A Modulation-demodulation Model of Speech Communication

Nobuaki Minematsu

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

mine@gavo.t.u-tokyo.ac.jp

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 *de*modulation. 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 very similar to 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 /aeiou/.

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 world. Between 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 content is often transmitted to receivers by changing one parameter of a carrier waveform in relation to that content. A sinusoid wave is often used as carrier wave. This transmission scheme is called modulation and, in [6], it is explained by using the performance of a musician as a metaphor. *A musician modulates the tone from a musical instrument by varying its volume, timing and pitch. The three key parameters of a carrier sine wave are its amplitude ("volume"), its phase ("timing") and its frequency ("pitch"), all of which can be modified in accordance with a content signal to obtain the modulated carrier.* We can say that a melody contour is a pitch-modulated (frequency-modulated) version of a carrier wave, where the carrier corresponds to the baseline pitch.

Figure 1: The tallest adult (7.9ft) and the shortest adult (2.4ft)

We speak using our instruments, i.e., vocal organs, not only by varying the above three parameters but also the most important parameter, called timbre or spectrum. From this viewpoint, it can be said that an utterance is generated by spectrum modulation [7]. The default shape and length of a vocal tube determines the speaker-dependent voice timbre and, by changing the shape dynamically, an utterance is produced as waveforms.

In a large number of previous studies of automatic speech recognition (ASR), the dynamic aspects of utterances were studied well and many dynamic features, such as modulation spectrum [8], RASTA (relative spectra) [9], spectro-tempral features [10], delta cepstrums [11], segment models [12] and so forth were proposed to improve the performance of ASR. In these studies, what they proposed are dynamic speech features for ASR. 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 *de*modulation?"

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, as a matter of course, he/she considers speech recognition as spectrum demodulation, which is a mathematical algorithm of removing the speaker-dependent voice timbre from an utterance and leaving its linguistic content to hearers intactly. FM generates a pitch contour and AM generates a power contour, both of which are often regarded as supra-segmental features. What about spectrum modulation? We consider that a spectrum (timbre) contour is another supra-segmental feature.

This paper argues that the method that we proposed in [4, 5] is a good candidate answer to the above question of spectrum *de*modulation and that this answer has good validity with re-

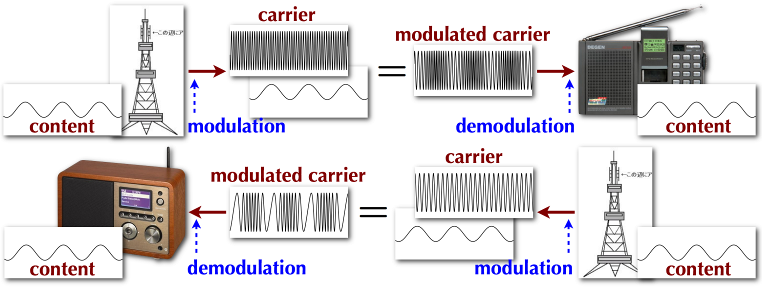

Figure 2: Frequency modulation and frequency demodulation

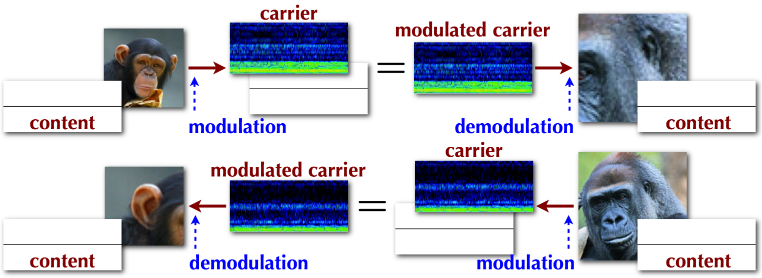

Figure 4: Spectrum modulation and demodulation with no content

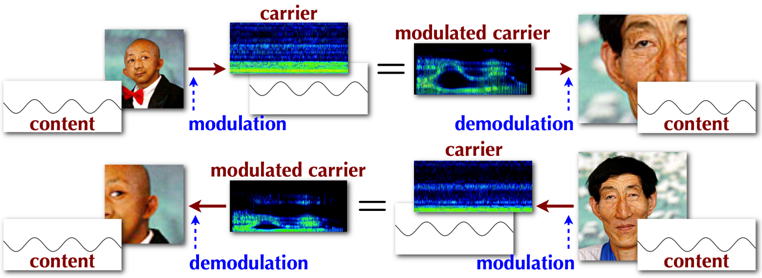

Figure 5: Spectrum modulation and demodulation with contents

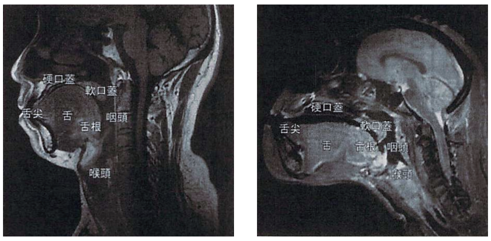

Figure 3: The vocal organs of a human and an ape [14]

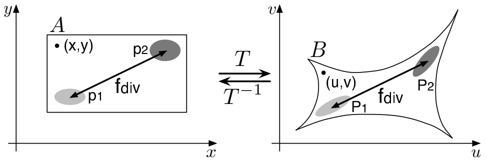

Figure 6: Transform-invariance of f-divergence

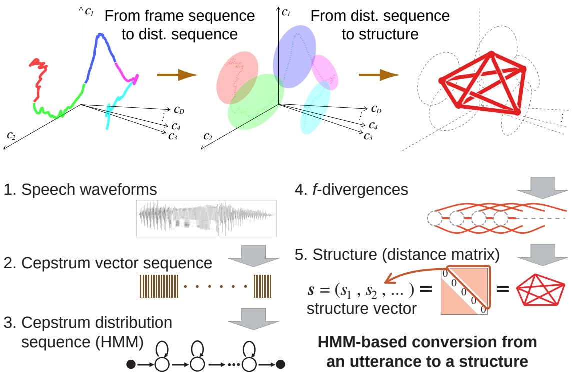
From frame sequence to dist. sequence
From dist. sequence to structure
1. Speech waveforms
2. Cepstrum vector sequence
3. Cepstrum distribution sequence (HMM)
4. f-divergences
5. Structure (distance matrix)
s = (s1 , s2 , ...) =
structure vector
HMM-based conversion from an utterance to a structure

Figure 7: Transform-invariant shape of an utterance

spect 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 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 *de*modulator exists only in the human organs of audition. Some findings in anthropology are shown, which indicate that apes don't have good *de*modulators.

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[1] 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.

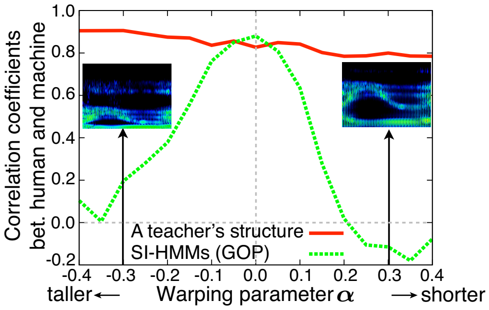

Figure 8: Structure-based assessment of pronunciation

We already applied this structural representation to speech recognition [24, 25], pronunciation assessment [26], dialect-based speaker classification [27], and speech synthesis [28]. In [24, 25], although the recognition task was relatively small, a speaker-independent speech recognizer was built only with several training speakers and without any explicit normalization or adaptation. It should be noted that our proposal is not for normalization but for removing carrier information, i.e., speaker information. In [26], a pronunciation structure was built from a male teacher, which was compared to those of students. Students of different vocal tract length were artificially prepared by using frequency warping [22]. Figure 8 shows the correlation between human and machine assessment. Even if speaker-independent HMMs were used to calculate GOP (Goodness Of Pronunciation) scores [29], the correlation easily dropped when a large mismatch existed between training and testing conditions. In this case, however, a pronunciation structure even of a single teacher was effectively used to assess those of students with no explicit normalization or adaptation. We claimed in [26] that GOP should stand for Goodness Of imPersonation.

A speech structure extracted from an utterance is a compact representation of the modulation pattern or the content in that utterance. In [28], the speaker-dependent voice timbre, i.e., carrier information, is given to a speech structure, i.e., a content, to generate its corresponding waveforms by modulating the spectrum. We call this framework structure-to-speech conversion.

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.

[1] $f_{div}(p_1, p_2) = \oint p_2(\boldsymbol{x}) \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} d\boldsymbol{x} = f_{div}(T(p_1), T(p_2))$

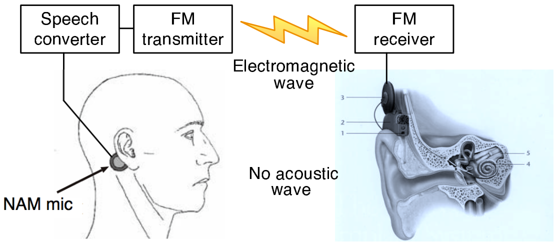

Figure 9: Wireless communication with hidden devices

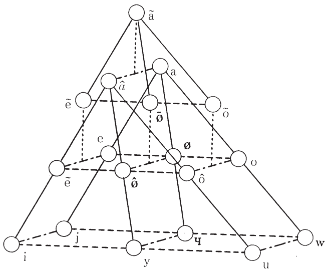

Figure 10: Jakobson's invariant shape of French vowels

The physical phenomena that they observed about the human communication were the spectrograms of electromagnetic waves because no acoustic wave was found. After observing different pairs of laryngectomized and deaf individuals, the anthropologists found a surprising fact that very different spectrogram patterns caused the same effect, i.e., the invariance and variability problem. We can explain easily that different sender-receiver pairs used different carrier frequencies. But the anthropologists did not know the mechanism of the communication. What kind of solution did they derive? A possible solution, which they considered reasonably naive, was to calculate a statistical model of the various spectrogram patterns causing the same effect, after collecting a huge number of samples. The anthropologists pondered well whether to take this solution.

4.2. Speech scientists and engineers on Earth

A long time ago, a kind of apes had stood up and begun 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 speech sounds. Figure 10 shows his invariant shape of French vowels and semi-vowels. We understand that our proposal is a computational implementation of his theory and we can claim that our proposal has high linguistic validity.

5. References

[1] J. S. Perkell and D. H. Klatt, *Invariance and variability in speech processes,* Lawrence Erlbaum Associates, Inc., 1986.

[2] R. Newman, "The level of detail in infants' lexical representations and its implications for computational models," Keynote speech in Workshop on Acquisition of Communication and Recognition Skills (ACORNS), 2009.

[3] S. Furui, "Generalization problem in ASR acoustic model training and adaptation," Keynote speech in IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2009.

[4] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," *Proc. ICASSP*, 889–892, 2005.

[5] Y. Qiao *et al.*,"A study on invariance of f-divergence and its application to speech recognition," *IEEE Transactions on Signal Processing*, 58, 2010 (to appear).

[6] http://en.wikipedia.org/wiki/Modulation

[7] S. K. Scott, "The neural basis of speech perception – a view from functional imaging," *Proc. INTERSPEECH*, 10–13, Keynote speech, 2007.

[8] S. Greenberg *et al.*,"The modulation spectrogram: in pursuit of an invariant representation of speech," *Proc. ICASSP*, 1647–1650, 1997.

[9] H. Hermansky *et al.*,"RASTA processing of speech," *IEEE Trans. SAP*, 2, 4, pp.578–589, 1994.

[10] "Special Session: Auditory-inspired spectro-temporal features," *Proc. INTERSPEECH*, 2008 (for example).

[11] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. ASSP*, 29, 3, 342–350, 1981.

[12] M. Ostendorf *et al.*,"From HMMs to segment models: a unified view of stochastic modeling for speech recognition," *IEEE Trans. on SAP*, 4, 5, 360–378, 1996.

[13] O. Sacks, *An anthropologist on Mars,* Vintage, 1996.

[14] S. Hayama, *The birth of human beings,* PHP Shinsho, 1999.

[15] H. Takemoto, "Morphological analyses and 3D modeling of the tongue musculature of the Chimpanzee," *American Journal of Primatology*, 70, 966–975, 2008.

[16] M. D. Hauser *et al.*, "The evolution of the music faculty: a comparative perspective," *Nature neurosciences*, 6, 663–668, 2003.

[17] K. Okanoya, "Birdsongs and human language: common evolutionary mechanisms," *Proc. Spring Meet. Acoust. Soc. Jpn.,* 1-17-5, 1555–1556, 2008.

[18] M. Kato, "Phonological development and its disorders," *J. Communication Disorders,* 20, 2, 84–85, 2003.

[19] S. E. Shaywitz, *Overcoming dyslexia,* Random House, 2005.

[20] M. Hayakawa, "Language acquisition and matherese," *Language*, 35, 9, 62–67, Taishukan pub., 2006.

[21] P. Lieberman, "On the development of vowel production in young children," in *Child Phonology vol.1,* Academic Press, 1980.

[22] M. Pitz *et al.*, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. SAP*, 13, 5, 930–944, 2005.

[23] D. Saito *et al.*, "Directional dependency of cepstrum on vocal tract length," *Proc. ICASSP*, 4485–4488, 2008.

[24] N. Minematsu *et al.*, "Implementation of robust speech recognition by simulating infants' speech perception based on the invariant sound shape embedded in utterances," *Proc. Speech and Computer (SPECOM)*, 35–40, 2009.

[25] Y. Qiao *et al.*, "A study of Hidden Structure Model and its application of labeling sequences," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 118–123, 2009.

[26] M. Suzuki *et al.*, "Sub-structure-based estimation of pronunciation proficiency and classification of learners," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 574–579, 2009.

[27] X. Ma *et al.*, "Dialect-based speaker classification of Chinese using structural representation of pronunciation," *Proc. Speech and Computer (SPECOM)*, 350–355, 2009.

[28] D. Saito *et al.* "Structure to speech – speech generation based on infant-like vocal imitation –," *Proc. INTERSPEECH*, 1837–1840, 2008.

[29] S. M. Witt *et al.*, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech Communications, 30, 95–108, 2000.

[30] R. Jakobson *et al.*, *The sound shape of language,* Mouton De Gruyter, 1987.

[31] T. Toda *et al.*, "Technologies for processing body-conducted speech detected with non-audible murmur microphone," *Proc. INTERSPEECH*, 632–635, 2009.