HUMAN SPEECH MODEL BASED ON INFORMATION SEPARATION — COLLECTION OR SEPARATION, THAT IS THE QUESTION. —

Nobuaki Minematsu

Graduate School of Information Science and Technology, The University of Tokyo
Email-Address : mine@gavo.t.u-tokyo.ac.jp

Abstract: This paper points out that no existing technically-implemented speech model is adequate enough to describe one of the most fundamental and unique capacities of human speech processing. Language acquisition of infants is based on vocal imitation [1] but they don't impersonate their parents and imitate only the linguistic and para-linguistic aspects of the parents' utterances. The vocal imitation is found only in a few species of animals: birds, dolphins, and whales, but their imitation is acoustic imitation [2]. How to represent exclusively what in the utterances human infants imitate? An adequate speech model for it should be independent of the extra-linguistic features and represents only the linguistic and paralinguistc aspects. We already proposed a new speech representation [3], called speech structure, which is proved mathematically to be invariant with any kind of transformation. Its extremely high independence of speaker differences was shown experimentally [4, 5, 6]. In this paper, by reviewing studies of evolutionary anthropology and those of language disorders, we discuss the theoretical validity of the new model to describe the human-unique capacity of speech processing.

1 Introduction

To build speech systems for speech recognition and/or speech synthesis, speech features are always extracted from speech waveforms and are used to realize these functions on machines. One of the most fundamental speech features is the spectrum envelope, which is compactly represented using cepstrum coefficients. Since the auditory characteristics of an ear are insensitive to changes of phase characteristics of a sound, scientists and engineers often focus only on (logarithmic) amplitude characteristics of speech. Further, considering the acoustic properties of producing speech sounds using the vocal cords and the vocal tract, the amplitude characteristics are often divided into two: the characteristics of source and filter, shown in Figure 1. Although the spectrum envelope is extracted after two steps of information separation, it is still easily affected by linguistic factors, para-linguistic factors, and extra-linguistic factors.

Take linguistic and extra-linguistic factors here for example. Two words are produced acoustically by a speaker using different articulatory movements. Then, the two words come to show two different temporal patterns of the spectrum envelope. The same word is produced by two speakers, who are supposed to have the different length and shape of the vocal tract. Also in this case, the two utterances come to show two different temporal patterns of the spectrum envelope. If one wants to build a speaker-independent word recognizer, the simplest way is to collect a large number of samples from different speakers for each word in the vocabulary and calculate a statistical model, such as HMM, for each *w*ord: $P(o|w)$. Similarly, the simplest way to built a text-independent speaker recognizer is to collect a large number of speech samples of different words for each speaker and calculate a statistical model, such as GMM, for each *s*peaker: $P(o|s)$. Since o heavily depends both on w and s, $P(o|w)$ and $P(o|s)$ should be rewritten as

$$P(o|w) = \sum_s P(o,s|w) = \sum_s P(o|w,s)P(s|w) \approx \sum_s P(o|w,s)P(s) \tag{1}$$

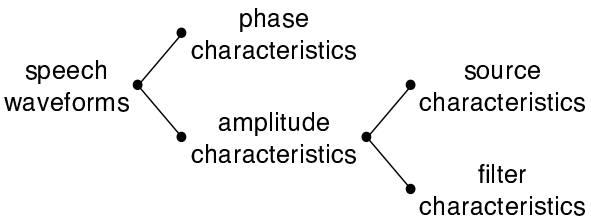

Figure 1 - Two steps of information separation

$$P(o|s) = \sum_w P(o,w|s) = \sum_w P(o|w,s)P(w|s) \approx \sum_w P(o|w,s)P(w). \quad (2)$$

Here, w and s are considered to be independent. Both equations are expectation operations of $P(o|w,s)$ but with regard to different variables: s and w. Expectation is a very useful tool to hide variables that are irrelevant to the target function, but it requires an extensive collection.

If speech feature o can be further separated into o_w and o_s, which are features of words and those of speakers, respectively, the collection is not required and $P(o_w|w)$ and $P(o_s|s)$ may be able to make word recognition and speaker recognition even simpler.

In this paper, we focus on Equation (1). By reviewing studies of evolutionary anthropology and those of language disorders, we discuss that this collection (expectation) approach may provide us with only a speech model of not normally developed individuals but severely impaired autistics, or a model of animals. In either case, normal acquisition of speech communication becomes difficult. How to build a human speech model with normal development? We claim that we have to add yet another information separation ($o \rightarrow o_w + o_s$) to Figure 1 and discuss the validity of our recently proposed model [3] to this aim.

2 Infant behaviors in language acquisition

Why is an extensive collection of samples required? The reason is simple and it is because o in $P(o|w)$ is heavily dependent also on s. For example, IBM once announced for advertisement that IBM collected speech samples from 350 thousands of speakers to build its ASR engine.

Every normally developed individual shows an extremely robust capacity for understanding spoken language. How does an infant acquire it? One obvious fact is that a majority of the utterances an infant hears come from its parents. After it begins to talk, about a half of the utterances it hears are its own. It can be claimed that the utterances an individual hears are strongly speaker-biased unless he or she has speaking disabilities. With this obvious fact, we claim that speech processing based on Equation (1) is unnatural and far from the human strategy.

Even with so-called speaker-independent HMMs, they are often adapted acoustically to new speakers to minimize acoustic mismatch between training and testing conditions. Basically speaking, in the current framework of speech recognition, the degree of *linguistic* equivalence (similarity) between two utterances is measured quantitatively through *acoustic* equivalence. Looking at the behaviors of infants in language acquisition, however, we can claim easily that this machine strategy seems reasonably unnatural and weird again.

Infants acquire language through active imitation of their parents' utterances called vocal imitation [1]. In this process, it is obvious that infants do not try to produce utterances acoustically matched with their parents' utterances. They do not impersonate their parents. They are very insensitive to extra-linguistic differences in language acquisition. A question is raised: how to represent what in the utterances human infants imitate? Researchers often represent it using a sequence of phoneme-like units [7], which certainly does not include any extra-linguistic features. Does this mean that, in the process of vocal imitation, infants represent a given utterance using phonemes and then convert each phoneme back to a sound? We can say "No" because in-

fants have no good phonemic awareness [8, 9]. Writing, including phonemic symbols, is merely a way of recording language by visible marks [10] and the capacity of using these marks requires children to learn written language, especially alphabets, for more than several years [11].

How then to represent *acoustically* what in the utterances human infants try to reproduce? What kind of *acoustic pattern* is underlying commonly between the utterance of a parent and the imitative response of an infant? This acoustic pattern has to be truly speaker-independent, different from so-called speaker-independent HMMs, which are often modified to new speakers. Some researchers of infant study explain this pattern using the terms of *holistic word form* [8], *word Gestalt* [12], and *related spectrum pattern* [13]. However, we could not find any mathematical formula for them. If a mathematical formula is made possible, $P(o_w|w)$ can be modeled with o_s excluded. Without this model, however, an extensive collection and acoustic match are always required although neither of them is needed for humans.

As far as we know, in some cases, the vocal imitation becomes like impersonation, where every aspect of a given utterance is to be reproduced. o in $P(o|w,s)$, not o_w in $P(o_w|w)$, is the target of imitation. This performance is found in severely impaired autistics [14, 15, 16], who have much difficulty in normal acquisition of speech communication. In [17], an autistic boy wrote that he could understand what his mother was saying but it was difficult for him to understand others. His mother said that it also seemed even difficult for him to understand her on a telephone line. We consider that $P(o_w|w)$ is a requisite model to realize the human-like processing on machines.

3 Cognitive differences between animals and humans

The performance of vocal imitation is rarely found in animals. Non-human primates do not perform it [18]. Only a few species do, such as birds, dolphins and whales [2]. But there exists a critical difference between the vocal imitation of humans and that of animals. Basically speaking, animals' imitation is acoustic imitation like impersonation [2]. A similar finding was obtained in studies of evolutionary anthropology [19]. Although humans easily perceive the equivalence between a melody and its transposed version, it was found to be difficult for monkeys to perceive it. It seems adequate to claim that animals keep absolute acoustic properties of input stimuli in memory and use them to judge to which one of the old stimuli a new stimulus is identical [20]. In other words, unlike humans, acoustic match is always required for equivalence. Temple Grandin, a professor of animal sciences who is herself autistic, described the similarity in information processing between animals and autistics [21]. The basic strategy of memory is to store every detailed aspect of incoming stimuli as it is in the brain.

Humans with normal development can perceive the equivalence between two messages conveyed by two speech streams not based on acoustic equivalence, not based on string-based equivalence but based on some other equivalence. By referring to a term used in developmental psychology, this equivalence should be called as Gestalt-based equivalence and the degree of this equivalence is considered to be able to be calculated quantitatively using $P(o_w|w)$. Without this model, the performance of a resulting system will have to resemble that of animals or severely impaired autistics. For example, as far as we know, almost all the types of speech synthesizers reproduce the voices of a training speaker acoustically and precisely. This is why possibility of using speech synthesizers to deceive a speaker verification system is discussed [22]. We consider these products not as human simulators but as parrot simulators.

4 Nature of perceptual constancy

Our perception is not only robust against speech variability but also against variability in other media. Psychologically speaking, robustness of perception is called perceptual constancy. Psy-

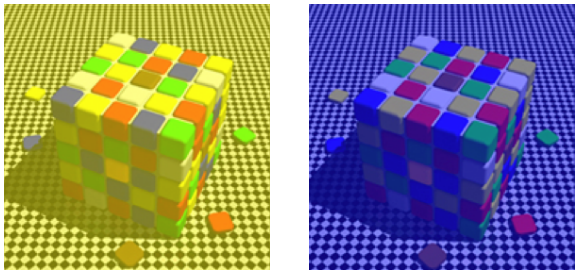

Figure 2 - The same Rubik's cube seen with two colored glasses

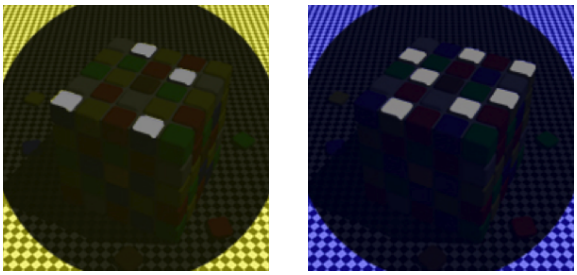

Figure 3 - Perception of colors without context [25]

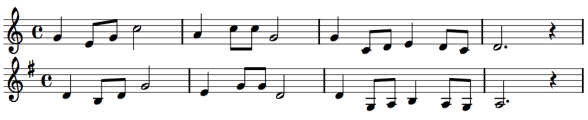

Figure 4 - A melody and its transposition

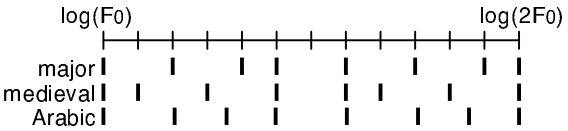

Figure 5 - Variation in tonal arrangement

chologists have discovered that, among different media, a similar mechanism functions [23, 24].

Figure 2 shows the appearance of the same Rubik's cube seen through differently colored glasses [25]. Although the corresponding tiles of the two cubes have objectively different colors, we label them identically. On the other hand, although we see four blue tiles on the top of the left cube and seven yellow tiles on the right cube, when their surrounding tiles are hidden, we suddenly realize that they have the same color, shown in Figure 3. We have to admit that different colors are perceived as identical and identical colors are perceived as different.

Similar phenomena can be easily found in tone perception. Figure 4 shows a melody and its transposed version. If listeners have relative pitch and can transcribe these melodies, they describe them as the same sequence of syllable names: So Mi So Do La Do Do So. The first tone of the upper sequence and that of the lower are different in fundamental frequency but listeners can name these tones as So. The first tone of the upper and the fourth of the lower are physically identical but the two tones are identified as different. Different tones are perceived as identical and identical tones are perceived as different. Similar to color perception, if a tone is presented without any surrounding tones, syllable name identification becomes impossible.

Both in colors and tones, it can be claimed that we perceive the value of each stimulus based on its relations to the surrounding stimuli and that perceptual constancy is realized because the relations are invariant to variability. As described in Section 3, perceptual constancy of tones is not observed even in monkeys but that of colors is found in animals including insects [26].

Figure 5 shows the scale of the major and this tonal arrangement is key-independent. Robust syllable name identification is possible due to this key-independence. However, this arrangement pattern can be changed for historical and regional reasons. In Figure 5, two other scales are also shown. One is a scale used in the medieval church music and the other is the Arabic scale. If a modern western melody is played with these scales, syllable name identification becomes difficult when listeners are not familiar with these scales[1]. It is the case with language.

Figure 6 shows variation in vowel arrangement (in part) of several regional accents of American English [27] by using the F_1/F_2- based vowel chart after vocal tract length normalization. Within a region, a fixed vowel arrangement pattern is observed independently of speakers. In other words, infants acquire not individual sounds acoustically in given utterances but the sound system underlying those utterances. It is natural that two speakers of different regional accents

[1]The author recommends readers to try the following two melodies: a western melody and its Arabic version.
http://www.gavo.t.u-tokyo.ac.jp/~mine/material/western.wav
http://www.gavo.t.u-tokyo.ac.jp/~mine/material/arabic.wav

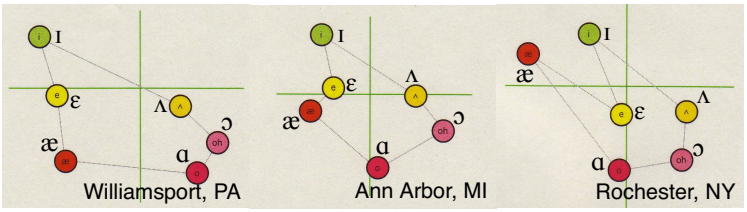

Figure 6 - Variation in American English vowel arrangement [27]

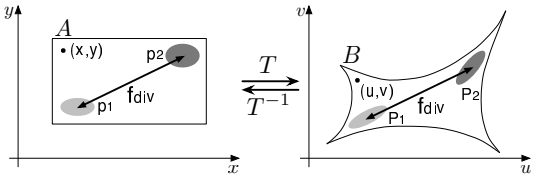

Figure 7 - Invariance of f-divergence

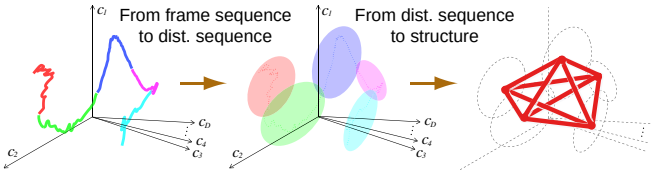

Figure 8 - Utterance to structure conversion

experience miscommunications if they are unfamiliar with the sound arrangement of each other. In classical studies of speech science, robust speech perception was discussed theoretically and experimentally based on invariant relational features among speech sounds [28, 29]. However, we could not find any mathematical formula to define the relations that are invariant against a very good variety of transformations representing speaker and environment differences.

5 Mathematical and technical solution

In [30], we proved that f-divergence[2] is invariant with any kind of invertible and differentiable transform (sufficiency) and that features invariant with any kind of transform, if any, have to be f-divergence (necessity). As shown in Figure 7, every event in a feature space has to be represented not as point but as distribution. By representing an utterance only with f-divergences, we can derive an invariant word Gestalt mathematically. Figure 8 shows its calculation procedure. A speech trajectory in a feature space is converted into a sequence of distributions. In [4], this process was implemented by applying the HMM training procedure. Between every distribution pair, f-divergence is calculated to form a distance matrix. We 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. [31] shows that this warping can be represented in the cepstrum domain as multiplication of a specific type of matrix by a cepstrum vector. In [32], we signified mathematically that this matrix is approximated as rotation matrix and demonstrated that the change of vocal tract length rotates a speech trajectory well. Directional components of the trajectory are strongly dependent on the speaker size. This is why we extract only scalar features in Figure 8.

We already applied this structural representation to speech recognition [4, 30], pronunciation assessment [6], dialect-based speaker classification [5], and speech synthesis [33]. In [4, 30], although the recognition task was small and artificial, a truly speaker-independent speech (word) 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 normalizing extra-linguistic features but for removing them, i.e., information separation illustrated in Figure 1.

In [6], a pronunciation structure was built from read sentences of a male teacher of American English. With the utterances, speaker-dependent phoneme HMMs were built. After select-

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

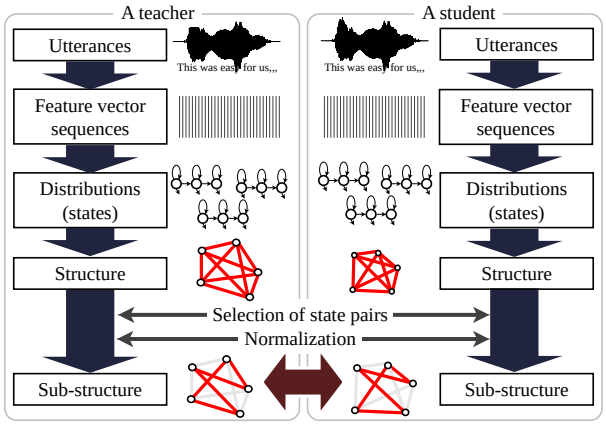

Figure 9 - State-based sub-structure

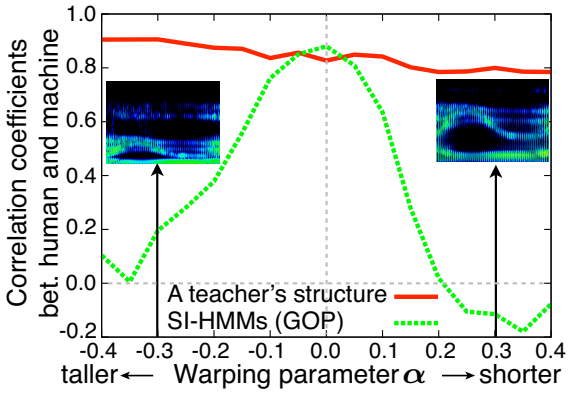

Figure 10 - Pronunciation assessment

ing an adequate set of states, a state-based structure (distance matrix) was calculated for that teacher (See Figure 9). From 26 Japanese students, their pronunciation structures were also estimated using the same set of states. By geometrically comparing the structure of the male teacher to that of each student, the pronunciation proficiency was estimated automatically. To test the robustness of pronunciation structure, the utterances of students were modified through frequency warping [31] to simulate extremely tall and short students. For comparison, GOP (Goodness Of Pronunciation) [34], which is a widely-used technique to assess the pronunciation using posterior probability of intended phonemes was also tested. Figure 10 shows the correlations between human assessment and machine assessment using structure and GOP. Although speaker-independent HMMs were used to calculate GOP scores, the correlation easily drops when a mismatch exists between training and testing conditions. However, a pronunciation structure even of a male teacher can be effectively used to evaluate students of any size.

Structural (geometrical) comparison between any pair of students leads to a distance matrix of the entire students, which shows distance between any pair of the students. With this matrix, they can be classified based on bottom-up clustering. Using a different definition of distance between two students, a different classification result is possible. Here, a new distance matrix was obtained by calculating acoustic distances of the corresponding states between any pair of the students. The former distance matrix is based on structural (relational) comparison between students and the latter is based on spectral (absolute) comparison between them. Figure 11 shows the structural classification and Figure 12 shows the spectral classification. Ward's method was used for bottom-up clustering. Alphabets mean 26 students, who are 14 males (black) and 12 females (red). $\overline{\mathrm{X}}$ and $\underline{\mathrm{X}}$ are taller (α=-0.3) and shorter (α=0.3) versions of student X, respectively. Clearly shown in these figures, the former is purely (linguistic) pronunciation classification with (extra-linguistic) speaker differences ignored and the latter is purely (extra-linguistic) speaker classification with (linguistic) pronunciation differences ignored. We can claim that information separation between the linguistic and extra-linguistic aspects is successfully realized.

6 Discussion and conclusions

Humans are insensitive to phase characteristics of a sound when hearing it. Humans are also insensitive to extra-linguistic features of utterances when acquiring spoken language. The former insensitivity was already technically realized but the latter seems not yet because researchers do not have a good model for that. Observation of the behaviors of animals and severely impaired autistics led us to consider that this insensitivity is one of the most fundamental and unique capacities of normally developed humans. In the conventional framework of speech processing,

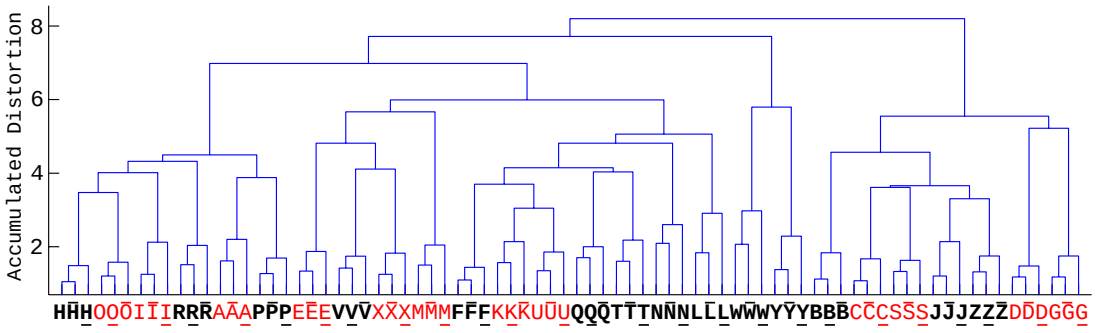

Figure 11 - Classification of Japanese students of the three sizes based on structural (relational) features

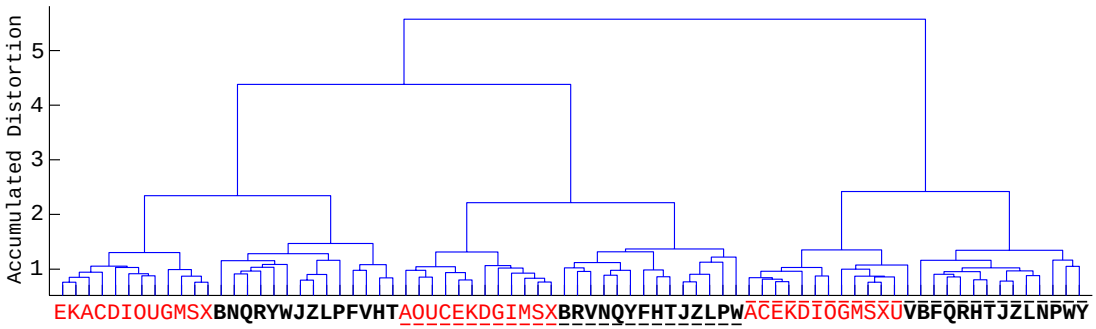

Figure 12 - Classification of Japanese students of the three sizes based on spectral (absolute) features

instead of pursuing a good model, for example, a statistical model of $P(o|w)$ was created by collection and expectation. In this paper, although no new experimental result was provided, we clearly pointed out that, without a good model, the performance of a resulting system has to resemble that of animals. Many speech synthesizers can be regarded as parrot simulators. If well-adapted HMMs are used for GOP, it can calculate scores highly correlated with human scores. In this case, however, what is assessed is not pronunciation but impersonation. Practically speaking, these systems will function well if good conditions are always prepared in advance. If one wants to develop not only outwardly appearing but also internally human-like speech systems, however, we believe that he/she has to pursue a good model for information separation. Certainly, collection and expectation is a powerful mathematical tool but careful consideration should be made on how to combine collection and separation. We can hardly claim that our proposal is the best or only solution but can claim that the speech community has to make good efforts to find a good human speech model.

References

[1] P.K. Kuhl, "Early language acquisition: Cracking the speech code," *Nature Reviews Neuroscience*, 5, 831–843, 2004

[2] K. Okanoya, "Birdsongs and human language: common evolutionary mechanisms," *Proc. Spring Meet. Acoust. Soc. Jpn.*, 1-17-5, 1555–1556, 2008 (including Q&A after his presentation)

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

[4] 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*, 35–40, 2009

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

[6] M. Suzuki *et al.*, "Sub-structure-based estimation of pronunciation proficiency and classification of learners," *Proc. ASRU*, 574–579, 2009

[7] T. Kobayashi *et al.*, "First 50 words in Japanese children: evidence from a web diary method," *Proc. European Conference on Developmental Psychology*, 2009

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

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

[10] L. Bloomfield, *Language,* New York: Henry Holt, 1933

[11] R. Port, "How are words stored in memory? Beyond phones and phonemes," *New Ideas in Psychology*, 25, 143–170, 2007

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

[13] P. Lieberman, "On the development of vowel production in young children," *Child Phonology*, edited by G. H. Yeni-Komshian, J. F. Kavanagh, and C. A. Ferguson, Academic Press, 1980

[14] K. Fukami, *A book of Hiroshi, vol.5*, Nakagawa Pub., 2006

[15] R. Martin, *Out of silence: an autistic boy's journey into language and communication,* Penguin, 1995

[16] U. Frith, *Autism: explaining the enigma,* Wiley-Blackwell, 2005

[17] N. Higashida *et al.*, *Messages to all my colleagues living on the planet*, Escor Pub., 2005

[18] W. Gruhn, "The audio-vocal system in sound perception and learning of language and music," *Proc. Int. Conf. on language and music as cognitive systems*, 2006

[19] M.D. Hauser and J. McDermott, "The evolution of the music faculty: a comparative perspective," *Nature neurosciences*, 6, 663–668, 2003

[20] D.J. Levitin *et al.*, "Absolute pitch: perception, coding, and controversies," *Trends in Cognitive Sciences*, 9, 1, 26–33, 2005

[21] T. Grandin *et al.*, *Animals in translation: using the mysteries of autism to decode animal behavior,* Scribner, 2004

[22] T. Masuko *et al.*, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," *Proc. INTERSPEECH*, 302–305, 2000

[23] R.B. Lotto *et al.*, "The effects of color on brightness," *Nature neuroscience*, 2, 11, 1010–1014, 1999

[24] T. Taniguchi, *Sounds become music in mind –introduction to music psychology–*, Kitaoji Pub., 2000

[25] http://www.lottolab.org/illusiondemos/Demo%2012.html

[26] A.D. Briscoe *et al.*, "The evolution of color vision in insects," *Annual review of entomology*, 46, 471–510, 2001

[27] W. Labov *et al.*, *Atlas of North American English,* Mouton and Gruyter, 2005

[28] L. Gerstman, "Classification of self-normalized vowels," *IEEE Trans. Audio Electroacoust.* AU-16, 78–80, 1968

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

[30] Y. Qiao *et al.*, "A study on invariance of f-divergence and its application to speech recognition," *IEEE Transactions on Signal Processing*, 58, 7, 3884–3890, 2010

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

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

[33] D. Saito *et al.*, "Optimal event search using a structural cost function – improvement of structure to speech conversion –," *Proc. INTERSPEECH*, 2047–2050, 2009

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