A MODULATION-DEMODULATION MODEL FOR SPEECH COMMUNICATION AND ITS EMERGENCE

NOBUAKI MINEMATSU

Graduate School of Info. Sci. and Tech., The University of Tokyo, Japan,
mine@gavo.t.u-tokyo.ac.jp

Perceptual invariance against large acoustic variability in speech has been a long-discussed question in speech science and engineering (Perkell & Klatt, 2002), and it is still an open question (Newman, 2008; Furui, 2009). Recently, we proposed a candidate answer based on mathematically-guaranteed relational invariance (Minematsu et al., 2010; Qiao & Minematsu, 2010). Here, transform-invariant features, f-divergences, are extracted from the speech dynamics in an utterance to form an invariant topological shape which characterizes and represents the linguistic message conveyed in that utterance. In this paper, this representation is interpreted from a viewpoint of telecommunications, linguistics, and evolutionary anthropology. Speech production is often regarded as a process of modulating the baseline timbre of a speaker's voice by manipulating the vocal organs, i.e., spectrum modulation. Then, extraction of the linguistic message from an utterance can be viewed as a process of spectrum *de*modulation. This modulation-demodulation model of speech communication has a strong link to known morphological and cognitive differences between humans and apes.

1. Introduction

Many speech sounds can be represented as standing waves in a vocal tube and their acoustic properties depend on the shape and length of the tube. The process of producing vowel sounds is very similar to that of producing sounds with a wind instrument. A vocal tube can be regarded as an instrument and, by changing its shape dynamically, sounds of different timbre such as [ɑːɛːiːɔːuː] can be generated.

The length and shape of the vocal tube varies among speakers. This is why voice quality differs among them and one can identify a speaker by hearing his voice. Figure 1 shows the tallest adult and the shortest adult in the world. There must be a very large gap in voice quality between the two, but they were able to communicate orally with no trouble the first time they saw each other. Human speech communication is very robust to acoustic variability caused by speaker differences. This is a good example of invariance and variability in speech processes.

In telecommunications, a message is transmitted to receivers by changing one parameter of a carrier wave in relation to that message. A sinusoidal wave is often used as carrier. This transmission scheme is called modulation and in (Wikipedia, 2011) it is explained by using the performance of a musician as a metaphor.

Figure 1. The tallest and shortest adults

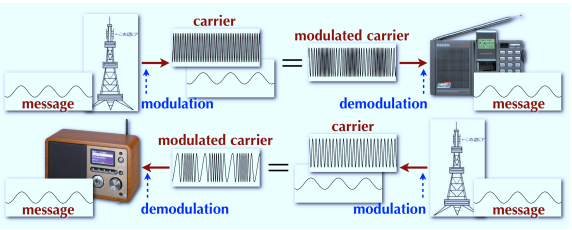

Figure 2. Frequency modulation and demodulation

> *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.

We speak using our instruments, i.e., vocal organs, by varying not only the above parameters, but also the most important parameter, called the timbre or spectrum envelope. From this viewpoint, it can be said that an utterance is generated by spectrum modulation (Scott, 2007). The default shape and length of a vocal tube determines speaker-dependent voice quality and, by changing the shape or modulating the spectrum envelope, an utterance is produced and transmitted.

In a large number of previous studies in automatic speech recognition (ASR), to bridge a gap between the ASR performance and the performance of human speech recognition (HSR), much attention was paid to the dynamic aspects of utterances, and many dynamic features were proposed (Greenberg & Kingsbury, 1997; Hermansky & Morgan, 1994; Furui, 1981; Ostendorf et al., 1996). Although these studies proposed new features for ASR, if one views speech production as spectrum modulation, he may point out that these proposals did not provide a good answer to a very fundamental question of ASR and HSR: "what is the algorithm for spectrum *de*modulation?"

In telecommunications, a transmitted message is received via demodulation. In (Wikipedia, 2011), demodulation is explained as a process of extracting the original message intact from a modulated carrier wave. In any form of modulation, AM, PM, or FM, methods exist that can be used to analytically extract the message component exclusively by removing the carrier component from a modulated carrier. Figure 2 illustrates the processes of transmitting and receiving a message via FM. The process of spectrum *de*modulation, which in this case refers to speech recognition, should therefore be realized by a mathematical algorithm that extracts the linguistic message exclusively by removing the speaker-dependent voice quality, i.e., removing speaker identity, from an utterance.

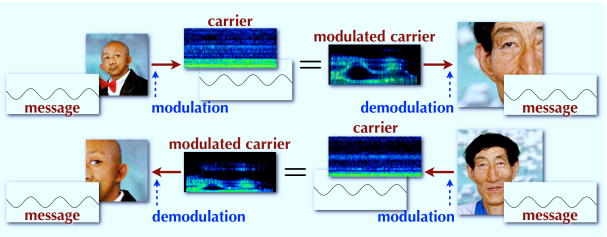

Figure 3. Spectrum modulation and demodulation with a message

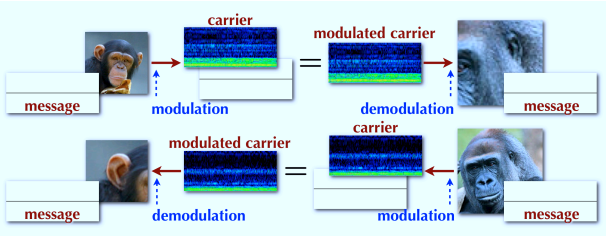

Figure 4. Spectrum modulation and demodulation with no message

This paper demonstrates that our proposal (Minematsu et al., 2010; Qiao & Minematsu, 2010) is a good candidate answer to the above question and that this answer is valid within the contexts of linguistics and evolutionary anthropology. Although this paper provides no new experimental results, we believe that it will provide a novel explanation of speech communication and its emergence.

2. Morphological and cognitive differences between humans and apes

2.1. *Morphological differences between humans and apes*

The proposed model regards a tongue as flexible modulator of the spectrum envelope. Here we discuss the morphological differences of the vocal organs between humans and apes (chimpanzees). Unlike humans, apes have two semi-independent tracts, one from the nose to the lung for breathing and the other from the mouth to the stomach for eating (Hayama, 1999). Apes can breathe and eat simultaneously but humans cannot. As a consequence, it seems reasonably difficult for apes to send an air flow directly from the lung to the vocal cavity. Further, apes have a much less flexible tongue (Takemoto, 2008). As a result, spectrum modulation is difficult. Why did humans gain this ability? One anthropological hypothesis that has been proposed is bipedal walking (Hayama, 1999), because of which the larynx dropped and the two tracts happened to develop an intersection.

Figure 3 illustrates the modulation-demodulation model of human speech communication. The two speakers have vocal tubes of different shape and length, and therefore their default voice timbre (carrier) is different. By dynamically ma-

nipulating the topology of the vocal tube, the timbre is modulated and a message is transmitted as a modulation pattern. Since receivers can efficiently demodulate an incoming modulated carrier, different body sizes do not cause problems.

Figure 4 depicts hypothetical voice communication between a small ape and a large ape for the purpose of illustrating our model. Like humans, they have their own unique voice timbre but unlike humans, it is reasonably difficult to embed messages in the voices via spectrum modulation. It is well-known that an ape can deliver some primitive messages to others using specialized calls. Our model does not deny their primitive vocal communication capabilities but assumes that their capabilities cannot be extended to a higher level of vocal communication, where thousands of different words are used and transmitted.

We have explained that a good and flexible timbre modulator is found only in humans' unique vocal tract morphology, but our model also claims that a good *de*modulator exists only in the human organs of hearing.

2.2. *Cognitive differences between humans and animals*

As explained in Section 1, a musical melody can be regarded as an FM version of a carrier wave. If we apply two different carriers to the same musical content, the result will be a melody and its transposition. From the two pieces, humans can extract the same musical content easily and this extraction is referred to as demodulation. But apes cannot perceive the equivalence between a melody and its transposed version (Hauser et al., 2003). It is difficult for apes to demodulate FM carriers. What about spectrum demodulation? Can they do this or not?

Human infants acquire spoken language via vocal imitation of the utterances of their parents. But they do not impersonate their parents. Their vocal imitation is not acoustic imitation. Vocal imitation is a trait that is rarely found in animals and humans are the only primates that exhibit it. We can find only a few other species that do, such as birds, dolphins and whales (Okanoya, 2008). But there exists a critical difference between the vocal imitation of humans and that of animals. Basically speaking, animals' imitation is acoustic (Okanoya, 2008). By interpreting this fact based on our model, we can say that animals imitate a modulated carrier, i.e., incoming signals themselves, not the message embedded in the carrier. It seems that animals also have difficulty in performing spectrum demodulation.

Which acoustic aspects of parents' utterances do infants imitate and which aspects do they ignore? One may assume that infants decompose the utterances into sequences of phonemes (text-like representation) and they realize each phoneme acoustically with their mouths. It seems however, that researchers of infant studies do not accept this assumption because infants do not have good phonemic awareness (Kato, 2003; Shaywitz, 2005). No infant acquires spoken language by reading text. In that case, what are they imitating? According to (Kato, 2003; Shaywitz, 2005; Hayakawa, 2006; Lieberman, 1980), infants are thought to extract holistic and speaker-independent speech patterns, called speech Gestalts, and

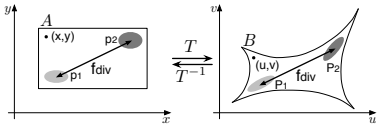

Figure 5. Transform-invariance of f-divergence

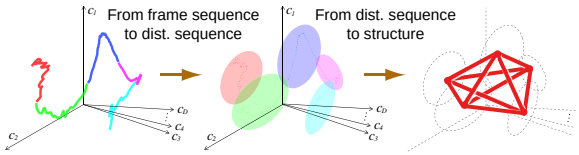

Figure 6. Transform-invariant shape of an utterance

they realize the patterns acoustically using their mouths. As far as we know, however, no researcher has proposed a physical definition for speech patterns that has the potential to lead to an algorithm of spectrum demodulation.

3. Implementation of the spectrum demodulation algorithm

3.1. *Removal of the speaker-dependent voice timbre*

Demodulation removes the carrier information and leaves the message intact. In (Minematsu et al., 2010; Qiao & Minematsu, 2010), we implemented this process on a computer using transform-invariant features. Speaker difference is often characterized as a transformation from one speaker's voice space to another's. This indicates that if one can represent an utterance using only transform-invariant features, that representation will contain no speaker-dependent features.

In (Qiao & Minematsu, 2010), we proved that f-divergence[a] is invariant with any kind of invertible and differentiable transform (sufficiency) and that the features invariant with any kind of transform have to be a function of f-divergence (necessity). Figure 5 illustrates the invariance of f-divergence and any speech event has to be characterized not as point but as distribution in a feature space.

3.2. *Computational implementation of speech Gestalts*

By representing an utterance only with f-divergence, we can obtain the speaker-independent speech pattern, i.e., the speech Gestalt. Figure 6 shows the extraction procedure. A speech trajectory in a feature space is converted into a sequence of distributions. Between every distribution pair, the f-divergence is calculated to form a distance matrix, which can specify a unique geometrical shape. This matrix is the mathematical expression for the speech Gestalt.

We have already applied this holistic representation to speech recognition (Minematsu et al., 2010; Qiao & Minematsu, 2010), pronunciation assessment (Suzuki et al., 2009) and dialect-based speaker clustering (Ma et al., 2009). If readers are interested in how effective the proposed holistic representation is at removing speaker identity from speech acoustics and extract the linguistic content exclusively, please refer to the above papers. Under the proposed method, two

[a] $f_{div}(p_1, p_2) \equiv \oint p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x} = f_{div}(T(p_1), T(p_2))$. Please refer to Figure 5.

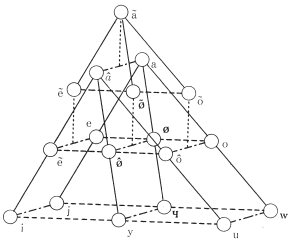

Figure 7. Jakobson's invariant system

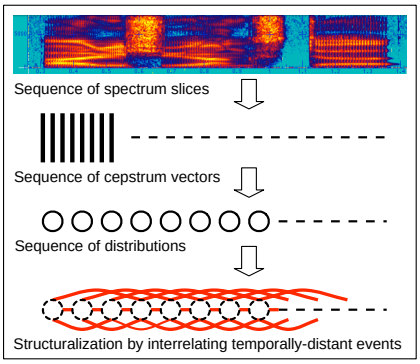

Figure 8. Extraction of a speech Gestalt

simulated utterances of the same message generated by the tallest man and the shortest man appear identical. Explicit model adaptation or feature normalization which are essential for the conventional ASR framework is not needed.

4. Discussion and conclusions

Similar claims are found in classical linguistics (Jakobson & Waugh, 1987; Ladefoged & Broadbent, 1957) and a recent study of speech science (Plummer et al., 2010). Figure 7 shows Jakobson's system of French vowels and semi-vowels and he wrote "we have to put aside the accidental properties of individual sounds and substitute a general expression that is the common denominator of these variables." In these studies however, no algorithm was proposed to remove the speaker-dependent features to extract an embedded message[b]. Further, speaker-independence (invariance) was not proved mathematically in a conclusive fashion.

Figure 8 shows the procedure of extracting a speaker-invariant pattern again. This figure and Figure 6 explain the same procedure from different viewpoints. A feature vector in Figure 8 corresponds to a point in the feature space in Figure 6. Clearly shown in Figure 8, invariant features exist as speech contrasts, which include not only local contrasts but also long-distant contrasts. As is well-known, all the textbooks on acoustic phonetics explain formant frequencies or spectrum envelopes as acoustic correlates with phonemes, i.e., linguistic messages. But it is also well-known that these features are strongly speaker-dependent. Our modulation-demodulation model claims that a linguistic message is transmitted mainly by speech contrasts, not just by local speech features observed at each time index. In Section 2.2, we explained apes' inability to perform demodulation, but apes can identify individuals very accurately by hearing the voices of the individuals. Considering these facts, for apes, formant frequencies and spectrum

[b]It should be noted that the method proposed in this paper aims not at normalizing speaker differences but at removing speaker identity from speech acoustics. As spectrum smoothing can remove pitch information from speech, our method can remove speaker information from speech.

envelopes are acoustic correlates with individuals and different acoustic features indicate different information (individuals). We speculate that it is through the acquisition of demodulation abilities that humans became able to extract the same message consistently from acoustically different sound streams. We argue that this invariant information is what we call language today.

In Section 2.2, we focused on the difference in vocal imitation between animals and humans. Although the following description may be out of the scope of this paper, we also found that in some cases acoustic imitation does become the default strategy for humans. Examples can be found in severely impaired autistics (Willey et al., 1999) and, in this case, the normal acquisition of speech communication often becomes difficult. Prof. Grandin, a professor of animal sciences, who is herself autistic, described the similarity in information processing between animals and autistics (Grandin et al., 2004). Autistics tend to memorize the details of input stimuli. An autistic boy wrote that he could understand what his mother was saying but that it was difficult for him to understand others (Higashida et al., 2005). He may not have a normal capability for demodulation.

In this paper, a modulation-demodulation model of speech communication is proposed to explain humans' capability for robust communication. This model however, claims nothing regarding the emergence of the syntactic structure of language, and it claims nothing either regarding why invariant speech patterns, which are merely topological patterns found in air particle vibrations around listeners' ears, can have very strong links with memory, i.e., what is stored as a topological pattern or cell assembly in the neural circuits of the listeners. But we believe that the ability to perform spectrum modulation and demodulation is at least a required condition for the emergence and development of spoken language.

References

Furui, S. (1981). Comparison of speaker recognition methods using statistical features and dynamic features. *IEEE Trans. ASSP*, *29*(3), 342-350.

Furui, S. (2009). Generalization problem in asr acoustic model training and adaptation. *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*.

Grandin, T., et al.. (2004). *Animals in translation.* Scribner.

Greenberg, S., & Kingsbury, B. (1997). The modulation spectrogram: in pursuit of an invariant representation of speech. *Proc. ICASSP*, 1647-1650.

Hauser, M. D., et al.. (2003). The evolution of the music faculty: a comparative perspective. *Nature neurosciences*, *6*, 663-668.

Hayakawa, M. (2006). Language acquisition and matherese. *Language*, *35*(9), 62-67.

Hayama, S. (1999). *The birth of human beings.* PHP Shinsho.

Hermansky, H., & Morgan, N. (1994). Rasta processing of speech. *IEEE Trans. SAP*, *2*(4), 578-589.

Higashida, N., et al.. (2005). *Messages to all my colleagues living on the planet.* Escor Pub.

Jakobson, R., & Waugh, L. R. (1987). *The sound shape of language.* Mouton De Gruyter.

Kato, M. (2003). Phonological development and its disorders. *J. Communication Disorders*, *20*(2), 84-85.

Ladefoged, P., & Broadbent, D. (1957). Information conveyed by vowels. *J. Acoust. Soc. Am.*, *29*, 98-104.

Lieberman, P. (1980). On the development of vowel production in young children. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology vol.1.* Academic Press.

Ma, X., et al.. (2009). Dialect-based speaker classification of chinese using structural representation of pronunciation. *Proc. Speech and Computer (SPECOM)*, 350-355.

Minematsu, N., et al.. (2010). Speech structure and its application to robust speech processing. *Journal of New Generation Computing*, *28*(3), 299-319.

Newman, R. S. (2008). The level of detail in infants' word learning. *Current directions in psychological science*, *17*(3), 229-232.

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

Ostendorf, M., et al.. (1996). From hmms to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Trans. on SAP*, *4*(5), 360-378.

Perkell, J. S., & Klatt, D. H. (2002). *Invariance and variability in speech processes.* Lawrence Erlbaum Associates, Inc.

Plummer, A. R., et al.. (2010). Learning speaker normalization using semisupervised manifold alignment. *Proc. INTERSPEECH*, 2918-2921.

Qiao, Y., & Minematsu, N. (2010). A study on invariance of f-divergence and its application to speech recognition. *IEEE Trans. Signal Processing*, *58*(7), 3884-3890.

Scott, S. K. (2007). The neural basis of speech perception – a view from functional imaging. *Proc. INTERSPEECH*, 10-13.

Shaywitz, S. E. (2005). *Overcoming dyslexia.* Random House.

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

Takemoto, H. (2008). Morphological analyses and 3d modeling of the tongue musculature of the chimpanzee. *American Journal of Primatology*, *70*(10), 966-975.

Wikipedia. (2011). http://en.wikipedia.org/wiki/Modulation.

Willey, L. H., et al.. (1999). *Pretending to be normal: living with asperger's syndrome.* Jessica Kingsley Publishers.