PARA-LINGUISTIC INFORMATION REPRESENTED AS DISTORTION OF THE ACOUSTIC UNIVERSAL STRUCTURE IN SPEECH

Nobuaki MINEMATSU†, Satoshi ASAKAWA†, and Keikichi HIROSE‡

†Graduate School of Frontier Sciences, The University of Tokyo
‡Graduate School of Information Science and Technology, The University of Tokyo
{mine,asakawa,hirose}@gavo.t.u-tokyo.ac.jp

ABSTRACT

Speech acoustics varies from speaker to speaker, microphone to microphone, etc. Recently, a novel method was proposed to separate these static non-linguistic features from speech as spectral smoothing can separate pitch information from speech[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 their external structure. This structure is called the acoustic universal structure and regarded as physical implementation of structural phonology because the structure is considered to represent only the linguistic and para-linguistic information. In this paper, the structural size is focused on and its correlation with the para-linguistic information is examined. Results showed that the size can be interpreted as magnitude of articulatory efforts made in speech production.

1. INTRODUCTION

Linguistics gives two definitions of the phoneme[2]. One definition is that "a phoneme is a class of phonetically similar sounds." It is clear that this definition brought about speaker independent acoustic models, each of which is trained by collecting sounds of the corresponding phoneme produced by many speakers, and that independently of the other models. The other definition is that "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 a contrastive and relative definition of the phoneme and it claims that what should be modeled acoustically is not the phonic entities but the phonic interrelations among the entities. The latter definition was derived from claim of Saussure, known as father of modern linguistics. "What defines a linguistic element is the relation in which it stands to the other elements in the linguistic system." This claim implies that the contrastive and relative definition is original and primary and the independent and absolute definition can be viewed as secondary.

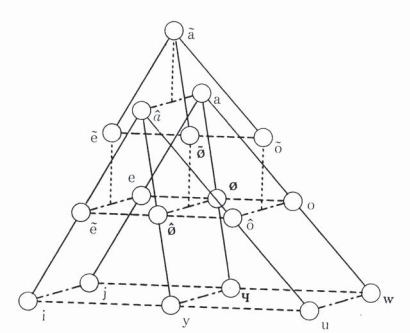

Fig. 1. *Jakobson's geometrical structure[3]*

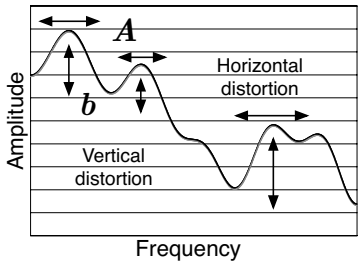

Fig. 2. *Two types of spectrum distortion caused by A and b*

Figure 1 shows Jakobson's structure of French vowels and semi-vowels[3]. In this external structure, an element is determined only by its interrelations to the others. Linguistics claims that this structure is universal with respect to speakers and the universality was verified mathematically in our previous study[1].

2. THE ACOUSTIC UNIVERSAL STRUCTURE

2.1. Mathematical modeling of the non-linguistic features

Differences of microphones and rooms are typical examples of convolutional distortion. GMM-based speaker modeling indicates that a part of speaker individuality is also regarded as convolutional distortion. If a speech event is represented by cepstrum vector c, then, the convolutional distortion is shown as $c'=c+b$.

Differences of vocal tract length are often modeled as frequency warping of the log spectrum and the warping is mathematically modeled as multiplication of matrix A by c[4]; $c' = Ac$.

Figure 2 shows the distortion caused by $c'=Ac+b$ schematically and these non-linguistic distortions are inevitable in speech. On the other hand, additive noise is not always inevitable because you can move to a quiet room if needed. In this paper, only the inevitable distortions are considered and they are eventually and simply modeled as $c'=Ac+b$, known as affine transformation.

2.2. Derivation of the acoustic universal structure

Let us consider a geometrical structure like Figure 1 in a cepstral space. A triangle can be determined uniquely by fixing length of all the three lines. Similarly, an n-point structure is determined by fixing length of all the ${}_nC_2$ lines including its diagonal lines. The length of all the lines is compactly represented as $n \times n$ distance matrix. Structural phonology claims that the n-point structure is invariant with affine transformation ($c'=Ac+b$) but this is impossible because the transformation always distorts a structure unless it is of a special form. Always variant structures can be invariant? Jakobson's structure is just an illusion mathematically? This problem can be solved easily by distorting the space surrounding the structure. We can introduce the following theorem.

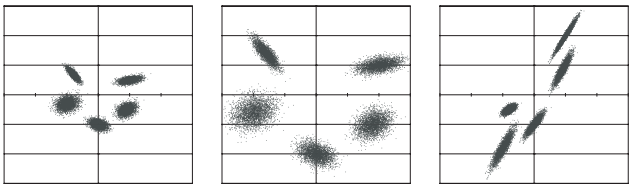

Fig. 3. *The invariant underlying structure of a data set*

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 unit name of BD is bit because BD can be regarded as self-information. Figure 3 shows three structures of five distributions. Any two of the three can be converted to each other by multiplying A. This fact means mathematically that the three structures (matrices) are evaluated as completely the same. Why this happens? Because BD calculation distorts the space where the distributions are observed.

2.3. Structuralization of an utterance

The theorem only requires events to be modeled not as points but as distributions. Then, their structure can be mathematically invariant with respect to a common affine transformation. This structuralization process can be applied simply to an utterance. Figure 4 shows its procedure. After a given utterance is converted into a sequence of distributions, only the interrelations (phonic differences) between any two of all the temporally-distant distributions are calculated to form a structure (distance matrix). This structural representation is invariant with respect to speakers. Our previous works showed experimentally that the structural acoustic models trained only with a *single* speaker outperformed HMMs trained with more than four thousand speakers although the task was very primitive and it was recognition of sequences of isolated vowels[5]. Surprisingly, the better performance was also obtained in recognizing the sequences in additive noise[6].

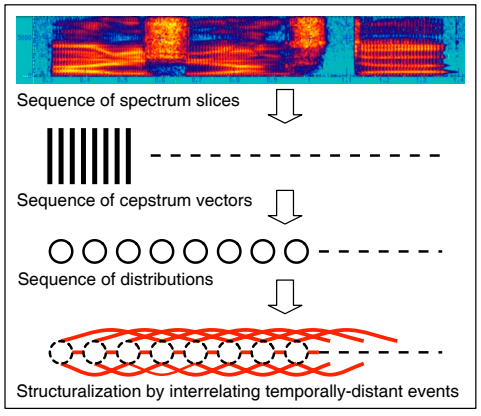

Fig. 4. *Structuralization of a single utterance*

3. STRESSED VOWELS AND UNSTRESSED VOWELS

3.1. Speech material used in the analysis

In order to discuss phonetic interpretation of the size of the structure, firstly in this paper, English vowels, stressed and unstressed, were focused on. A TIMIT-based phoneme-balanced sentence set were read by two American speakers (male and female). Conditions of the acoustic analysis are shown in Table 1 and a single-Gaussian distribution was adopted to train the vowel models with a full variance-covariance matrix. For the speech material, phonemic and stress labels were assigned by a semi-automatic method. PRONLEX dictionary was referred to for determining the initial labels and they were modified with speaker-dependent acoustic models trained with the above speech material. The acoustic models and the phonemic and stress labels were simultaneously trained and adjusted. In the rest of the paper, æ1 and æ0 mean stressed and unstressed æs, respectively.

3.2. Monophthongs of American English

Figure 5 shows the vowel chart of the monophthongs of American English[8]. Distance matrix was calculated from the monophthong acoustic models using $\sqrt{\text{BD}}$ as distance measure between two models. The reason of using $\sqrt{\text{BD}}$, not BD, is that $\sqrt{\text{BD}}$ can approximately satisfy a certain geometrical condition which is always satisfied by Euclid distance and not satisfied by BD[7]. Figure 5 visualizes the distance matrix of the female American based on Multi Dimensional Scaling (MDS). Here, `isoMDS` in MDS software of R was used. Although the MDS chart was expected to depend on phonemic environments of the individual vowel instances, rather good correspondence is found between the two charts. As is well-known in phonetics, schwa is the most fundamental vowel in that it is located at the center of the vowel chart, the articulatory center, and that it is located at the center of the MDS chart, the acoustic center. It is also known that schwa

Table 1. *Acoustic conditions for the analysis*

sampling	16bit / 16kHz
window	25 ms length and 10 ms shift
parameters	Improved cepstrum (1∼12)
speakers	Two Americans (a male and a female)
training data	746 & 709 sentences for the male & the female
HMMs	speaker-dependent, context-independent, and 1-mixture monophones with full matrices
topology	3 states and 1 distribution per HMM (GM)
monophones	monophthongs of American English i, ɪ, u, ʊ, ɛ, æ, ʌ, ɑ, ɔ, ə, ɚ

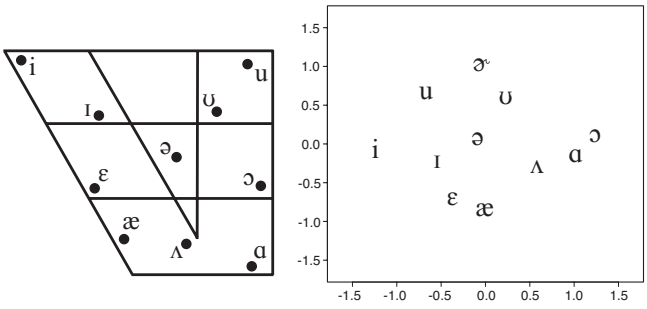

Fig. 5. *The vowel chart and an MDS chart of vowel examples*

is acoustically generated with a sound tube of a uniform cross-sectional area, which indicates that schwa is produced with the least articulatory effort. As for unstressed vowels, it is often said that if vowels are reduced, they get much acoustically closer to schwa[8]. These considerations directly led to acoustic and articulatory hypothesis on the size of the structure; the larger the size is, the larger the effort is.

3.3. Size of the vowel structure

The size of the vowel structure was calculated with the speech material of the two speakers. Only with a distance matrix, it is possible to calculate what geometrically corresponds to the radius of the structure. To show the radius adequately, in this section, a tree diagram is drawn to visualize a distance matrix with Ward's method. In this method, two elements are merged into one sequentially so that the accumulated distortion should be minimized. The accumulated distortion is represented by the height of the tree grown so far. Finally, all the elements are merged into a single element (centroid) and the height of the final tree is equal to VQ (Vector Quantization) distortion when all the data is represented by a centroid. This quantity can be regarded as the radius of the structure.

Figure 6 shows three tree diagrams; vowels, stressed vowels, and unstressed vowels of the female American. In the vowel tree, about 60 % of the vowels were stressed ones. For a few vowels, a very strong bias between occurrences as stressed and those as unstressed was found and these vowels were deleted. i, ɪ, u, ʊ, ɛ, æ, ʌ, ɑ, and ɚ were used in the tree diagrams. The vowel tree is lower than the stressed vowel tree and higher than the unstressed vowel tree. The stressed tree is 1.4 times higher than the unstressed one. Although the shape of the tree is similar between the vowel tree and the stressed one, some differences are found between the unstressed tree and the other two trees. We consider that some unstressed vowels were acoustically realized completely as schwa sounds. The same characteristics was found with the male American. The stressed tree is 1.2 times higher than the unstressed tree, which seems to have some structural distortion compared to the vowel tree and the stressed vowel one.

Another similar analysis was done with Japanese short vowels (a, i, u, e, and o) and long vowels. The two structures are visualized by MDS and shown in Figure 7. The long-vowel structure is larger than the other. This is because longer vowels have more stable spectrums and they make their variances smaller and their inter-vowel distances longer. The structural representation can also contain durational information in speech.

The above experimental results support our interpretation of the size of the structure. Considering the acoustic and articulatory fact that the central sound is the least energetic sound, the structural size is regarded as articulatory effort with high validity.

4. JAPANESE VOWELS OF VARIOUS SPEAKING STYLES

Do other factors change the size of the structure? It is easily assumed that some speaking styles change the size and, in this section, speech samples of the same linguistic content with various speaking styles were analyzed with respect to their sizes.

4.1. Speech material used in the analysis

Isolated vowels of Japanese, a, i, u, e, and o, were recorded by a professional voice actress with the following 12 speaking styles or

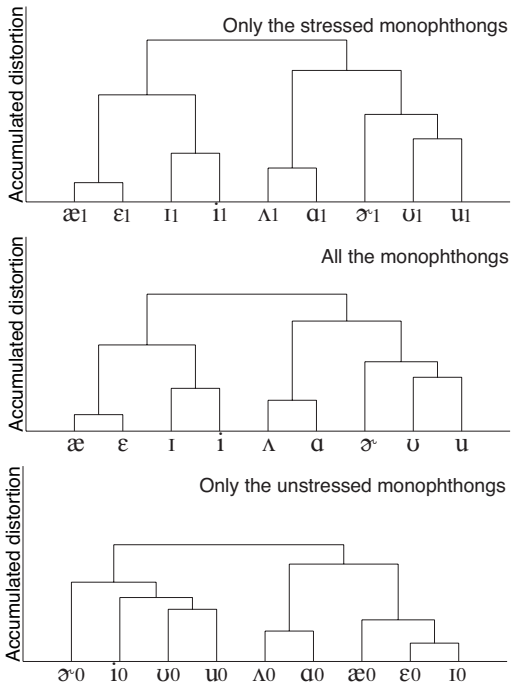

Fig. 6. *Tree diagrams of American English vowels*

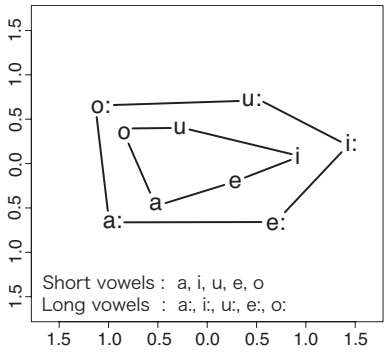

Fig. 7. *Two structures of Japanese short vowels and long vowels*

situations. The recording was repeated 5 times for each.

1) unintelligible (indistinct), 2) with a sigh, 3) scared, 4) whisper, 5) too surprised to speak aloud, 6) with wobbles, 7) intelligible (distinct), 8) without strong intension, 9) the loudest, 10) with full energy, 11) ashamed, and 12) proud

The aim of this recording was just for collecting 5-vowel utterances with many different styles and it should be noted that appropriateness of her vocal expression as the designated style is not focused on in this analysis. In the recording, we sometimes gave her detailed instructions about the desired situation.

4.2. Size of the vowel structure

The size of the vowel structure was acoustically estimated for each of the utterances. Since the recording was repeated 5 times, 5 distance matrices were obtained for each of the styles and they were averaged. Figure 8 shows the averaged vowel diagrams of 8 styles out of the 12. Clearly seen in the figure, the size changes according to the style or situation. However, separation between the front

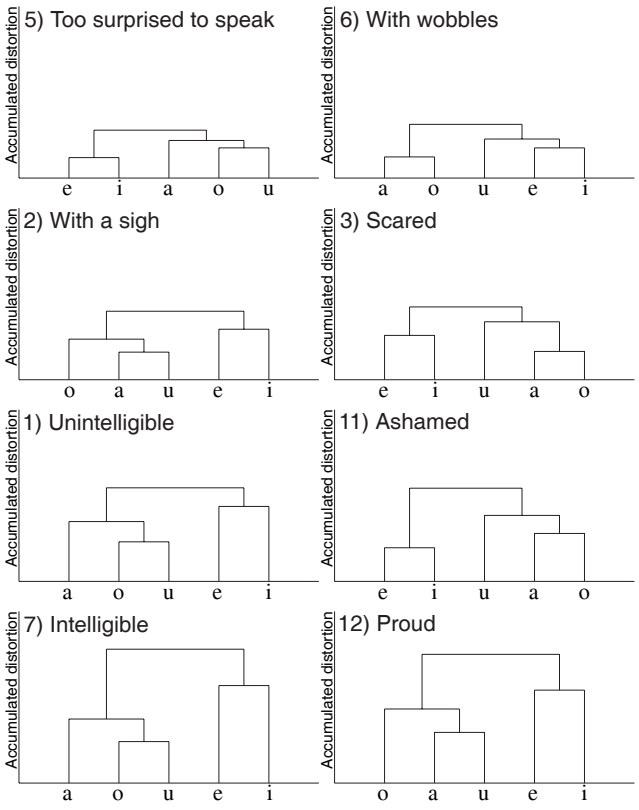

Fig. 8. *Japanese 5 vowels with different styles*

vowels and the others at the top of the tree is commonly found except for 6). Figure 9 shows the size of the structure for each style (white bars). Large variability can be clearly seen. In the following section, it is examined whether the acoustically-measured size of the structure can be a good measure of the perceptual distinction among the vowels when hearing the vowels of each style.

4.3. Comparison with human perception of the distinction

A listening test was done with 5 university students of Japanese with normal hearing. The task was to score the magnitude of distinction among the 5 vowels after hearing each vowel set. A 5-degree scale was used, where 1 and 5 meant the least and the most distinct, respectively. Averaged perceptual distinction scores over the subjects were standardized over the styles to have the same mean and the same variance that the sizes of the structures (white bars) have in Figure 9. The standardized scores are shown as gray bars in Figure 9. Very good correspondence is found between the two quantities and their correlation is 0.903. However, rather large a difference is found in the case of 9); the loudest. The 5 vowels in this style were uttered with high energy and physical efforts made for speech production were perceived well. The vowels were uttered so loudly that their durations were relatively short. It is considered that, in this case, stability of the spectrogram is reduced, and therefore, BD between two vowels became smaller. Another reason is possible. It is expected that speech production with extremely high energy will make it difficult to control the articulators accurately and to increase the acoustically-defined distinction. Still in this case, the energy is well transmitted to listeners and the perceptually-defined distinction is easily increased. If the case of

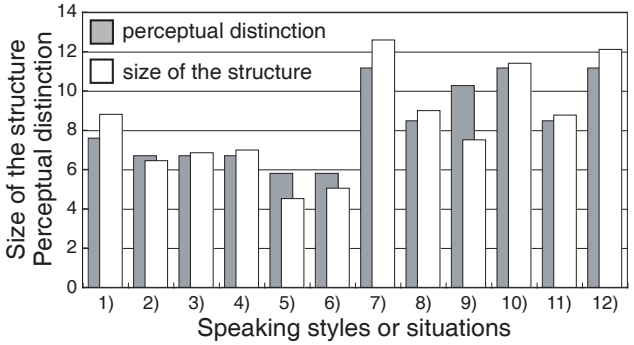

Fig. 9. *Size of the structure and perceptual distinction among the vowels*

9) can be ignored, the correlation between the size of the structure and the perceptual distinction is 0.978.

Although only the relative acoustic properties were focused on in this paper, these properties are considered just one aspect of speech acoustics. The absolute properties, conventionally used in speech processing, are naturally supposed to have some functions to transmit para-linguistic information. Integration of the relative properties with the absolute ones is a future work.

5. CONCLUSIONS

Firstly in this paper, a novel method of acoustic representation of speech, structuralization of speech, was introduced, where the dimensions to indicate the static non-linguistic features are effectively suppressed. Then, by focusing on the size of the structure, its correlation to articulatory efforts was examined using vowel samples with various speaking styles. Results of the experiments showed that extremely high correlation was found between the structural size and the perceptual distinction among the vowels. Our previous studies showed that vowel sequences can be recognized perfectly only with their structural representation[5, 6] and the current study showed that some of the para-linguistic features can be detected also from the structural representation. We're planning to do additional analysis on the *local* structural distortion and its correlation with para-linguistic information and are interested in structure-based automatic estimation of the information.

6. REFERENCES

[1] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," Proc. ICASSP, pp.889–892 (2005)
[2] H. A. Gleason, An introduction of descriptive linguistics, New York: Holt, Rinehart & Winston (1961)
[3] R. Jakobson *et al.*, Notes on the French phonemic pattern, Hunter, N.Y. (1949)
[4] 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)
[5] T. Murakami *et al.*, "Japanese vowel recognition based on structural representation of speech," Proc. EUROSPEECH, pp.1261–1264 (2005)
[6] T. Murakami, *et al.*, "Japanese vowel recognition using external structure of speech," Proc. ASRU, pp.203–208 (2005)
[7] N. Minematsu, "Pronunciation assessment based upon the phonological distortions observed in language learners' utterances," Proc. ICSLP, pp.1669-1672 (2004)
[8] J. Clark and C. Yallop, An introduction of phonetics and phonology, 2nd edition, Blackwell Publishers Inc. (1995)