

What Is a Speech Chain and How Can This Concept Be Applied to the Various Areas of Speech Communication in an Intelligent Society?

Takayuki Arai^a

Dept. of Information & Communication Sciences, Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo, Japan

Keywords: Speech Chain, Speech Communication, Speech Production, Speech Perception, Vocal-tract Models, My Voice Project.


Abstract: The concept of “Speech Chain” introduced by Denes and Pinson is widely used to interpret speech communication systems. The concept was originally aimed at human speech communication: a speaker first forms a message in his/her brain, the message is transformed into an acoustic signal that is sent to a listener, and the listener decodes the signal back into the original message. This simple situation can be extended to many scenarios. The acoustic signal can be fed into a telephone and transmitted over a telephone network. In human-computer communication, the speaker can be a speech synthesis system or the listener can be an automatic speech recognition system. For people who have lost the ability to talk, a speech synthesis system can improve their quality of life, and for people who have impaired hearing, an automatic speech recognition system can be a saviour. Communication with others is crucial as we live with other people in a society. As societies transform into intelligent societies, it is even more important to investigate speech communication systems from a scientific point of view and develop relevant applications in accordance with scientific findings. In this talk, the speech production mechanism will first be reviewed by using a set of vocal-tract models. Then, Speech Chain variations will be introduced for various areas in speech communication. Finally, application of the Speech Chain concept to an intelligent society through our “My Voice” project will be shared.

1 INTRODUCTION

The “Speech Chain” concept was originally introduced by Denes and Pinson (1993), whose book features a drawing of two people communicating with each other via acoustic signals (i.e., speech sounds). Such a human speech communication system contains several levels. First, a speaker forms a message in his/her brain, and words are selected and ordered to express the message. This is called the “linguistic level” because the speaker uses linguistic knowledge such as vocabulary and grammar. Based on the word sequence composed at the linguistic level, vocal and speech organs are moved to phonate and articulate speech sounds via the nervous system and muscles. This is called the “physiological level.” Then, speech sounds are produced and propagated from the speaker and reach the listener. This is called

the “acoustic level.” In the peripheral auditory process, the sounds are converted from mechanical vibrations into nerve signals on the physiological level. The nerve signals are transmitted to the listener’s brain, and the original message is constructed or decoded by applying linguistic knowledge on the linguistic level. The speaker’s ears also receive the speaker’s own speech sounds. In other words, the speaker is always monitoring their own output sounds. Thus, all events are chained; therefore, they are collectively a speech chain.

Communication with others is crucial as we live with other people in a society. As societies transform into intelligent societies, it is even more important to investigate speech communication systems from a scientific point of view and develop relevant applications in accordance with scientific findings. Therefore, first, we will review the speech production mechanism using a set of vocal-tract models. Next,

^a <https://orcid.org/0000-0003-1084-8083>

we will introduce speech chain variations for various areas of speech communication. Finally, we will share an example of utilizing the speech chain concept in an intelligent society through our “My Voice” project.

2 VOCAL-TRACT MODELS

In this section, we review human speech production by introducing physical models that imitate various phenomena that occur as we produce speech sounds (Arai, 2007, 2012, 2016).

2.1 Lung Models

As the speech chain concept, a message is formed in a speaker’s brain, and a word sequence is composed based on the speaker’s linguistic knowledge. Speech sounds are produced for the word sequence, and speech production is usually done during exhalation, which is part of the base human activity of breathing with lungs.

The bottom part of Fig. 1 shows a lung model with balloons (Arai, 2007). The phase of inhalation is when the pink diaphragm changes the volume of the thoracic cavity, and the air pressure inside the cavity becomes negative when the diaphragm is pulled down because the volume of the sealed cavity increases. During this phase, the two balloons sitting inside the cavity are inflated to compensate for the gap between the air pressures inside and outside the cavity. However, during the phase of exhalation when the diaphragm is pushed up, the volume of the cavity decreases, the air pressure inside the cavity becomes positive, and the air inside the balloon goes out through the larynx.

2.2 Artificial Larynx

In Fig. 1, the blue arrow indicates the location of an artificial larynx (Arai, 2007). During exhalation, the air goes through the artificial larynx and produces a glottal sound by vibrating the membrane of the larynx (i.e., “phonation”). Figure 2 shows a reed-type artificial larynx (Arai, 2012). In Fig. 2, the air coming from the pump goes through the reed-type larynx and causes the reed to vibrate. That also makes a similar glottal sound.

2.3 Vocal-Tract Models

The top part of Fig. 1 shows a head-shaped model (Arai, 2007) that produces vowel /a/ because its

cavity forms a vocal tract when we produce vowel /a/. The gray part in Fig. 2 is a straight tube imitating a vocal-tract area function along the length of the human vocal tract (Arai, 2012). Because the area function in this case is based on the vowel /o/, the output sound through the vocal-tract model in Fig. 2 sounds like /o/.

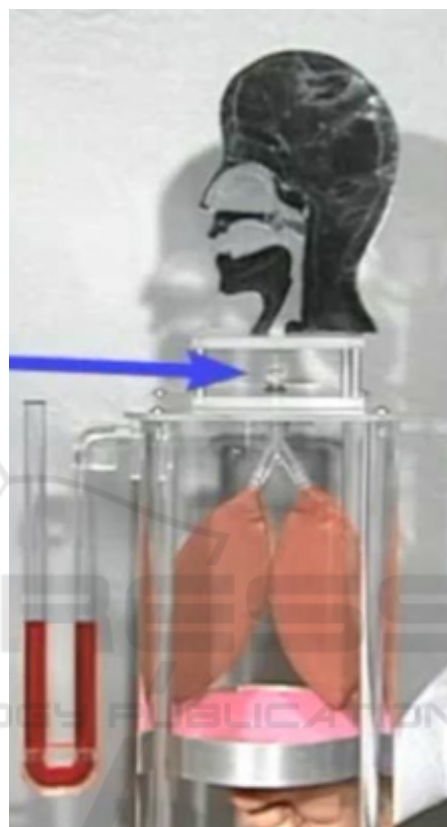


Figure 1: The lung model, an artificial larynx (blue arrow) and a head-shaped model.



Figure 2: The air pump (in blue), a reed-type sound source (left) and a vocal-tract model (right).

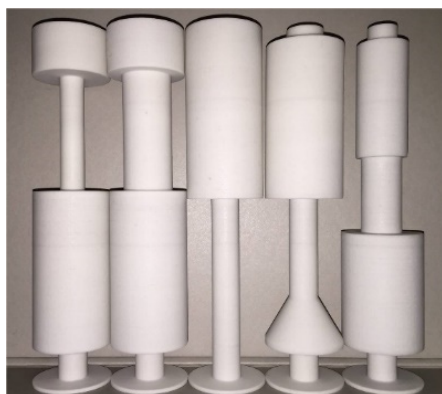


Figure 3: The vocal-tract models, VTM-T20. Vowels /i/, /e/, /a/, /o/, and /u/ (from left to right).



Figure 4: A model of the human vocal tract with a flexible tongue and a movable mandible.

The picture in Fig. 3 shows a set of vocal-tract models called VTM-T20 (Arai, 2016). When glottal sounds are fed into the bottom end of each tube, different vowels can be heard from the top end. While the shape of the vocal tract mainly determines what vowel sound is produced (i.e., “articulation”), the glottal sound mainly determines the pitch (height) of the voice and the voice quality. Thus, vocal tracts act as resonators, and their shapes produce different speech sounds.

A real vocal tract changes its shape in time. Therefore, our model in Fig. 4 has a flexible tongue and a movable mandible (Arai, 2020). By manipulating the tongue configuration and adjusting the jaw opening, we can produce different speech sounds dynamically with this model.

3 SPEECH CHAIN IN DIGITAL ERA

Thus, the speech chain describes how the human speech communication system works as each event is connected as a chain. The original speech chain shows us a simple situation, but it can be extended to many scenarios. When we talk on a telephone network, acoustic signals are fed into the telephone, converted into electric signals, and transmitted over the network. In human-computer communication, the speaker can be a speech synthesis system, or the listener can be an automatic speech recognition system. A speech synthesis system can improve the quality of life of people who have lost the ability to talk, and an automatic speech recognition system can be a great help to people who have impaired hearing.

In the following section, I will explain the “My Voice” project that I was involved with a patient as an example of a speech chain in the digital era.

3.1 What Is My Voice?

We lose our voice for various reasons, such as amyotrophic lateral sclerosis (ALS). When an ALS patient has difficulty breathing, they may receive a tracheotomy, which causes the patient to lose the ability to speak. A laryngectomy is another procedure that causes people to lose their voice. “My Voice” is free, widely used Japanese speech synthesis software that gives patients the ability to use their voice after surgeries that take away their ability to speak.

My Voice is widely used for a few reasons in particular. First of all, it is free. Also, it is easy to use. The recording time can be minimized for patients and reduce their workload, as the main recording is only for basic Japanese syllable units and words can be concatenated from the recorded syllable units. Furthermore, its software GUI is well-designed, so a patient’s therapist, family members, and friends can also use it with simple training. By recording words and/or phrases before surgery, patients can keep communicating using their voice with technology even after losing that ability physically.

3.2 Why My Voice?

Commercial speech synthesizers usually aim for clarity and intelligibility in synthesized speech sounds. However, My Voice uses a speaker’s voice and vocal characteristics. When an ALS patient has difficulty breathing as their disease progresses, they must choose to have a tracheotomy and lose their

voice or gradually lose the ability to breathe altogether. My Voice gives them the option to have a tracheotomy and continue using their voice.

4 CONCLUSIONS

We reviewed the speech chain concept and the speech production mechanism by using a set of vocal-tract models. After extending the speech chain concept into the digital era through various forms of speech communication, we discussed its application to an intelligent society through our “My Voice” project.

Even though society is becoming increasingly intelligent and speech communication is evolving over time, speech communication itself will always exist because it is a part of being human. One of the United Nations’ Sustainable Development Goals (SDGs) is about health and well-being (UNs, 2019). We need to explore speech communication more and develop more applications for sustainable societies.

Our vocal-tract models are also used to test the airstream of human breath and droplets from producing speech. One way to visualize the aerosol and the droplets from speech production is by passing a laser sheet over a human speaker. However, a laser beam can damage human beings, and therefore, our vocal-tract models are now being used to simulate our speech communication (Arai, 2021). It is also another way of contributing our technology towards the SDG goals.

ACKNOWLEDGEMENTS

I would like to thank the organizers of the conference for inviting me. I would also like to thank the people involved in the My Voice project (especially Kenji Fujimoto and his family members), Takaki Yoshimura, Musashi Homma, and Shigeto Kawahara, and members of Arai Laboratory (Sophia University). This work was partially supported by JSPS KAKENHI Grant Numbers 18K02988 and 21K02889.

REFERENCES

Arai, T. (2007). ‘Education system in acoustics of speech production using physical models of the human vocal tract’. *Acoustical Science and Technology* 28(3) 190–201.

Arai, T. (2012). ‘Education in acoustics and speech science using vocal-tract models’. *The Journal of the Acoustical Society of America* 131(3) 2444–2454.

Arai, T. (2016). ‘Vocal-tract models and their applications in education for intuitive understanding of speech production’. *Acoustical Science and Technology* 37(4) 148–156.

Arai, T. (2020). ‘Two different mechanisms of movable mandible for vocal-tract model with flexible tongue’. *Proc. of INTERSPEECH* 1366–1370.

Arai, T. (2021). ‘Vocal-tract models to visualize the airstream of human breath and droplets while producing speech’. *Proc. of INTERSPEECH*.

Denes, P. B, and Pinson, E. N. (1993). *The Speech Chain: The Physics and Biology of Spoken Language*, 2nd edition, W. H. Freeman, New York.

Kawahara, S., Homma, M., Yoshimura, T., and Arai, T. (2016). ‘My Voice: Rescuing voices of ALS patients’. *Acoustical Science and Technology* 37(5) 202–210.

United Nations. (2019). The Sustainable Development Goals Report, <https://doi.org/10.18356/55eb9109-en>.