Speaking Rhythm Extraction and Control by Non-negative Temporal Decomposition

Sadao Hiroya

Abstract

Speaking rhythm plays an important role in speech production and the perception of non-native languages. This article introduces a novel method for extracting and controlling speaking rhythm from speech signals using non-negative temporal decomposition.

Keywords: articulatory movements, non-negative temporal decomposition, speaking rhythm

1. Introduction

Speech communication using non-native languages is difficult for many people both in speaking and listening to speech. By way of example, most native Japanese speakers have difficulty understanding what native English speakers are saying and therefore cannot communicate well in English with them. There are two major differences between Japanese and English: pronunciation (e.g., the number of vowels and the /R-L/ contrast) and rhythm. Pronunciation is very important for communication in English using words and short sentences (e.g., "Coffee, please." and "Where is the toilet?"). For long sentences, on the other hand, rhythm is more important than pronunciation. However, most Japanese learners of English regard pronunciation as important, rather than rhythm. As a result, native Japanese speakers have trouble communicating in English using long sentences with native English speakers.

In this article, I introduce a novel method of automatically correcting the halting English rhythm of native Japanese speakers by approximating the natural rhythm of native English speakers (**Fig. 1**).

2. Speaking rhythm

Rhythm generally refers to a pattern in time. In linguistics, languages can be categorized into two rhythms: stress-timed rhythm (e.g., English) and syllable-timed rhythm (e.g., Japanese). Chen et al. explained this as follows: "Stress-timed rhythm is determined by stressed syllables, which occur at regular intervals of time, with an uneven and changing number of unstressed syllables between them. Syllable-timed rhythm is based on the total number of syllables since each syllable takes approximately the same amount of time," [1] (**Fig. 2**). A syllable-timed rhythm is thus simpler than a stress-timed rhythm.

Humans follow a rhythm in various situations: speaking, playing musical instruments, clapping hands, walking, etc. Therefore, the definition of rhythm is not limited to only the temporal structure of sounds. In this study, I define speaking rhythm as a temporal pattern of movements made by articulatory organs such as the lips, jaw, tongue, and velum (soft palate); that is, not as sounds, but as articulatory movements. Some readers might question whether the definition of rhythm should be used for articulatory movements since most articulatory organs are inside the mouth, where speech is produced. However, I measured articulatory movements using an electromagnetic articulography (EMA) system and magnetic resonance imaging (MRI) [2] (**Fig. 3**) and estimated articulatory movements from speech signals [3]. My previous study indicated that articulatory movements are suitable for defining speaking rhythm [4]. Specifically, I analyzed the articulatory parameters

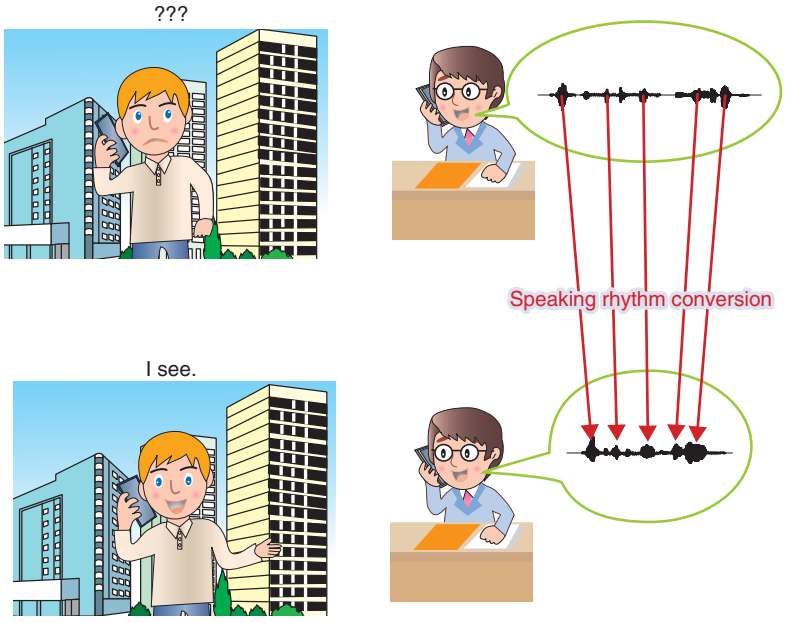


Fig. 1. Example of speaking rhythm conversion.

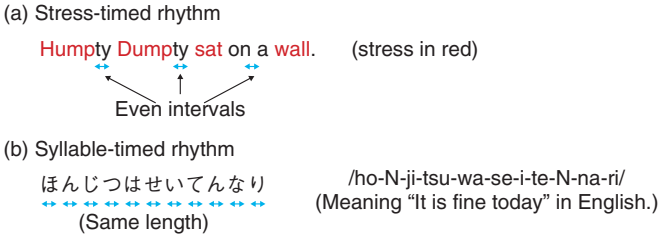
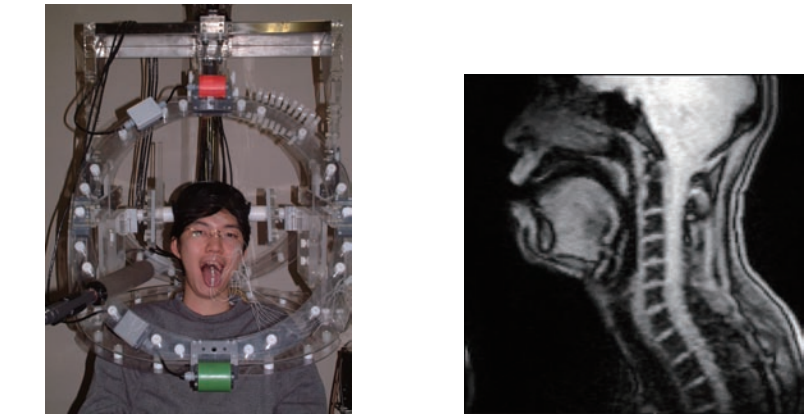


Fig. 2. Examples of stress-timed rhythm and syllable-timed rhythm.



(a) Electromagnetic articulography system (b) Magnetic resonance imaging*

* In collaboration with Konan University, Hyogo, Japan.

Fig. 3. Methods use to measure articulatory movements.

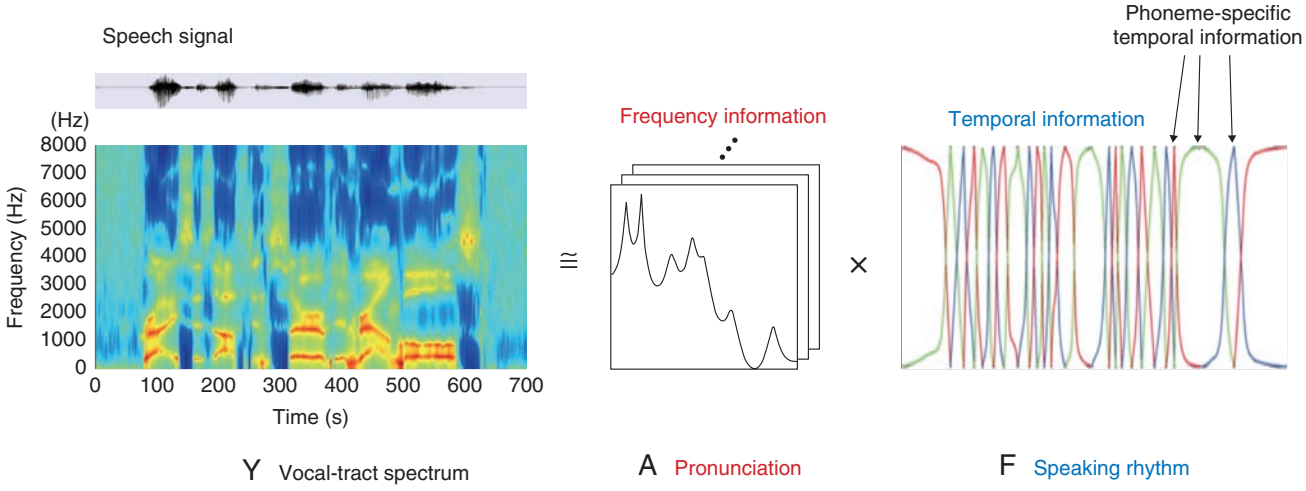


Fig. 4. Non-negative temporal decomposition.

represented by the vertical and horizontal positions of six receiver coils, which were placed on the lower incisor, the upper and lower lips, and the tongue (three positions), which were measured by the EMA system during speech production. The results revealed that articulatory parameters can be represented by articulatory positions at the central point of each phoneme and by linear interpolation. That is, a sparse representation of articulatory movements is suitable for obtaining speaking rhythm. This finding is related to articulatory phonology [5] and recent findings on the neural mechanism of speech production [6]. Also, this indicates that articulatory parameters change smoothly.

Consequently, treating articulatory movements as speaking rhythm should make it possible to easily extract and control speaking rhythm.

3. Non-negative temporal decomposition

Speech signals contain both frequency and temporal information. In audio signal processing, non-negative matrix factorization (NMF) can be applied to decompose audio signals into frequency and temporal information [7]. However, the NMF algorithm does not introduce articulatory-specific restrictions. Thus, it is not guaranteed that the temporal information will have a bell-shaped velocity profile, which is characteristic of human articulatory movements, and that only phonemes adjacent to the temporal information will affect it.

To overcome this problem, I developed a non-negative temporal decomposition (NTD) method to extract the speaking rhythm (temporal information) from speech signals under articulatory-specific restrictions.

NTD decomposes a vocal-tract spectrum (e.g., a line spectral pair), which is associated with articulatory organs, into a set of temporally overlapped phoneme-dependent event functions F and corresponding event vectors A under articulatory-specific restrictions (**Fig. 4**). Temporal information F introduces the phoneme-specific model and is affected only by adjacent phonemes. The NTD algorithm is as follows. First, a vocal-tract spectrum is calculated from speech signals. Then, an event function, which is restricted to the range [0,1], is determined by minimizing the squared Euclidean distance between the input and the estimated vocal-tract spectrum based on the multiplicative update rules in the NMF algorithm. Multiplicative update rules make it possible to obtain non-negative values of event functions and unimodal event functions without any penalty functions [4]. In fact, the multiplicative update rules would be more effective for improving the bell-shaped velocity profiles than a smoothing method with a penalty function introduced to NMF.

In NTD, the event timings need to be known. In this study, the timings were modified by minimizing the squared Euclidean distance by utilizing dynamic programming (DP). Thus, NTD can be considered a constrained NMF with DP. The only input for NTD is speech signals, but NTD can extract the speaking rhythm of articulatory movements due to the

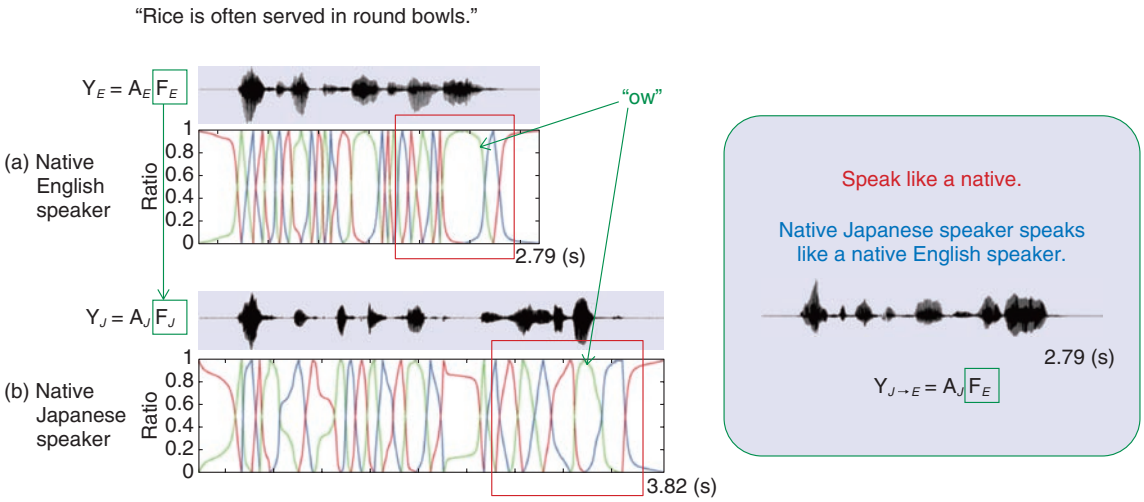


Fig. 5. Speaking rhythm conversion method.

articulatory-specific restrictions. Therefore, NTD is expected to be useful for acoustic-to-articulatory inversion [4].

4. Control of speaking rhythm

In this section, I explain how the speaking rhythm of an English sentence spoken by a native Japanese speaker is converted into the rhythm of a native English speaker (**Fig. 5**). First, both native Japanese and native English speakers read the same English sentence (e.g., "Rice is often served in round bowls".) Next, NTD is applied to extract frequency information A_J and temporal information F_J from the vocal-tract spectrum of the native Japanese speaker and to extract A_E and F_E from that of the native English speaker. I substitute F_E for F_J to obtain a vocal-tract spectrum with the pronunciation of native Japanese speaker A_J and the rhythm of native English speaker F_E. Finally, speech signals are generated from the vocal-tract spectrum and source signals. The generated speech signal in Fig. 5 appears to be time-compressed speech, in which the temporal characteristics of the speech signal are altered by reducing its duration without affecting the frequency characteristics. However, the temporal pattern in "bowls" (red square in Fig. 5) differs between the Japanese and English native speakers; the duration of "ow" for the English speaker is much longer than that for the Japanese speaker. This indicates that the technique is effective for controlling the English speaking rhythm of the native Japanese speaker. Feedback from native English speakers indicated that this speaking-rhythm-controlled speech signal using a personal computer was easier to understand.

5. Future prospects

Speech translation systems using another person's voice can also assist native Japanese speakers when they are communicating verbally in English. However, the opportunities for speech communication in English using one's own voice rather than another person's voice are expected to increase owing to the fact that English has been a required subject in elementary schools since 2011 in Japan.

This technique will be useful in practical applications such as communicating in English via teleconferences, public speaking, and using mobile phones. However, the technique cannot change the speaking rhythm of a sentence unless there is already a sample of the same sentence that has been read before. Thus, to become widely used, it will be necessary to model event functions between languages. I hope that this technique will eventually alleviate the burden involved in communication using non-native languages.

References

[1] C. Chen, C. Fan, and H. Lin, "A New Perspective on Teaching English Pronunciation: Rhythm," Proc. of the 4th International Symposium on English Teaching, pp. 24–41, Kaohsiung, Taiwan, 1996.

[2] S. Hiroya and T. Kitamura, "Generation of a vocal-tract MRI movie based on sparse sampling," Proc. of International Seminar on Speech Production (ISSP) 2011, pp. 1–8, Montreal, Canada.

[3] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," IEEE Trans. on Speech and Audio Processing, Vol. 12, No. 2, pp. 175–185, 2004.

[4] S. Hiroya, "Non-negative temporal decomposition of speech parameters by multiplicative update rules," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 21, No. 10, pp. 2108–2117, 2013.

[5] C. P. Browman and L. Goldstein, "Articulatory phonology: an overview," Phonetica, Vol. 49, No. 3-4, pp. 155–180, 1992.

[6] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation," Nature, Vol. 495, No. 7441, pp. 327–332, 2013.

[7] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, Vol. 401, No. 6755, pp. 788–791, 1999.

Sadao Hiroya

Senior Research Scientist, NTT Communication Science Laboratories.

He received the B.S. degree from Tokyo University of Science in 1999 and the M.E. and Ph.D. degrees from Tokyo Institute of Technology in 2001 and 2006, respectively. He joined NTT Communication Science Laboratories in 2001. From 2001 to 2003, he was also a researcher in the CREST project of the Japan Science and Technology Agency. From 2007 to 2008, he was a visiting scholar at Boston University, MA, USA. In 2006, he received the 1st Itakura Prize Innovative Young Researcher Award from the Acoustical Society of Japan (ASJ). His current research interests include the links between production and perception of speech, functional brain imaging, and acoustic-to-articulatory inversion problems. He is a member of ASJ, the Institute of Electronics, Information and Communication Engineers, the Society for Neuroscience, and IEEE.