Speech Structure and Its Application to Robust Speech Processing

Nobuaki MINEMATSU, Satoshi ASAKAWA, Masayuki SUZUKI, and Yu QIAO

The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN

{mine,asakawa,suzuki,qiao}@gavo.t.u-tokyo.ac.jp

Received 31 May 2009

Abstract Speech communication consists of three steps: production, transmission, and hearing. Every step inevitably involves acoustic distortions due to gender differences, age, microphone- and room-related factors, and so on. In spite of these variations, listeners can extract linguistic information from speech as easily as if the communications had not been affected by variations at all. One may hypothesize that listeners modify their internal acoustic models whenever extralinguistic factors change. Another possibility is that the linguistic information in speech can be represented separately from the extralinguistic factors. In this study, inspired by studies of humans and animals, a novel solution to the problem of intrinsic variations is proposed. Speech structures invariant to these variations are derived as transform-invariant features and their linguistic validity is discussed. Their high robustness is demonstrated by applying the speech structures to automatic speech recognition and pronunciation proficiency estimation. This paper also describes the immaturity of the current implementation and application of speech structures.

Keywords Speech Structures, Extralinguistic Features, Invariance, f-divergence, ASR, CALL, Robustness

§1 Introduction

Every normally developed individual shows an extremely robust capacity for understanding spoken language. Even a young child can understand the words of a caller on a mobile phone despite hearing the caller's voice for only the first time. The voices of some animated characters sound unrealistic because they are artificially created using speech technologies, but children can easily understand what they say. A TV show hosting the world's tallest and shortest adults demonstrates the ability of these individuals to communicate orally with no difficulty, despite the largest possible gap in voice timbre between the two. Why is our perception so robust? Linguistic messages in speech are regarded as the information encoded in a speech stream[1)]. What, then, is the human algorithm for decoding this information so robustly[2)]?

Our perception is not only robust against speech variability but also against variability in other sensory media. Psychologically speaking, robustness of perception is called perceptual constancy. A visual image is modified in shape by viewpoint changes but our perception remains constant. As for color, a flower in daylight and the same one at sunset present us with objectively different color patterns but we properly perceive the equivalence between them. When a man and a woman hum a tune, the tones differ in fundamental frequency but we can tell that the melody is the same. Male voices are deeper in timbre than those of females but perception is invariant between a father's "good morning!" and that of a mother. Although the above stimuli are presented via different media, all the variations are commonly caused by static biases.

In this paper, discussions of psychologists on perceptual constancy are reviewed with respect to evolution and development. Following this review, we describe our proposed theory of speech structure: a speaker-invariant contrastive and dynamic representation of speech. After that, we apply the structure to realize highly robust Automatic Speech Recognition (ASR) systems and Computer-Aided Language Learning (CALL) systems.

§2 Nature of perceptual constancy

Psychologists have discovered that among different media a similar mechanism functions to cancel static biases and realize invariant perception[3), 4), 5)]. The left-hand side of Figure 1 shows the appearance of the same Rubik's cube seen through differently colored glasses[6)]. 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

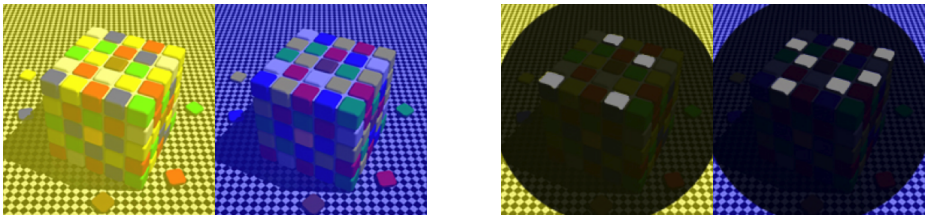

Fig. 1 Perception of colors with and without context[6].

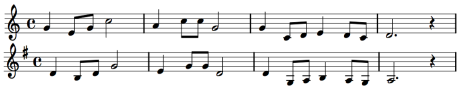

Fig. 2 A melody and its transposed version.

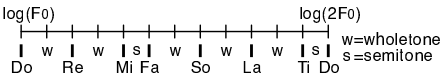

Fig. 3 Tonal arrangement of the major key.

tiles on the right, when their surrounding tiles are hidden, we suddenly realize that they have the same color (See the right-hand side of Figure 1). Different colors are perceived as identical and identical colors are perceived as different.

Similar phenomena can be found in tone perception. Figure 2 shows two sequences of musical notes. The upper corresponds to the humming of a female and the other to the same melody hummed by a male. If listeners 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 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 sequence and the fourth of the lower are physically identical but the two tones are identified as being different. Different tones are perceived as identical and identical tones are perceived as different.

Researchers have found that the invariant perception of colors and tones occurs through contrast-based information processing[3, 4, 5]. To some degree, this invariant perception is guaranteed by the invariant relationship of the focused stimulus to its surrounding stimuli. For individuals with relative pitch, a single tone is difficult to name but tones in a melody are easy to identify and transcribe. If a melody in a major key includes two tones that are three whole tones apart in pitch (and possibly temporally distant), these tones must be Fa and Ti according to the tonal arrangement (scale) of the major key (See Figure 3). This arrangement is invariant against key changes and, using this arrangement as a constraint, key-invariant tone identification is made possible.

As entomological studies have shown, invariant color perception occurs in butterflies and bees[7]. In contrast, anthropologists have found that invariant

tone perception is difficult even for monkeys[8]. It is not that monkeys cannot transcribe a melody, but rather that they cannot perceive the equivalence between a melody and its transposed version[8]. Thus, invariant color perception seems to have evolved much longer ago than invariant tone perception.

§3 Human development of spoken language

How do infants acquire the capacity for robust speech processing? Recent research, especially in the field of artificial intelligence, has focused on infants' acquisition and development of cognitive abilities[9, 10, 11] to realize robust speech processing on machines. 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 speech. It can be claimed that the utterances an individual hears are strongly speaker-biased unless he or she has speaking disabilities. Current ASR technology tries to solve the speech variability problem by collecting a huge number of samples and often adapting the resulting statistical models if necessary. We believe, however, that the problem should not be solved by extensive sample collection if a human-like speech processor is the goal of research.

Infants acquire language through active imitation of their parents' utterances, called vocal imitation. But they do not 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 sequence of phonemes and that each phoneme is reproduced acoustically. But researchers of infant studies deny this claim because infants do not have good phonemic awareness[12, 13].

An alternative answer, also derived from infant studies, involves a holistic sound pattern embedded in an utterance[12, 13], called word Gestalt[14] or a related spectral pattern[15]. This holistic pattern has to be speaker-invariant because, no matter who speaks a specific word to an infant, its imitative responses are similar acoustically. Another question is then raised: what is the physical definition of the speaker-invariant holistic patterns underlying individual utterances? As far as we know, psychologists have yet to demonstrate a mathematical formula. In this paper we describe our own proposal.

Vocal imitation is rare in animals[16], and non-human primates scarcely imitate the utterances of others[17]. This behavior 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 acoustic imitation,

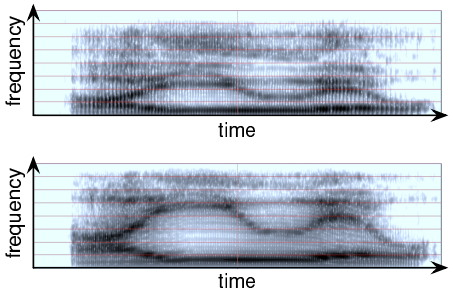

Fig. 4 /aiueo/s produced by a tall speaker (above) and a short speaker (below).

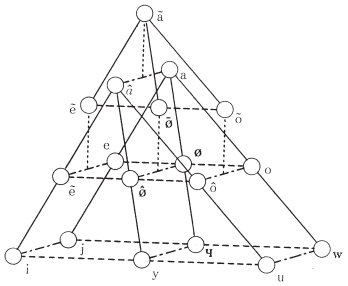

Fig. 5 Jakobson's invariant system of French vowels and semi-vowels[21].

similar to impersonation[16]. Considering monkeys' lack of invariant tone perception, again, acoustic variability seems to be an insoluble problem for animals.

§4 Natural solution of speaker variability

As for speech, changes in vocal tract shape and length result in changes of timbre. Basically speaking, dynamic morphological changes of the vocal tract generate different phonemes acoustically. However, static morphological differences of the vocal tract among speakers cause speaker variability. Figure 4 shows the same linguistic messages generated by a tall speaker and a short one.

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, several proposals have been made[18, 19, 20] but speaker variability was always modeled simply as $\hat{f}=\alpha f$(f:frequency, α:constant). In this case the proposed invariance depends strictly on this simple model. Many studies of voice conversion have adopted other sophisticated transforms, indicating that this simple model will be inadequate in characterizing speaker variability. Further, we should note that all these proposals have tried to find invariant features in individual speech sounds, not in holistic patterns composed only of speech contrasts or relations.

As shown in [8], perceptual constancy of colors is found in butterflies. As far as we know, however, no researcher has claimed that a butterfly acquires statistical models of individual colors by looking at all the colors through thousands of differently colored glasses. Further, naming individual colors (elements) is not needed in order to perceive the equivalence between a flower in daylight and the same flower at sunset. In contrast, the most popular method of acoustic

modeling of conventional ASR is the statistical modeling of individual phonemes (elements) using thousands of speakers (differently shaped vocal tubes). As we discussed in Section 3, we consider this strategy to be unnatural and, if a human-like speech processor is the goal, robust speech processing should be implemented on machines based on holistic patterns composed of speech contrasts or relations.

A similar claim can be found in classical linguistics[21]. Jakobson has proposed a theory of acoustic and relational invariance called distinctive feature theory. He repeatedly emphasizes the importance of relational, systemic, and morphological invariance among speech sounds. Figure 5 shows his invariant system of French vowels and semi-vowels. In a classical study of acoustic phonetics, the importance of relational invariance was experimentally verified in word identification tests[22]. It should be noted that Ladefoged discussed the significant similarity between the perception of vowels and that of colors[22]. A good survey of vowel perception based on relational invariance is found in [23].

Recently in [24], Hawkins proposed a memory-prediction theory to explain intelligence from the point of view of a neuroscientist. "I believe a similar abstraction of form is occurring throughout the cortex. Memories are stored in a form that captures the essence of relationships, not the details of the moment. The cortex takes the detailed, highly specific input and converts it to an invariant form. Memory storage and recall occur at the level of invariant forms."

From an engineering viewpoint, if a developed system works well for a given task, a natural solution might not be needed. If one wants to develop not only outwardly appearing but also internally human-like speech systems[9, 10, 11], however, we believe that he or she has to develop computational algorithms that are in accordance with findings in the human sciences. In the following section we describe our proposal, but can hardly claim that this is the best or only solution. After demonstrating some experimental results, we also describe the immaturity of the current implementation of speech structures.

§5 Mathematical solution of the variability

In [25], we proved that f-divergence[26] between two distributions is invariant with any kind of invertible and differentiable transforms (sufficiency). Further, we also proved that features, which are invariant with any transform, have to be, if any, f-divergence (necessity). f-divergence is a family of divergence

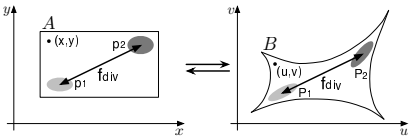

Fig. 6 Invertible deformation of shapes. p_1 and p_2 are transformed to P_1 and P_2.

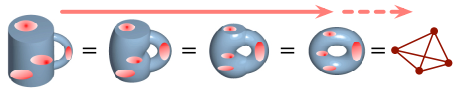

Fig. 7 Complete topological invariance. An f-divergence-based distance matrix is completely invariant with invertible transforms.

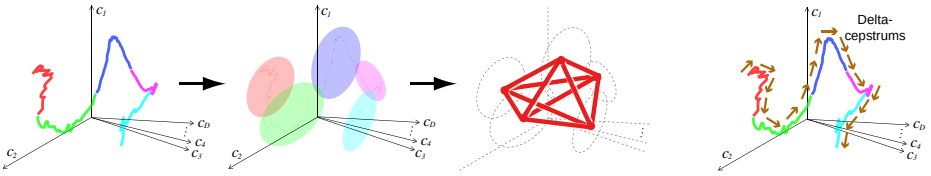

Fig. 8 Utterance structure composed only of f-divergences. A feature trajectory is converted into a distribution sequence. From the distributions, an invariant distance matrix is formed.

measures and it is defined as

$$f_{\mathrm{div}}(p_1, p_2) = \oint p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x}, \tag{1}$$

where $g(t)$ is a convex function for $t > 0$. If we take $t\log(t)$ as $g(t)$, f_{div} becomes KL-divergence. When $\sqrt{t}$ is used for $g(t)$, $-\log(f_{\mathrm{div}})$ becomes Bhattacharyya distance. Figure 6 shows two shapes and they 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. Here, the invariance of f-divegence is always satisfied[25].

$$f_{\mathrm{div}}(p_1, p_2) \equiv f_{\mathrm{div}}(P_1, P_2) \tag{2}$$

Figure 7 shows a famous example of deformation from a mug to a doughnut, often used to explain topology, where two shapes are treated as identical if they can be transformed continuously. Suppose that a number of events exist as distributions on the surface of the mug. When the mug is deformed in varying degrees into the doughnut, f-divergences between any event pair cannot change. An f-divergence-based distance matrix is completely invariant quantitatively.

In our previous studies[25, 27, 28, 29], we have been using the Bhattacharyya distance (BD) as our f-divergence. Figure 8 shows the procedure whereby an input utterance is represented only by BDs. The utterance in a feature space is a sequence of feature vectors and it is converted into a sequence of distribu-

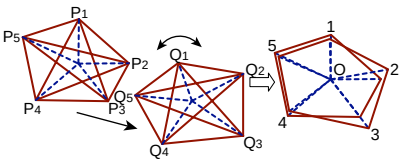

Fig. 9 Structure matching. Similarity is calculated after shifting and rotating two structures to obtain the best overlap.

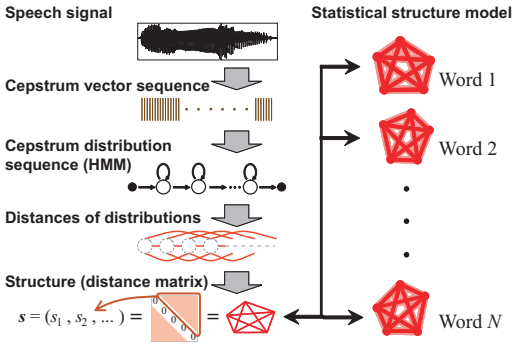

Fig. 10 Structure-based isolated word recognition.

tions, i.e., automatic segmentation. Here, any speech event is characterized as a distribution. The BDs are then calculated from every distribution pair, including temporally distant ones, to form a BD-based matrix. As a distance matrix in a Euclidean space can specify a unique shape, we call the matrix a speech structure. Here, we should note that velocity vectors, relative changes at each point in time (See the right-hand side of Figure 8), are not good candidates for speaker-invariant features. The reason is explained in the following section.

Once two utterances are represented as two speech structures, how does one calculate the similarity between the two? We have already proposed a very simple answer[28, 29]. Since a distance matrix is symmetric, we can form a vector composed of all the elements in the upper triangle of the matrix. This vector is henceforth called a structure vector. As shown in Figure 9, similarity between two structures is defined as the minimum summation of distances between the corresponding two points (events) after one structure is shifted and rotated so that the two structures overlap as completely as possible. The Euclidean distance between the two structure vectors can approximate the minimum summation[28, 29]. In a cepstrum space, rotation approximately represents cancelation of differences in vocal tract length[30] and shift cancels microphone differences. This means that structure matching will give us acoustic similarity between two utterances after global speaker and microphone adaptation. But no explicit adaptation is needed because adaptation is implicitly performed during the structure matching process. In other words, structure matching is a computational shortcut. Chemically speaking, this matching scheme is called Root Mean Square Deviation (RMSD)[31], where a distance matrix represents the shape of a molecule. RMSD is often used to calculate structural differ-

ences between two molecules without explicit estimation of a mapping function to transform one molecule into the other. If absolute positions of individual events in a (parameter) space are used as observation and model parameters, however, the mapping function must always be estimated. As far as we know, conventional adaptation methods in ASR, such as Maximum Likelihood Linear Regression (MLLR), are based on this strategy. This is why acoustic models must be updated whenever extralinguistic or environmental factors change.

Figure 10 shows the basic framework of isolated word recognition with speech structures. To convert an utterance into a distribution sequence, the Maximum a Posteriori (MAP)-based training procedure of Hidden Markov Models (HMMs) is adopted. Then, the BDs between every distribution pair are obtained. After calculating the structure, absolute properties such as spectrums are discarded. The right-hand side of the figure shows an inventory of word-based statistical structure models (Gaussian models) for the entire vocabulary. The candidate word showing the maximum likelihood score is a result of recognition.

§6 Isolated word recognition

6.1 Two problems and their solutions

The proposed speech structure is invariant with any kind of transforms. This had led us to expect that two different words could be evaluated as identical and our preliminary experiments showed that this expectation was correct. As both speaker and phoneme differences are basically differences in timbre, it is not complete invariance but adequately constrained invariance that is needed. We have to strike the proper balance between invariance and discrimination.

To realize this balance, we have modeled the variability due to vocal tract length differences mathematically and, based on the model, we have introduced a new technique. Speech modification due to vocal tract length difference is often modeled as frequency warping[32, 33, 34]. In [32, 33], it was shown that this warping can be modeled in the cepstrum domain by multiplying cepstrum vector $\boldsymbol{c}$ by matrix $\boldsymbol{A}$ ($\boldsymbol{c'}=\boldsymbol{Ac}$). BD is completely invariant with any kind of $\boldsymbol{A}$ and this invariance is too strong. $\boldsymbol{A}$ in [32, 33] is a band matrix and our goal is the invariance of only band matrices. Here, we divide a cepstrum stream into two substreams, where $\boldsymbol{c}_{i,j}$ means a substream from i-th to j-th dimension.

$$\begin{pmatrix} \boldsymbol{c}'_{1,n} \\ \boldsymbol{c}'_{n+1,N} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{c}_{1,n} \\ \boldsymbol{c}_{n+1,N} \end{pmatrix} + \begin{pmatrix} \boldsymbol{b}_{1,n} \\ \boldsymbol{b}_{n+1,N} \end{pmatrix}, \qquad (3)$$

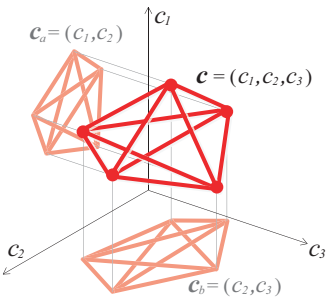

Fig. 11 Projection of a structure into two subspaces.

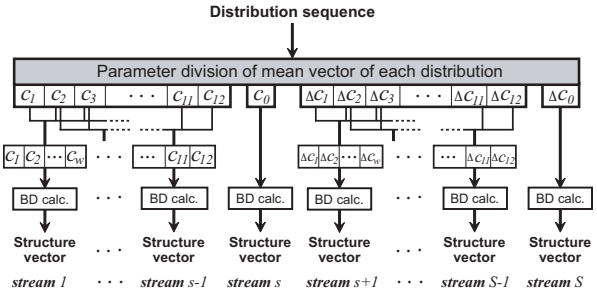

Fig. 12 Multiple Stream Structuralization (MSS). Mean vector of each distribution is divided into subvectors. Using a sequence of subvectors and their variances, a substructure is formed for each substream.

where $\boldsymbol{b}$ is a static bias vector representing microphone difference. If we assume independence between the two substreams in speech modification, $\boldsymbol{A}_{12}$ and $\boldsymbol{A}_{21}$ are zero matrices. If we consider more than two substreams, $\boldsymbol{A}$ more closely resembles a band matrix. Speech modification with a band matrix approximately indicates that each cepstrum substream is modified reasonably independently of the others. We expect that adequately constrained invariance can be obtained by structure matching after dividing a stream into multiple substreams. We call this technique Multiple Stream Structuralization (MSS).

If a three-dimensional stream is divided into two substreams, the resulting substructures are shown in Figure 11. Structure matching is performed in each subspace. Figure 12 shows a general MSS procedure. An input utterance is converted into an HMM, a set of distributions. For the mean vector of each distribution, w adjacent cepstrums form a subvector and w adjacent Δcepstrums form another. Here, we have S subvectors totally. One subvector and that adjacent to it partially overlap (See Figure 12). Using these subvectors and their corresponding variances, a substructure is constructed in each subspace. The final similarity score is obtained by summing up the scores in the subspaces.

The second problem is that the parameter dimension is increased with $O(n^2)$, where n is the number of distributions in an utterance. In this case, the number of edges (contrasts) in a structure becomes ${}_nC_2$ (See Figure 8). Then, the total number of dimensions is $S\,{}_nC_2$. To simultaneously reduce the number of dimensions and increase discriminability, a widely used method is adopted here. Linear Discriminant Analysis (LDA) is introduced twice. Figure 13 demonstrates the procedure. After MSS, LDA is carried out for each substream (sub-

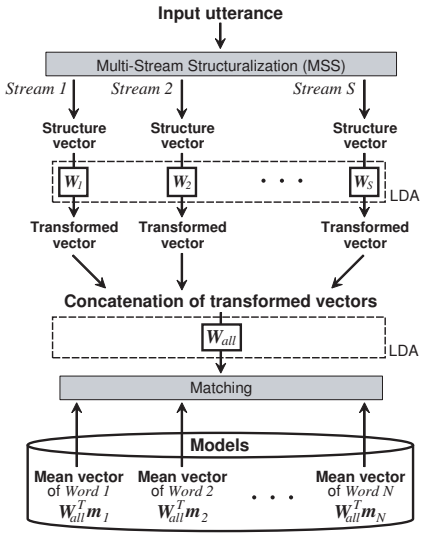

Fig. 13 Structure matching through two-stage LDA.

structure), which is the first LDA. $\boldsymbol{W}_i (i{=}1...S)$ are its transform matrices. Then, all the transformed substructure vectors are concatenated to form a single integrated vector. This vector is transformed again with $\boldsymbol{W}_{all}$, which is the second LDA. The resulting vector is used for matching with pre-stored templates.

6.2 Two word sets used in the experiments

Two word sets were prepared. One was an artificial word set, where each word consisted of a five-vowel sequence such as /eoiau/. Since Japanese has only five vowels, the vocabulary size was 120 ($={}_5P_5$). The other set was a Japanese phoneme-balanced natural word set[35], which is often used in the Japanese ASR community to verify the effectiveness of new techniques. The word length in terms of phonemes varied from 3 to 10 and the vocabulary size was 212. Considering that vowel sounds are more speaker-dependent than some consonant sounds such as unvoiced plosives and fricatives, it was reasonably expected that our proposal would be more appropriate for the first word set.

With matrix $\boldsymbol{A}$, various kinds of non-linear frequency warping can be applied to the word utterances. Considering the fact that the tallest adult in the world is 257 cm high and the shortest is 74 cm high, the warping was done to cover this range and these warped data were used for testing the proposed technique. In real situations, however, we would hardly see such tall or short

Table 1 Acoustic analysis conditions.

window	25 ms length and 10 ms shift
parameters	FFT-CEP(1 to 16) + ΔCEP(1 to 16) + ΔPower
distribution	1-mixture Gaussian with a diagonal matrix
	20 distributions for each vowel word (n=20)
	25 distributions for each balanced word (n=25)
estimation	MAP (for extracting a structure from an utterance)
	ML (for training an HMM from multiple utterances)

speakers, although it might not be uncommon to hear them on television. As described in Section 1, the voices of some animated characters are created by transforming real human voices. Although they sound unrealistic as human voices, children can easily understand what the characters are saying. How does this compare to the current speech recognition systems?

6.3 Experimental conditions

The acoustic analysis conditions are shown in Table 1. For comparison, word-based HMMs were built with the same training data. As shown in Figure 8, the structures captured only the relational features of speech contrast but the HMMs captured mainly absolute spectrogram characteristics. In the latter, relative features of Δcepstrums, which are velocity vectors in the cepstrum space, are often used in addition (See the right-hand side of Figure 8). Δcepstrums are invariant with static bias vector $\boldsymbol{b}$ in Equation (3), meaning that they are invariant with microphone differences. However, we mathematically showed in [30] that $\boldsymbol{A}$ in [32, 33] is approximated as a rotation matrix. This claims that differences in vocal tract length change the direction of a timbre trajectory in Figure 8. Experimental verification of this claim was performed in [30]. This is why we stated in Section 5 that Δcepstrums are not good invariant features.

FFT-cepstrums, not Mel-Frequency Cepstrum Coefficients (MFCC), were used here for two reasons. One is that analytical matrix representation of frequency warping was shown in [32, 33] using FFT-cepstrums. The other is that a Mel transform is a frequency warping and corresponds to shortening of the vocal tract. One would expect this effect, then, to be cancelled due to the invariance of structures. Some characteristics of MFCC, such as the use of overlapped triangular windows and DCT, may improve the performance. However, since what we want to discuss here is the performance difference between absolute features

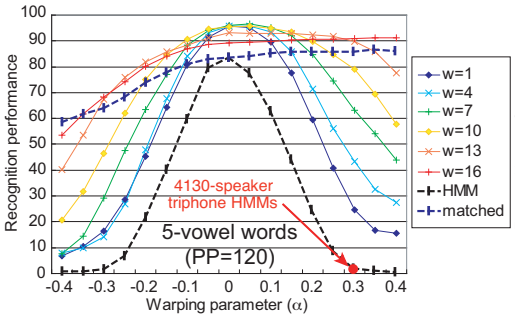

Fig. 14 Performance with vowel words.

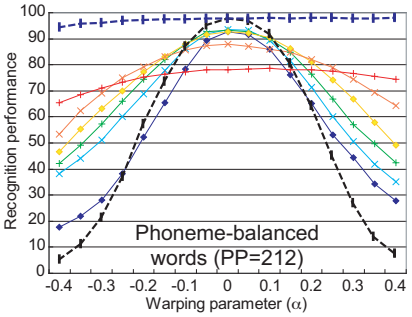

Fig. 15 Performance with balanced words.

and relational ones used to build statistical models, we adopted FFT-cepstrums.

Both for structures and HMMs, the number of distributions, n, was set to 20 for each vowel word and 25 for each balanced word. For the vowel words, four males and four females were used for training and a different four males and four females were for testing. For the phoneme-balanced words, 15 males and 15 females were used for training and another 30 speakers were for testing. For the former set, each speaker uttered the word set five times and each reference structure and each HMM were trained with 40 samples. For the latter set, each speaker uttered the word set once and each template was built with 30 samples.

6.4 Experimental results

Figures 14 and 15 show the performances with vowel words and balanced words. w is the width of a subvector (See Section 6.1) and it varies from 1 to 16. The X-axis represents warping parameter α[32, 33]. Positive and negative values of α indicate shortening or lengthening of the vocal tract length, respectively. The length is approximately halved when α=0.4 and doubled when α=−0.4. **HMM** in the figures signifies the performance of the word-based HMMs trained using the same data (original utterances) that were used in training the structures. **Matched** indicates the performance of 17 sets of word-based HMMs. They were separately trained using the training data warped with each value of α and tested with the testing utterances warped with the same α value. In other words, **matched** shows the performance with no mismatch.

We had expected that if the implicit adaptation mechanism worked well in structure matching, the performance of a single set of structures would be comparable to that of the 17 matched sets of HMMs. In the case of vowel words, we can say that our expectation was totally correct. The performance

of structures with w=16 is comparable to or even higher than that of the 17 matched sets of HMMs. The improvement over the matched HMMs is thought to be due to the LDA-based parameter reduction. Even if LDA is used with HMMs, however, drastic improvement in robustness is difficult to realize because the training data do not include a very wide variety of speakers.

We found that w worked to strike a balance between invariance and discrimination. As expected in Section 6.1, larger values of w tended to enhance invariance and reduce discrimination. The performance of w=1 in the matched condition (α=0) was better than that of $w \gg 1$. However, the performance of w=1 in the mismatched conditions was much worse that that of $w>1$.

In another experiment, speaker-independent tied-state triphone HMMs, which were trained with 4,130 adult speakers, were tested with the utterances warped using α=0.3. This triphone HMM set is distributed by the Japanese academic ASR community and often used as the baseline triphone set[36]. In contrast to the previous experiments, MFCCs and Cepstrum Mean Normalization (CMN) were used for acoustic analysis because the triphone HMMs used them. Using this HMM set, an isolated word recognition system was built using a dictionary of the 120 words mentioned above. The recognition performance was 1.4%, lower by far than that of the structures (91.0%, w=16), which were trained with only eight adult speakers. Although the triphone set is distributed as a speaker-independent model set, it did not work at all with the data warped at α=0.3. It is true that these utterances do not sound like real human voices but rather like the voices of animated characters such as small animals or insects. As described in Section 1, humans, even children, can understand their utterances easily, but it seems that the current speech recognition system cannot at all. If the HMMs are adapted and modified adequately using some warped utterances, as indicated in Figure 14, the same performance as that of the structure models should be obtained because, in a sense, the proposed method resembles the conventional methods. The difference lies in which adaptation strategy to adopt, implicit or explicit. In the former, no additional processing is needed for a new environment, while in the latter, modification of model parameters is always required for new environments. Technically speaking, we consider that this difference is very critical and significant.

In the balanced set, shown in Figure 15, the performance of structures was worse than that of HMMs in the matched condition (α=0). We consider that the use of a constant number of distributions (n=25) is inadequate for words

consisting of different numbers of phonemes. However, the current implementation of structure matching allows us only to compare two utterances composed of the same number of distributions. Further, as we expected, since unvoiced consonants are less speaker-dependent, the results imply that absolute spectrum features are necessary to represent these sounds. To solve these two problems, we tentatively propose combining relational features with absolute ones to enhance speech structure and make possible the flexible alignment between two structures. We suggest that interested readers should refer to [37].

Even with the current implementation of speech structures, however, in the mismatched conditions a high robustness was shown with larger values of w (w=10, 13). We found again that w functioned to balance invariance and discrimination. As noted before, speaker difference and word difference are in a large part attributable to spectrum difference.

Structure-based ASR is also possible with domains other than cepstrums. For example, spectrum-based structures are feasible because a spectrum envelope is obtained by linearly transforming cepstrums, i.e., FFT. We consider that spectrum-based MSS structures are similar to modulation spectrums[38] and RelAtive SpecTrA (RASTA)[39]. All of these capture only the dynamic aspect of speech but our structure uniquely grasps it in a mathematically speaker-invariant way. This invariance is obtained by removing the directional features of a speech trajectory because they are strongly speaker-dependent[30] and by modeling only the resulting speech contrasts, including temporally distant ones (see Figure 8).

§7 Pronunciation proficiency estimation

7.1 Urgent requirement for highly robust technologies

One of the main ASR applications is CALL, where pronunciation errors are detected or pronunciation proficiency is estimated automatically for foreign language learners using ASR technologies[40, 41]. English education in Japan is supposed to encounter a turning point soon. The Japanese government decided to introduce oral English communication lessons in every public primary school starting in 2011, but we do not yet have a sufficient number of English teachers. The government expects class teachers, many of whom did not receive an education adequately preparing them to teach English, to play an important role in these lessons. Given this situation we anticipate that various technical solutions may be introduced to classrooms. Automatic estimation of pronunciation

proficiency will be one of the key technologies, and it requires high robustness because users include adult (tall) teachers and young (short) children.

7.2 Use of speech structures as pronunciation structures

The assessment of each sound instance in an utterance can be viewed as a phonetic assessment and that of the entire system of the instances can be regarded as a phonological assessment. In the former, the question is whether each sound has the proper acoustic features, while in the latter, it is whether an adequate sound system underlies a learner's pronunciation. Jakobson, who proposed a theory of relational invariance[21], claimed that in language acquisition, children acquire not individual sounds but the entire sound system.

In implementing machine-based phonological assessment, we have already applied speech structures to the automatic assessment of English vowels produced by learners. For example in [42, 43], from the utterances of 11 English monosyllabic words, each including one of the 11 American English monophthongs, the vowel structure of a learner was calculated. The structure contains almost no extralinguistic features and characterizes the accentedness of that learner's pronunciation well. We should note that with only a vowel structure (distance matrix), it is impossible to estimate the spectrum envelope pattern for the individual vowels. On the contrary, the vowel structure could be used effectively for pronunciation error detection and pronunciation proficiency estimation[42, 43]. Further, the pronunciation structures successfully made it possible to classify learners not based on their gender and age but based on their foreign accentedness[43]. In this paper, we examine experimentally the robustness of the structure-based proficiency estimation against large speaker variability.

7.3 Experimental conditions

Our phonetic assessment method adopted Goodness Of Pronunciation (GOP), which was originally proposed in [41] and is widely used today. By using speaker-independent phoneme HMMs, phoneme-based GOP is calculated as the posterior probability of the intended phonemes given input utterances.

$$GOP(o_1,...,o_T,p_1,...,p_N) = \log(P(p_1,...,p_N|o_1,...,o_T))$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{D_{p_i}}\log\left\{\frac{P(o^{p_i}|p_i)}{\sum_{q\in Q}P(o^{p_i}|q)}\right\} \approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{D_{p_i}}\log\left\{\frac{P(o^{p_i}|p_i)}{\max_{q\in Q}P(o^{p_i}|q)}\right\} \quad (4)$$

T is the length of a given observed sequence and N is the number of the intended

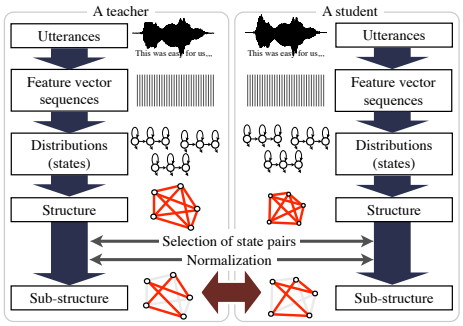

Fig. 16 State-based sub-structure extraction. From a learner's speaker-dependent HMMs, adequate state pairs are selected.

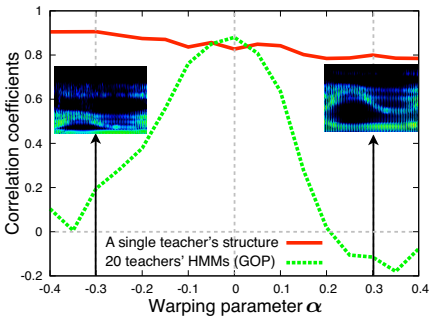

Fig. 17 Correlations between human and machine scores calculated using acoustically warped utterances.

phonemes. o^{p_i} is a speech segment for p_i obtained by 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[42, 43]. By comparing a learner's structure with that of a teacher (See Figure 9), the pronunciation proficiency of that learner is estimated. In [42, 43], after training speaker-dependent vowel HMMs for each learner, a vowel-based structure was built individually. In this paper, after training HMMs of all the English phonemes, state-based substructures were calculated through adequate selection of HMM state pairs. Figure 16 displays how to extract a pronunciation substructure from a teacher or learner's utterances. As explained shortly, set-6 of the English Read by Japanese (ERJ) database[44] was used to evaluate the performances of the GOP and the structure. Using the other seven sets of the database, selection of state pairs had been performed incrementally and greedily so that the correlation between human and machine scores was maximized. Using the optimal definition of the substructure, the Euclidean distance between a learner's substructure and that of a teacher was calculated and its negative value was used as a structure-based proficiency score. MSS was not done here.

The ERJ contains English sentences read aloud by 200 randomly selected Japanese university students (100 males and 100 females) and the same sentences read aloud by 20 native speakers of American English. The database also contains proficiency scores for the 200 students as rated by five native teachers of American English. As we described above, the pronunciation proficiency of 26 learners of set-6 was estimated using the two methods. Each learner read common 75 sentences. Two types of the machine scores, the GOP and the struc-

ture, were compared with the proficiency scores as rated by the five teachers. For estimating GOP, speaker-independent monophone HMMs were trained using all the 20 native speakers, while for structures, a pronunciation structure of M08 (a male native speaker) was used as a model structure. As in Section 6, frequency warping was also performed to simulate both tall and short learners.

Unlike previous word recognition experiments, MFCC (1 to 12), ΔMFCC (1 to 12), and ΔPower were used as acoustic features for both methods of the GOP and the structure. Frame length and shift were the same as in Table 1.

7.4 Experimental results

Figure 17 shows the correlations between the teachers' scores and the two types of machine scores. The X-axis represents warping parameter α applied only to testing utterances. In the figure, two speech segments of α=±0.3 are shown. Frequency warping resulted in a drastic acoustic modification. Despite it, extreme robustness of the structure is shown. On the other hand, extreme weakness of the GOP is obviously indicated at the same time. We can say that even a single teacher's structure can be used effectively for learners of any size.

As GOP is based on posterior probability, it possesses the inherent function of canceling the acoustic mismatch between teachers' HMMs and learners' utterances. But this function only works when forced alignment (the GOP numerator) and continuous phoneme recognition (the GOP denominator) perform well. With a large mismatch, however, these processes will probably fail. To avoid this, teachers' HMMs are often adapted to learners. If one wants to prepare the most suitable HMMs to estimate the proficiency of a specific learner, one has to train them with that learner pronouncing the target language correctly.

This technical requirement leads us to consider that GOP might have to stand for not Goodness Of Pronunciation, but 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. No male student tries to produce female voices when asked to repeat what a female teacher said. No child learner tries to produce a deep voice to repeat what a tall male teacher said. If Jakobson's claim is correct, a learner extracts a speaker-invariant sound system underlying a given utterance and tries to reproduce that system orally. But the inevitable difference in size and shape of the vocal organs between a learner and a teacher has to result in acoustic differences between their utterances. However, learning to pronounce is not affected at all by these differences.

§8 Discussion and conclusions

In this paper, we first reviewed psychologists' consensus on perceptual constancy and then discussed the evolutionary and developmental mechanism underlying the acquisition of perceptual constancy by referring to old and new findings in evolutionary anthropology and developmental psychology. Considering these discussions, a new framework of speech representation, called speech structure, was proposed, wherein an utterance is represented only with speaker-invariant speech contrasts. The invariance of speech contrasts is based on f-divergence, which is mathematically guaranteed to be invariant with any kind of invertible transforms. The speech structure was applied to ASR and CALL and its high robustness against speaker variability was successfully verified.

For ASR applications, adequately constrained invariance could be achieved using MSS. With it, the utterances warped by band matrices were recognized correctly with no explicit model adaptation. However, one of the two recognition tasks, vowel permutation, was artificial and the speaker variability was simulated variability. Evaluation of the proposed method using real data is needed to build real-world applications. With this goal, as described in Section 6.3, we are tentatively enhancing the speech structure. Further, as part of our future work we are planning to generalize MSS. The adequately constrained invariance realized in this paper depends on the form of $\boldsymbol{A}$, i.e., band matrices. If real speaker variability is better modeled using a different finite set of transforms, we must derive the adequately constrained invariance for them. Mathematically speaking, if the constrained invariance can be obtained for a given arbitrary set of transforms, we expect that it can be applied directly to processing of media other than speech.

For CALL applications, not only vowel sounds but also consonant sounds were used and a new technique of substructure formation was proposed to improve the robustness. This method did not involve MSS. Although the data used in the experiment were derived from actual learners, speaker variability was created artificially, again using $\boldsymbol{A}$. We consider that evaluation of the proposed method using actual children's data is a necessary prerequisite to introducing our CALL systems to classrooms.

To conclude this paper, we want to emphasize again that we are aiming at building human-like speech processors. To this end, technically speaking, we have proposed a completely different framework of speech representation and acoustic matching. This paper focused on the difference in vocal imitation between animals and humans. Animals' imitation is acoustic in nature whereas hu-

mans reproduce an underlying pattern embedded in a given utterance. Although the following description may be out of the scope of this paper, we found that, in some cases, acoustic imitation becomes the default strategy of humans' imitation. This performance can be found in severely impaired autistics[45, 46] and, in this case, the normal acquisition of speech communication becomes difficult. Prof. Grandin, a professor of animal sciences who is herself autistic, described the similarity in information processing between animals and autistics[47]. An autistic boy wrote that he could understand what his mother was saying but it was difficult for him to understand others[48]. His mother said that it seemed difficult for him to understand her on a telephone line. Looking at Figures 14 and 15, we can find a behavioral similarity to the HMM performance without adaptation, i.e., absolute processing. In this paper we explained our proposal to build human-like speech processors by referring to human factors from a broader viewpoint of human sciences. Although our understanding of humans may still be immature, we believe that the target processors are reasonably impossible to build without aiming for a comprehensive understanding of humans.

References

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

2) M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi and C. Wellekens, "Automatic speech recognition and speech variability: A review," 49, 763–786, 2007

3) R. B. Lotto and D. Purves, "An empirical explanation of color contrast," *Proc. the National Academy of Science USA*, 97, 12834–12839, 2000

4) R. B. Lotto and D. Purves, "The effects of color on brightness," *Nature neuroscience*, 2, 11, 1010–1014, 1999

5) T. Taniguchi, *Sounds become music in mind –introduction to music psychology–*, Kitaoji Pub., 2000

6) http://www.lottolab.org/illusiondemos/Demo%2012.html

7) A. D. Briscoe and Lars Chittka, "The evolution of color vision in insects," *Annual review of entomology*, 46, 471–510, 2001

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

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

10) Human Speechome Project http://www.media.mit.edu/press/speechome/

11) Infants' Commonsense Knowledge Project
http://minny.cs.inf.shizuoka.ac.jp/SIG-ICK/

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

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

14) M. Hayakawa, "Language acquisition and matherese," *Language*, 35, 9, 62–67, Taishukan pub., 2006

15) 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

16) 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)

17) 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

18) S. Umesh, L. Cohen, N. Marinovic, and D. J. Nelson, "Scale transform in speech analysis," *IEEE Trans. Speech and Audio Processing*, 7, 1, 40–45, 1999

19) T. Irino and R. D. Patterson, "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

20) A. Mertins and J. Rademacher, "Vocal trace length invariant features for automatic speech recognition," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 308–312, 2005

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

22) P. Ladefoged and D. E. Broadbent, "Information conveyed by vowels," *Journal of Acoust. Soc. Am.*, 29, 1, 98–104, 1957

23) T. M. Nearey, "Static, dynamic, and relational properties in vowel perception," *Journal of Acoust. Soc. Am.*, 85, 5, 2088–2113, 1989

24) J. Hawkins and S. Blakeslee, *On intelligence,* Henry Holt, 2004

25) Y. Qiao and N. Minematsu, "A study on invariance of f-divergence and its application to speech recognition," *IEEE Transactions on Signal Processing*, 58, 2010 (to appear).

26) I. Csiszar, "Information-type measures of difference of probability distributions and indirect," *Stud. Sci. Math. Hung.*, 2, 299–318, 1967

27) N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," *Proc. Int. Conf. Acoustics, Speech, & Signal Processing*, 889–892, 2005

28) N. Minematsu, T. Nishimura, K. Nishinari, and K. Sakuraba, "Theorem of the invariant structure and its derivation of speech Gestalt," *Proc. Int. Workshop on Speech Recognition and Intrinsic Variations*, 47–52, 2006

29) N. Minematsu, "Pronunciation assessment based upon the phonological distortions observed in language learners' utterances," *Proc. Int. Conf. Spoken Language Processing*, 1669–1672, 2004

30) D. Saito, R. Matsuura, S. Asakawa, N. Minematsu, and K. Hirose, "Directional dependency of cepstrum on vocal tract length," *Proc. Int. Conf. Acoustics, Speech, & Signal Processing*, 4485–4488, 2008

31) I. Edihammer, "Structure comparison and structure patterns," *Journal of Computational Biology,* 7, 5, 685–716, 2000

32) M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech and Audio Processing*, 13, 5, 930–944, 2005

33) T. Emori and K. Shinoda, "Rapid vocal tract length normalization using maximum likelihood estimation," *Proc. EUROSPEECH*, 1649–1652, 2001

34) M. Naito, L. Deng, and Y. Sagisaka, "Model based speaker normalization methods for speech recognition," *IEICE Trans.* J83-D-II, 11, 2360–2369, 2000.

35) *Tohoku university – Matsushita isolated Word database (TMW)*, http://research.nii.ac.jp/src/eng/list/detail.html#TMW

36) T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, "Recent progress of open-source LVCSR engine Julius and Japanese model repository," *Proc. Int. Conf. on Spoken Language Processing*, 3069–3072, 2004

37) Y. Qiao, M. Suzuki, and N. Minematsu, "A study of Hidden Structure Model and its application of labeling sequences," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp.118–123, 2009

38) S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," *Proc. Int. Conf. Acoustics, Speech, & Signal Processing*, pp.1647–1650, 1997

39) H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, 2, 4, pp.578–589, 1994

40) M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, 51, 10, 832–844, 2009

41) S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, 30, pp.95–108, 2000

42) N. Minematsu, S. Asakawa, and K. Hirose, "Structural representation of the pronunciation and its use for CALL," *Proc. IEEE Int. Workshop on Spoken Language Technology*, 126–129, 2006

43) N. Minematsu, "Training of pronunciation as learning of the sound system embedded in the target language," *Proc. Int. Symposium on Phonetic Frontiers*, CD-ROM, 2008

44) N. Minematsu, *et al.*, "Development of English speech database read by Japanese to support CALL research," *Proc. Int. Conf. Acoustics*, 577–560, 2004

45) U. Frith, *Autism: explaining the enigma,* Wiley-Blackwell, 2003

46) L. H. Willey and T. Attwood, *Pretending to be normal: living with Asperger's syndrome,* Jessica Kingsley Publishers, 1999

47) T. Grandin and C. Johnson, *Animals in translation: using the mysteries of autism to decode animal behavior,* Scribner, 2004

48) N. Higashida and M. Higashida, *Messages to all my colleagues living on the planet*, Escor Pub., Chiba, 2005