Speech Recognition Only with Supra-segmental Features — Hearing Speech as Music —

Nobuaki MINEMATSU†, *Tazuko NISHIMURA*‡, *Takao MURAKAMI*∗, *Keikichi HIROSE*∗

†Graduate School of Frontier Sciences, The University of Tokyo
‡Graduate School of Medicine, The University of Tokyo
∗Graduate School of Information Science and Technology, The University of Tokyo
{mine, murakami, hirose}@gavo.t.u-tokyo.ac.jp, nt-tazuko@ams.odn.ne.jp

Abstract

This paper proposes a novel paradigm of speech recognition where only the supra-segmental features are utilized. Absolute properties of speech events such as formants and spectrums are completely discarded and only the relative and differential properties of the events are extracted as phonic contrasts. The phonic contrasts are considered as supra-segmental features and they are mathematically shown not to carry non-linguistic features such as speaker, age, gender, etc. This fact leads us to expect that speaker-independent speech recognition should be possible with the reference models built only with a single speaker's speech. Experiments of isolated vowel sequence recognition show that this expectation is correct and that the performance of the new paradigm is better than that of the conventional one using more than four thousand speakers, even in the case of noisy speech. Hearing sounds through capturing only their contrasts and their structure is often done when hearing musical sounds, indicating that the proposed paradigm hears speech as music.

1. Introduction

It is known that children can acquire their native language even though only a small amount of stimulus of the language is available, i.e., poverty of stimulus. It is also known that children can recognize speech produced by any speaker although speaker differences exposed to them, especially infants, are remarkably limited. This phenomenon can be regarded as another problem posed by poverty of stimulus. Some researchers may claim that people adapt their ears whenever they hear different speakers. This claim implies that people at a party have to adapt their ears to many speakers simultaneously. Is speech the busiest media or the easiest media? Enjoying your conversation with a good wine may be difficult using the adaptation-based model.

To solve the problem of acoustic variability in speech, the first author proposed a novel acoustic representation of speech, called the acoustic universal structure [1]. Absolute properties of speech events, such as formants and spectrums, are completely discarded and only the phonic differences or contrasts between the events are extracted to form an external structure. Based on a mathematical model to represent the static non-linguistic factors in speech, the external structure, composed as a set of the phonic contrasts, is shown to be invariant with differences in speakers, microphones, etc. As spectrum smoothing is used to separate pitch information from speech, the proposed method can remove the non-linguistic factors from speech.

In this paper, we will discuss the experimental results obtained so far by using the acoustic universal structure and the homogeneity between music and speech in terms of relativism.

2. Acoustic matching done by humans

Before describing the experimental results, we want to address a simple question about human judgment of similarity between two utterances. A young girl is mimicking an unknown word in her father's speech, which is one of the most common scenes in families with daughters. Here, we want to ask "*What is mimicked acoustically by the girl?*" It is obvious that she is not mimicking the word as it is. In repeating her father's speech, she does not try to produce *her father* in her voice. Some readers may answer that she decomposes the speech into a sequence of phonemes and each phoneme is then generated by her mouth. This answer is not good because young children do not have good phonemic awareness and it is difficult for them to decompose an utterance into phonemes. Why doesn't a girl try to produce her father by the mouth although the acoustic matching technique based on spectral comparison requires her to do so?

Speech is often modeled as combination of linguistic, paralinguistic and non-linguistic aspects. Based on this model, it can be said that the girl is mimicking only the linguistic and paralinguistic aspects by separating the non-linguistic aspect such as speaker individuality. In speech science and engineering, however, the acoustic separation of the non-linguistic aspect has not been discussed well and many researches have been done on the separation of the para-linguistic aspect, i.e., source-filter model. We have never seen a girl mimicking only the linguistic and non-linguistic aspects and consider that removal of speaker individuality should have been discussed before F_0 removal by spectral smoothing. Speech recognition is a technique to extract only the linguistic aspect. But speech science and engineering have provided only a naive method for that, i.e., data collection, although the young girl seems to need no additional data.

$$g(\text{linguistic}) = \sum_{\text{non-linguistic}} f(\text{linguistic}, \text{non-linguistic}) \quad (1)$$

In visual neurosciences, it is a classical but well-established model that the visual attributes of an object are perceived in two pathways on the cerebral cortex; the so-called "how" and "what" pathways. While the former pathway exists in the dorsal region and is associated with motion and location of the object, the latter exists in the ventral region and shape, color, and texture of the object are perceived [2]. Recently in auditory neurosciences, researchers began to propose anatomical and functional models of the auditory cortex [3]. In [4], a model was proposed where the dorsal pathway is involved in perceiving spectral dynamics or motions to extract the verbal message contained in an utterance and the ventral pathway is responsible for identifying the speaker. Namely, dynamic and static features

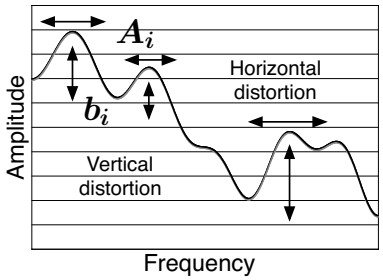

Figure 1: *Spectrum modifications caused by A_i and b_i*

are separated and perceived in the dorsal and ventral pathways, respectively. Although experimental confirmation is required, findings in auditory neurosciences imply that the linguistic and non-linguistic aspects can be separated. If good confirmation is obtained experimentally, the collection-based approach for the separation may be regarded as weird because humans can suppress speaker individuality as the young girl is supposed to do so. It is interesting that the model claims that the dorsal pathway similarly processes the melody of an instrumental piece and the ventral one recognizes the instrument by its timbre.

3. The acoustic universal structure

3.1. Mathematical modeling of the non-linguistic features

In speech recognition, three types of distortions or noises, additive, multiplicative, and linear transformational, are often discussed. Background noise is a typical example of additive noise but this is not inevitable because a speaker can move to a quiet room if needed. In this paper, as we want to focus only on the inevitable distortions or noises, additive noise is ignored.

The distortions caused by microphones and lines are typical examples of multiplicative distortion. GMM-based speaker modeling assumes that a part of the individuality is regarded as this type. These distortions are inevitable because speech has to be produced by a certain human and recorded by a certain acoustic device. If a speech event is represented by cepstrum vector c, the distortion is 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 is often modeled as frequency warping of the spectrum and formant shifts are well approximated. Hearing characteristics difference is another frequency warping. Any monotonous frequency warping in the spectral domain can be converted into multiplication of matrix A in the cepstral domain [5]; $c'=Ac$.

Although various distortion sources can be found in speech communication, the eventual distortion due to the *inevitable* sources, A_i and b_i, is simply modeled as $c'=Ac+b$, i.e., affine transformation. Figure 1 schematizes the spectrum distortions due to A_i and b_i, which are horizontal and vertical ones, respectively. In MLLR speaker adaptation, multiple matrices are used for a mixture-based bottom-up clustering of triphones [6]. Triphones are trained with many speakers who read different sentences, implying that different parts of the triphones have different speaker individuality. 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 effective [7]. However, a single and global matrix may not be so effective to model the entire non-linguistic factors. Some preprocessing will be examined.

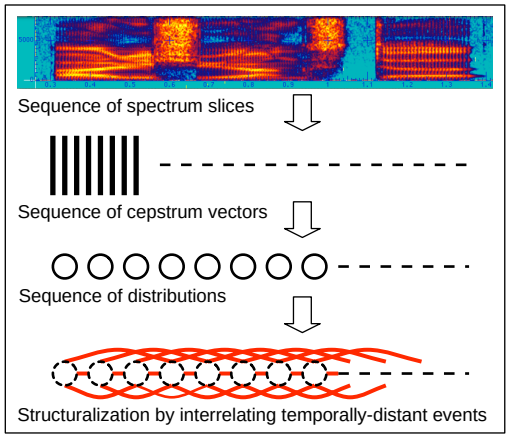

Figure 2: *Structuralization of an utterance*

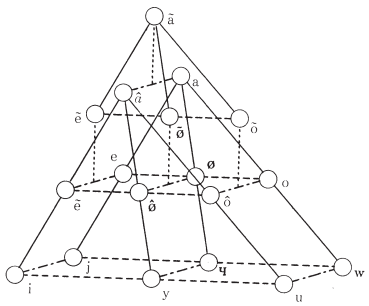

Figure 3: *Jakobson's geometrical structure of French vowels*

3.2. Derivation of the acoustic universal structure

Figure 2 shows a method to form the acoustic universal structure of an utterance. The utterance is converted into a cepstrum vector sequence, and then, into a sequence of cepstrum distributions, each of which is modeled as Gaussian mixture. After that, Bhattacharyya distance between any two distributions is calculated and then, all the absolute properties of the distributions are discarded. The obtained contrasts between any speech events are considered as a full set of spectral motions and the contrasts are mathematically invariant with the static non-linguistic features represented as affine transformation. Conversion from a cepstrum sequence to a distribution sequence is a similar process of training an HMM with a single utterance. In training HMMs for speech synthesis, almost all the initial HMMs are trained only with a single example of the target phoneme. This is because the number of contextual attributes of an HMM is remarkably larger than that used in speech recognition although the size of training data, often less than 1,000 sentences, is remarkably smaller than that for speech recognition. But the training procedure of HMMs for speech synthesis cannot be used directly in forming the acoustic universal structure because the number of distributions cannot be given. Therefore, in this paper, the acoustic universal structure is formed from a sequence of *isolated* vowels for easy estimation of the number. As described above, the obtained contrasts are invariant with the static non-linguistic features characterized by affine transformation. But what these contrasts mean? In the following discussion, linguistic interpretation of the phonic contrasts is described.

Figure 3 shows Jakobson's geometrical structure of French vowels, where phonic differences between two vowels are represented by the style of lines indicating difference of the distinctive features [8]. In structural phonology, it is claimed that this structure is invariant with respect to speakers. What is the

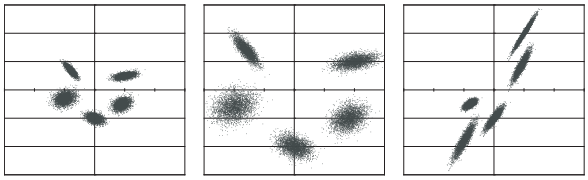

Figure 4: *The invariant underlying structure of a data set*

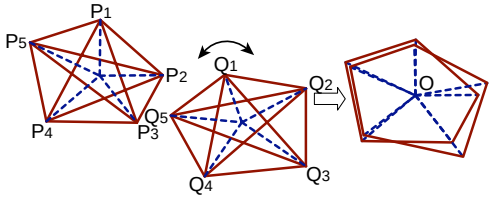

Figure 5: *Acoustic matching after shift (b) and rotation (A)*

most compact representation of an n-point structure? Geometrically speaking, an n-point structure is determined uniquely by fixing length of all the ${}_nC_2$ segments including the diagonal lines. If n distributions are given in an acoustic space and all the ${}_nC_2$ distances are calculated as distance matrix, the matrix uniquely determines shape of the structure expanded by the n distributions. In Figure 2, distributions are given in temporal sequence and the obtained contrasts determines the matrix and the structure. As described above, this structure is invariant with $c'=Ac+b$. If this simple mathematical model is accepted, we consider that Jakobson's structure is mathematically correct.

As is well-known, affine transformation distorts a structure unless it is of a special form. Figure 4 shows 3 sets of 5 distributions, each set of which is transformed into another by multiplying A. It is true that the 3 sets provide 3 different structures on an euclidean plane. If distances are calculated as Bhattacharyya distances, however, the distance matrix is the same among them. This mathematically means that calculation of Bhattacharyya distances distorts a space where distributions are found and this distorted (non-euclidean) space is manifold [9].

It is known that Jakobson was inspired by Saussure, father of modern linguistics [10]. "*Language is a system of only conceptual differences and phonic differences.*" "*What defines a linguistic element 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.*" It is very interesting that the neuroanatomical and functional model of the auditory cortex described in Section 2 and the famous classical linguist propose very similar claims. The former assumes that a verbal message can be extracted only by spectral motions and the latter claims that a word can be distinguished from the others only by phonic differences. If speech recognition is shown experimentally to be possible only with the structural representation, we consider that these claims are valid enough.

Clearly shown in Figure 2, the phonic differences or contrasts can be obtained by considering acoustic features covering speech duration longer than a phone. This means that the phonic contrasts can be viewed as supra-segmental features.

4. Speech recognition only with supra-segmental features

4.1. Acoustic matching between two structures

Considering geometrical properties of affine transformation of a structure, multiplication of A and addition of b in the distorted space are interpreted as rotation and shift of the structure, respectively. Acoustic matching between two n-point structures can be done by shifting (b) and rotating (A) a structure so that the two structures can be overlapped the best, shown in Figure 5. As global affine transformation is the simplest realization of MLLR speaker adaptation [6], then, the minimum of the total distance between the corresponding two points is regarded as acoustic matching score after speaker adaptation. Suppose two n-point structures in an N-dimensional *euclidean* space, where A, representing rotation, is an orthogonal matrix. In this case, the minimum distance is simply formulated as

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

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 or estimation of A and b. This mathematical fact implies possibility of speech recognition only based on the phonic differences. In other words, the proposed method points out a mathematical shortcut for speech recognition. But the above quantity cannot be used directly because triangular inequality is not always satisfied in the distorted space. Then, some approximate solution only with the two distance matrices should be prepared. In [11], it was shown experimentally that Equation 2 is proportional to euclidean distance between the two distance matrices, regarded as two vectors.

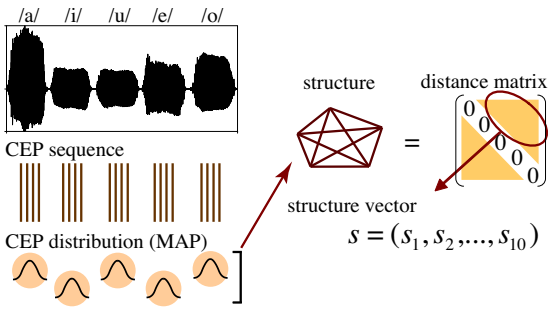

Figure 6: *Parameter extraction to calculate a structure vector*

4.2. Automatic recognition of clean 5-vowel utterances

To discuss the fundamental characteristics of the proposed method, a very simple recognition task was adopted; recognizing sequences of *isolated* vowels [12]. Since the non-linguistic factors were expected to be suppressed effectively, only a single speaker's speech samples were used to train reference models.

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 ${}_5P_5=120$. After cepstrum calculation, each vowel was represented as distribution by using its central portion only (140ms). Since only a small number of frames is used to estimate a distribution, not ML (Maximum Likelihood) criterion but MAP (Maximum A-Posteriori) criterion was adopted. A structure vector, i.e., the elements in the upper triangle of a distance matrix, was obtained to represent the input utterance holistically. The procedures are shown in Figure 6. As described in Section 4.1, euclidean distance between two structure vectors can approximate an acoustic matching score after the adaptation with A and b.

From the training speaker, a structural and statistical model was trained for each of the 120 words. An input utterance, after

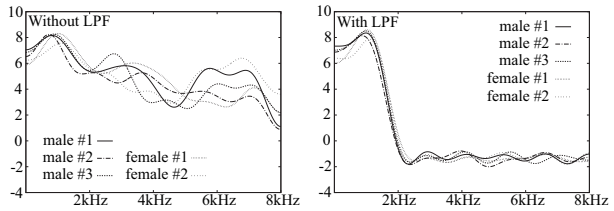

Figure 7: *Spectrum modifications by LPF as preprocessing*

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

being structurally represented, was matched with these models. 4 male and 4 female speakers were used as testing speakers and the total number of testing samples was 25,000. Since the non-linguistic factors were simply modeled as a global affine transformation, its effectiveness was considered to be restricted. A previous study showed that speaker differences are much likely to be observed in upper bands of spectrum [13] and, following this finding, lowpass filtering (LPF) was examined as preprocessing. Figure 7 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 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 done with HMMs trained only with 2kHz LPF speech of the training speaker. Results showed 88.8% performance for 2kHz LPF speech samples of the 8 test speakers. This indicates that 2kHz LPF cannot delete the non-linguistic factors completely and the remaining factors are considered to be suppressed effectively by structuralizing an utterance.

It is interesting that the 2kHz LPF speech is acoustically similar to *the first speech*; the speech of the mother which an unborn infant listens to continually for several months before birth. In [14], it is shown that, up to 2kHz, there is no difference between two kinds of vowel samples; one recorded in front of the mouth and the other recorded in water in the stomach. Since the inner ear is the first sensory system to fully develop in the womb, we wonder whether this listening experience may affect some inherent characteristics of human hearing of sounds.

Although the adopted task is very primitive and some problems about continuous speech including consonant sounds re-

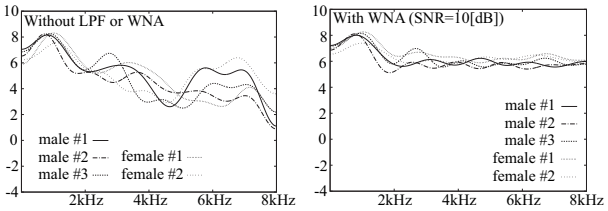

Figure 8: *Spectrum modifications by WNA as preprocessing*

Table 3: *Recognition rates of the proposed method in noise [%]*

cut-off	full band	2kHz
∞	70.3	100.0
20[dB]	92.9	99.8
10[dB]	99.1	86.7
0[dB]	87.0	85.1

Table 4: *Recognition rates of the three methods in noise [%]*

SNR	HMM(260)	HMM(4,130)	Proposed(1)
∞	100.0	100.0	100.0
20[dB]	100.0	98.8	99.8
10[dB]	94.3	97.2	99.1
0[dB]	83.0	86.8	87.0

main to be solved, we consider that the experimental results show the very high potential of the proposed representation and that the similar claims proposed by auditory neurosciences and classical linguistics are sufficiently valid.

4.3. Automatic recognition of noisy 5-vowel utterances

In the previous section, it was experimentally shown that unintelligible speech is recognized more accurately than intelligible speech because the latter carry speaker differences more clearly. Although suppression of speaker differences was realized by LPF, which modified spectrum envelopes at upper bands to be uniform at low power level, similar uniformization of the spectrum can be done at high power level, shown in Figure 8. This modification can be easily realized by white noise addition (WNA) and Figure 8 shows the effectiveness of WNA to reduce speaker differences in speech. In Section 3.1, additive noise was ignored because it is not inevitable. Then, the acoustic universal structure cannot separate this type of noise from speech mathematically. However, Figure 8 shows clearly that additive noise has a definite function to suppress speaker differences [15]. Can noisy speech be recognized more accurately than clean speech?

In the previous section, LPF was carried out as preprocessing both for training speech samples and testing ones. Similarly, by adding the same level of white noise commonly in training and testing, the performance was easily expected to be improved effectively. But it is difficult to know in advance the noise level in the testing condition. Considering that the proposed method needs speech samples of only a single speaker for reference, however, an interesting discussion is possible about the mismatch problem with respect to additive noise. If a system has an extremely high-quality text-to-speech synthesizer to build reference patterns on-line and the system can detect the level of the environmental noise correctly, then the system can generate the reference patterns on-line according to the noise level of the actual environment. In the case of HMMs trained with thousands of speakers, complete re-training of the HMMs using the whole speech samples with the matched noise takes a very long time. Parameter-level adaptation of the HMMs is considered to work worse compared to the HMMs generated through the complete re-training. Since the proposed method

uses only a single speaker to generate the reference patterns, the complete re-training is possible enough if a perfect text-to-speech synthesizer exists. At least, a human listener has a perfect synthesizer if he is not handicapped. We consider that what is discussed here is regarded as structure-based motor theory.

WNA was carried out for every vowel sample of the 8 testing speakers (SNR=0, 10 or 20[dB]) and two types of LPF were examined (cut-off=2 or 8[kHz]). Both of WNA and LPF were conducted commonly in training and testing. Table 3 shows the recognition performance of the proposed method. Since the proposed method has to estimate a distribution from a small number of frames, the estimation was done based on MAP. In the table, the best performance is listed among the weighting factors examined (w=10, 1, 0.1, and 0.01). As expected, the accuracy was drastically improved by WNA, clearly indicating that WNA has a definite function of speaker difference suppression. However, the performance with 2kHz LPF got worse in noisy environments (SNR=10 and 0[dB]). That with full band also got worse in the most noisy condition (SNR=0[dB]). We are interested in the optimal combination of LPF and WNA to suppress speaker differences as preprocessing.

The noisy testing speech samples were recognized by the conventional methods of full band HMMs with spectral subtraction (SS). Estimation of the power spectrum in noisy segments was done by averaging the spectrum of the beginning portion (300[ms]) of each utterance. Table 4 shows the performance of the conventional methods (with SS) and the proposed method. In the case of lower SNRs, the performance of the complete and on-line re-training with a single speaker is superior to that of the conventional methods with SS, although the re-training is currently done with natural speech. These results also indicate the surprisingly high potential of the proposed method.

5. Discussions

5.1. What is speech as linguistic sounds to human hearers?

We believe that the results obtained in the previous sections raise some interesting issues. Speech with LPF or WNA is recognized more accurately than clean speech and, as Figures 7 and 8 show, this result is physically reasonable. This leads us to wonder where on earth one could find clean speech. Before birth, every human continues to hear speech with LPF for several months and, after birth, he or she always hears speech with some additive noise. Clean speech can be found in soundproof rooms today but it could not be found on earth at all a thousand years ago. Ecologically speaking, clean speech may be the most unnatural and artificial speech. It could be obtained for the first time by separating speech completely from its environment. Why do many speech researchers regard clean speech as *natural* speech? The reason is simple. Most of them define speech as what is produced by a mouth to many ears. Under this definition, what is observed acoustically immediately after speech production is the most natural target for research.

In the framework where speech is defined as what is produced by a mouth, it is very natural that much attention was paid to the physical mechanism of speech production. Here, a single speaker's production of speech is often observed and separation of glottal source and vocal tract is usually discussed. As mentioned in Section 2, this framework provided researchers only with the naive method to suppress the static non-linguistic aspect. Individual phones are acoustically modeled using absolute properties of speech such as formants and spectral envelopes and the suppression is realized by collecting data. It might be rather impossible to devise an idea of removing speaker individuality when researchers' mind is largely occupied with observing the process of a speaker's production of speech, not the process of hearing multiple speakers through various channels.

In some anatomical and functional models of the auditory cortex, motions in speech are often focused on. Physiologically speaking, this is because motions in stimuli often have significant values for life. For example, humans can see outer objects because they move or change. If they are fixed spatially and temporally to eyes, they disappear in ten seconds. If everything is fixed, humans can see nothing. The reason of human ability of seeing a non-moving object is for involuntary movement of eyes. It was experimentally shown that if everything is fixed relatively to the moving eyes, they are gone in ten seconds [16]. Considering these properties, researchers of neurosciences often focus on relative contrasts or differences in stimuli. Even if only the absolute properties are captured by brain, robustness of any process executed by the brain has to be reduced naturally because the environment is full of intrinsic variations.

In the framework where speech is defined as what is transmitted to an ear through various channels from various mouths, clean speech is naturally viewed as the most unnatural. The most natural speech here is the speech with some inevitable acoustic distortions and additive noise created by the environment. This paper devised a novel technique that can treat speech based on this framework considering the classical theories of linguistics and the well-accepted consensus of neurosciences. First, a method was proposed to separate the non-linguistic factors from speech based on a simple mathematical model representing the factors. Then, to compensate for the simplicity, speech modification, which is very natural for humans, was introduced as preprocessing, i.e., WNA and/or LPF.

5.2. Underlying homogeneity between speech and music

Hearing sounds through capturing only their contrasts and their structure is often done when hearing instrumental sounds. Individual constituents or notes are not identified and, only with their relative patterns, most people can enjoy music. In theoretical studies of musicology, the sound structure of music is often discussed. In [17], purely geometrical structure of music, which is defined as distance matrix calculated from a sequence of notes, is introduced as a model characterizing the musical structure. In [18], a similar structuralization method was used to implement a musical application software as we also used the acoustic universal structure for CALL applications [19].

We believe that the proposed framework considers speech as music without identifying individual phonemes. It is very interesting that people with absolute pitch say that they cannot understand why people without absolute pitch can enjoy music because they cannot imagine musical activity without notes in mind. Absolute sense of musical sounds is absolute pitch and that of speech sounds is phonemic awareness. After the experiments, we wondered whether some real people enjoy speech without phonemic awareness although we cannot imagine speech activity without phonemes in mind. Then, we found surprisingly that these people really exist; dyslexics, who are considered to lack in phonemic awareness but can communicate orally. They have great difficulty in reading and writing. Many musicians say that absolute pitch is not required to enjoy music and it is required only to transcribe a tune as sequence of notes. Dyslexia indicates that phonemic awareness may not be required for speech communication and it is required only to transcribe an utterance as sequence of symbols. Dyslexics are

very good at capturing things as a holistic entity, a Gestalt, but very bad at capturing things separately and independently [20]. It is said that they cannot see the trees for the wood. We consider that other types of people don't have phonemic awareness but start enjoying speech communication. They are infants, who might encounter speech as music in the womb as described in Section 4.2. This might be the very reason why they can solve another problem of poverty of stimulus so easily after birth.

5.3. Other findings in studies of the handicapped

Some people show completely a reverse pattern of behaviors to those of dyslexics. They are autistics and have great difficulty in perceiving things as Gestalt and it is often said that they cannot see the wood for the trees. For example, they are 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, and much worse at associating an element with others to capture the holistic quality. In [21], it is explained that autism consists of a lack of drive towards central coherence and that autistics live in a fragmented world. It is also known that speech is the most difficult media for them although it is the easiest for the others.

Musicians with extreme absolute pitch are known to have some troubles in performing music. Physical realization of the note A above middle C depends on orchestras and it is sometimes 442 or 445 Hz. For those whose note A is 440 Hz, they naturally have troubles to cooperate with other members in the orchestra because the note A of the orchestra is not his/her A. A Japanese autistic boy wrote a book using PC about his daily experiences and he wrote that he can recognize his mother's speech but *cannot* recognize the others', even his father's [22]. He hears every sound normally but the sounds from which he can extract linguistic messages are only his mother's. We wonder whether he has extreme absolute sense of speech sounds.

In the conventional acoustic modeling framework, when the language has N phonemes, the acoustic space is fragmented into N^3 sub-spaces and the observations in each sub-space are modeled independently of those in the others, called *triphones*. In some studies, 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. The above Japanese boy resembles a speaker-dependent speech recognizer. It is well-known that, in the 90's, AI researchers found the robots they built had behavioral similarity to autistic children [23]. Both were extremely weak at small environmental changes, known as the frame problem. Some AI researchers and autism therapists are collaborating together [23]. Speech engineers may have to face the same problem that AI researchers had and still have.

6. Conclusions

This paper proposed a novel paradigm of speech recognition using only the phonic differences or the spectral motions based on the classical theories of linguistics and the well-accepted consensus of neurosciences. The proposed structural representation of speech can hardly have dimensions indicating the static non-linguistic factors and it is considered to capture speech as music. Experiments of recognizing sequences of isolated vowels showed the high validity of the proposed method. Furthermore, we pointed out the behavioral and strategic similarity between autistics and the current speech recognizers. However, we don't deny the conventional paradigm because we can identify an isolated phone using its absolute acoustic properties. We consider that the conventional paradigm has focused on just one aspect of speech and that the other aspect should be investigated more intensively and integrated with the conventional method. If readers have any interest in the relations of the acoustic universal structure to para-linguistic features, they should refer to [24] because the features can also be represented structurally.

7. References

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

[2] L. G. Ungerleider, "Two cortical visual systems," in *Analysis of Visual Behavior* (edited by David J. Ingle), pp.549–586, MIT Press (1982)

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

[4] P. Belin and R. J. Zatorre, "'What', 'where' and 'how' in auditory cortex," *Nature neuroscience*, vol.3, no.10, pp.965–966 (2000)

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

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

[7] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A context clustering technique for average voice models," *IEICE Trans. Inf. & Syst.*, vol.E86-D, no.3, pp.534–542 (2003)

[8] R. Jakobson and J. Lotz, *Notes on the French phonemic pattern*, Hunter, N.Y. (1949)

[9] N. Minematsu, T. Nishimura, K. Nishinari, and K. Sakuraba, "Theorem of the invariant structure and its derivation of speech Gestalt," *Proc. Int. Workshop on Speech Recognition and Intrinsic Variations* (2006, submitted)

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

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

[12] T. Murakami, K. Maruyama, N. Minematsu, and K. Hirose, "Japanese vowel recognition based on structural representation of speech," *Proc. EUROSPEECH*, pp.1261–1264 (2005)

[13] T. Kitamura, K. Honda, and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoustical Science and Technology*, vol.26, no.1, pp.16–26 (2005)

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

[15] T. Murakami, K. Maruyama, N. Minematsu, and K. Hirose, "Japanese vowel recognition using external structure of speech," *Proc. ASRU*, pp.203–208 (2005)

[16] A. L. Yarbus, *Eye movements and vision,* Prenum Press (1967)

[17] C. L. Krumhansl, "The geometry of musical structure: a brief introduction and history," *ACM Computers in Entertainment*, vol.3, no.4 (2005)

[18] M. Goto, "SmartMusicKIOSK: Music listening station with chorus-search function," *Proc. Annual ACM Symposium on User Interface Software and Technology*, pp.31–40 (2003)

[19] S. Asakawa, N. Minematsu, T. I. Jaakkola, and K. Hiorse, "Structural representation of the non-native pronunciations," *Proc. EUROSPEECH*, pp.165–168 (2005)

[20] R. D. Davis and E. M. Braun, *The Gift of Dyslexia*, Perigee (1997)

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

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

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

[24] N. Minematsu, S. Asakawa, and K. Hirose, "Para-linguistic information represented as distortion of the acoustic universal structure in speech," *Proc. ICASSP* (2006, accepted)