocal organs and, at the same time, by clarifying the motor control mechanisms for speech that are executed by the brain [6].

References

[1] M. Honda, NTT CS Laboratories, Speech synthesis technology based on speech production mechanism, How to observe and mimic speech production by human, Journal of the Acoustical Society of Japan, Vol. 55, No. 11, pp. 777-782, 1999 (in Japanese).

[2] S. Saito and K. Nakata, Fundamentals of Speech Signal Processing, Ohm Publishing, 1981 (in Japanese).

[3] M. Honda, H. Gomi, T. Ito and A. Fujino, NTT CS Laboratories, Mechanism of articulatory cooperated movements in speech production, Proceedings of Autumn Meeting of the Acoustical Society of Japan, Vol. 1, pp. 283-286, 2001 (in Japanese).

[4] T. Kaburagi and M. Honda, NTT CS Laboratories "A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes," J. Acoust. Soc. Am. Vol. 99, pp. 3154-3170, 1996.

[5] S. Suzuki, T. Okadome and M. Honda, NTT CS Laboratories, "Determination of articulatory positions from speech acoustics by applying dynamic articulatory constraints," Proc. ICSLP98, pp. 2251-2254, 1998.

[6] M. Honda, NTT CS Laboratories, "Task planning mechanism of speech motor control, Japan Science and Technology Corporation, CREST Project," Journal of the Acoustical Society of Japan, Vol. 56, No. 11, pp. 771, 2000 (in Japanese).

Masaaki Honda

He received the B. E. M. E. and D. Eng. degrees from Waseda University, Japan, in 1973, 1975, and 1978, respectively. Since 1978, he has been with NTT Basic Research Laboratories, where he has been involved in research on digital speech coding and speech signal modeling from 1978 to 1988, and on speech production mechanism and the computational model from 1989 to 2003. Currently, he is the former group leader of Speech and Motor Control research group of Human Information Science Laboratory. Since 1999, he has also been a research group leader of CREST project of Japan Science and Technology, where he has been involved in research on motor planning of speech production.