A Modulation-demodulation Model of Speech Communication

Nobuaki Minematsu

Graduate School of Information Science and Technology, The University of Tokyo

Abstract

Perceptual invariance against a large amount of acoustic variability in speech has been a long-discussed question in speech science and engineering [1] and it is still an open question [2, 3]. Recently, we proposed a candidate answer to it based on mathematically-guaranteed relational invariance [4, 5]. Here, completely transform-invariant features, $f$-divergences, are extracted from speech dynamics of an utterance and they are used to represent that utterance. In this paper, this representation is interpreted from a viewpoint of telecommunications and evolutionary anthropology. Speech production is often regarded as a process of modulating the baseline timbre of a speaker’s voices by manipulating the vocal organs, i.e., spectrum modulation. Then, extraction of the linguistic content from an utterance can be viewed as a process of spectrum demodulation. This modulation-demodulation model of speech communication has a good link to known morphological and cognitive differences between humans and apes. The model also claims that a linguistic content is transmitted mainly by supra-segmental features.

Index Terms: speech recognition, invariant features, spectrum demodulation, evolutionary anthropology, language acquisition

1. Introduction

Many speech sounds exist as standing waves in a vocal tube and their acoustic properties mainly depend on the shape of the vocal tube. A process of producing vowel sounds is the same as that of producing sounds with a wind instrument. A vocal tube is an instrument and, by changing its shape dynamically, sounds of different timbre are generated such as /æiou/.

The length and shape of the vocal tube is also different among speakers. This is the reason why the voice timbre is different among them and one can identify a speaker by hearing his/her voices. Figure 1 shows the tallest adult and the shortest one of the two, there must be the largest gap of the voice timbre. But they could communicate orally with no trouble just after they saw each other for the first time. This is a good example of invariance and variability in speech processes.

In telecommunications, a transmitted content is received and interpreted via demodulation. In [6], demodulation is explained as a process of extracting the original content intactly from a modulated carrier wave. In any case of amplitude modulation (AM), phase modulation (PM), and frequency modulation (FM), one is able to use methods that can mathematically remove the carrier information exclusively from the modulated carrier to leave the original content to receivers intactly.

Figure 2 shows processes of transmitting and receiving the contents via FM. If one views speech production as spectrum modulation, he/she may point out that these proposals do not give an answer to a fundamental and essential question of ASR, that is “What is an algorithm of spectrum demodulation?”

In telecommunications, a transmitted content is received and interpreted via demodulation. In [6], demodulation is explained as a process of extracting the original content intactly from a modulated carrier wave. In any case of amplitude modulation (AM), phase modulation (PM), and frequency modulation (FM), one is able to use methods that can mathematically remove the carrier information exclusively from the modulated carrier to leave the original content to receivers intactly.

Figure 2 shows processes of transmitting and receiving the contents via FM. If one views speech production as spectrum modulation, he/she may point out that these proposals do not give an answer to a fundamental and essential question of ASR, that is “What is an algorithm of spectrum demodulation?”

In telecommunications, a transmitted content is received and interpreted via demodulation. In [6], demodulation is explained as a process of extracting the original content intactly from a modulated carrier wave. In any case of amplitude modulation (AM), phase modulation (PM), and frequency modulation (FM), one is able to use methods that can mathematically remove the carrier information exclusively from the modulated carrier to leave the original content to receivers intactly.

Figure 2 shows processes of transmitting and receiving the contents via FM. If one views speech production as spectrum modulation, he/she may point out that these proposals do not give an answer to a fundamental and essential question of ASR, that is “What is an algorithm of spectrum demodulation?”

In telecommunications, a transmitted content is received and interpreted via demodulation. In [6], demodulation is explained as a process of extracting the original content intactly from a modulated carrier wave. In any case of amplitude modulation (AM), phase modulation (PM), and frequency modulation (FM), one is able to use methods that can mathematically remove the carrier information exclusively from the modulated carrier to leave the original content to receivers intactly.

Figure 2 shows processes of transmitting and receiving the contents via FM. If one views speech production as spectrum modulation, he/she may point out that these proposals do not give an answer to a fundamental and essential question of ASR, that is “What is an algorithm of spectrum demodulation?”

In telecommunications, a transmitted content is received and interpreted via demodulation. In [6], demodulation is explained as a process of extracting the original content intactly from a modulated carrier wave. In any case of amplitude modulation (AM), phase modulation (PM), and frequency modulation (FM), one is able to use methods that can mathematically remove the carrier information exclusively from the modulated carrier to leave the original content to receivers intactly.

Figure 2 shows processes of transmitting and receiving the contents via FM. If one views speech production as spectrum modulation, he/she may point out that these proposals do not give an answer to a fundamental and essential question of ASR, that is “What is an algorithm of spectrum demodulation?”

In telecommunications, a transmitted content is received and interpreted via demodulation. In [6], demodulation is explained as a process of extracting the original content intactly from a modulated carrier wave. In any case of amplitude modulation (AM), phase modulation (PM), and frequency modulation (FM), one is able to use methods that can mathematically remove the carrier information exclusively from the modulated carrier to leave the original content to receivers intactly.

Figure 2 shows processes of transmitting and receiving the contents via FM. If one views speech production as spectrum modulation, he/she may point out that these proposals do not give an answer to a fundamental and essential question of ASR, that is “What is an algorithm of spectrum demodulation?”

In telecommunications, a transmitted content is received and interpreted via demodulation. In [6], demodulation is explained as a process of extracting the original content intactly from a modulated carrier wave. In any case of amplitude modulation (AM), phase modulation (PM), and frequency modulation (FM), one is able to use methods that can mathematically remove the carrier information exclusively from the modulated carrier to leave the original content to receivers intactly.
pect to anthropology and linguistics. Further, we carry out an interesting thought experiment of “anthropologists from Mars” [13]. Although this paper provides no new experimental result, we believe that it gives a novel model of speech communication.

2. Morphological and cognitive differences between humans and apes

2.1. Morphological differences between humans and apes

The proposed model of speech communication regards a tongue as a flexible modulator of the spectrum. Figure 3 shows the vocal organs of a human and those of an ape (chimpanzee) [14]. Different from humans, apes have two independent tracts, one is from the nose to the lung for breathing and the other is from the mouth to the stomach for eating. Apes can breathe and eat at the same time although humans cannot. Therefore, it is generally difficult for apes to send an air flow from the lung directly to the vocal cavity. Further, apes have a much lower degree of freedom of deforming the tongue shape compared to humans [15]. Then, spectrum modulation is reasonably difficult. Why is it easy for humans? Anthropologically speaking, the reason is bipedal walking [14]. A long time ago, a kind of apes had stood up and begun to walk on foot. Then, the larynx fell down due to the gravity and the two tracts happened to have an intersection. Figure 4 shows voice communication between a small ape and a large ape. They have their own internal cavities but their sizes are very different. This reasonably causes spectral (timbre) differences between their voices. Because of the reason stated above, it is very difficult for them to embed some contents (messages) in the voices via spectrum modulation.

In humans, however, a short human and a tall one can transmit some contents (messages) via spectrum modulation (See Figure 5). Because of the size difference, their modulated carriers are very different acoustically. Independently of their own voice timbre, however, they can transmit the contents intactly. The modulation-demodulation model of speech communication assumes that a good demodulator exists only in the human organs of audition. Some findings in anthropology are shown, which indicate that apes don’t have good demodulators.

2.2. Cognitive differences between humans and animals

As explained in Section 1, a musical melody can be characterized as FM version of a carrier wave. If we apply two different carriers to the same musical content, they will become a melody and its transposed version. From the two, humans can extract the same musical content easily and this extraction is demodulation. But apes cannot perceive the equivalence between a melody and its transposed version [16]. It is difficult for apes to demodulate FM carriers. What about spectrum demodulation?

Human infants acquire spoken language via vocal imitation of utterances from their parents. But they don’t impersonate their parents. Their vocal imitation is not acoustic imitation.
Here is a simple question. What acoustic aspects of the utterances do infants imitate and what aspects do they ignore?

The performance of vocal imitation is rarely found in animals. No other primate than human does not perform it. Only a few species do, such as birds, dolphins and whales [17]. But there exists a critical difference between the vocal imitation of humans and that of animals. Animals’ imitation is acoustic imitation, i.e., impersonation [17]. If we consider their vocal communication from a viewpoint of the proposed model, we can say that animals imitate the modulated carriers, not the contents.

What acoustic aspects of parents’ utterances do infants imitate? One may assume that infants decompose the utterances into sequences of phonemes (text-like representation) and they realize each phoneme acoustically with their mouths. But researchers of infant studies deny this assumption because infants do not have good phonemic awareness [18, 19]. No infant acquires spoken language by reading text out. Then, what do they imitate? A good candidate answer is found in [18, 19, 20, 21]. Infants extract holistic and speaker-independent speech patterns, called word Gestalts, and they realize the patterns acoustically using their small mouths. However, no researcher in infant studies has provided a mathematical formula of the speech patterns, which is considered to lead to an algorithm of spectrum demodulation or extraction of the embedded contents.

3. Implementation of spectrum demodulation

3.1. Removal of the speaker-dependent voice timbre

Demodulation removes the carrier information and leaves the content intactly to receivers. In [4, 5], we implemented this process using transform-invariant speech features. Speaker difference is often characterized by transformation from a speaker’s voice space to another’s. This indicates that, if an utterance can be represented only with transform-invariant features, that representation comes to have no speaker-dependent features.

In [5], we proved that $f$-divergence is invariant with any kind of invertible and differentiable transform (sufficiency) and that the features invariant with any kind of transform, if any, have to be $f$-divergence (necessity). Here, as shown in Figure 6, any event in a space has to be represented as distribution.

3.2. Computational implementation of speech Gestalts

By representing an utterance only with $f$-divergence, we can calculate an invariant speech Gestalt mathematically. The upper side of Figure 7 shows its extraction procedure. A speech trajectory in a feature space is converted into a sequence of distributions. In [4, 5], as shown in the lower side of Figure 7, this process was implemented by applying the HMM training procedure. Between every distribution pair, $f$-divergence is calculated to form a distance matrix. We also call it speech structure.

If one wants to focus on the dynamic aspect of an utterance, he/she may calculate a velocity vector at each point in time, i.e., delta cepstrum. We have to claim, however, that this strategy is inadequate. Spectrum modification caused by vocal tract length difference is often modeled as frequency warping. [22] indicates that this warping can be represented in the cepstrum domain as multiplication of a specific type of matrix by a cepstrum vector. In [23], we showed mathematically that this matrix is approximated as rotation matrix and showed experimentally that the change of vocal tract length rotates a speech trajectory. This is why we extract only scalar features in Figure 7.

\[ f_{\text{div}}(p_1, p_2) = \int p_2(x)p_1(T(x))dx = f_{\text{div}}(T(p_1), T(p_2)) \]

4. Discussions and conclusions

4.1. Anthropologists from Mars [13]

The proposed model claims that speech recognition should be based on removing the speaker-dependent aspects of utterances. Here, we discuss this claim using a thought experiment. Let us assume that anthropologists, who came from Mars, observed human-to-human communications on Earth for the first time. Fortunately or unfortunately, their observations were done for communications between laryngectomized individuals and deaf individuals. They communicated with each other using some special devices but the anthropologists did not notice them.

NAM (Non-Audible Murmur) microphones are applied to support laryngectomized patients [31]. They have difficulty in speaking aloud but they can generate murmurs. NAM is too weak to be detected as air vibrations but can be sensed as skin vibrations by a NAM microphone. These skin vibrations are converted to normal speech [31], which is transmitted via FM.

It is technically possible enough to connect an FM receiver to a cochlea implant. If a deaf individual uses an FM-connected cochlea implant, he/she can receive messages from a laryngectomized patient with an FM-connected NAM microphone. This is what the anthropologists saw, shown in Figure 9.
4.2. Speech scientists and engineers on Earth

A long time ago, a kind of apes had stood up and began to walk only on foot. Then, the larynx fell down due to the gravity. This enabled them to speak [14], in other words, to modulate their baseline timbre in various ways. Infants acquire spoken language not by reading text (a sequence of sound symbols) out. The proposed model explains that infants’ language acquisition may start with learning how to modulate their baseline timbre. About seventy years ago, speech scientists and engineers on Earth observed the spectrograms of utterances visually for the first time. Today, as many speech researchers consider speech as spectrum modulation [8, 9, 10], it is definitely true that the spectrograms are modulated carriers, not contents. It is also true that calculation of a statistical model of the modulated carriers will not leave the contents to receivers. We consider that what speech researchers on Earth have to do is not to collect samples but to clarify the internal mechanism of speech communication. The modulation-demodulation model is our proposal for that.

In [30], Jakobson proposed a theory of acoustic and relational invariance, called distinctive feature theory. He repeatedly emphasizes the importance of relational and morphological invariance among sound speech sounds. Figure 10 shows its invariant shape of French vowels and semi-vowels. We understand that our proposal is computational implementation of his theory and we can claim that our proposal has high linguistic validity.

5. References