Speaker-invariant and rhythm-sensitive representation of spoken words

Nobuaki Minematsu*, Yousuke Ozaki†, Keikichi Hirose†, and Donna Erickson‡
* Graduate School of Engineering, The University of Tokyo, Tokyo, Japan
E-mail: mine@gavo.t.u-tokyo.ac.jp
† Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
E-mail: {ozaki,hirose}@gavo.t.u-tokyo.ac.jp
‡ Showa University of Music, Kanagawa, Japan
E-mail: ericksondonna2000@gmail.com

***Abstract*—It is well-known that human speech recognition (HSR) is much more robust than automatic speech recognition (ASR) [1], [2]. Given that HSR's robustness to large acoustic variability is extremely high, it is reasonable for researchers to assume that humans are able to extract invariant patterns underlying input utterances [3]. Recently in developmental psychology, it was found that infants are very sensitive to distributional properties in the sounds of a language [4], [5]. Following this finding, the first author proposed a speaker-independent or invariant speech representation of each utterance, formed by using distributional properties in the sounds of that utterance [6], [7], [8]. This representation is called speech structure and was tested in isolated word recognition experiments [7], [8]. This paper introduces another kind of sensitivity into speech structure, that is sensitivity to language rhythm. Sonority-based syllable nucleus detection is implemented and we extract local and syllable-based structures as well as conventional global and holistic structures. Isolated word recognition experiments show that the recognition performance is improved with rhythm-sensitive and local speech structures.**

I. Introduction

In developmental psychology and speech science, a great deal of knowledge regarding first language acquisition (LA) has been accumulated. However, the underlying principles of the learning process are not well-known yet. This is considered to be because language is an extremely complex phenomenon. In this situation, computational models play an important role to deepen researchers' understanding of LA [9] because researchers can verify their own models quantitatively as well as qualitatively through computer simulation. However, these models can often explain only certain aspects of LA, not all.

One of the classical but still open questions in LA is about how humans acquire the ability of super robust speech recognition [2], [9]. Especially, [2] claimed that the current ASR model lacks the ability to generalize, something that human infants acquire easily. In training automatic speech recognizers, a large speech corpus with high speaker variability is often needed. In the case of infants' LA, however, a majority of speech samples exposed to infants are from their parents and caretakers. We can say that ASR needs speaker-*balanced* speech samples but it seems true that HSR has no problem with speaker-*biased* speech samples. Another approach to training ASR models with a relatively small corpus is speaker-adaptive training [10]. In this approach, input speech features are normalized and transformed to be features of the imaginary reference speaker who is assumed to have an average vocal tract length. Recently, new computational models were proposed for LA in [11], [12], where acoustic mapping or manifold alignment was investigated among infants and adults. Speaker-adaptive training in ASR models and these new computational models in HSR are similar in that both approaches try to transform speech features of a source speaker to those of a target speaker. In speaker adaptive training, the target speaker is the imaginary reference speaker and in the LA models, the target speaker is an adult and the source speaker is an infant.

With regard to the speech variability problem in this paper, we investigate a totally different speech model or speech representation where the acoustic aspect of an utterance which corresponds to speaker identity is removed effectively from the speech acoustics. What remains can reasonably become a speaker-independent or speaker-invariant representation, called speech structure [6], [7], [8][1]. Unlike the proposals made in [11], [12], our model does not need any explicit mapping or alignment between different speakers. Mapping is required in [11], [12] because these models represent speech acoustics with speaker identity still included. In acoustic analysis of speech, the spectrum envelope is often extracted from the power spectrum by removing pitch harmonics. So, the envelope pattern is pitch-independent. Similarly in our model, speaker identity is removed from speech acoustics based on mathematical equations to represent speaker identity in speech.

In our previous study [13], [14], we discussed this speech model by associating its invariant properties to infants' good abilities to generalize because this model seems to be much in accordance with recent findings of infants' performance in

[1]The term of "speaker-independent" is often used to indicate *statistical* independence. Since the theory of probability defines $P(a) = \sum_b P(a,b)$, through collecting samples, any variable can be treated as a *hidden* variable. Speaker-independent HMMs are trained by this strategy using a speaker-*balanced* corpus to hide speaker identity in the distribution. In this paper, we focus on another kind of independence, which should be referred to as *physical* independence, where a variable can disappear by physically removing or separating it from observations [16]. In speech analysis, phase and pitch are often removed physically from observations. In this paper, we claim that speaker identity can also be removed.

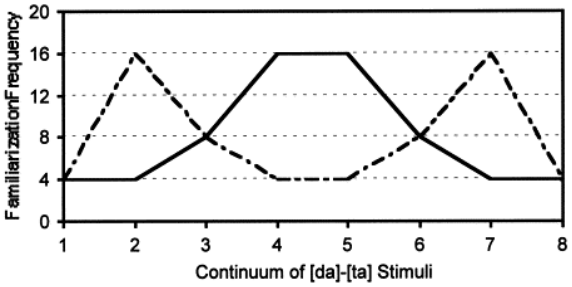

Fig. 1. Two different distributions of language sounds [4]

LA. Infants were found to be very sensitive to distributional properties in the sounds of a language [4], [5]. In [13], [14], we investigated this model as a model of infants' spoken word acquisition and in the current paper, we attempt to incorporate into our model yet another parameter in infants' sensitivity, that is sensitivity to language rhythm [15]. Our new structural model is evaluated in spoken word recognition experiments and it shows better performance than our old model.

This paper is organized as follows. In Sect. II, recent findings of infants' sensitivity to distributional properties of language sounds are explained and some related facts and discussions found in dialectology and classical theory of phonology are described. Our invariant speech structure model is introduced in Sect. III and in Sect. IV, sonority peak detection is technically implemented based on prior research. Our proposal of syllable-based and local speech structures are explained in Sect. V and they are tested in the task of isolated word recognition experiments in Sect. VI Finally, Sect. VII concludes this paper.

II. Infants' sensitivity to distributional properties in speech

Americans can discriminate [r] and [l] easily but Japanese generally do not discriminate these two sounds. If infants in a language environment can discriminate two sounds of x and y and those in another language environment do not, one can claim that infants' language acquisition is affected by that environmental difference. In [4], it was found that infants' performance of sound discrimination is easily affected by environmental differences in terms of distributional properties of language sounds in the environments.

In [4], two groups of infants were placed in two different sound environments just for several minutes. One environment is characterized by a unimodal frequency distribution of speech sounds and the other is by a bimodal distribution. Fig. 1 shows both distributions. In this study, speech stimuli are a continuum of [da] to [ta] and intermediate stimuli were generated by a speech synthesizer. After several minutes' familiarization to these sound environments, the same sound pair of stimuli, that are 1 and 8 in Fig. 1, were presented to the two groups of infants. Experimental results showed that only the infants in the bimodal distribution environment can discriminate these two sounds and the authors concluded that infants are very

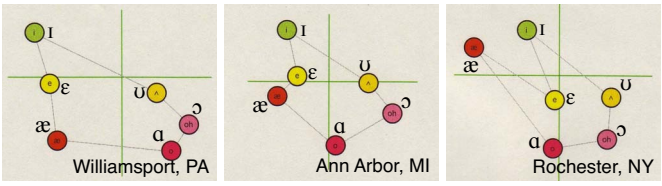

Fig. 2. Dialect-specific vowel distributions in American English [17]

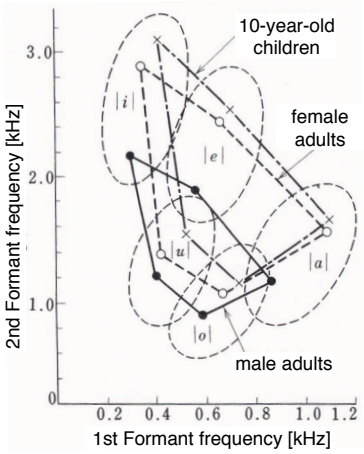

Fig. 3. Vowel distribution patterns in three groups of Japanese speakers [18]

sensitive to distributional properties of language sounds in the environment. In [5], infants' sensitivity to distributional properties was investigated again.

Sensitivity of infants to distributional properties of language sounds when acquiring a language is easy to understand when we consider dialectal differences in a language. Fig. 2 shows several distribution patterns of six vowels of American English dialects [17]. The vowels are plotted on an F1/F2 plane after vocal tract length normalization. It is well-known that different dialects show different vowel distribution patterns. When infants are born and brought up in an geographical area, they inevitably acquire the dialect pronunciation in that area. It is reasonably evident that infants are very sensitive to the sound distribution pattern in their process of LA.

On the other hand, to which aspect of speech are infants insensitive? Fig. 3 shows Japanese vowel distributions of male adults, female adults, and 10-year-old children [18]. It is also well-known that formant frequencies strongly depend on the vocal tract length of a speaker. Adults have lower formant frequencies and children have higher formant frequencies. In infant studies, it is often said that infants' language acquisition is based on their vocal learning, which includes vocal imitation of utterances of their parents or caretakers. It should be noted that their vocal imitation is not acoustic imitation. Infants do not impersonate their parents or caretakers but they do learn the sound distribution pattern. Considering this performance, we can say that infants are not sensitive to absolute properties in speech sounds but sensitive to relational or distributional properties in them. Putting it in another way, infants are sensitive to the sound system of a language.

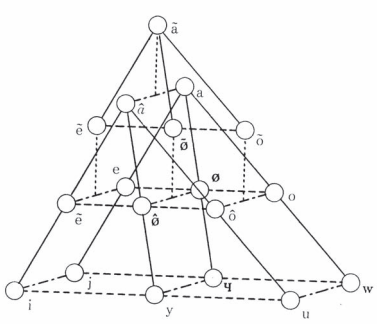

Fig. 4. Jakobson's structure of the French vowels and semi-vowels [20]

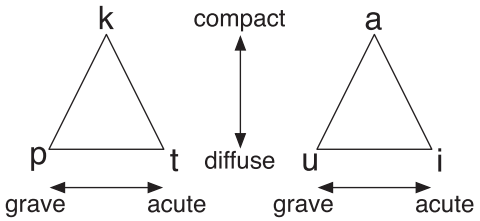

Fig. 5. Consonant and vowel triangles and related features [21]

Similar discussions can be found in classical phonological literature. R. Jakobson proposed a theory of relational invariance, called distinctive feature theory. In [19], he repeatedly emphasized the importance of relational and systemic invariance among speech sounds. "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." In [20], he drew the invariant system of French vowels and semi-vowels, shown in Fig. 4. The distinctive features were proposed to describe the relation or difference between sounds. Two simple sound systems and their related features are shown in Fig. 5.

What is the simplest definition of a (sound) system? Geometrically speaking, the shape of a three-point structure (a triangle) can be defined by the length of the three edges of the triangle. What about an n-point structure? In this case, the length of all the edges including the diagonal edges can define the shape of that structure, shown in Fig. 6. The distance matrix extracted from an n-point structure is the simplest definition of the shape of that structure. If the distance matrix representing the sounds generated by a speaker and the matrix representing the sounds of the same message generated by another speaker is the same, we can say that those matrices are speaker-invariant or speaker-independent and that infants seem to be sensitive to the matrix properties because, geometrically speaking, the matrix is one of the simplest definitions of distributional properties. How can one measure the length of an edge of an n-point structure in a speaker-invariant way? This question is dealt with in the following section.

Recently in ASR studies, some researchers pay special attention to the distinctive feature theory [22], [23], [24]. Here,

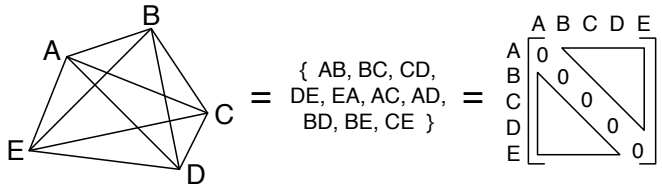

Fig. 6. The distance matrix of a structure defines its shape.

features are often referred to as "attributes" [22], [23] and the researchers tried to define the features acoustically and detectors of the features or the attributes are used in the front-end of ASR systems. R. Jakobson introduced the features to describe the invariant relational properties qualitatively. In our study, instead of searching for a good acoustic definition of the features, the invariant shape of sounds which underlies an input utterance is extracted and used for speech processing. We understand that the features were introduced by R. Jakobson to describe this invariant shape.

III. Invariant speech structure

A. Derivation of invariant speech structure

How to calculate distance between two sounds in a speaker-invariant way? In this section, a mathematical solution proposed in [6], [7], [8] is explained. Speaker difference is modeled mathematically as space mapping in studies of voice conversion. In [8], we proved that f-divergence between two distributions is invariant with any kind of invertible and differentiable transforms (sufficiency) and that any invariant measure with respect to two distributions has to be written in the form of f-divergence (necessity). f_{div} is a distance metric between two distributions, p_1 and p_2, and it is formulated 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 $t > 0$ [25]. 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. Fig. 8 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 are closed manifolds and the invariance of f-divergence is always satisfied [8].

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

In [6], [7], [8], we have been using the Bhattacharyya distance (BD) as one of the f_{div} measures. If an input utterance is represented as a BD-based distance matrix by using only the distributions found in that utterance, the matrix is an invariant representation of that utterance. Fig. 7 shows a procedure of representing an input utterance only by BD. The utterance in a feature space, such as cepstrum space, is a sequence of feature vectors and it is converted into a sequence of distributions

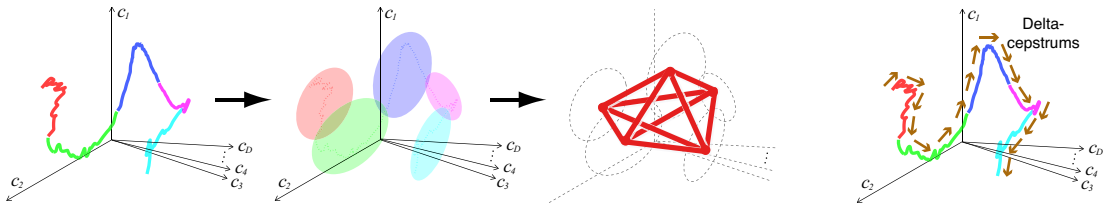

Fig. 7. Utterance structure composed only of f-divergence

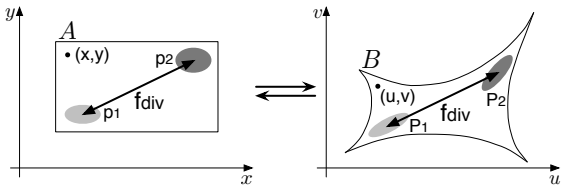

Fig. 8. Invariance of f-divergence against topological deformation of manifolds (shapes)

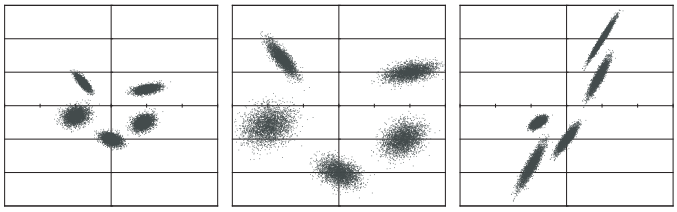

Fig. 9. Feature modification by linear transformation. A common and invariant underlying structure exists among these three sets of distributions.

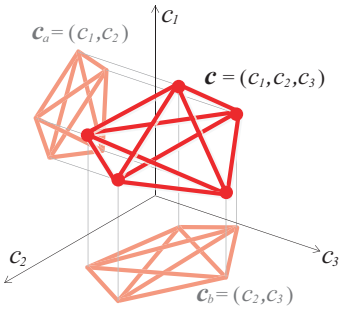

Fig. 10. Multiple Stream Structuralization (MSS) [7]

through automatic and unsupervised segmentation. Here, any speech event is characterized as a distribution. The BD is calculated from any pair of distributions and we can get an invariant distance matrix for that utterance. This matrix-based invariant representation is a speech structure.

Here, we should note that velocity vectors, relative changes at each point on the time axis (See the right-hand side of Fig. 7), are not good candidates for speaker-invariant features. This is because vocal tract length normalization can be appropriated as rotating a feature trajectory [26], [27] and the direction of velocity vectors are strongly dependent on the vocal tract length [27]. When the acoustic feature of interest is a one-dimensional feature, such as fundamental frequency, since rotation is geometrically impossible, velocity vectors can become invariant features. Perception of relative and directional changes in fundamental frequency is often called relative pitch in musicology and, due to this, one can perceive syllable names, not pitch names, of Do, Re, Mi,..., in a key-invariant way. We can say that an f_{div}-based distance matrix is an extended concept of relative pitch, that will be relative *timbre*. Because pitch is one-dimensional and timbre is multi-dimensional, invariance can be found as directional and local changes in the former but in the latter, it can be found only as local contrasts and distant contrasts between acoustic events. We can say our invariant structure is a general solution of finding invariance in dynamics of multi-dimensional features.

Invariance and discrimination have a trade-off relation. Models that are too invariant will reduce the performance of discriminating different words. For example, [26] found that change of cepstrum features by lengthening or shortening the vocal tract can be approximated as a linear transformation. If we assume a single Gaussian for each acoustic event in Fig. 7, invariance is satisfied only with linear transformations. Fig. 9 shows several examples of linearly transformed distributions. Among these three sets of distributions, a common and invariant underlying structure exists. Among three sets of vowel distributions in Fig. 3, we can find differences in their sizes, which can be cancelled in our structure model, shown in Fig. 9. Further, [26], [27] show that the transformation matrix for vocal tract length change becomes a band matrix. To obtain invariance only with band matrix transformations, we proposed multiple stream structuralization (MSS) [7], where a feature stream is divided into several sub-streams and for each sub-stream, an f_{div}-based distance matrix was calculated and used for recognition. The structure for a sub-stream will be called sub-structure. In Fig. 10, a three-dimensional structure is decomposed into two two-dimensional sub-structures.

[9] gives a good survey of studies of computational models of LA, where the models are divided into two paradigms, phonetic learning and lexical learning. In the former, it is claimed that phonetic categories are first acquired and then used for lexical learning. In the latter, word acquisition is hypothesized to come first. In our speech structure, an utterance has to be divided into several events by automatic and unsupervised segmentation but they do not have to be classified. Therefore, our speech structure is regarded as a model of lexical learning. Based on findings in [4], [5], however, the distributional properties can be associated with phonetic category learning.

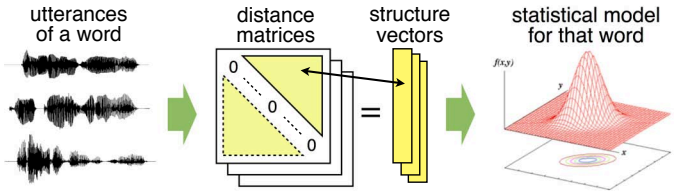

Fig. 11. Statistical structure model

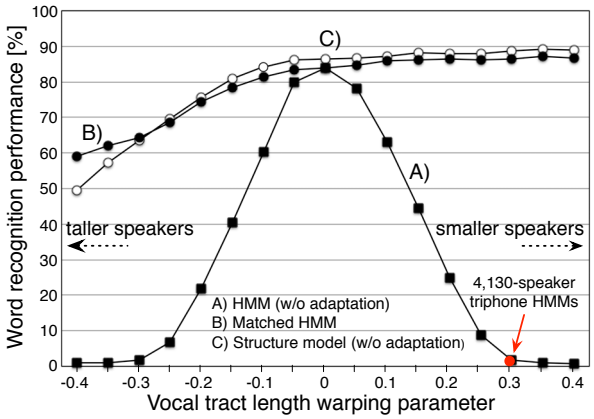

Fig. 12. Results of word recognition experiments [8]

In [28], it is shown that acoustic details, which include speaker information, are effectively used in human perception of spoken words. Our structure model is speaker-independent, where speaker identity is supposed to be removed, and this is supported by other cognitive science studies. [29] shows that, although infants pay attention to acoustic details of input utterances at the very early stage, they start shortly to generalize acoustic variations in the utterances, where they can treat linguistically identical but acoustically very different spoken utterances as identical messages. Further, [30] explains interesting performances of an autistic boy. He can understand easily what his mother says but it is difficult to understand what anybody else says. It seems that he maps the acoustic details of his mother's utterances to meaning. It is widely known that autistic individuals have difficulty in generalizing sensory stimuli but have extremely good memory of detailed aspects of the stimuli [31].

B. Use of speech structure for isolated word recognition

In our previous study [8], we tested our structure model in isolated word recognition experiments which used a vocabulary artificially designed for that experiment. The Japanese five vowels (/a/,/i/,/u/,/e/,/o/) were arranged variously to produce 120 five-vowel sequences (words) such as /eoaui/. Conversion from a feature sequence into a distribution sequence was realized as HMM training with only the input utterance. The 20-state HMM was adopted as reference topology and any word utterance was characterized as a 20-distribution sequence. Then, a 20×20 distance matrix was calculated and it was used as a feature vector (structure vector) for that input word utterance. Parameter estimation becomes very unstable when only a single sample is used for HMM training. So, the parameters were estimated with MAP adaptation. For word template models, we built a statistical model for each word by using structure vectors of that word, shown in Fig. 11. The likelihood scores of an input structure vector were calculated for the template models and used for word recognition. The detail experimental setup is found in [8]. Results of word recognition experiments are shown in Fig. 12, where word utterances of very tall speakers and very small speakers were artificially prepared as testing samples by using vocal tract length warping [26]. The x-axis of Fig. 12 indicates the value of the warping parameter. If it is negative, an input speaker becomes taller and if it is positive, he becomes smaller. Three methods were compared. A) word HMMs without model adaptation, B) 17 sets of word HMMs trained in matched conditions and C) structure models without model adaptation. The baseline word HMMs are extremely weak when acoustic mismatch takes place between training and testing conditions. This performance degradation can be avoided by modifying model parameters so that they are matched well with input speakers. The performance in the case of B), matched conditions, is drastically improved from that of word HMMs without adaptation. Although the model parameters of our invariant structure model are constant for any value of the warping parameter, our model shows extremely robust performance, which is comparable to that of the matched HMMs. We can say that our structure models have a good ability to generalize.

After our initial trial of isolated word recognition using speech structures, the structures were integrated into HMM-based continuous digit recognition [32] and large vocabulary speech recognition [33]. In these works, multiple hypotheses were generated from the baseline recognizer and these hypotheses were re-ranked discriminatively by using structural likelihood scores. Structural re-ranking improved the performance in both cases. However, the development of the baseline system required a huge amount of speech samples. Modeling infants' process of word acquisition based on their ability to generalize, however, we focus on our invariant structure from a different viewpoint, not hastily combining the structure with the current ASR framework.

In the following sections, we modify our structure model so that it will become more in accordance with experimental facts found in infant studies. Here, we focus on language rhythm.

IV. Syllable Nucleus Detection Using Waveform Envelopes

In this paper, we use the term of "rhythm" to indicate a regulated succession of strong and weak elements [34]. If one applies this definition directly to language, one will find a well-known and language-universal principle of language rhythm, the sonority sequencing principle [35], [36]. Sonority is an auditory phonetic term describing the overall loudness of a sound relative to others of the same pitch, stress and duration [37]. Acoustically speaking, it is considered to be related to

the quality of being resonant. Each phone is considered to have its own sonority value and [38] proposed a universal sonority scale that categorizes phones according to their distinctive features. Vowels have higher sonority and consonants have lower sonority. Different sonority values are assigned to vowels and unvoiced fricatives have minimum sonority. It is clear that a syllable has a sonority peak at its syllable nucleus and its onset and coda have lower sonority values. It is a language-universal that any utterance is composed of a sequence of syllables, and that each utterance has a sonority modulation pattern.

Following this principle, we attempt to detect sonority peaks (syllable nuclei) and extract local or syllable-sized structures around the detected syllable nuclei. These peak-sensitive structures are called rhythm-sensitive structures in this work and will be used in isolated word recognition experiments. Although [38] defined the sonority value of each phone theoretically, we cannot use this definition. Since our structure model is a model of lexical learning, not phonetic learning, phonetic categories should not be used to estimate sonority of each segment in a given utterance. In previous works [39], [40], [41], estimation of sonority or detection of sonority peaks (valleys) were investigated directly from raw acoustic features without phonetic classification. We follow this strategy. In [42], unsupervised syllable boundary detection methods were compared. One of the methods was proposed in [43], which uses waveform envelopes for unsupervised syllable boundary detection. By slightly modifying this method, we implemented a method of syllable nucleus detection using waveform envelopes. The procedure is as follows, explained in Fig. 13 schematically.

1) Speech signals are input to a BPF (500 Hz to 1,500 Hz).
2) Then, full wave rectification and LPF (50Hz) are done to obtain waveform envelopes.
3) Envelope peaks are detected as nucleus candidates.
4) Any candidate that does not satisfy the following conditions are removed.
   - The amplitude of the candidate is larger than a fixed threshold.
   - The candidate has the maximum amplitude among the candidate peaks within a fixed time interval.
5) The resulting candidates are adopted as syllable nuclei.

Using 210 utterances of American English speakers (8 males and 12 females) in the ERJ database [44], our syllable nucleus detector was tested. Recall and precision of the detected nuclei were calculated objectively and subjectively. Objective calculation was done by using the vowel boundaries obtained through forced alignment with an HMM-based speech recognizer and the transcripts attached to the ERJ database. Because alignment errors were inevitable and some vowels were reduced and unvoiced, which should have been judged not to have a syllable nucleus, we asked a native speaker of American English, the fourth author, to locate syllable nuclei that she perceived in the utterances. These perceptual nuclei were used as reference in subjective assessment. Tab. I shows the performance of our detector. The detected syllable nuclei will be used in the following sections to define and extract rhythm-sensitive and local speech structures.

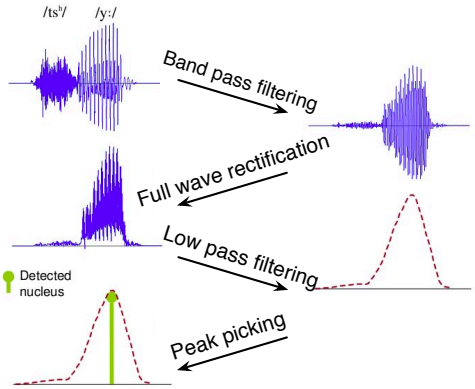

Fig. 13. Syllable nucleus detection using waveform envelopes

TABLE I
PERFORMANCES OF OUR SYLLABLE NUCLEUS DETECTOR [%]

	recall	precision	F-value
objective	74.2	80.7	77.3
subjective	85.7	92.3	88.9

V. SYLLABLE-BASED LOCAL SPEECH STRUCTURE

A. Use of the detected syllable nuclei as landmarks

Fig. 7 and Fig. 11 show how to extract the invariant structure from an utterance and how to obtain a statistical structure model from a set of utterances of one and the same word, respectively. The (i, j) element of the matrix is a speech contrast between the i-th event and the j-th event in the utterance. Since structure extraction is done in an unsupervised way, the linguistic instance of the i-th event can vary among the utterances even when all of them are the same word. In other words, how a feature sequence is aligned to its distribution sequence can vary for each utterance. In this work, we attempt to use syllable-based and rhythm-sensitive structures to make our structure model more in accordance with infants' behaviors. This attempt can be interpreted technically as a solution of misalignment problem between a feature sequence and its distribution sequence by using detected syllable nuclei as salient landmarks.

B. Two kinds of syllable-based structures

After converting a feature sequence into a distribution sequence, by using syllable nucleus detection results, it is possible to detect distributions that have a syllable nucleus, which will be referred to as nucleus distribution. For each of the nucleus distributions, we can form a syllable-based and local structure by using K $(3 \leq K \leq 5)$ adjacent distributions with its center being the nucleus distribution. This syllable-based structure is called hereafter *intra*-syllable structure. It is possible to define another kind of syllable-based structure, which will be called *inter*-syllable structure. In this case, f_{div}

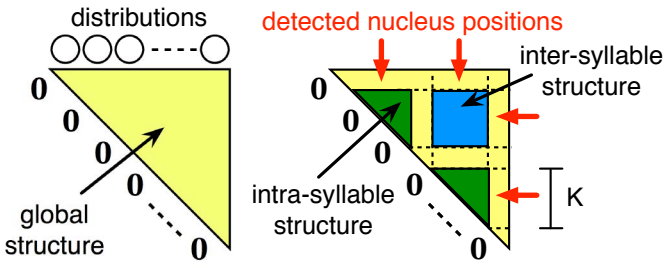

Fig. 14. Global and local structures

is measured between two distant syllables. Fig. 14 shows comparison between a conventional and global speech structure and the two kinds of syllable-based and local speech structures, intra-syllable and inter-syllable structures. The number of intra-syllable structures in an utterance depends on the number of automatically detected syllable nuclei. In the case of N nuclei in an utterance, one will have N intra-syllable structures and $\binom{N}{2}$ inter-syllable structures in that utterance.

VI. Isolated Word Recognition Experiments

A. Soft-decision on the number of syllables

For word recognition experiments, as in Fig. 11, a statistical global word model was built for each word and a statistical local syllable-based model was built for each syllable and each syllable pair in the word. It should be noted that the number of detected nuclei can vary depending on acoustic properties of input utterances even when they are the same word. This makes it difficult to carry out matching an input syllable-based structure vector with syllable-based structure models. We took the following solution to this problem.

In training the statistical structure model, global and local, for each word, since the number of syllables in that word is known, we adjusted thresholds of syllable nucleus detection so that the number of automatically detected syllable nuclei becomes identical to the number of syllables found in the phonemic transcription of that word. Then, we can obtain a word-based global structure model and syllable-based local structure models for each word.

In testing the statistical structure models, since the number of syllables of an input utterance is unknown, we adopted the strategy of a variable number of syllables in that utterance. For any input utterance, the number of nuclei was set to two to five[2]. By adjusting the nucleus detection thresholds, we detected two to five syllables in that utterance[3].

[2]The number of syllables per word in the database used in our experiments varies from two to five. This fact is given to our system.

[3]Syllable nucleus detection performances in Tab. I were obtained using a fixed value for each threshold. We can run word recognition experiments using these fixed values, where the number of syllables in an utterance is determined before word recognition. Preliminary experiments, however, showed that the number of syllables should not be treated deterministically.

TABLE II
Experimental conditions

sampling	16 bit / 16 kHz
window	25 ms length / 10 ms shift
features	MFCC (12dim) + ΔMFCC
distribution	single diagonal Gaussian
parameter estimation	MAP adaptation
#states per HMM	20
#words in the vocabulary	212
#training speakers	15 males and 15 females
#testing speakers	other 15 males and 15 females
#distributions in a local structure (K)	3 or 5
Width of a sub-stream (L)	1 or 2

B. Fusion of global and local likelihood scores

In the experiments, we use both a global structure and local structures for any utterance. Then we can use multiple likelihood scores. Fusion of these scores is done by

$$V = S + \omega \left(\frac{1}{N} \sum_{n=1}^{N} T_n + \frac{1}{M} \sum_{m=1}^{M} U_m \right), \qquad (3)$$

where S is the likelihood score from the global structure, and T_n and U_m correspond to the scores of the n-th intra-syllable structure and the m-th inter-syllable structure, respectively. N is the number of hypothesized syllables and M is the number of hypothesized syllable pairs, $\binom{N}{2}$. ω is a weight of the syllable-based likelihood scores to the global likelihood score. The word that maximizes V is a result of word recognition.

C. Experimental conditions

In the isolated word experiments, we used a more realistic set of word utterances, the phoneme-balanced 212 Japanese word set [45], which is often used as spoken word samples in the Japanese speech research community. Conditions of acoustic analysis and word recognition are shown in Tab. II. The number of distributions per word was set to 20, irrespective of the number of syllables (morae) in the input utterances. The statistical structure model of a word, global or local, was built as a Gaussian distribution estimated from structure vectors (see Fig. 11 and Fig. 14) extracted from multiple utterances of that word. To constrain invariance of the structure models, we adopted multiple stream structuralization (MSS) [7]. Here, a 12-dimensional cepstrum stream was divided into multiple sub-streams of smaller dimensions, where dimensional overlap was allowed between adjacent sub-streams. In the experiments, we set the width of a sub-stream, L, to 1 or 2. The number of distributions considered around a hypothesized syllable nucleus, K, was set to 3 or 5.

D. Results and discussion

Results of the isolated word experiments are shown in Fig. 15 as a function of ω, where larger ω means higher sensitivity to local structures extracted around syllable nuclei. ω=0 means word recognition only using global structures, which corresponds to using our previous structure model. The top figure shows the results using a global structure and both intra-syllable and inter-syllable structures for each input

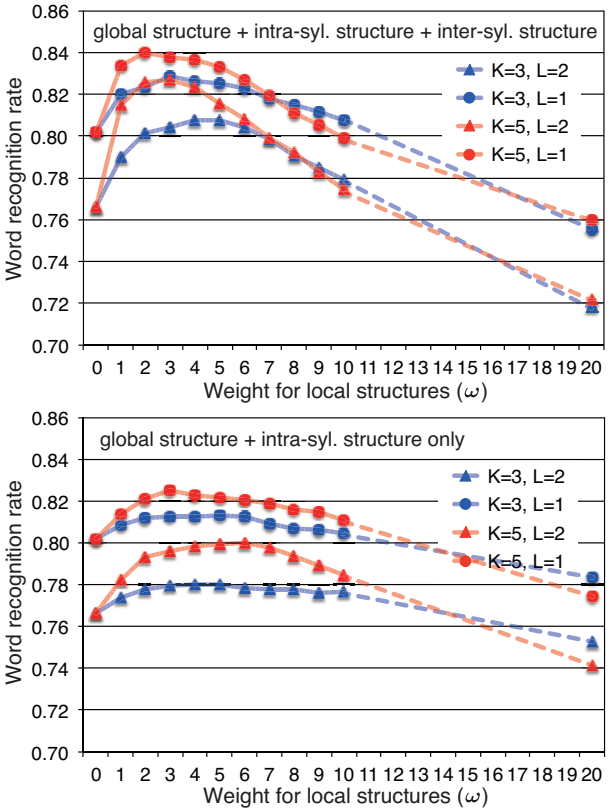

Fig. 15. Word recognition results using global and local structures

utterance; the bottom shows the results using a global structure and only intra-syllable structures. Colors indicate difference of K and shapes of the marks indicate difference of L.

In every case, by introducing syllable-based structures with adequate weights, the recognition performance is improved. This clearly shows that modification of our structure model based on rhythm-sensitivity or nucleus-sensitivity can improve discrimination. We consider that our modified structure models accord more with infants' behaviors. When we interpret this improvement from a technical viewpoint, rhythm-sensitivity appears to solve the misalignment problem to some degree.

By comparing the two figures, it is very interesting that not only intra-syllable structures but also inter-syllable structures can contribute to performance improvement. As far as we know, such long-distance contrasts are not used in the conventional ASR framework but our results show that they are indeed beneficial. Logically speaking, long-distance contrasts or relations are defined only by capturing input stimuli holistically. In the current ASR, both acoustic models and language models can capture acoustic or linguistic phenomena only locally. Extraction of holistic features or patterns will be one of the key issues to make ASR closer to HSR.

However, weights that are too large (ω=20) for local structures lead to performance degradation. This indicates that the elements in a structure matrix which are not related to syllable-based structures can certainly contribute to discrimination. Comparison between the case of L=1 and that of L=2 shows that better performances are obtained in the case of L=1. This result is very reasonable because, in MSS, smaller width of a sub-stream can increase discriminability while decreasing invariance. Very similar results were obtained in [7]. Discrimination and invariance have a trade-off relation. As for K, the case of K=5 shows better results compared to the case of K=3. The optimal value of K should depend on the number of distributions used for modeling word utterances. In this experiment, it is 20.

In this paper, isolated word recognition was carried out by using only structural features. It is very interesting that acoustic features corresponding to spectrum envelopes such as MFCC are not used for recognition. As shown in Fig. 7, only speech contrasts, some of which are distant contrasts, are extracted as invariant features from speech dynamics. Our model has a good ability to generalize but is not able enough to discriminate because, as shown in Fig. 15, the recognition rate is about 84%. If one wants to improve performance for performance sake, the simplest solution is to combine our structure model with the current ASR model [32], [33]. As told in Sect. III-B, we consider that this strategy is not adequate if one wants to simulate infants' process of word acquisition through gaining ability to generalize. On the other hand, it is also true that some speech sounds, such as unvoiced consonants, are much less speaker-dependent than voiced and resonant sounds, and these show smaller speaker variability. Considering this fact, we will introduce the current ASR framework only for processing speech segments with sonority valleys. This is one of the future works.

As already mentioned in Sect. I, a computational model can explain only certain aspects of LA. Our structure model attempts only to explain human performance of robust speech processing in the domain of isolated word recognition. In this paper, the words were treated as given in a *supervised* way. We're planning to introduce our speech structure model to *unsupervised* word discovery.

VII. Conclusions

Speech structure was originally proposed as a *physically*, not statistically, speaker-independent representation of speech and is implemented as an f_{div}-based distance matrix among feature distributions found in a given utterance. Since this representation seems to be in accordance with recent findings of infants' sensitivity to distributional properties in a given language, we introduced to speech structure yet another sensitivity of infants, which is to language rhythm. Syllable nucleus detection was implemented by using waveform envelopes and from this, we defined two kinds of syllable-based local structures, intra-syllable and inter-syllable structures. Experiments showed that a combination of global and local structures can improve the performance. We conclude that our new structure model accords more with infants' behaviors.

REFERENCES

[1] R. K. Moore, "A comparison of the data requirements of automatic speech recognition systems and human listeners," *Proc. INTERSPEECH*, 2581–2582, 2003.

[2] S. Furui, "Generalization problem in ASR acoustic model training and adaptation," *Proc. ASRU*, 1–10, 2009.

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

[4] J. Maye, J. F. Werker, L. Gerken, "Infant sensitivity to distributional information can affect phonetic discrimination," *Cognition*, 82, B101–B111, 2002.

[5] J. F. Werker, F. Pons, C. Dietrich, S. Kajikawa, L. Fais, S. Amano, "Infant-directed speech supports phonetic category learning in English and Japanese," *Cognition*, 103, 147–162, 2007.

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

[7] N. Minematsu, Y. Qiao, S. Asakawa, M. Suzuki, "Speech structure and its application to robust speech processing," *Journal of New Generation Computing*, 28, 3, 299–319, 2010.

[8] Y. Qiao, N. Minematsu, "A study on invariance of f-divergence and its application to speech recognition," *IEEE Trans. on Signal Processing*, 58, 7, 3884–3890, 2010.

[9] O. Räsänen, "Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions," *Speech Communication*, 54, 975–997, 2012.

[10] T. Anastasakos *et al.* "A compact model for speaker-adaptive training," *Proc. ICSLP*, 1137–1140, 1996.

[11] A. R. Plummer, M. E. Beckman, M. Belkin, E. Fosler-Lussier, B. Munson, "Learning speaker normalization using semisupervised manifold alignment," *Proc. INTERSPEECH*, 2918–2921, 2010.

[12] G. Ananthakrishnan, G. Salvi, "Using imitation to learn infant-adult acoustic mappings," *Proc. INTERSPEECH*, 765–768, 2011.

[13] N. Minematsu, S. Asakawa, Y. Qiao, D. Saito, T. Nishimura, "Implementation of robust speech recognition by simulating infants' speech perception based on the invariant sound shape embedded in utterances," *Proc. Speech and Computer (SPECOM)*, 35–40, 2009.

[14] N. Minematsu and T. Nishimura, "Consideration of infants' vocal imitation through modeling speech as timbre-based melody," in *New Frontiers in Artificial Intelligence*, LNAI4914, 26–39, Springer, 2008.

[15] R. Mazuka, "The rhythm-based prosodic bootstrapping hypothesis of early language acquisition: Does it work for learning for all languages?" *Gengo Kenkyu*, 132, 1–15, 2007.

[16] N. Minematsu, "Human speech model based on information separation and its application to speech processing," *Proc. Int. Symposium on Chinese Spoken Language Processing*, 477–482, 2010.

[17] W. Labov, S. Ash, C. Boberg, *Atlas of North American English*, Mouton and Gruyter, 2005.

[18] S. Nakagawa, Y. Tohkura, and K. Shikano, *Speech, hearing and neural network,* Ohmsha, Tokyo, 1990.

[19] R. Jakobson, L. R. Waugh, *The sound shape of language*, Mouton de Gruyter, 2002.

[20] R. Jakobson, J. Lotz, *Notes on the French phonemic pattern*, Hunter, N.Y. 1949.

[21] R. Jakobson, M. Halle, *Preliminaries to speech analysis,* MIT Press, Cambridge, MA, 1952.

[22] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: a new speech research paradigm for next generation automatic speech recognition," Proc. ICSLP, 2004.

[23] C.-H. Lee, M. A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.-H. Juang, L. R. Rabiner, "An overview on automatic speech attribute transcription (ASAT)," *Proc. INTERSPEECH*, 1825–1828, 2007.

[24] T. Fukuda, T. Nitta, "Orthogonalized Distinctive Phonetic Feature Extraction for Noise-robust Automatic Speech Recognition," The Institute of Electronics, Information and Communication Engineers (IEICE) Transactions on Information and Systems, E87-D, 5, 1110–1118, 2004.

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

[26] 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.

[27] D. Saito, N. Minematsu, K. Hirose, "Rotational properties of vocal tract length difference in cepstral space," *Journal of Research Institute of Signal Processing*, 15, 5, 363–374, 2011.

[28] L. Lachs, K. McMichael, D. B. Pisoni, "Speech perception and implicit memory: Evidence for detailed episodic encoding of phonetic events," In J. Bowers and C. Marsolek (eds.) *Rethinking Implicit Memory,* Oxford Univ Press, 2000.

[29] R. S. Newman, "The level of detail in infants' word learning," *Current Directions in Psychological Science*, 17, 3, 229–232, 2008.

[30] N. Higashida, M. Higashida, *Messages to all my colleagues living on the planet*, Escor Pub., Chiba, 2005. (in Japanese)

[31] U. Frith, *Autism: explaining the enigma,* Wiley-Blackwell, 2003.

[32] M. Suzuki, G. Kurata, M. Nishimura, N. Minematsu, "Continuous digits recognition leveraging invariant structure," *Proc. INTERSPEECH*, 993–996, 2011.

[33] M. Suzuki, G. Kurata, M. Nishimura, N. Minematsu, "Discriminative reranking for LVCSR leveraging invariant structure," *Proc. INTERSPEECH*, 2012.

[34] http://en.wikipedia.org/wiki/Rhythm

[35] E. Selkirk, "On the major class features and syllable theory," In Aronoff and Oehrle (eds.) *Language Sound Structure: Studies in Phonology,* 107–136, MIT Press, 1984.

[36] G. N. Clements, "The role of the sonority cycle in core syllabification," In J. Kingston and M. E. Beckman (eds.) *Papers in Laboratory Phonology I: Between the grammar and the physics of speech,* 283–333, Cambridge University Press, 1990.

[37] D. Crystal, *A dictionary of linguistics and phonetics,* 4th edition, Blackwell Publishers, Oxford, 1997.

[38] J. Blevins, "The syllable in phonological theory," in John Goldsmith ed. *The handbook of phonological theory* Blackwell Publishers, Cambridge MA, 1995.

[39] G. Kawai, J. v. Santen, "Automatic detection of syllabic nuclei using acoustic measures," *Proc. IEEE workshop on speech synthesis*, 39–42, 2002.

[40] A. Galves, J. Garcia, D. Duarte, C. Galves, "Sonority as a basis for rhythmic class discrimination," *Proc. Speech Prosody*, 2002.

[41] A. Cros, D. Demolin, A. G. Flesia, A. Galves, "On the relationship between intra-oral pressure and speech sonority," *Proc. INTERSPEECH*, 2165–2168, 2005.

[42] R. Villing, T. Ward, J. Timoney, "Performance limits for envelope based automatic syllable segmentation," IET Irish Signals and Systems Conference, 521–526, 2006.

[43] P. Mermelstein, "Automatic segmentation of speech into syllable units," *JASA*, 58, 880–883, 1975.

[44] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, S. Makino, "Development of English speech database read by Japanese to support CALL research," *Proc. Int. Conf. Acoustics*, 557–560, 2004.

[45] Tohoku univ. and Matsushita 212 phonemically balanced word corpus. http://research.nii.ac.jp/src/TM212.html