Speech Structure: a new framework of speech processing inspired from infants' behaviors and animals' behaviors

Nobuaki Minematsu

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

Abstract: Speech communication has several steps of production (encoding), transmission, and hearing (decoding). In every step, acoustic and static distortions are involved inevitably by differences of gender, age, microphone, room, line, auditory characteristics, etc. In spite of these variations, human listeners can extract linguistic information from speech so easily as if the variations do not disturb the communication at all. One may hypothesize that listeners modify their internal acoustic models whenever either of a speaker, a room, a microphone, or a line is changed. Another one may hypothesize that the linguistic information in speech can be represented separately from the extra-linguistic factors. In this study, being inspired from infants' behaviors and animals' behaviors, our solution to the intrinsic and inevitable variations in speech is described [1,2,3]. Speech structures, invariant to these variations, are derived as completely transform-invariant features [4] and their linguistic and psychological validity is discussed here. Further, some speech applications of ASR [3] and CALL [5] using the structures are shown, where extremely robust performance with speaker variability can be obtained with speech structures.

Index Terms: Speech structures, extra-linguistic feature, vocal imitation, invariance, f-divergence, ASR, and CALL

1 Introduction

Every normally developed individual shows extremely robust performance of speech processing. A five-year-old boy can understand on a mobile phone what a caller says even when he hears the voices of that caller for the first time. In a TV show, the tallest man and the shortest one in the world communicate orally against the largest gap of voice quality existing between the two. Why is our perception so robust? Linguistic messages in speech are the information encoded in a speech stream [6]. Then, what is the human robust algorithm of decoding that information?

Our perception is not only robust with speech variability but also robust with variability in other media. Generally and psychologically speaking, the robustness of perception is called perceptual constancy. For example, a visual image is modified in its shape by viewpoint changes but our perception is constant. As for color, a flower in broad daylight and the same one at sunset give us different color patterns but we perceive the equivalence between them. Humming by a male and that of the same melody by a female often differ in fundamental frequency but we easily perceive the equivalence. Male voices are deeper in timbre than female ones but the invariant perception is easy between a father's "hello!" and a mother's. Although the above stimuli are presented as different media, all the changes are caused commonly by static biases.

In this paper, psychologists' discussions on the perceptual constancy are overviewed from an engineering viewpoint. Then, a focus is put on how robust animals' perception is and humans' perception is, and then, what kind of difference is found between them. Following these discussions, we describe our proposal of speech structure: a speaker-invariant contrastive and dynamic representation of speech. After that, we survey what we did so far using the speech structures.

2 Nature of perceptual constancy

It seems that researchers of psychology found that, among different media, a similar mechanism is working to cancel the static biases and realize the invariant perception [7,8,9]. Figure 1 shows the look of the same Rubik's cube seen through differently colored glasses. Although the corresponding tiles of the two cubes have different colors absolutely, we name them using the same labels. On the other hand, though

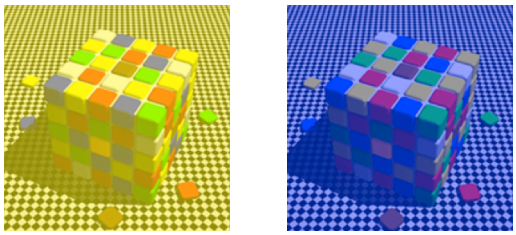

Figure 1: The same Rubik's cube seen with two colored glasses

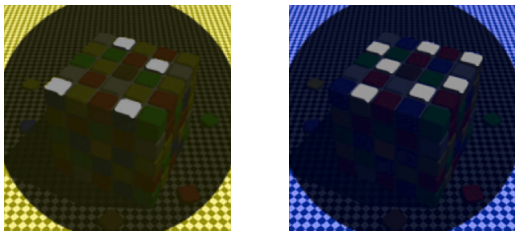

Figure 2: Perception of colors without context

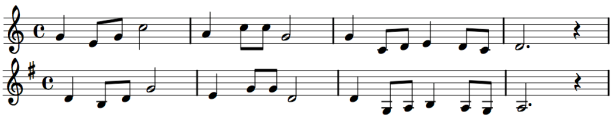

Figure 3: A musical melody and its transposed version

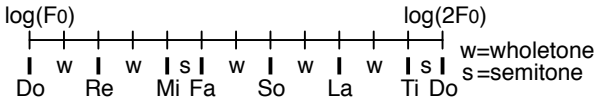

Figure 4: Tonal arrangement (scale) of the major key

we see four *blue* tiles on the top of the left cube and seven *yellow* tiles on the right, when their surrounding tiles are hidden, we suddenly see that they have the same color (See Figure 2). Absolutely different colors are perceived as identical and absolutely identical colors are perceived as different.

Similar phenomena are found in tone perception. Figure 3 shows two sequences of musical notes. The upper corresponds to humming by a female and the other to that of the same melody by a male. If hearers have relative pitch and can transcribe these melodies, they convert the two melodies into the same sequence of syllable names (So Mi So Do La Do Do So). The first tone of the upper and that of the lower are different absolutely but they name these tones using the same label. The first tone of the upper and the fourth of the lower are identical absolutely but they claim that the two tones are different. Similar to colors, absolutely different tones are perceived as identical and absolutely identical tones are perceived as different.

Researchers found that the invariant perception, colors and tones, occurs primarily based on contrast-based information processing [7,8,9]. In other words, our invariant perception of colors and tones is guaranteed by the invariant relationship of the focused stimulus to its surrounding stimuli. For individuals with relative pitch, a single tone is hard to name but tones in a melody are easy to transcribe. If two tones in a melody of the major key, which can be temporally distant, are three wholetones apart in pitch, they must be Fa and Ti according to the tonal arrangement (scale) of the major key (See Figure 4). This arrangement is invariant with key and, using this arrangement as constraint, the key-invariant tone identification can occur.

As was found in ecology, the invariant color perception occurs even to butterflies and bees [10]. It is extremely old evolutionarily. In contrast, researchers of anthropology found that the invariant tone perception is difficult even for monkeys [11]. What they claim is not that monkeys cannot transcribe a melody but that monkeys cannot perceive the equivalence between a melody and its transposed version [11]. The relative pitch perception is very new evolutionarily.

3 Human development of spoken language

How can infants acquire the ability of robust speech processing? Recently, in the field of AI, there is a research trend to focus on infants' acquisition and development of cognitive abilities [12,13,14]. One obvious fact is that a major part of the utterances an infant hears are from its parents. After it begins to talk, about a half of the utterances it hears are its own speech. It can be said that the utterances an individual hears are strongly speaker-biased unless he/she has speaking disabilities. The speech variability problem should be solved not by collecting samples if one really wants to realize a human-like speech processor.

Infants acquire spoken language through imitating their parents' utterances actively, called vocal imitation. But they don't impersonate their parents. A question is raised: what acoustic aspect of the voices do infants imitate? One may claim that infants decompose an utterance into a phoneme sequence and each phoneme is realized acoustically by their mouths. But researchers of infant studies deny this claim because infants don't have good phonemic awareness [15,16]. Then, what is imitated?

An answer from infant studies is the holistic sound pattern embedded in an utterance [15,16] called otherwise as word Gestalt [17] and related spectral patterns [18]. The holistic pattern has to be speaker-invariant because, whoever speaks a specific word to an infant, its responses of imitation are similar acous-

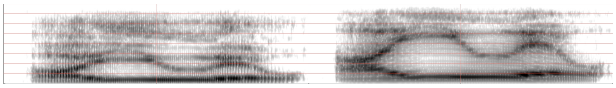

Figure 5: /aiueo/ produced by a tall speaker and a short one

tically. Another question is raised: what is the physical definition of the speaker-invariant holistic pattern?

The vocal imitation is rare in animals [19] and non-human primates scarcely imitate others' utterances [20]. This performance can be found in only a few species of animals: birds, whales, and dolphins. But there is a critical difference between humans and animals. Animals' imitation is basically the imitation of sounds like impersonation [19]. Take myna birds for example. They imitate the sounds of cars, dogs as well as human voices. Hearing a very good myna bird say something, one can guess its human owner [21] but cannot guess the parents of an infant by hearing its voices. Considering that the same pitch contours (intonation patterns) of different keys (genders or speakers) are equivalent for humans but different for monkeys and that the same linguistic content acoustically generated by different speakers are equivalent for humans but probably different for animals [22], the ability of extracting an invariant and abstract pattern from a variable sound stream might be unique to humans.

4 Natural solution of speaker variability

As for speech, changes in vocal tract shape and length cause timbre changes. Basically speaking, dynamic and morphological changes of the vocal tract generate different phonemes acoustically. Static differences of the vocal tract shape and length among speakers cause speaker variability. Figure 5 shows the same message generated by a tall speaker and a short one. What is the speaker-invariant holistic pattern?

Speaker difference is often modeled mathematically as space mapping in studies of voice conversion. This means that if we can find some transform invariant features, they can be used as speaker-invariant features. Recently, some proposals have been done [23,24] but speaker variability was always modeled simply as $\hat{f} = \alpha f$ (f=frequency, α=constant). Many studies of speaker conversion adopted more sophisticated transforms, indicating that $\hat{f} = \alpha f$ cannot characterize speaker variability well enough. Further, it should be noted that all of these proposals tried to find invariant features in individual speech sounds, not in a holistic pattern only composed of speech contrasts.

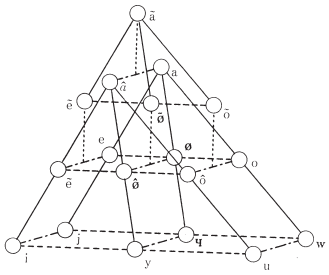

Figure 6: The invariant system of French vowels

As shown in [10], the perceptual constancy of colors is found in butterflies. But no researcher claims that a butterfly has statistical models of individual colors which are acquired by looking at the colors through thousands of differently colored glasses. Further, naming individual colors (elements) is not needed to perceive the equivalence between a flower in broad daylight and the same one at sunset. In contrast in ASR, acoustic and statistical modeling of individual phonemes (elements) using thousands of speakers (differently shaped tubes) is the most popular approach. From an ecological and evolutionary viewpoint, this strategy is remarkably weird and the invariant speech perception should be implemented on machines based on processing holistic patterns composed of invariant contrasts or relations.

A similar claim can be found in classical linguistics [25]. Jakobson proposed a theory of acoustic and relational invariance, called distinctive feature theory. He repeatedly emphasizes the importance of relational and systemic invariance among speech sounds by referring to phrases of other scholars such as Klein (topologist), Baudouin, and Sapir (linguists). Figure 6 shows his invariant system of French vowels and semi-vowels. Figure 4 is the key-invariant tonal arrangement in melody and Figure 6 is the speaker-invariant timbre arrangement in vowel sounds. Considering that pitch is one-dimensional but timbre is multi-dimensional, what has to be implemented on machines is a mechanism of relative timbre perception, where invariant and multi-dimensional timbre contrasts are used to determine the value of individual sounds. In a classical study of acoustic phonetics, the importance of relational invariance was experimentally verified in word identification tests [26]. It should be noted that Lagefoged discussed a very good similarity between perception of vowels and that of colors [26].

5 Mathematical solution of the variability

4.1 Mathematically guaranteed complete invariance

Are there any invariant and contrastive features (measures) with respect to any linear or non-linear invertible transforms? In [4], we proved that f-divergence between two distributions is invariant with any kind of invertible and differentiable transforms (sufficiency). We also proved that any completely invariant measure with respect to two distributions has to be written in the form of f-divergence (necessity), which is formulated as

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

Figure 7 shows two spaces (shapes) which are deformed into each other through an invertible and differentiable transform. An event is described not as point but as distribution. Two events of p_1 and p_2 in A are transformed into P_1 and P_2 in B. The invariance of f-divergence is always satisfied [4].

$$f_{div}(p_1, p_2) \equiv f_{div}(P_1, P_2)$$

In a series of our previous studies [1,2,3,4,5], we have been using Bhattacharyya distance (BD) as one of the f-divergence measures. Figure 8 shows a procedure of representing an input utterance only by BD. The utterance in a feature space is a sequence of feature (cepstrum) vectors and it is converted into a sequence of distributions through automatic segmentation. Here, any speech event is modeled as a distribution. Then, the BDs are calculated from any pair of distributions to form a BD-based invariant distance matrix. As a distance matrix can specify a unique geometrical shape, we call the matrix as speech structure. Individual speech sounds can change but their entire system cannot change at all.

4.2 Isolated word recognition

Figure 9 shows the basic framework of isolated word recognition based on speech structures. To convert an utterance into a distribution sequence, the MAP-based HMM training is adopted. Then, the BD between any pair of the distributions is obtained. After calculating a structure, its structure vector is formed by using all the elements in the upper triangle. This vector is a holistic and speaker-invariant

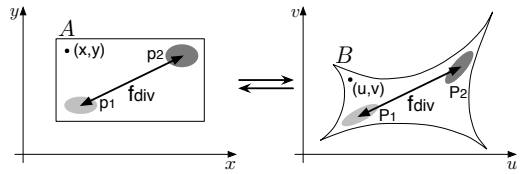

Figure 7: Topological deformation of manifolds (shapes)

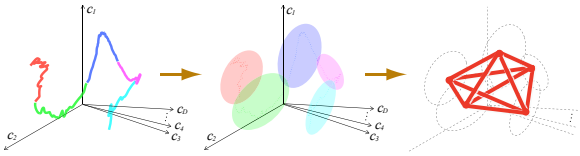

Figure 8: An utterance structure composed of f-divergence

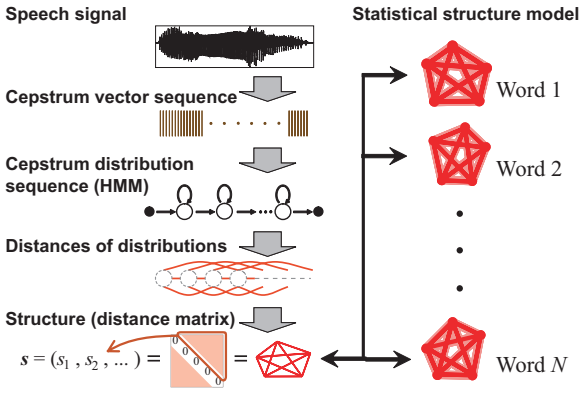

Figure 9: Framework of structure-based word recognition

representation of a word utterance. The right-hand side of the figure shows an inventory of word-based statistical structure models for the entire vocabulary. The candidate word showing the maximum likelihood score is a result of recognition.

The speech structure is invariant with any kind of invertible transforms. This indicates that two different words can be evaluated as the same. To solve this problem, we introduced good constraints called Multiple Stream Structuralization (MSS) [27] so that we could obtain the invariance only with respect to speaker variability. Due to the limit of space, MSS is not explained in details in this paper but interested readers should refer to [3,27].

In [3,27], structure-based isolated word recognition was compared to substance-based word recognition. The former used the proposed structure (contrast) models and the latter used the conventional word HMMs trained with spectrum-based (substance-based) features. Two word sets were used. In a set, a word was artificially composed of five vowels such as /eauoi/ and /uoaie/. As Japanese has only five vowels, PP=120. The other set was a Japanese phoneme-balanced word set and PP=220

Table 1: Comparison between HMMs and structures [%]

α	-0.40	-0.35	-0.30	-0.25	-0.20	-0.15	-0.10	-0.05	0.00
HMMs	0.92	0.94	1.75	6.83	21.8	40.5	60.2	80.0	83.9
matched	58.9	62.1	64.3	68.5	74.3	78.3	81.5	83.5	83.9
Structures	53.6	61.9	68.3	74.3	80.1	84.0	86.9	88.8	89.1
α	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	
HMMs	78.2	63.1	44.5	24.8	8.85	1.88	1.00	0.67	
matched	84.7	85.8	86.3	86.3	86.3	86.4	87.2	86.6	
Structures	89.5	89.8	90.5	90.6	90.9	91.0	91.2	91.3	
α	-0.40	-0.35	-0.30	-0.25	-0.20	-0.15	-0.10	-0.05	0.00
HMMs	5.33	11.2	21.5	37.4	57.6	74.1	87.6	95.8	98.3
matched	94.9	96.4	96.6	97.4	97.8	98.0	97.9	98.3	98.3
Structures	46.7	55.3	63.1	69.9	77.4	83.2	88.0	91.6	92.6
α	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	
HMMs	97.5	92.3	81.2	64.6	45.6	27.2	14.0	7.65	
matched	98.4	98.5	98.4	98.5	98.5	98.3	98.4	98.6	
Structures	92.1	90.6	86.3	81.0	74.0	66.9	58.0	49.3	

[28]. To investigate the robustness with respect to mismatch between training and testing conditions, frequency warping was applied to testing samples to simulate speech samples generated by very tall and very short speakers. Table 1 summarizes the results.

α is a warping parameter and varied from -0.4 to 0.4 at 17 steps. α=-0.4/+0.4 indicates doubling/halving the vocal tract length. Both HMMs and structures used no speaker adaptation technique. The number of distributions per word is 25 for the vowel words and 30 for the balanced words. In the figure, matched shows the results of using 17 sets of matched conditioned HMMs. In the vowel word set, a single set of structures shows almost the same or higher performance compared to the 17 matched HMM sets. In the phoneme balanced set, however, the performance of the structures is lower than that of HMMs at α=0.0 although the robustness of the structures is shown at α>0.15. This is because unvoiced consonants are less speaker-dependent and absolute features are needed for them. Currently, we're integrating both the models for compensation. Detailed description of the experiments is in [3,27].

4.3 Pronunciation proficiency estimation for CALL

Acoustic assessment of individual sounds in a learner's utterances can be viewed as *phonetic* assessment and that of the entire system of the sounds can be regarded as *phonological* assessment. In the former, it is assessed whether each sound has proper acoustic features, while in the latter, it is examined or not whether an adequate sound system underlies a learner's pronunciation. In [29], it was discussed which strategy can provide learners with a more robust framework of proficiency estimation.

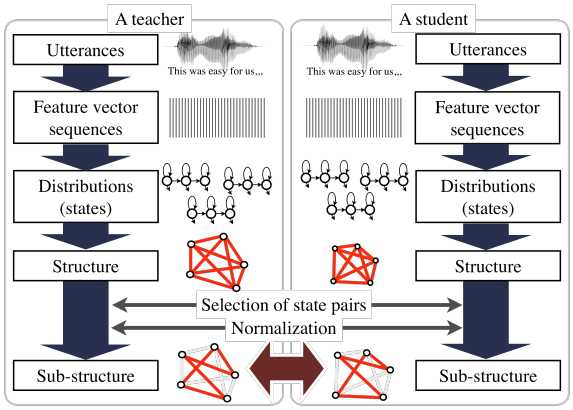

Figure 9: Sub-structure extraction for a teacher and a learner

For phonetic assessment, we adopted GOP (Goodness Of Pronunciation), which was originally proposed in [30] and is a widely-used technique. By using speaker-independent phoneme HMMs, it is calculated as a posterior probability of the intended phonemes given input utterances.

$$
\begin{aligned}
&GOP(o_1,...,o_T,p_1,...,p_N)\\
&=P(p_1,...,p_N|o_1,...,o_T)\\
&\approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{D_{p_i}}\log\left\{\frac{P(o^{p_i}|p_i)}{\max_{q\in Q}P(o^{p_i}|q)}\right\}
\end{aligned}
$$

T is the length of a given observation sequence and N is the number of the intended phonemes. o^{p_i} is the speech segment obtained for p_i through forced alignment and D_{p_i} is its duration. $\{o^{p_1}...o^{p_N}\}$ correspond to $\{o_1...o_T\}$. Q is the phoneme inventory.

For phonological assessment, we used pronunciation structure analysis [5]. Based on comparison between a learner's structure and a teacher's one, pronunciation proficiency of that learner was estimated. In [5], after training speaker-dependent monophone HMMs for each learner, a phoneme-based structure was estimated for that learner. In [29], however, state-based structures were estimated after adequate selection of the states. Figure 9 displays how to extract a pronunciation sub-structure from a teacher's utterances or from a learner's utterances. Euclidian distance between the two sub-structures was calculated and its negative value was used as proficiency score. Physical meaning of the Euclidian distance between two structures is described well in [3,29]. Interested readers should refer.

Pronunciation proficiencies of 26 learners in set-6 of ERJ (English Read by Japanese) database [31] were estimated using GOP and structures. For GOP, monohpone HMMs were trained using all the

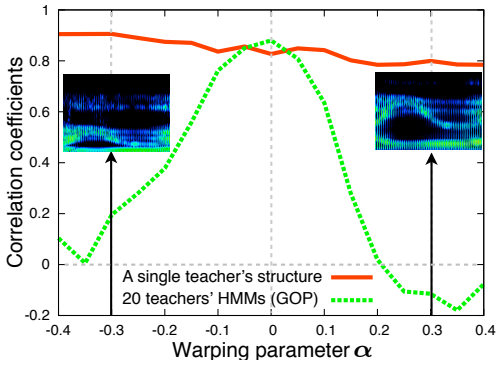

Figure 10: Correlations with GOP and structures

20 native teachers of American English, while for structures, M08's pronunciation structure was used. Similar to Section 4.2, frequency warping was also done here to simulate utterances of very tall learners and very short learners. Figure 10 shows the correlations between human scores and the two kinds of machine scores. Extreme robustness of the structures and extreme weakness of the GOP are shown. We can say that even a single teacher's structure can be used very effectively for learners of any size.

As GOP is a posterior probability, it internally has a function of canceling acoustic mismatch between HMMs and learners. But this function only works when forced alignment (numerator of GOP) performs well. With a large mismatch, however, this process fails. To avoid this, teachers' models (HMMs) are often adapted to learners. If one wants to prepare the most adequate models for a specific learner, one has to train the models with that learner who would pronounce the target language correctly.

This technical requirement leads us to consider that GOP should stand for Goodness Of imPersonation, which quantifies how well a learner can impersonate the model speaker. But learning to pronounce is not learning to impersonate at all. Young learners don't impersonate their teachers but they acquire remarkably well the sound system underlying their teachers' utterances. In contrast, the current framework of speech processing including ASR often captures and models given sounds directly as they are. Considering old and new findings in animal and infant studies, we can claim definitely that this strategy is much more animal than human. If one wants to build a human-like processor, what is needed is a method for good abstraction. What children discard should be discarded also by machines. Speech structure is our answer for abstraction.

6 Conclusion

This paper describes speech structure with its underlying philosophy and its several applications. We hope that our proposal will help researchers to approach the ultimate goal of understanding how we decode messages encoded in a speech stream.

Reference

[1] N. Minematsu. "Mathematical evidence of the acoustic universal structure in speech," *Proc. ICASSP*, 889-892 (2005)

[2] N. Minematsu *et al.*, "Theorem of the invariant structure and its derivation of speech Gestalt," *Proc. Int. Workshop on Speech Recognition and Intrinsic Variations*, 47-52 (2006)

[3] N. Minematsu *et al.*, "Implementation of robust speech recognition by simulating infants' speech perception based on the invariant sound shape embedded in utterances," *Proc. SPECOM* (2009, to appear)

[4] Y. Qiao *et al*., "f-divergence is a generalized invariant measure between distributions," *Proc. INTERSPEECH*, 1349-1352 (2008)

[5] N. Minematsu, "Training of pronunciation as learning of the sound system embedded in the target language," *Proc. Int. Symposium on Phonetic Frontiers*, CD-ROM (2009)

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

[7] R. B. Lotto *et al*., "An empirical explanation of color contrast," Proc. the National Academy of Science USA, 97, 12834-12839 (2000)

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

[9] T. Taniguchi, *Sounds become music in mind -introduction to music psychology-*, Kitaoji Pub. (2000)

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

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

[12] Acquisition of Communication and Recognition Skills Project (ACORNS) http://www.acorns-project.org/

[13] Human Speechome Project http://www.media.mit.edu/press/speechome/

[14] Infants' Commonsense Knowledge Project http://minny.cs.inf.shizuoka.ac.jp/SIG-ICK/

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

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

[17] M. Hayakawa, "Language acquisition and matherese," *Language*, 35, 9, 62-67, Taishukan pub. (2006)

[18] P. Lieberman, "On the development of vowel production in young children," *Child Phonology vol.1*, edited by G. H. Yeni-Komshian, J. F. Kavanagh, and C. A. Ferguson, Academic Press (1980)

[19] 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)

[20] 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)

[21] K. Miyamoto, *Making voices and watching voices*, Morikawa Pub. (1995)

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

[23] T. Irino *et al*., Segregating information about the size and shape of the vocal tract using a time-domain auditory model: the stabilised wavelet-Mellin transform, Speech Communication, 36, 181-203 (2002)

[24] A. Mertins *et al.* "Vocal trace length invariant features for automatic speech recognition," *Proc. ASRU,* 308-312 (2005)

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

[26] P. Ladefoged *et al*., Information conveyed by vowels, J Acoust. Soc. Am. 29, 1, 98-104 (1957)

[27] S. Asakawa *et al*., "Multi-stream parameterization for structural speech recognition," *Proc. ICASSP*, 4097-4100 (2008)

[28] Tohoku university & Matsushita isolated Word database (TMW), http://research.nii.ac.jp/src/eng/list/detail.html¥#TMW

[29] M. Suzuki et al., "Improved structure-based automatic estimation of pronunciation proficiency," *Proc. SLaTE* (2009, submitted)

[30] S. M. Witt, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, 30, 95-108 (2000)

[31] N. Minematsu *et al*., "Development of English speech database read by Japanese to support CALL research," *Proc. ICA*, 577-560 (2004)