Implementation of Robust Speech Recognition by Simulating Infants' Speech Perception Based on the Invariant Sound Shape Embedded in Utterances

N. Minematsu, S. Asakawa, Y. Qiao, D. Saito, T. Nishimura, and K. Hirose

The University of Tokyo, Tokyo, Japan

{mine,asakawa,qiao,dsk_saito,hirose}@gavo.t.u-tokyo.ac.jp, nt-tazuko@ams.odn.ne.jp

Abstract

Recently, a novel and structural representation of speech was proposed [1, 2], where inevitable acoustic biases caused by static extra-lingusitic factors are completely removed from speech. This speech structure is composed of only transform-invariant (topologically invariant) speech contrasts or dynamics [3] with no use of absolute and static acoustic features such as spectrums. Although this framework posed a problem of so strong invariance that two different words could be evaluated as the same, our previous study successfully introduced good constraints to the invariant features [4]. We realized the invariance only with respect to speaker variability through Multiple Stream Structuralization (MSS) [4]. In this paper, after introduction of our proposed representation, we describe that it can be a good mathematical model of infants' speech perception [5, 6, 7, 8], perceptual constancy of speech [9, 10, 11, 12], and Jakobson's classical theory of relational invariance [13, 14, 15]. Next, we show new experimental results using the proposed speech representation. The high robustness is verified again by using frequency warped utterances, which simulate the utterances of very tall speakers and very short ones. Further, we also investigate this representation using a phoneme-balanced word set because, in the previous study, we used only an artificial word set comprised of vowel sequences such as /aeoui/. Results are very promising.

1. Introduction

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. One may hypothesize that listeners adapt their internal acoustic models whenever either of a speaker, a room, a microphone, or a line is changed. Another may hypothesize that the linguistic information in speech can be represented separately from the non-linguistic factors. Recent studies of brain sciences proposed neuroanatomical models of the auditory cortex, where the linguistic features and the non-linguistic features in speech are separately processed in different regions of the human brain [16]. These findings seem to support the second hypothesis of human speech perception.

How can infants acquire the ability of robust speech processing? To implement this ability on machines, engineers have collected speech samples from a huge number of speakers and trained speaker-independent acoustic models statistically. Are those speech samples needed for infants? Recently, especially in the field of artificial intelligence, there is a research trend to focus on infants' acquisition and development of cognitive abilities [17, 18, 19]. 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. We can say definitely that the utterances an individual hears are strongly speaker-biased unless he/she has speaking disabilities. The variability problem should be solved not by collecting samples if one wants to realize a human-like speech processor.

How to separate the linguistic features and the non-linguistic features, both of which are existing in a single speech stream? How to implement the ability of robust speech processing on machines, that infants acquire easily? Psychologically speaking, the question we have is called perceptual constancy of speech. Not only human perception of speech sounds but also that of other stimuli such as colors and tones are very robust although these stimuli inevitably vary due to various environmental factors. In the following section, we briefly introduce our proposal of speaker-invariant representation of speech [1, 2] and, after that, we discuss that it can be a good mathematical model of infants' speech perception, perceptual constancy of speech, and Jakobson's classical theory of relational invariance.

2. Speech representation based on the complete topological invariance

Variability in speech and invariance in its perception [12], we consider that this is one of the very classical and still open questions in speech science and engineering. Many studies proposed models to explain this mystery. In [10] and [11], the models are classified into two procedures, intrinsic (internal) and extrinsic (external) normalization. While, in the former, the relationships among formant frequencies (and fundamental frequency) in a given speech sound is used for normalization, in the latter, those among different sounds, e.g. the entire vowel system, are used for normalization. In both approaches, however, an observation is reformulated to have a different and normalized value.

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. Here, an observation needs no normalization. A critical question to note is how well the transform can characterize real speaker variability. In the previous proposals of the invariant features [20, 21, 22], however, speaker variability was modeled simply as $\hat{f}=\alpha f$ (f=frequency, α=constant). But 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 the entire system of given sounds.

Are there any invariant features with respect to any linear or non-linear invertible transforms? The answer is yes. In [3], we proved that f-divergence [23] between two distributions is invariant with any kind of invertible and differentiable transforms (sufficiency). Further, we also proved that any invariant measure with respect to two distributions has to be written in the form of f-divergence (necessity), which is formulated as

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

Figure 1 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. Generally speaking, the two spaces (shapes) are closed manifolds and the invariance of f-divegence is always satisfied [3].

$$f_{div}(p_1(x,y), p_2(x,y)) \equiv f_{div}(P_1(u,v), P_2(u,v)) \qquad (2)$$

Figure 2 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 are transformed continuously. So, the mug and the doughnut in Figure 2 are identical topologically. Suppose that some 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 pair of the events cannot change at any degree of the deformation. This means that f-divergence-based distance matrix is completely invariant quantitatively. Individual events can change but their system cannot change at all. Suppose that an event is an electron cloud and $g(x)$ in Equation (1) is $\sqrt{x}$. Then, the invariant distance matrix becomes what is called overlap matrix in quantum chemistry, which is one of the measures used to calculate the geometrical shape of molecules or proteins [24].

In a series of our previous studies [1, 2, 3, 4], we have been using Bhattacharyya distance (BD) as one of the f-divergence measures. Figure 3 shows a procedure of representing an input utterance only by BD. The utterance in a feature space is a sequence of feature 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 fix a unique geometrical shape, we call the matrix as speech structure.

Once two utterances are represented as two speech structures, how should these be compared to each other? How to calculate similarity between the two? We already showed a very simple answer [2]. As a distance matrix is symmetric, we can form a vector from the matrix, which is composed of all the elements in the upper triangle of the matrix. This vector is called structure vector, henceforth. As shown in Figure 4, similarity between two structures is defined as the minimum of the total distance between the corresponding two points (events) after one structure is rotated and shifted so that the two structures are overlapped the best. Euclidean distance between the two structure vectors can approximate well the minimum of the total distance [2]. In a cepstrum space, rotation represents cancelation of vocal tract length difference [25] and shift represents cancelation of microphone difference. This means that the structure matching will give us an acoustic similarity score between two utterances after speaker/microphone adaptation. But no explicit adaptation is needed because an adaptation process is embedded internally in the structure matching. In other words, it gives us a useful mathematical shortcut. Chemically speaking, this simple and powerful matching scheme is called Root Mean Square Deviation (RMSD) method [26]. It is often used to calculate structural difference between two proteins without explicit estimation of a mapping function to transform a structure into the other. If absolute positions of individual events in a (parameter) space are used as observation, however, the function has to be estimated. As far as we know, the conventional speaker adaptation/normalization methods are based on this strategy and this is why acoustic models have to be updated whenever either of a speaker, a room, a microphone, or a line is changed.

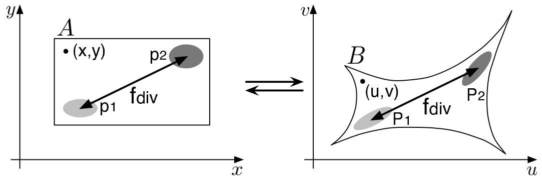

Figure 1: Topological deformation of manifolds (shapes)

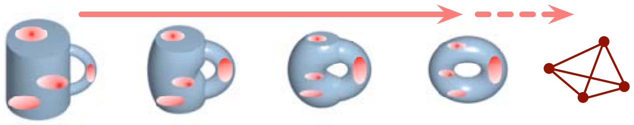

Figure 2: Complete topological invariance

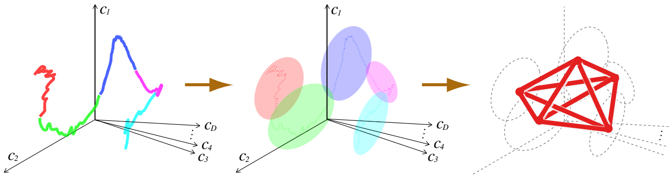

Figure 3: Utterance structure composed only of f-divergence

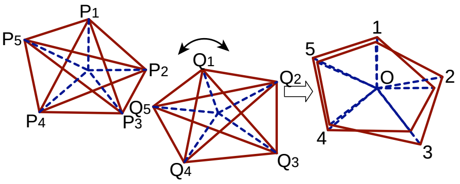

Figure 4: Structure matching after shift and rotation

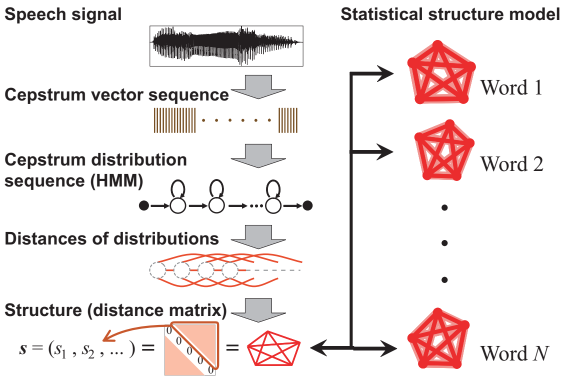
Speech signal
Statistical structure model
Cepstrum vector sequence
Cepstrum distribution sequence (HMM)
Distances of distributions
Structure (distance matrix)
s = (s1 , s2 , ...) =
=
Word 1
Word 2
Word N

Figure 5: Basic framework of structure-based word recognition

Figure 5 shows the basic framework of isolated word recognition based on speech structures. To convert an utterance into a distribution sequence, the MAP(Maximum A Posteriori)-based training procedure of HMMs is adopted. Then, the BD between any pair of the distributions is obtained. After calculating the structure, absolute properties of speech such as spectrums are completely discarded. The right-hand side of the figure shows an inventory of word-based statistical structure models (Gaussians) for the entire vocabulary. The candidate word showing the maximum likelihood score is a result of recognition.

In the following section, this framework is evaluated from viewpoints of developmental psychology, cognitive science, and linguistics. After that, some new experimental results are shown.

3. Psychological and linguistic interpretation of the speech structure

3.1. Link to infants' ability of oral communication

As discussed in Section 1, it seems that infants acquire the ability of robust speech processing by hearing strongly speaker-biased utterances. And infants imitate their parents' utterances actively, called vocal imitation, but they don't impersonate their parents. Here, we have a question. 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 small mouths. But researchers of infant study deny this claim because infants don't have good phonemic awareness [6, 7]. Then, what is imitated acoustically?

An answer from infant studies is the holistic sound pattern embedded in an utterance [6, 7], called otherwise as word Gestalt [8] and related spectral patterns [5]. The holistic pattern has to be speaker-invariant because, whoever speaks a specific word to an infant, its responses of imitation are similar acoustically. We consider that the speech structure in Figure 3 is mathematical and acoustic implementation of the word Gestalt [2].

The vocal imitation is rare in animals [28] and non-human primates don't imitate the utterances of others [27]. This performance can be found in only a few species of animals, i.e. birds, whales, and dolphins. But there is a critical difference between humans and animals. The vocal imitation of animals is the imitation of sounds [28]. Take myna birds for example. They imitate the sounds of cars, dogs as well as human voices. Hearing a myna bird say something, one can guess its human owner [29] but cannot guess the parents of an infant by hearing its voices. The ability of extracting an abstract and scale-invariant sound pattern from a sound stream might be unique to humans.

3.2. Link to perceptual constancy of non-speech stimuli

As discussed in Section 1, not only speech sounds but also colors, tones, and so forth are very variable actually but our perception of these stimuli is robust and constant. Although we perceive them through different physical media, it seems that researchers found that a similar mechanism is working to cancel static biases and realize the perceptual constancy [30, 31, 32]. Figure 6 shows the look of the same Rubik's cube through differently colored glasses. Although the corresponding tiles of the two cubes have different colors absolutely, we name them using the same labels. In contrast, although different colors are perceived for the four *blue* tiles on the top of the left cube and the seven *yellow* tiles on the right, when their surrounding tiles are hidden, we can find easily that they have the same color (See Figure 7). Absolutely different colors are perceived as identical and absolutely identical colors are perceived as different.

Similar phenomena can be found in sounds. Figure 8 shows two melodies. One is a transposed version of the other. If hearers have relative pitch and can transcribe these two melodies, their transcriptions using syllable names are identical between the two (So Mi So Do...). The first tone of the upper and that of the lower are different absolutely but they name these tones by 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. Absolutely different tones are perceived as identical and absolutely identical tones are perceived as different.

Researchers of psychology found that the perceptual constancy of colors and tones occurs commonly based on contrast-based information processing [30, 31, 32]. In other words, our constant perception of colors and tones is guaranteed by the invariant relations of the focused stimulus to its surrounding stimuli. As was found in ecology, the constant color perception occurs even to butterflies and bees [33]. This color perception is very old evolutionarily. In contrast, researchers of anthropology found that the constant tone perception is difficult even for monkeys. Non-human primates can hardly perceive the equivalence between a melody and its transposed version [34]. Relative pitch perception is very new. Considering the discussion in Section 3.1, animals seem to be good at dealing with visual deformation but poor at acoustic deformation. This is one of the reasons why most of the research trials to teach a human language to chimpanzees had to adopt visual signs, not oral ones. The human voices are too difficult to deal with adequately [35].

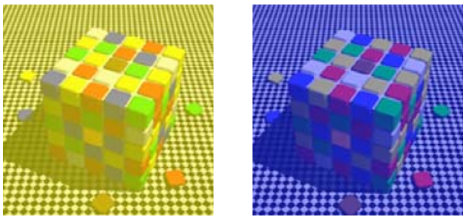

Figure 6: The same cube seen through two colored glasses

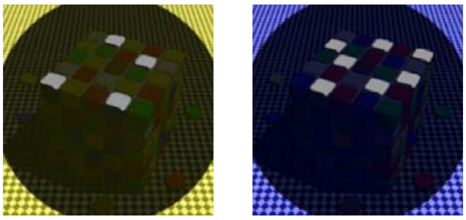

Figure 7: Perception of colors without context

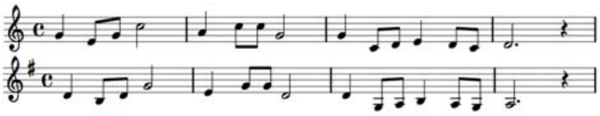

Figure 8: A musical melody and its transposed version

3.3. Link to Jakobson's classical theory of language

As is well-known, Jakobson proposed a theory of relational invariance, called distinctive feature theory. In [13], 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). "The 'given' is a multiplicity and a transformational group; the patterns to which this multiplicity is related have to be investigated with respect to those properties which remain unaffected by the transformations of the group." "Physiologically identical sounds may possess different values in conformity with the whole sound system, i.e. with their relations to the other sounds." "We have to put aside the accidental properties of individual sounds and substitute a general expression that is the common denominator of these variables." A difference between a language sound and an acoustic sound is that the former is "placed with reference to other sounds", i.e. "the relational gaps between the sounds of a language". Jakobson also denies a concept of absolute invariance and a similar claim is found in [14].

In a classical study of phonetics, the importance of relational invariance was experimentally verified as for vowel perception [9]. It is interesting that Lagefoged discussed a good similarity between perception of vowels and that of colors.

In the following sections, new experimental results are shown after solving two problems of our invariant speech structures.

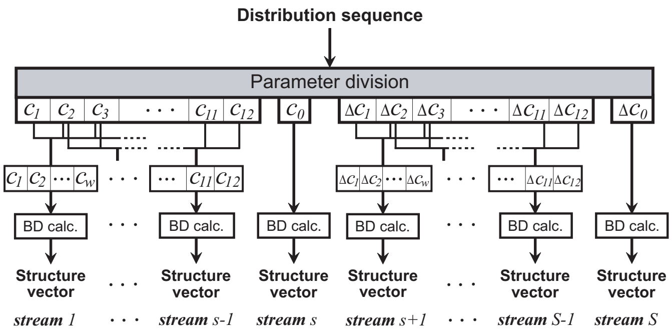

Figure 9: Multiple Stream Structuralization (MSS)

4. Two problems and their solution

4.1. Too strong invariance of speech structures

The proposed speech structure is invariant with any kind of invertible transform. This led us to expect that two different words can be evaluated as the same and this expectation was correct. To solve this problem, we introduced good constraints [4] so that we could obtain the invariance only with speaker variability. Vocal tract length difference causes non-linear frequency warping and [25, 36] showed that this warping can be modeled in the cepstrum domain approximately as multiplying cepstrum vector c by matrix A ($c'\!=\!Ac$). BD is completely invariant with any kind of A and this invariance is too strong. A in [25, 36] is a band matrix and what we want is the invariance only for band matrices. This constrained invariance was obtained successfully by Multiple Stream Structuralization (MSS) [4]. Figure 9 shows its procedure. For a mean vector of a distribution of an HMM converted from an input utterance, w consecutive cepstrums form a sub-vector and w Δcepstrums form another one. Here, we have S sub-vectors (sub-streams) totally. Using a sequence of sub-vectors, a sub-structure is constructed. Geometrically speaking, a speech structure in the entire space is projected into sub-spaces and, in each sub-space, sub-structure matching is done. The final similarity score is obtained by summation of the scores of the individual sub-spaces. Detailed description of MSS is found in [4], where it was tested with a set of vowel sequence words spoken by adult speakers. In this paper, MSS is tested with both of the vowel word set and a phoneme-balanced word set generated through a much larger speaker variability.

4.2. Too high dimensionality of speech structures

The other problem is that the dimension of parameters is increased with $O(n^2)$, where n is the number of distributions in an utterance and the number of edges in a structure is ${}_nC_2$ (See Figure 3). Then, the total number of edges is S_nC_2. To reduce the dimension and to increase discriminability simultaneously, in this paper, widely-used Linear Discriminant Analysis (LDA) is introduced in two stages. Figure 10 shows the procedure. After MSS, LDA is carried out for each sub-stream (sub-structure), which is the 1st stage LDA. $W_i(i\!=\!1...S)$ is a transform matrix. Then, all the transformed sub-structure vectors are concatenated to form a single integrated vector. This vector is transformed again with W_{all}, which is the 2nd stage LDA. The resulting vector is used for matching with pre-stored templates.

4.3. Use of inter-stream distance as additional feature

In Figure 10, the individual sub-streams are treated independently. It is reasonable to consider that the use of distance between a pair of sub-streams (sub-structures) will improve the performance. Then, another vector is formed for S sub-structures, which is composed of ${}_SC_2$ inter-sub-structure distances. This new vector is concatenated to the integrated vector in Section 4.2. The resulting vector is transformed in the 2nd stage LDA. The effectiveness of Inter-Stream Distance (ISD) is examined shortly.

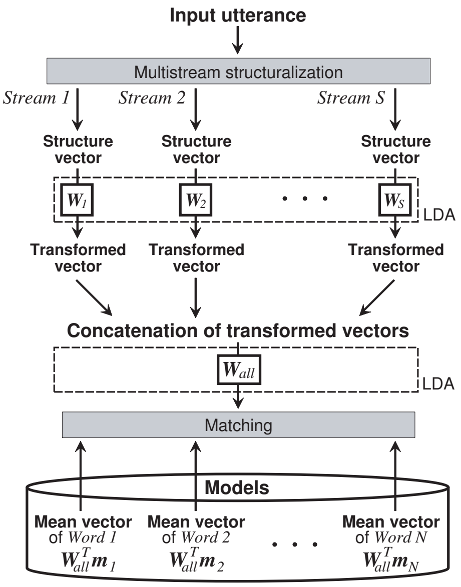

Figure 10: Structure matching through 2-stage LDA

5. Isolated word recognition experiments

5.1. Vowel sequence words and phoneme-balanced words

We prepared two word sets. One is a vowel sequence word set, where each word is a five-vowel sequence such as /eoiau/. Since Japanese has only five vowels, the vocabulary size is 120 ($={}_5P_5$). The other set is a Japanese phoneme-balanced word set [37], which is often used in the Japanese ASR community to verify the effectiveness of new techniques. The word length in phonemes varies from 3 to 10 and the vocabulary size is 212. Considering that vowel sounds are much more dependent on speakers than consonants, it is reasonable to expect that the proposed technique is more appropriate for the first word set.

5.2. Simulated speaker variability for robustness evaluation

With matrix A, various kinds of non-linear frequency warping were applied to the word utterances. Considering the fact that the tallest adult in the world is 257 cm high and the shortest adult is 74 cm high, the warping was done to cover this body height range. In real situations, however, we can hardly see such tall or short speakers but we hear them on TV not rarely. The voices of some animation characters are created by transforming real human voices and children can understand their utterances easily. What about the current speech recognizers?

5.3. Experimental conditions

The acoustic analysis condition for structure extraction and matching in the case of unwarped (original) utterances is shown in Table 1. For comparison, word-based HMMs were built with the same training data. The condition for the HMMs is shown in Table 2. In the case of warped utterances, however, the condition was slightly changed. FFT-cepstrums (0 to 16 dim) were used both for structures and HMMs. This is because Mel trans-

Table 1: Conditions for structures with unwarped utterances

sampling	16bit / 16kHz (vowel word set)
	12bit / 16kHz (balanced word set)
window	25 ms length and 10 ms shift
parameters	Mel-Cepstrum (0 to 12) + Δ (0 to 12)
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

Table 2: Conditions for HMMs with unwarped utterances

sampling	16bit / 16kHz (vowel word set)
	12bit / 16kHz (balanced word set)
window	25 ms length and 10 ms shift
parameters	MFCC (1 to 12) + Δ (1 to 12) + ΔP
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	ML

formation is just a frequency warping and corresponds to shortening the vocal tract length. The effect of Mel transformation is expected to be cancelled due to the invariance of structures.

Both for structures and HMMs, the number of distributions, n, was set to 20 for each vowel word and it was 25 for each balanced word. For the vowel words, the number of speakers for training was 8 (4 males and 4 females) and that for testing was 8 (other 4 males and 4 females). For the phoneme-balanced words, 30 speakers (15 males and 15 females) were used for training and other 30 speakers were for testing. For the former, each speaker uttered the word set five times. Then, each structure was trained with 40 samples. For the latter, each speaker uttered once. Each structure was built with 30 samples.

5.4. Results of the experiments

Figure 11 shows the performance for unwarped (original) utterances of the two word sets. The upper figure is for the vowel set and the lower is for the balanced set. In both figures, the X-axis shows block size w. The baseline performance, which is obtained by word-based HMMs, is 98.3% and 99.6%. As for the speech structures, the performance with/without Δ parameters is shown. With Δ, the number of streams, S, is $2(14-w)$ and, without it, S=$14-w$. In MSS, the number of edges for each sub-stream is ${}_{20}C_2$=190 for the vowel set and it is ${}_{25}C_2$=300 for the balanced set. In both cases, in the 1st stage LDA, the number of parameters is reduced to 20. In the 2nd LDA, it is 119 and 211 (the vocabulary size -1). The best performance of the structures is 98.9% for the vowel set (w=3 with MSS+LDA+Δ+ISD) and is 96.4% for the balanced set (w=1 with MSS+LDA+Δ+ISD).

Figure 12 shows the performance for warped utterances. All the results were obtained with MSS+LDA+Δ+ISD and w varied from 1 to 16[1]. The X-axis represents warping parameter α [25, 36]. With positive/negative values of α, the vocal tract length is shortened/lengthened. α=0.4 halves the length and α=-0.4 doubles it approximately. **HMM** in the figures means the performance of the word-based HMMs trained with the same training data (unwarped utterances) that were used in training the structures[2]. **matched** means the performance of 17 sets of word-based HMMs, which were separately trained with each value of α and separately used for recognizing the warped utterances of the corresponding value of α. In other words, **matched** shows the performance with no mismatch. We expected that the performance of a single set of structures would be comparable to that of the 17 matched sets of HMMs.

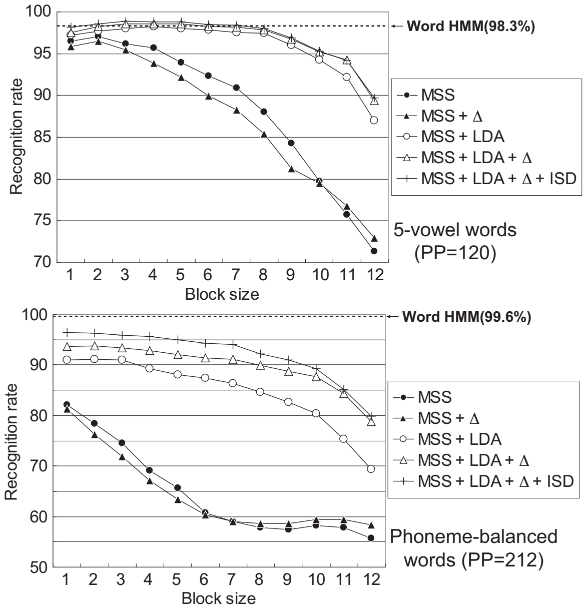

Figure 11: Word recognition rates for unwarped utterances

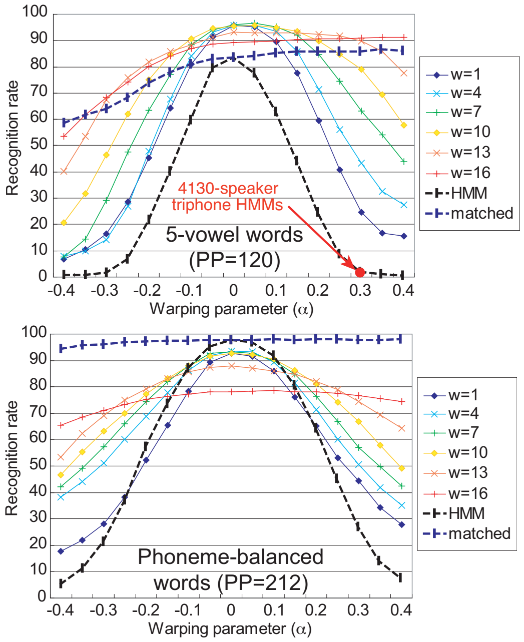

Figure 12: Word recognition rates for warped utterances

5.5. Discussions

In Figure 11, LDA is effective especially for the balanced set. This is partly because speech structures are formed in a somewhat inappropriate way for words including speaker-independent sounds, e.g. unvoiced fricatives and plosives. As for Δs, they are harmful without LDA but, with it, they increase the performance. ISD is more effective in the balanced set.

Figure 12 clearly shows the high robustness of structure-based speech recognition to speaker variability. In the vowel set, the performance of a single set of structures (w=16) is com-

[1] Cepstrum coefficients of 0-th to 16-th dimensions were used here.

[2] In this experiment, both the structures and the word-based HMMs were trained and tested with FFT-cepstrums as explained in Section 5.3.

parable to or higher than that of the 17 matched sets of HMMs.

We carried out another experiment. Speaker-independent triphone HMMs, which were trained with 4,130 speakers and are used as the standard HMMs by the Japanese academic ASR community [38], were tested against the utterances warped with α=0.3[3]. The language model was CFG allowing only the 120 words. The performance was 1.4%, by far lower than that of the structures (91.0%, w=16), trained only with 8 speakers.

In the balanced set, the performance of structures is worse than that of HMMs in the matched condition (α=0). In the mismatched conditions, however, the high robustness of structures is shown (w=10, 13). In both of the word sets, although larger values of w increase speaker-invariance, they induce misrecognition. This tendency is very natural because a major part of speaker difference and word difference are commonly attributed to timbre (spectrum) difference. Speaker-invariance and word discrimination are trade-off and w can control it.

Structure-based ASR is possible in other domains than cepstrums. For example, spectrum-based structures are very 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 and RASTA. All of these capture only the dynamic aspect of speech but only the structure grasps it in a speaker-invariant way. This invariance is obtained basically by removing the directional features of a speech trajectory because they are strongly speaker-dependent [25] and by extracting and modeling only the speech contrasts, including temporally distant contrasts (see Figure 3).

6. Conclusions

In this paper, after briefly describing our previous proposal of the invariant speech structure, we clarify the linkage of the structure to infants' (and maybe human-unique) oral communication ability, perceptual constancy of non-speech stimuli, and Jakobson's classical theory of language sounds. Here, considering new findings of animal sciences and evolutionary anthropology, we discuss a difference in sound perception between humans and animals. Animals are weak at acoustic deformation. After that, isolated word recognition experiments are carried out with the speech structures and the HMMs. The difference between them is what is modeled acoustically. In the structures, speech contrasts are modeled and, in the HMMs, speech substances are modeled. Experimental results show merits and demerits of using the structures, i.e. the high robustness to speaker variability but somewhat reduced discrimination among unvoiced consonant sounds. We are interested in integrating both the models for compensation because humans are very relative in their perception but it is also true that we had evolved from animals.

7. References

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

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

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

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

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

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

[9] P. Ladefoged *et al.*, "Information conveyed by vowels," *J Acoust. Soc. Am.* 29, 1, 98–104, 1957

[10] W. Ainsworth, "Intrinsic and extrinsic factors in vowel judgments," in *Auditory Analysis and Perception of Speech,* edited by G. Fant and M. Tatham, Academic Press, 1975

[11] T. M. Nearey, "Static, dynamic, and relational properties in vowel perception," *J Acoust. Soc. Am.,* 85, 5, 2088–2113, 1989

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

[13] R. Jakobson and L. R. Waugh, *The sound shape of language,* Mouton De Gruyter, 1987

[14] G. Fant, "The role of speech research in the advance of speech technology," *Quarterly progress and status report*, Dept. speech, music and hearing, KTH, 1990

[15] R. B. Sangster, *Roman Jakobson and beyond – the quest for the ultimate invariants in language,* Mouton De Gruyter, 1983

[16] S. K. Scott *et al.*, "The neuroanatomical and functional organization of speech perception," *Trends in Neurosciences,* 26, 2, 100–107 (2003)

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

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

[19] Infant Commonsense Knowledge Project http://minny.cs.inf.shizuoka.ac.jp/SIG-ICK/

[20] S. Umesh *et al.*, "Scale transform in speech analysis," *IEEE Trans. Speech and Audio Processing*, 7, 1, 40–45, 1999

[21] 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

[22] A. Mertins *et al.* "Vocal tract length invariant features for automatic speech recognition," *Proc. ASRU*, 308–312, 2005

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

[24] Y. Harada, *Quantum chemistry,* Shokabo Pub., 1975

[25] D. Saito *et al.*, "Decomposition of rotational distortion caused by VTL difference using eigenvalues of its transofmation matrix," *Proc. Interspeech*, 1361–1364, 2008

[26] I. Edihammer, "Structure comparison and structure patterns," *J. Computational Biology,* 7, 5, 685–716, 2000

[27] 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

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

[29] K. Miyamoto, *Making voices and watching voices,* Morikawa Pub., 1995

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

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

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

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

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

[35] S. Kojima, *A search for the origins of human speech – auditory and vocal functions of the chimpanzee,* Trans Pacific Press, 2003

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

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

[38] T. Kawahara *et al.*, "Recent progress of open-source LVCSR engine Julius and Japanese model repository," *Proc. ICSLP*, 3069–3072, 2004

[3] In this experiment, MFCCs were extracted and used with CMN.