THEOREM OF THE INVARIANT STRUCTURE AND ITS DERIVATION OF SPEECH GESTALT

Nobuaki MINEMATSU†, Tazuko NISHIMURA†, Katsuhiro NISHINARI†, and Kyoko SAKURABA‡

† The University of Tokyo ‡Kiyose-shi Welfare Center for the Handicapped

mine@gavo.t.u-tokyo.ac.jp, nt-tazuko@ams.odn.ne.jp
tknishi@mail.ecc.u-tokyo.ac.jp, sakuraba@mtd.biglobe.ne.jp

ABSTRACT

Speech communication has several steps of production, encoding, transmission, decoding, and hearing. In every step, acoustic distortions are involved inevitably as differences of vocal tract length, gender, age, microphone, room, line, hearing characteristics, etc. These are static non-linguistic factors and completely irrelevant to speech recognition. Although the spectrogram always carries these factors, almost all the speech applications have been built on this *noisy* representation. Recently, the first author proposed a novel representation of speech, called the acoustic universal structure[1, 2]. What is represented here is only the interrelations among speech events and their absolute properties are discarded completely. It is very interesting that the non-linguistic factors can be removed effectively from speech as cepstrum smoothing of the spectrogram can remove pitch information from speech. The first author already used this new representation in some speech applications[3, 4] and, in this paper, its theoretical background is described in detail from the viewpoints of linguistics, psychology, acoustics, and mathematics with some results of recognition experiments and perceptual experiments. It is shown that the new representation can be viewed as speech Gestalt.

1. INTRODUCTION

Speech is very variant due to acoustic distortions caused by the non-linguistic factors. In spite of the variations, human listeners can extract linguistic information from speech so easily as if the variations can never disturb the communication at all. 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 acoustically and 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[5]. The acoustic universal structure was derived as invariant acoustic properties based on a mathematical model of the speech variations due to the non-linguistic factors.

Most of the current speech recognizers are based on phone-based HMMs; speech is modeled as a linear string of phones, like *beads on a string*[6]. In these systems, a phone has no explicit *internal* structure beyond the HMM topology. To improve the robustness of speech recognition, some previous studies investigated the use of the internal structure of a phone based on distinctive features[7, 8, 9]. In this approach, a phone or phoneme is regarded as a bundle of features and the features are considered as the minimum units of speech. It is well-known that the distinctive features, acoustic and/or articulatory, were introduced into structural phonology by Jakobson[10].

Though the acoustic universal structure was partly inspired from structural phonology, it focuses on the *external* structure of speech. This is because the distinctive features were originally introduced to represent the external structure. Putting it another way, although the features were firstly used to describe differences or contrasts between phonemes, they were eventually used to define the individual phonemes absolutely and independently as bundles of features.

2. EXTERNAL STRUCTURE OF SPEECH

"*Language is a system of only conceptual differences and phonic differences.*" This is a famous phrase of Saussure, father of modern linguistics[11]. "*What defines a linguistic element, conceptual or phonic, is the relation in which it stands to the other elements in the linguistic system.*" "*The important thing in the word is not the sound alone but the phonic differences that make it possible to distinguish this word from the others.*" Being inspired by these claims, Jakobson introduced the distinctive features originally to describe the phonic differences. Figure 1 shows a consonant triangle and a vowel triangle proposed by Jakobson[10]. In these triangles, the differences are represented by two features of compact/diffuse and grave/acute. Figure 2 shows his geometrical structure of French vowels and semi-vowels[12], where the phonic differences are represented by the features. Although the phoneme was initially defined only by its feature-based interrelations to the others, it seems that Jakobson eventually defined the individual phonemes absolutely and independently as bundles of features. Then, he proposed the famous mapping table between the features and the phonemes. We consider that these two definitions of the phonemes, relative and absolute, have significant difference and that it is obvious that the original definition corresponds directly to Saussure's claims of the language.

These two definitions of the phonemes can be found in a textbook of descriptive linguistics[13]. "*A phoneme is a class of sounds that are phonetically similar and show certain characteristic patterns of distribution in the language or dialect under consideration.*" This is the absolute and independent definition of the phonemes and it is clear that this definition brought about HMMs. Many of /a/ sounds from multiple speakers can create a speaker-independent HMM of /a/. However, it represents only the averaged distribution of /a/ sounds and it can easily have outlier speakers acoustically. For them, speaker adaptation techniques are often required. Speaker-independent models require speaker adaptation techniques, which means that the models are not really speaker-independent.

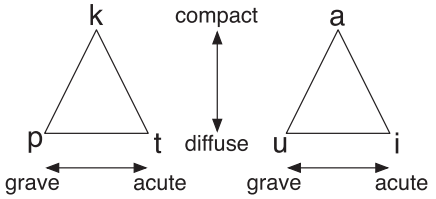

Fig. 1. Consonant and vowel triangles proposed by Jakobson

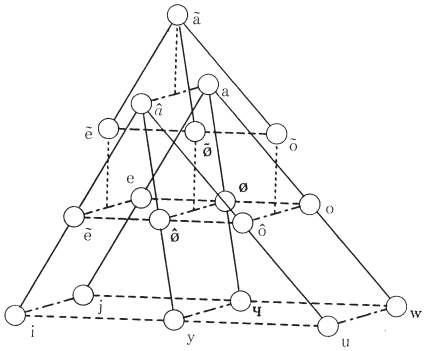

Fig. 2. Jakobson's structure of the French vowels and semi-vowels

"*A phoneme is one element in the sound system of a language having a characteristic set of interrelations with each of the other elements in that system.*" This is the interrelational or contrastive definition of the phonemes, which corresponds better to Saussure's claim. In the textbook, some additional properties of the phonemes are described. "*The phoneme cannot be acoustically defined.*" "*The phonemes of a language are a set of abstractions.*" In this definition, the averaged distribution of /a/ may make no sense. Two speakers offer two different acoustic realizations of /a/ and it is meaningless to consider which of the two samples is closer to the ideal instance of /a/. Both the samples are ideal equally no matter how large acoustic differences they show. In this definition, what to model acoustically is not the phonic entities but the phonic differences or contrasts because linguists consider that they are invariant with speakers.

Let us consider the external geometrical structure of speech, similar to Jakobson's structure in Figure 2, in a cepstrum space. It should be noted that, in this study, the phonic differences are treated not qualitatively as features but quantitatively. If a phone is represented as a point in the space, n phones naturally form an n-point structure. A 3-point structure, triangle, can be determined fully and uniquely by fixing length of all the three lines. Similarly, an n-point structure can be determined uniquely by fixing length of all the ${}_nC_2$ lines including the diagonal lines. All the ${}_nC_2$ differences can be represented compactly as $n \times n$ distance matrix of the n points. To sum up, a geometrical structure as in Figure 2 is mathematically equivalent to its distance matrix. Then, we consider that the distance matrix of speech events can be the simplest mathematical interpretation of Saussure's claim of "*system of only phonic differences.*" Linguists claim that the external structure is invariant with speakers. In the following sections, after a mathematical model to represent acoustic distortions caused by the non-linguistic factors is devised, it is examined whether the external structure can be observed as invariant with the non-linguistic factors, i.e., whether the distance matrix is invariant mathematically with these factors.

3. INEVITABLE NON-LINGUISTIC FACTORS

In speech recognition, three types of distortions or noises, additive, multiplicative (convolutional), and linear transformational, are often discussed. Background noise and music are typical examples of additive noise, often observed in actual environments. But this is not inevitable because a speaker can turn off a radio or move to a quiet room if needed. In this paper, this type of distortion is ignored.

The distortions caused by microphones, rooms, and lines are typical examples of multiplicative distortion. GMM-based modeling of speaker identity assumes that a part of the individuality is also

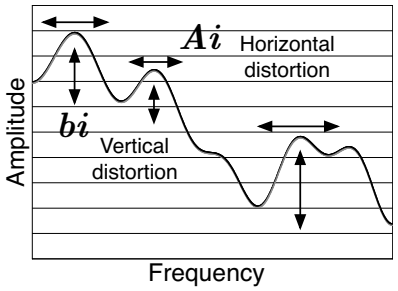

Fig. 3. Spectrum distortions caused by A_i and b_i

regarded as this type. This distortion is inevitable because speech has to be produced by a certain human, transmitted by a certain media, and recorded by a certain acoustic device. If a speech event is represented by cepstrum vector c, the distortion of this type can be modeled as addition of vector b; $c'=c+b$.

Two speakers have different vocal tract shapes and two listeners have different hearing characteristics. Mel or Bark scaling is just the average pattern of the hearing characteristics. These are typical examples of linear transformational distortion, which is naturally inevitable. Vocal tract length difference causes formant shifts, which are often modeled as frequency warping of the spectrum. Hearing characteristics difference causes another frequency warping of the spectrum. Any monotonous frequency warping of the spectrum can be well approximated as multiplication of matrix A[14]; $c'=Ac$.

Although various distortion sources are found in speech communication, the total distortion due to the *inevitable* sources, A_i and b_i, is simply modeled as $c'=Ac+b$, i.e., affine transformation. Figure 3 schematizes the spectrum distortions due to A_i and b_i, which are horizontal and vertical ones, respectively. In MLLR adaptation, multiple matrices are used for a mixture-based bottom-up clustering of triphones[15]. Triphones are trained with many speakers who read different sentences. This implies that different parts of the triphones take on different speaker individuality and this is a main reason why multiple matrices are required. In MLLR adaptation in HMM-based speech synthesis, i.e., adaptation from one speaker to another, a smaller number of matrices can be used effectively. However, a single and global matrix may not be so effective to model the entire non-linguistic factors. Some preprocessing will be examined later.

The static non-linguistic factors are modeled simply as a global affine transformation. It is well-known that an affine transformation functions as operator of rotation, shift, contraction, expansion, shear, or their combination of a geometrical structure. Among many kinds of affine transformations, rotation and shift are the only transformations which don't change the shape of the structure. If the non-linguistic factors can be modeled as the special forms of affine transformation, i.e., rotation and/or shift, then, it can be said that the factors cannot change the structure. However, it is shown in [14] that this assumption is not valid because the non-linguistic factors need more general forms of affine transformation. This mathematically means that the shape of an n-point structure in a cepstrum space has to be distorted and variant by the non-linguistic factors. We wonder whether Jakobson's structure is an illusion mathematically. Or, is it possible to make always-variant structures invariant?

4. STRUCTURAL REPRESENTATION OF SPEECH

The solution of this problem is given by using a kind of mathematical trick, that is the use of an noneuclidean (distorted) space so that the structure can become invariant. We introduce the following theorem.

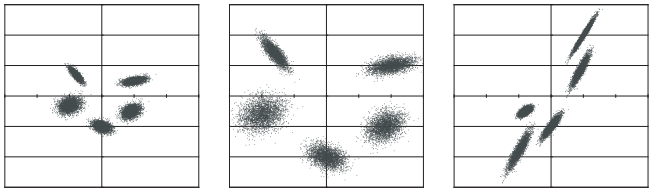

Fig. 4. The invariant underlying structure of a data set

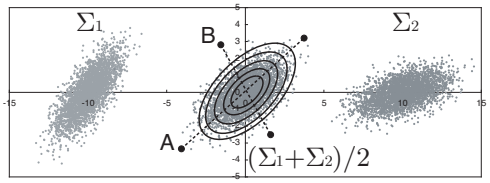

Fig. 5. Two distributions and their averaged one

THEOREM OF THE INVARIANT STRUCTURE

N events are observed and every one is described not as point but as distribution. Distance between any two events is calculated as Bhattacharyya or Kullback-Leibler distance, which is based on information theory. A single and common affine transformation cannot change the distance matrix, i.e., the structure.

Distribution means a Gaussian mixture. Bhattacharyya distance was adopted here because it can be interpreted as normalized cross correlation between two PDFs $p_1(x)$ and $p_2(x)$.

$$BD(p_1(x), p_2(x)) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx, \qquad (1)$$

where $0.0 \leq \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}dx \leq 1.0$ and the name of unit of BD is bit because BD can be regarded as self-information. If the two distributions are Gaussian, BD is formulated as follows.

$$BD(p_1(x), p_2(x)) = \frac{1}{8}\mu_{12}^T \left(\frac{\sum_1 + \sum_2}{2}\right)^{-1} \mu_{12} + \frac{1}{2}\ln \frac{|(\sum_1 + \sum_2)/2|}{|\sum_1|^{\frac{1}{2}}|\sum_2|^{\frac{1}{2}}} \qquad (2)$$

μ_{12} is $\mu_1 - \mu_2$. Figure 4 shows three structures of five distributions. Any two of the three can be converted to each other by multiplying matrix A, meaning that the three structures (matrices) are completely the same. Why this happens? Because BD calculation distorts the space where the distributions are observed. The first term of the righthand side of Equation 2 is in the form of Mahalanobis distance where the covariance matrix is calculated by averaging Σ_1 and Σ_2. In this term, the unit distance changes according to direction of μ_{12}, as shown in Figure 5. The unit distance with μ_{12} lying in the direction of A is longer than that in the direction of B. The structural invariance is obtained with a good combination of a distribution function representing an event, a distance function between two events, and a function to transform the entire events. Other distributions than Gaussians may show the structural invariance with adequate distance and transformation functions. This distorted space can be analyzed with differential geometry. If distribution is characterized by $p(x)=p(\mu,\sigma)=\mathcal{N}(\mu,\sigma)$ then, $d(BD)$ is obtained as follows.

$$BD(\mu_1, \sigma_1, \mu_2, \sigma_2) = \frac{1}{4}\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} + \frac{1}{2}\ln \frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2} \qquad (3)$$

$$d(BD) = M_{\mu\mu}d\mu^2 + 2M_{\mu\sigma}d\mu d\sigma + M_{\sigma\sigma}d\sigma^2 \qquad (4)$$

$$M_{\mu\sigma} = \frac{1}{8}\int_{-\infty}^{\infty} p \frac{\partial \ln p}{\partial \mu}\frac{\partial \ln p}{\partial \sigma}dx \qquad (5)$$

M is a metric and the metric obtained here is called Fisher metric, indicating that the distorted space obtained is a manifold defined in information geometry[16]. Existence of the invariant structure was

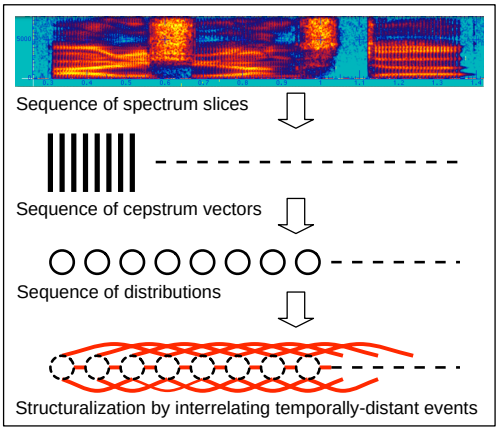

Fig. 6. Structuralization of a single utterance

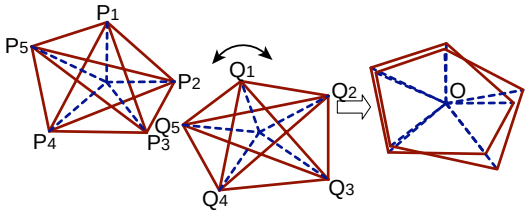

Fig. 7. Acoustic matching after shift (b) and rotation (A)

verified by a distorted (noneuclidean) space. Mathematically speaking, what is discussed here is similar to Einstein's theory of general relativity, where any massive object can distort spacetime around it and the interaction between the spacetime distortions by two objects causes gravity between the two. The metric of the distorted spacetime is calculated by solving the well-known Einstein's equations.

The structural representation can be applied to a *single* utterance, shown in Figure 6. After a given utterance is converted into a sequence of distributions, only the interrelations (phonic differences) of any two of all the temporally-distant distributions are calculated to form a structure (distance matrix). This is called the acoustic universal structure in speech[1, 2]. After that, absolute properties of the individual events such as spectrums and formants are discarded completely. Since matrix A cannot change the distance matrix, any A is interpreted as rotation. For example, human growth is regarded as very slow rotation of the structure, which takes about 15 years.

Acoustic matching between two n-point structures can be done by shifting (b) and rotating (A) a structure so that the two can be overlapped the best, shown in Figure 7. Suppose that there are two n-point structures in an N-dimensional *euclidean* space, where a matrix representing rotation only is an orthogonal matrix. Here, the minimum of the total distance of the corresponding two points after the adaptation with respect to A and b is formulated as

$$\sum_{i=1}^{n} \overline{OP_i}^2 + \overline{OQ_i}^2 - 2\sum_{i=1}^{N}\sqrt{\alpha_i}, \qquad (6)$$

where O is the common gravity center of the two structures P and Q. α_i is the i-th eigen value of $N \times N$ matrix S^tTT^tS. S and T are $(\vec{OP}_1, ..., \vec{OP}_n)$ and $(\vec{OQ}_1, ..., \vec{OQ}_n)$ respectively. It should be noted that the acoustic matching score after the adaptation can be calculated only with two distance matrices, without explicit calculation of A and b. Equation 6 is considered as mathematical shortcut to calculate the acoustic matching score. This implies possibility of speech recognition where only the phonic differences are used. But Equation 6 cannot be adopted directly because triangular inequality is not always satisfied in the distorted space. Some approximate solution only with the two distance matrices has to be prepared. In

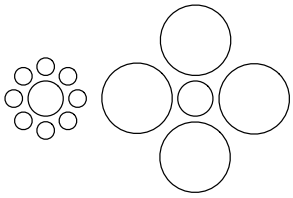

Fig. 8. Visual illusion invoked by Gestalt perception

[1], it was experimentally shown that the minimum of the total distance after the adaptation in Figure 7 is proportional to euclidean distance between the two distance matrices, where the upper-triangle elements form a vector. This approximation will be used hereafter.

5. PSYCHOLOGICAL INTERPRETATION OF THE STRUCTURAL REPRESENTATION OF SPEECH

When people hear music, many of them cannot identify its individual notes. But it is possible to identify the name of the music. Why this happens? This is because music perception is done by perceiving the relative patterns of the notes, not the individual notes separately. It is known that transposition of music cannot affect the identity of the music and the same melody is perceived after the transposition. To explain this effect, Christian von Ehrenfels introduced Gestalt as the holistic quality invariant to transposition[17]. After that, Gestalt came to be used widely to explain various perceptual phenomena. Another famous example of Gestalt perception is visual illusion, shown in Figure 8. The two central circles are physically the same in size but, with some figures around them, they come to look different in size. The visual illusion inevitably happens to humans because visual perception is done by capturing not the individual figures separately but the holistic quality generated by the interrelations among all the figures. One of the major theories to explain the visual illusion assumes that humans distort the space in their brains where the objects are observed[18]. In this theory, the distorted space is analyzed using Einstein's equations directly and Schwarzschild's solution is used to derive the distorted space. It is known that Ehrenfels was influenced by Ernst Mach, who claimed that the elements can become sensations only in the connection and relation and that the physiological space is non-homogeneous (noneuclidean). It is also well-known in science history that Mach influenced Einstein greatly.

The proposed method extracts all the interrelations from speech events in a noneuclidean space. Then, the structurally-represented utterance can become invariant with the static and inevitable non-linguistic factors. This mathematical fact led us to regard a structurally-represented utterance as speech Gestalt. It is interesting that Trubetzkoy, a senior colleague of Jakobson's, claimed the following[19]. "*The phonemes should not be considered as building blocks out of which individual words are assembled. Each word is a phonic entity, a Gestalt, and is also recognized as such by the hearer.*" "*As a Gestalt, each word always contains something more than the sum of its constituents (phonemes), namely, the principle of unity that holds the phoneme sequence together and lends individuality to a word. Yet in contrast with the individual phonemes, it is not possible to localize this principle of unity within the word entity.*" We interpret that Saussure's system of the phonic differences, Jakobson's geometrical structure, and Trubetzkoy's principle of unity indicate the same mathematical entity. The questions are whether a spoken word can be identified correctly only with its phonic differences, and whether human hearers use the differences in speech communication. Some experimental results will be shown in the following section.

If readers have good knowledge both on linguistics and physics,

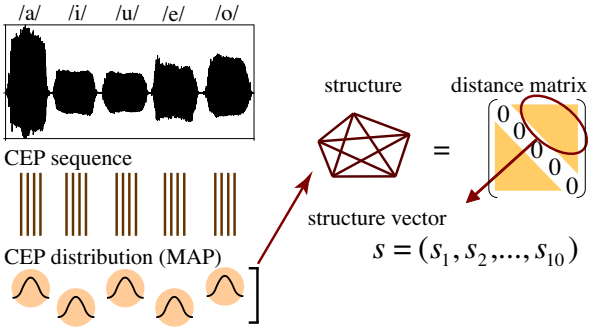

Fig. 9. Parameter extraction to calculate a structure vector

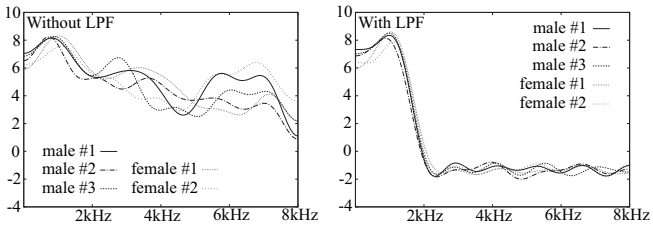

Fig. 10. Spectrum modifications by LPF as preprocessing

they should know the philosophical similarity between Saussure's theory and Einstein's one. The both theories claim that an element cannot have the absolute value by itself. Saussure found this philosophy in language and Einstein found it in spacetime. As told above, Ehrenfels found this philosophy in mind. Jakobson pointed out the *philosophical* similarity between Saussure and Einstein [20] and this paper points out the *mathematical* similarity between them.

6. SOME EXPERIMENTAL FACTS

6.1. Automatic recognition of 5-vowel utterances

The first author applied the new representation to speech recognition[4]. To discuss the fundamental characteristics of the method, a very simple recognition task was adopted; recognition of isolated vowel sequences. Since the non-linguistic factors were expected to be suppressed, only a single speaker's speech samples were used to train the acoustic models. It should be noted that the absolute acoustic entities of speech, spectrum envelopes, were not directly used at all.

The sequence was V_1-V_2-V_3-V_4-V_5, where $V_i \neq V_j$. Since Japanese has five vowels, the vocabulary size is 120. After cepstrum calculation, each vowel was represented as distribution by using its central portion only (140ms). As shown in Figure 9, a structure was composed of the five distributions and a structure vector was obtained to represent the input utterance. As described in Section 4, euclidean distance between two structure vectors can approximate the acoustic matching score after the adaptation between the two utterances.

From the training speaker, a structural and statistical model was trained for each of the 120 words. An input utterance, structurally represented, was matched with these models. 4 male and 4 female speakers were used as testing speakers. The total number of testing samples of the 5-vowel utterances was 25,000. Since the non-linguistic factors were simply modeled as a global affine transformation, the effectiveness was considered to be restricted. A previous study showed that speaker differences are much likely to be observed in upper bands of spectrum[21] and, following this finding, lowpass filtering (LPF) was examined as preprocessing. Figure 10 shows two kinds of spectrum of /a/; clean samples of 5 speakers and those with LPF. The upper portions are modified to show little differences

Table 1. Recognition rates as function of cut-off frequencies

cut-off [kHz]	8.0	4.0	3.5	3.0	2.5	2.0
accuracy [%]	43.0	62.8	81.8	96.9	80.0	100.0

Table 2. Recognition rates of the three methods [%]

methods	full-band	telephone band	2kHz LPF
HMM(260)	100.0	93.8	72.3
HMM(4,130)	100.0	95.2	87.5
Proposed(1)	100.0	100.0	100.0

among the speakers. Table 1 shows the results. With 2kHz cut-off LPF, the recognition performance was raised up to 100%. Since the LPF speech showed the perfect performance, the proposed method was expected to show higher robustness than the conventional methods. This is because, most of the cases, input speech of different acoustic conditions is able to be converted to the LPF speech with 2kHz cut-off. For comparison, two sets of HMMs were prepared, 4,130-speaker and 260-speaker gender-independent models, both of which were trained with full-band MFCC and CMN for acoustic mismatch cancellation. The network grammar allowing only the 120 words was used as language model. Table 2 shows the performance for full-band, telephone band, and 2kHz LPF speech. The parenthesized numbers are those of training speakers. 2kHz LPF was always done as preprocessing in the proposed method. It is clearly shown that the proposed method outperforms the conventional HMMs with CMN. Another experiment was carried out. The HMMs trained only with 2kHz LPF speech of the training speaker showed 88.8% performance for 2kHz LPF speech of the testing speakers. This indicates that 2kHz LPF cannot delete the non-linguistic factors completely and the remaining factors can be removed by structuralization. The performance of the proposed method with noisy speech is described in detail in [4], to which interested readers should refer.

It is very interesting that the 2kHz LPF speech is acoustically similar to *the first speech*; the speech of the mother which an unborn baby listens to continually for several months before birth. In [22], it is shown that, up to 2kHz, there is almost no acoustic difference between two kinds of vowel samples; one recorded in front of the mouth and the other recorded in water in the stomach.

Although the adopted task is very primitive and some problems about continuous speech with consonant sounds remain to be solved, we consider that the potential of the proposed representation is extremely high and that Saussure's claim that a word can be recognized only with its phonic differences was experimentally verified.

6.2. Perception of speaker-variable speech

If an utterance is structuralized, the non-linguistic factors can hardly be seen there. However, it should be noted that the universal acoustic structure is based on a premise that these factors are static. Human speakers cannot change their identity while speaking but technologies can do that. Speaker-variable speech, whose speaker identity changes along the time axis, can be generated by HMM-based speech synthesis. It is obvious that the phonic difference between two speech events will take an abnormal value if the events have different speaker identity. Can a listener extract linguistic information correctly from the speaker-variable speech? If he perceives it in a piecewise manner, the extraction will be done normally. If he perceives it holistically as Gestalt, the extraction will be damaged to some extent. A speech recognizer with speaker-independent HMMs will show the constant performance irrespective of the speaker change.

HMMs of 7 male speakers were trained separately for speech synthesis. Meaningless sequences of 8 morae were synthesized as stimuli. 6 different intervals of the speaker change were realized; 8, 4, 2, 1 mora(e), a phone, and a state. If the speaker changes state by state, the change occurs 5 times in a phone. The most frequent change was expected to be impossible to perceive because spectrum smoothing is done when generating a spectrum sequence. As for prosody generation, a fixed F_0 pattern was always assigned to the sequences, which was LHHHLLLL (type-4 word accent). 25 sequences of randomly selected 8 morae were prepared for each interval. The so-called *tokushuhaku*, such as a choked sound and a syllabic nasal, were ignored. The total number of different morae used in the experiment, i.e., mora-based perplexity, was 43.

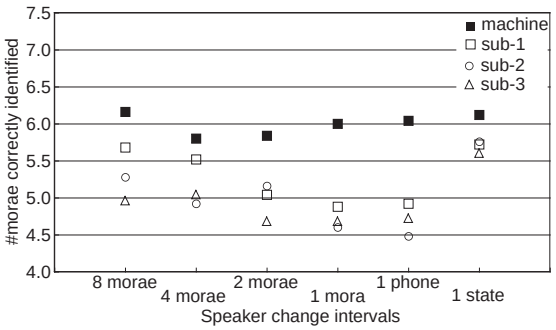

Fig. 11. Mora identification rates by students of a law school

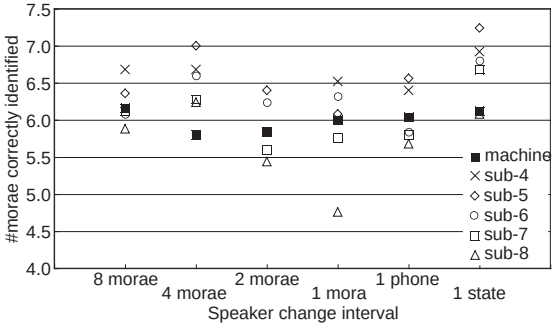

Fig. 12. Mora identification rates by students of a speech lab.

Three types of subjects joined the experiment; 5 students of a speech lab., 3 students of a law school, who never joined a listening test, and a speech recognizer with speaker-independent HMMs. Although the number of human subjects is small, that of morae presented to a subject at each interval of the speaker change is 200 and statistical analysis is possible enough separately for each subject.

Each stimulus was presented twice and the subjects were asked to fill in the 8 blanks on the web. Absence of the *tokushuhaku* was known to the subjects in advance. After the listening test, each sequence was recognized by HVite. The network grammar was used allowing only the 8-mora sequences with the *tokushuhaku* excluded.

Figures 11 and 12 show results of the law students and those of the lab. students. X-axis and Y-axis represent the speaker change interval and the averaged number of morae correctly recognized. Full black rectangles are the performance of the speech recognizer. Significant difference of the machine performance was not found between any two cases of the speaker change interval. However, the identification performance is very different between the two student groups. The law school students are always worse than the recognizer but the speech lab. students are better than the recognizer in most cases. This is considered due to difference of familiarity with synthetic speech and this fact is not focused on in this paper. What is focused on is difference between the two student groups in the performance change along with decrease of the speaker change interval.

The performance of the law students is degraded with decrease

of the change interval. At the shortest interval, as expected, the performance was drastically increased. Except for this increase, significant difference of the performance ($<10\%$) was found at 8m-2m (p=7.54%), 8m-1m (p=3.56%), and 8m-1p (p=5.46%) of subject-1 and 8m-1m (p=6.04%), 8m-1p (p=1.58%), and 2m-1p (p=5.81%) of subject-2. m and p represent mora and phone, respectively. The performance of the lab. students is not degraded except for subject-8. Significant differences are found only at 8m-1m (p=1.79%), 4m-2m (p=5.67%), and 4m-1m (p=0.35%) of subject-8.

Both the figures imply that the lab. students listened to the stimuli piecewise and that the law students captured the holistic quality of each stimulus, which is composed of the phonic differences. Although a quantitative analysis was not done yet, we consider that some speaker changes were perceived as phoneme changes. These effects were much to be expected because speakers and phonemes are represented by the same acoustic feature, i.e., spectrum envelopes.

7. FINDINGS IN STUDIES OF THE HANDICAPPED

Trubetzkoy claimed that a hearer recognizes an input word as Gestalt. We know that some people have great difficulty in perceiving things as Gestalt. They are much less likely to experience visual illusion, much more likely to have absolute pitch, much less likely to show the McGurk effect, much better at memorizing semantically unrelated words such as birth dates and telephone numbers. They are much better at processing sensory elements but much worse at relating an element to others to capture the holistic and coherent quality. They are autistics. Uta Frith claims that autism consists of a lack of drive towards central coherence and explains that autistics live in a fragmented world[23]. It is also known that speech is the most difficult media for them although it is the easiest for the others.

In the conventional acoustic modeling paradigm, when the language has N phonemes, the entire acoustic space is fragmented into N^3 sub-spaces and the observations in each sub-space are modeled basically independently of those in the others, called *triphones*. In some studies[6, 7, 8, 9], even smaller fragments or units are examined, called *features*. We cannot help considering strategic similarity of processing speech between autistics and the current speech recognizers, namely, the reductionism. It is well-known that, in the 90's, AI researchers found the robots they built had behavioral similarity to autistic children[24]. Both were extremely weak at small environmental changes, known as the frame problem. Some AI researchers and autism therapists are collaborating together[24]. The current recognizers are also weak at small environmental changes such as speaker change to children. Speech engineers may have to face the same problem that AI researchers had and still have. We don't deny the conventional methods because humans can identify an isolated phone. We consider that the conventional methods have focused on just one aspect of speech and that the other aspect should be investigated intensively and both the paradigms should be integrated.

8. CONCLUSIONS

This paper proposed a novel method of acoustic modeling of speech and its theoretical backgrounds were described in detail from linguistic, psychological, acoustic, and mathematical points of view. Linguistically speaking, however, we consider that the proposed method is the most classical approach of acoustic modeling of speech. The proposed method is directly based on the original definition of the phonemes, namely, Saussure's system of only the phonic differences, Jakobson's geometrical structure, and Trubetzkoy's principle of unity. Some experimental results showed the high potential of the proposed method. Similarity between autistics and speech recognizers was also discussed. Speech scientists and engineers may have to revisit the classical theories of linguistics, where the word was not treated as just a temporal string of some linguistic and independent elements.

Some readers may have noticed that the proposed method regards speech as music because it captures only the relative patterns. The underlying equality of speech and music is discussed in [25].

9. REFERENCES

[1] N. Minematsu, "Yet another acoustic representation of speech sounds," Proc. ICASSP, pp.585–588 (2004)

[2] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," Proc. ICASSP, pp.889–892 (2005)

[3] N. Minematsu, "Pronunciation assessment based upon the phonological distortions observed in language learners' utterances," Proc. ICSLP, pp.1669–1672 (2004)

[4] T. Murakami *et al.*, "Japanese vowel recognition using external structure of speech," Proc. ASRU, pp.203–208 (2005)

[5] S. K. Scott *et al.*, "The neuroanatomical and functional organization of speech perception," Trends in Neurosciences, vol.26, no.2, pp.100–107 (2003)

[6] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," Proc. ASRU, pp.79–84 (1999)

[7] A. Gutkin *et al.*, "Structural representation of speech for phonetic classification," Proc. ICPR, pp.438–441 (2004)

[8] L. Deng *et al.*, "Production models as a structural basis for automatic speech recognition," Speech Communication, vol.33, no.2-3, pp.93–111 (1997)

[9] T. Fukuda *et al.*, "Orthogonalized distinctive phonetic feature extraction for noise-robust automatic speech recognition," IEICE transactions, vol.E87-D, no.5, pp.1110–1118 (2004)

[10] R. Jakobson *et al.*, Preliminaries to speech analysis, MIT Press, Cambridge, MA (1952)

[11] F. Saussure, Cours de linguistique general, publie par Charles Bally et Albert Schehaye avec la collaboration de Albert Riedlinge, Lausanne et Paris, Payot (1916)

[12] R. Jakobson *et al.*, Notes on the French phonemic pattern, Hunter, N.Y. (1949)

[13] H. A. Gleason, An introduction to descriptive linguistics, New York: Holt, Rinehart & Winston (1961)

[14] M. Pitz *et al.*, "Vocal tract normalization equals linear transformation in Cepstral space," IEEE Trans. Speech and Audio Processing, vol. 13, pp.930–944 (2005)

[15] C. J. Leggetter *et al.*, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol.9, pp.171–185 (1995)

[16] S. Amari and H. Nagaoka, Methods of Information Geometry, Oxford University Press (2000)

[17] B. Smith and C. Ehrenfels (ed.), Foundations of Gestalt Theory, Philosophia (1988)

[18] T. Goto *et al.*, Handbook of the science of illusion, The University of Tokyo Press (2005, in Japanese)

[19] N. S. Trubetzkoy, Principles of phonology, Univ. of California Press (1969)

[20] E. Holenstein, Roman Jakobson's approach to language: phenomenological structuralism, Indiana Univ. Press (1977)

[21] T. Kitamura *et al.*, "Individual variation of the hypopharyngeal cavities and its acoustic effects," Acoustical Science and Technology, vol.26, no.1, pp.16–26 (2005)

[22] I. Yamanouchi *et al.*, "The transmission of ambient noise and self-generated sound into human body," Acta Paediatrica Japonica, vol.32, no.6, pp.615–624 (1990)

[23] U. Frith, Autism: Explaining the Enigma, Blackwell Pub (1992)

[24] J. Nade, "The developing child with autism: evidences, speculations and vexed questions," Tutorial Session of IEEE International Conference on Development and Learning (2005)

[25] N. Minematsu *et al.*, "Speech recognition only with supra-segmental features – hearing speech as music –," Proc. Int. Conf. Speech Prosody (2006, accepted)