The Acoustic Universal Structure in Speech and Its Correlation to Para-linguistic Information in Speech

Nobuaki Minematsu, Satoshi Asakawa, and Keikichi Hirose

University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
e-mail: {mine,asakawa,hirose}@gavo.t.u-tokyo.ac.jp
http://www.gavo.t.u-tokyo.ac.jp/

Keywords: the acoustic universal structure, para-linguistic information, speech communication

Abstract

Speech acoustics inevitably carries non-linguistic information such as age, gender, speaker, microphone, line, room, and so on. These kinds of information is nothing but noise if linguistic and/or para-linguistic information in speech is considered. A novel method of acoustic representation of speech is proposed in this paper, which has no dimensions of the static and inevitable non-linguistic information. The representation is realized by mathematically interpreting a claim of Saussure, father of linguistics and structuralism, based on information theory and by implementing structural phonology, a part of linguistics, on physics. Speech events are modeled probabilistically as distributions, distance between two events is calculated based on information theory, and an entire set of the events are captured relatively as a structure. The resulting structure is shown to have no dimensions of the non-linguistic information because multiplicative and linear transformational distortions are geometrically interpreted as shift and rotation of the structure, respectively. Therefore, the structure is considered to transmit only the linguistic and para-linguistic information. In this paper, size of the structure is investigated and several experiments show that the size can be interpreted as magnitude of articulatory efforts made during speech production.

1 Introduction

Speech communication has several steps of production, encoding, transmission, decoding, and hearing. In every step, multiplicative or linear transformational distortions are inevitably involved such as differences of vocal tract shape, gender, age, microphone, room, line, hearing characteristics, and so on. In spite of the various non-linguistic distortions, humans can extract linguistic and/or para-linguistic information from speech so easily as if the distortions cannot disturb the communication at all and it is a fact that speech is still the easiest communication media for humans. To explain this somewhat abnormal fact, one may hypothesize that the linguistic and/or para-linguistic information in speech is acoustically represented in brains where no dimensions of the above distortions exist, namely, mental abstraction.

In every speech application, speech sounds are modeled based on acoustic phonetics, where a speech sound is modeled independently of the others. But a speech sound is easily distorted by the above various factors and this causes the co-called "mismatch problem". As far as the authors know, all of the previous studies tried to solve the problem by either of adaptation or normalization. With these methods, however, every speech recognizer still has "sheep and goats" and it means that the complete solution is almost impossible. The authors believe that the most essential reason for the problem is that every speech system is built on an assumption that the system has to have acoustic models of the individual sounds. Under this assumption, even after normalization, every sound model has certain acoustic properties with regard to each dimension of the non-linguistic distortions. Strictly speaking, the phonetics-based models of speech sounds cannot solve this problem completely. The complete solution can be done only by finding acoustic representation of the linguistic and para-linguistic information in speech where no dimensions of the inevitable non-linguistic distortions exist, namely, physical implementation of the abstraction.

Readers may well claim that it should be impossible. But this paper mathematically shows that there exists the acoustic universal structure in speech. The structure is shown to be yet another acoustic representation of speech and have no dimensions of the inevitable multiplicative and linear transformational distortions. The abstraction is not mental but physical. This representation is realized by mathematically interpreting a claim of Saussure, father of linguistics and structuralism, based on information theory and by implementing structural phonology, a part of linguistics, on physics. Speech events are modeled probabilistically as distributions, distance between two events is calculated based on information theory, and an entire set of the events are captured relatively as a structure. Once speech events are structuralized, the multiplicative and linear transformational distortions are geometrically interpreted as shift and rotation of the structure, respectively. This structure is named as the acoustic universal structure, which is assumed to transmit the linguistic and para-linguistic information only. In this paper, size of the structure is also focused on and its correlation to the para-linguistic information is examined.

2 Inevitable Acoustic Distortions in Speech

What kind of distortions are involved in speech communication and which ones are inevitable? The authors consider three types of distortions; additive, multiplicative, and linear transformational. Background noise and music are typical examples of the additive distortion (noise). But this is *not* inevitable because a speaker can turn off a TV set if he wants. If he cannot for some reasons, he and a listener can move to the next quiet room to obtain an environment for clean speech communication.

Acoustic distortions caused by microphones, rooms, and lines are typical examples of the multiplicative distortion. GMM-based speaker modeling assumes that speaker individuality is well-represented by an average pattern of a long-term series of the log-spectrum of the speaker. This indicates that a part of speaker individuality is also regarded as the multiplicative distortion. This distortion is inevitable because speech has to be produced by a human and recorded by an acoustic device. If a speech event is represented by cepstrum[1] vector c, the multiplicative distortion is converted to addition of vector b and the resulting cepstrum is represented as $c' = c + b$.

Two speakers have different vocal tract shapes and two listeners have different hearing characteristics. Mel or Bark scaling is just an average pattern of the hearing characteristics. These are typical examples of the linear transformational distortion, which is naturally inevitable. Vocal tract length difference is often modeled as frequency warping of the log spectrum, where formant shifts are well approximated. Hearing characteristics difference is another frequency warping of the log spectrum. According to [1], any monotonous frequency warping of the log spectrum can be mathematically converted into multiplication of matrix A in cepstrum domain. The resulting cepstrum is depicted as $c' = Ac$.

Various distortion sources are found in speech communication. But the total distortion of speech caused by the *inevitable* sources, A_i and b_i, is eventually modeled as $c' = Ac + b$, known as affine transformation. Different speakers or environments will cause different A_i or b_i. Acoustic phonetics claims that every speech is distorted and one can obtain distortion-free speech only by stopping speaking and stopping listening. The distortion-free speech is called *silence.* The authors believe that this is the most essential reason why every speech system has the so-called "sheep and goats" problem inevitably.

3 Physical Implementation of Structural Phonology

3.1 Saussure's Claim on Language and Structural Phonology

As mentioned in Section 1, the complete solution is considered possible only by finding acoustic representation of speech with no dimensions of the inevitable non-linguistic distortions. Acoustic phonetics is,

[1] Cepstrum coefficients are calculated through inverse Fourier transform of the log-spectrum.

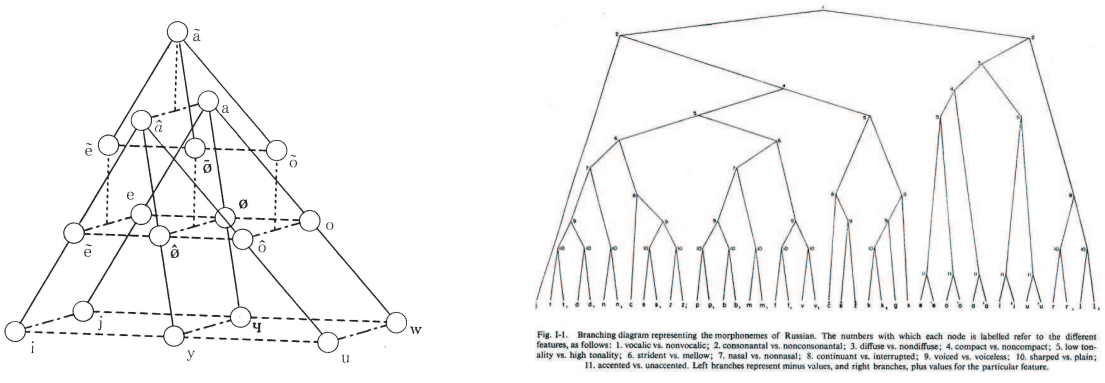

Figure 1: Jakobson's geometrical structure of some French phonemes (lefthand) and Halle's tree diagram of the Russian phonemes (righthand)

however, unable to provide the representation by itself. In this work, another speech science is focused on, which is phonology, a part of linguistics. In phonology, the inevitable distortions are mentally ignored in linguists' brains and speech sounds are represented as abstract entities named phonemes. Phonology is intended to clarify a system or structure embedded in a set of the phonemes of a language or in sequences of phoneme instances in actual utterances. Here, structural phonology, i.e., structure in a set of the phonemes, is focused on. The other phonology is, in turn, is called generative phonology.

Inspired by Saussure's claim[2], Jakobson, Halle, and others discussed a system of the phonemes embedded in a language by using distinctive features[3], which were originally proposed by Jakobson. Figure 1 shows Jakobson's geometrical structure proposed for some French phonemes[4] and Halle's tree diagram of the Russian phonemes[5]. They claimed that the structure is invariant and independent of speakers. Their structuralization of the phonemes is based on distinctive features of the phonemes and, for example, differences in the shape of line segments between two phonemes in Jakobson's structure represent differences of distinctive features of the corresponding two phonemes. In this paper, however, the distinctive features are ignored because different linguists claim different sets of the features. Here, only the linguists' consciousness of the phonemic structure is focused on and the consciousness was raised by a single claim of Saussure on language; "*Language is a system of conceptual differences and phonic differences.*"

The authors are interested in the acoustic aspect of languages and only the phonic differences are considered here. Geometrically speaking, Saussure's claim that language is a system of phonic differences can be interpreted as a very simple definition of a structure. In Euclidean space, an n-point structure is uniquely represented by a set of lengths of its ${}_nC_2$ diagonal lines, i.e., a set of all the possible differences among the n points. The differences are formulated by a distance matrix of the n points and, with a bottom-up clustering algorithm, the matrix can produce a tree diagram of the structure. These considerations lead to the following. The distance matrix among the n phonemes in an acoustic space can be regarded as mathematical and physical interpretation of the Saussure's claim and the matrix is geometrically equivalent to the structure itself and can produce the tree diagram shown in Figure 1. Viewing the n elements as a structure indicates that the elements are observed only relatively. The structure extraction can be regarded as a process of ignoring some information in a set of the elements. If it is possible to embed all the sources of the inevitable non-linguistic distortions in the ignored information, the resulting structure is expected to be the acoustic representation which the authors pursue.

3.2 Necessary and Sufficient Condition for the Physical Implementation

The above section discusses how to interpret Saussure's claim on language mathematically and one answer was provided where a set of all the differences among the phonemes, i.e., a distance matrix of the phonemes, may correspond to his insight into language. Linguists claim that the structurally-represented phonemes are invariant and universal with regard to speakers, ages, genders, microphones, rooms, lines, and so on. Now, it is possible to derive a necessary and sufficient condition to implement structural phonology on physics. Let phoneme x be represented as point c_x in a cepstrum space. If n phonemes are found in the space, an n-point structure is defined. Structural phonology claims that the n-point structure should not be distorted by affine transformation of $c' = Ac + b$ because the transformation represents the inevitable and non-linguistic distortions. But it is well-known that affine transformation distorts a structure such as warping and scaling. Specific forms of the transformation, rotation and shift, cannot change the structure. But matrix A proposed in [1] shows that it is not in these forms. The authors wonder whether it is proved that structural phonology is just an illusion mathematically.

3.3 Physical Implementation of Structural Phonology Based on Information Theory

This section mathematically shows that structural phonology can be implemented on physics by proving that any affine transformation cannot change the structure if it is composed of speech events[6, 7, 8]. As is shown below, the physical implementation requires information theory.

In the previous section, phoneme was regarded as point in a cepstrum space which represents a single spectrum slice. In this paradigm, it was shown that structural phonology has to be just an illusion physically and mathematically. Every speech researcher knows that complete repetitions of a single pitch waveform, even extracted from *natural* speech, sound like a buzzer. Acoustic perturbations are inevitably observed in speech and a single spectrum slice cannot represent this essential characteristics of speech. Human speech production is incomplete because of the inevitable acoustic perturbations and the complete production cannot generate natural speech but generate buzzer-like sounds only. Let phoneme x be represented as distribution $d_x(c)$ in a cepstrum space. Since an n-point structure can be determined uniquely by a set of lengths of its ${}_nC_2$ diagonal lines, a necessary and sufficient condition for the physical implementation is that distance between any two elements (distributions) should not be changed by any affine transformation. It there any distribution-to-distribution distance measure satisfying this condition?

Bhattacharyya distance (BD) measure satisfies the condition. BD between two probability density functions, $d_x(c)$ and $d_y(c)$, is formulated as follows.

$$BD(d_x(c), d_y(c)) = -\ln \int_{-\infty}^{\infty} \sqrt{d_x(c) d_y(c)} dc, \tag{1}$$

where $0.0 \leq \theta = \int_{-\infty}^{\infty} \sqrt{d_x(c) d_y(c)} dc \leq 1.0$. Originally, this measure was introduced based on geometrical investigation of two discrete probability distributions[9, 10]. If θ is able to be assigned a specific meaning as probability, however, BD is interpreted as amount of self-information of θ. For example, it may be possible to suppose that θ is probability of occurrence of an event where two sets of samples randomly extracted from two populations of x and y are recognized by observers as those extracted from a *single* population. If the two distributions follow Gaussians, the following is obtained.

$$BD(d_x(c), d_y(c)) = \frac{1}{8} \mu_{xy}^T \left(\frac{\sum_x + \sum_y}{2} \right)^{-1} \mu_{xy} + \frac{1}{2} \ln \frac{|\sum_u + \sum_v|/2}{|\sum_u|^{\frac{1}{2}} |\sum_v|^{\frac{1}{2}}} \tag{2}$$

μ_x and Σ_x are the average vector and the variance-covariance matrix of $d_x(c)$, respectively. μ_{xy} is $\mu_x - \mu_y$. Although affine transformation of $c'=Ac+b$ modifies $\mathcal{N}(\mu, \Sigma)$ into $\mathcal{N}(A\mu + b, A\Sigma A^T)$, BD between $d_x(c)$

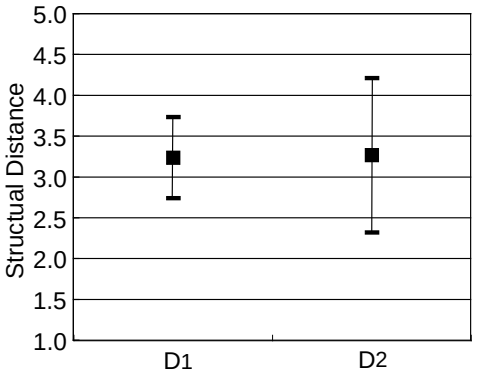

Figure 2: Good cancellation of the inevitable distortions from speech

and $d_y(c)$ is not changed.

$$BD(A\mu_x + b, A\Sigma_x A^T, A\mu_y + b, A\Sigma_y A^T) = BD(\mu_x, \Sigma_x, \mu_y, \Sigma_y) \tag{3}$$

These facts mean that BD between two distributions (phonemes) is not changed by any affine transformation and that the structure composed of the n phonemes is not changed. Multiplication of A and addition of b are geometrically interpreted as rotation and shift of the structure, respectively. For example, acoustic change of speech caused by increase of vocal tract length, i.e., human growth, is mathematically regarded as very slow rotation of the structure which takes about 15 years. Even when $d_x(c)$ and $d_y(c)$ are modeled as Gaussian mixtures, the invariance is still valid.

It should be noted that another distribution-to-distribution distance measure satisfies the condition of the invariability with regard to affine transformation and the measure is symmetrized Kullback-Leibler distance measure, which is derived from information theory more directly and is formulated as follows.

$$KL(d_x(c), d_y(c)) = \int_{-\infty}^{\infty} d_x(c) \ln \frac{d_x(c)}{d_y(c)} dc + \int_{-\infty}^{\infty} d_y(c) \ln \frac{d_y(c)}{d_x(c)} dc$$

This section showed that Jakobson's geometrical structure of phonemes, i.e., the universal and essential structure of speech, exists not only in his insight into a language but also in pure acoustics of speech. In the current study, this physical structure is defined as the acoustic universal structure in speech.

3.4 Cancellation of the Inevitable Distortions from Speech

A simple experiment was designed and carried out to verify how well the inevitable distortions are canceled from speech by extracting the structure. Isolated vowels of Japanese, /a/, /i/, /u/, /e/, and /o/, were recorded from 2 male and 2 female Japanese adults. They repeated the recording three times. From the vowel utterances, 12 5-vowel structures were obtained, 3 structures for each speaker. Each vowel is represented as a single Gaussian. Distance between two structures, P and Q, is defined as

$$D = \sqrt{\frac{1}{M^2} \sum_{i<j} (\overline{P_i P_j} - \overline{Q_i Q_j})^2}. \tag{4}$$

i is vowel index and M is the number of vowels. D is approximately equal to average distance between corresponding vowels of P and Q after full adaptation with regard to A and b of the inevitable distortions[6]. If the inevitable distortions are canceled very well by extracting the structure, intra-speaker structural

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

distance, D_1, and inter-speaker structural distance, D_2, should be the same. Figure 2 shows D_1 and D_2 with no differences between the two distances. While the maximum structural distance was found as inter-speaker distance, the minimum was also found as inter-speaker distance. The acoustic universal structure was experimentally shown to exist physically.

The following sections investigate correlation of the acoustic universal structure to para-linguistic information included in speech.

4 Correlation of the Structure to Stressed and Unstressed Vowels

4.1 Speech Material Used in the Analysis

To discuss phonetic interpretation of size of the structure, firstly in this paper, English vowels, stressed and unstressed, were considered. Two kinds of reading material were prepared; a TIMIT-based phoneme-balanced sentence set and another sentence set extracted from several English textbooks[11]. They were read by two Americans (male and female) and 746 and 709 sentences were recorded from the male and the female speakers, respectively. Acoustic conditions of the analysis is shown in Table 1 and a single-Gaussian model was used to characterize each of the vowels with a full variance-covariance matrix. To build the acoustic models, phonemic and stress labeling was required and this was done 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[12]. The acoustic models and the phonemic and stress labeling were simultaneously trained and adjusted. In the rest of the paper, æ1 and æ0 mean stressed and unstressed æs, respectively.

4.2 Monophthongs of American English

Figure 3 shows the vowel chart of the monophthongs of American English[13, 14]. Distance matrix was calculated from the monophthong acoustic models using $\sqrt{\mathrm{BD}}$ as distance measure between two models. The reason of using $\sqrt{\mathrm{BD}}$, not BD, is that $\sqrt{\mathrm{BD}}$ can approximately satisfy a certain geometrical condition which is always satisfied by Euclid distance[8]. Figure 4 visualizes the distance matrix of the female American based on Multi Dimensional Scaling (MDS). Here, `isoMDS` in MDS software of R[15] was used. Since the phonemic, not phonetic, labeling was done for the training data, strictly speaking, the MDS chart should be drawn with phonemic, not phonetic, symbols. For easy comparison with the vowel chart, however, phonetic symbols are used. Although the MDS chart depends on phonemic environments of the individual vowel instances in the training samples, rather good correspondence is found between

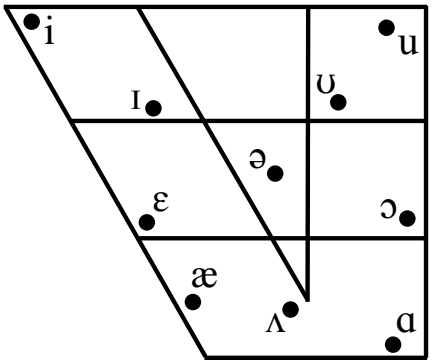

Figure 3: The vowel chart of American English

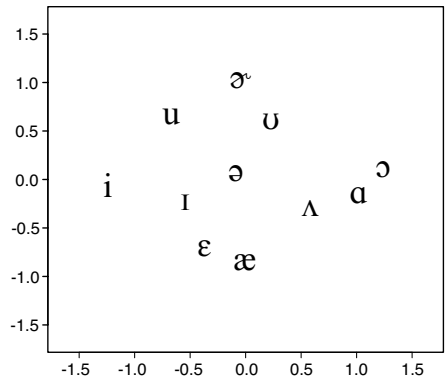

Figure 4: Visualization of the matrix with MDS

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 is acoustically generated with a sound tube of a *fixed* 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[13, 14]. These considerations directly lead to acoustic and articulatory hypothesis on size of the acoustic universal structure; the larger the size is, the larger the effort is.

4.3 Size of the Vowel Structure

Size of the vowel structure was quantitatively calculated with the speech material of the two speakers. Only with a distance matrix, it is possible to calculate what geometrically corresponds to radius of the structure. A tree diagram is drawn to visualize a distance matrix and some hierarchical clustering algorithms are widely used, one of which is Ward's method. Two elements are merged into one sequentially so that the accumulated distortion should be minimized. The accumulated distortion is represented by height of the tree grown so far. Finally, all the elements are integrated into a single element (centroid) and height of the final tree is equal to VQ (Vector Quantization) distortion when all the data is represented by the centroid. This quantity can be regarded as radius of the structure.

Figure 5 shows three tree diagrams; vowels, stressed vowels, and unstressed vowels of the female American. In the vowel tree, about 60 % of the vowels in the training data were stressed ones. The vowel tree and the MDS chart visualize the same distance matrix in different ways except for one thing. 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 shape of the tree is very similar between the vowel tree and the stressed one, some differences are found between the unstressed tree and the other two trees. As described in Section 4.2, phonemic environments are easily expected to have some effects on shape of the tree. It is also likely that some unstressed vowels were acoustically realized completely as schwa sounds. These are considered as reasons for the differences. 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.

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

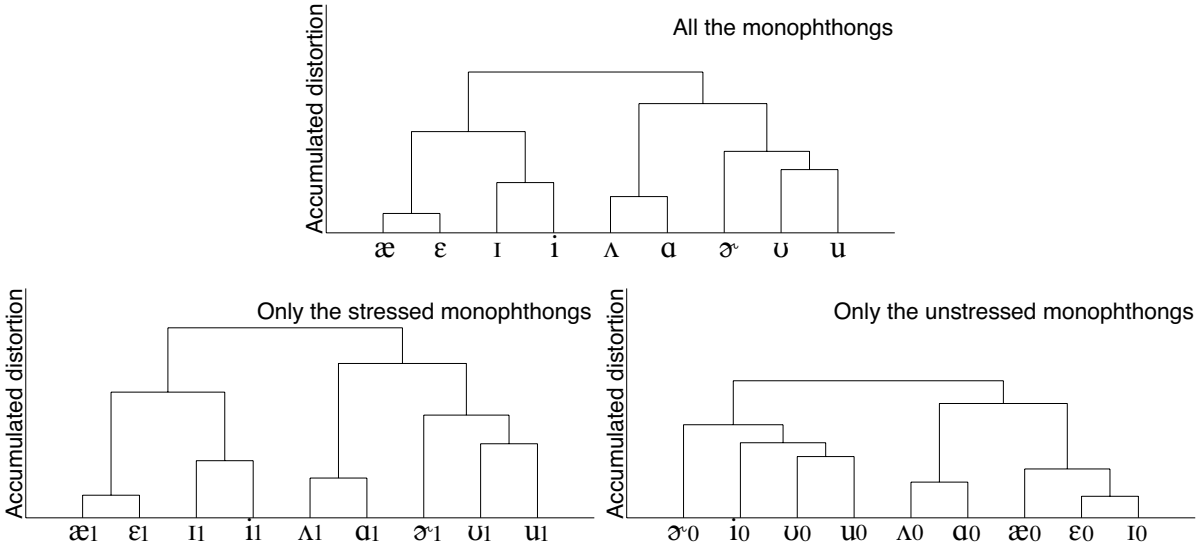

Figure 5: Tree diagrams of American English vowels

5 Correlation to Japanese Vowels of Various Speaking Styles

Do other factors change 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.

5.1 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 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 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, the authors sometimes gave her specific situations as instructions and asked her to utter the 5 vowels suited for the situations.

5.2 Size of the Vowel Structure

Size of the vowel structure was acoustically estimated for each of the 5-vowel utterances. Since the recording was repeated 5 times, 5 distance matrices were obtained for each of the styles and they were averaged. Figure 6 shows the averaged vowel diagrams of 8 styles out of the 12 ones. Clearly seen in the figure, size of the structure changes according to the style or situation. However, separation between front vowels and the others at the top of the tree is commonly found except for 6). Figure 7 shows size of the structure for each style (white bars). Large variability can be clearly seen in the figure. In the following section, it is examined whether size of the structure can be a good measure of distinction perceived by humans when hearing the 5 vowels of each style.

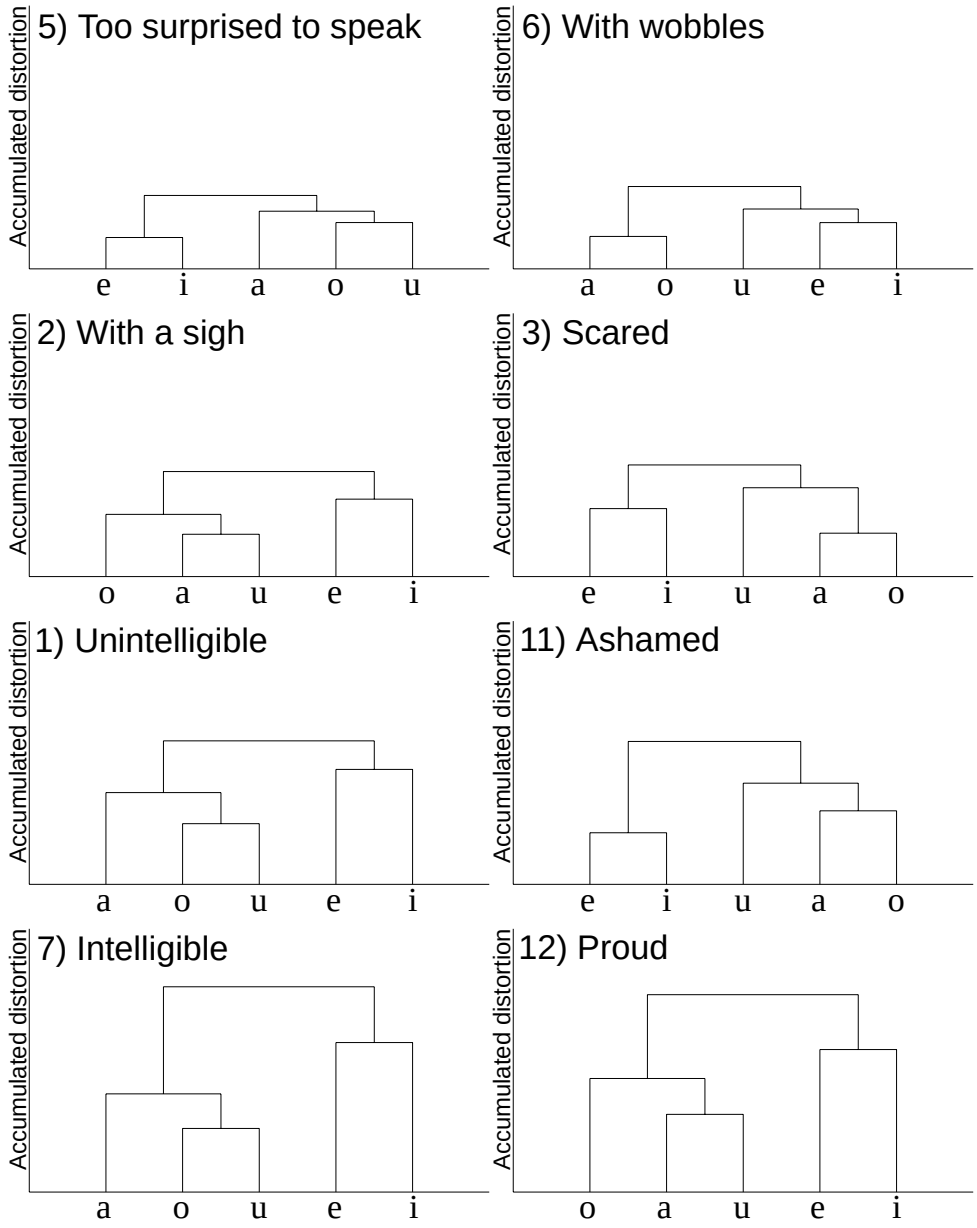

Figure 6: Japanese 5 vowels with different styles

5.3 Comparison with Human Perception of Distinction

A listening test was done with 5 university students of Japanese with normal hearing. The task was to score the magnitude of distinction of the 5 vowels after hearing each vowel set. A 5-degree scale was used for the scoring, 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 size of the structure (white bars) has in Figure 7. The standardized scores are shown as gray bars in Figure 7. 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

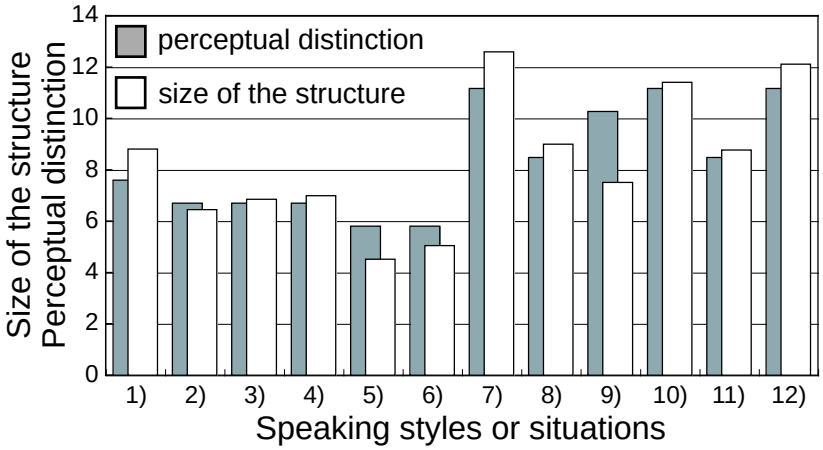

Figure 7: Size of the structure and perceptual distinction

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 becomes 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 acoustically-defined distinction. Still in this case, the energy is well transmitted to listeners and perceptually-defined distinction is easily increased. In this paper, only the relative acoustic properties were focused and these properties are considered just one aspect of speech acoustics. The absolute properties are naturally supposed to have some functions to transmit para-linguistic information. Integration of the relative properties with the absolute ones, including prosodic features, is a future work. If the case of 9) can be ignored, correlation between size of the structure and the perceptual distinction is 0.978.

6 Conclusions

Firstly in this paper, a novel method of acoustic representation of speech, the acoustic universal structure of speech, was introduced, where dimensions to indicate the static non-linguistic information are completely lost. The new representation is realized by mathematically interpreting Saussure's claim and physically implementing structural phonology. In other words, this study has introduced a more than half-century-old discussion of linguistics into mathematics and physics. Then, by focusing on size of the structure, its correlation to articulatory efforts was experimentally examined using speech samples with various speaking styles. Results of the experiments showed that extremely high correlation was found between size of the structure and perceptual distinction. Conventional speech engineering is based on acoustic phonetics and it claims that every speech is distorted and that distortion-free speech can be obtained only by stopping speaking and stopping listening. The proposed acoustic representation of speech implies possibility of yet another speech engineering based on structural phonology implemented on physics. It may claim that speech cannot be distorted if it is produced by a human speaker and that distorted speech can be obtained only by adding para-linguistic information on speech.

Although only the correlation of the acoustic universal structure to para-linguistic information was examined in this paper, the authors are working on identifying kinds of the para-linguistic information by using the structure with some prosodic features. In the symposium, some new results will be presented.

References

[1] M. Pitz and H. Ney, "Vocal tract normalization as linear transformation of MFCC," Proc. EUROSPEECH, pp.1445–1448 (2003)

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

[3] R. Jakobson, G. Fant, and M. Halle, "Preliminaries to speech analysis: the distinctive features and their correlates," MIT Press, Cambridge (1952)

[4] R. Jakobson and M. Halle, "Fundamentals of language," The Hague: Mouton (1975)

[5] M. Halle, "The sound patterns of Russian: a linguistic and acoustical investigation," The Hague: Mouton (1959)

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

[7] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," Proc. ICASSP (2005, submitted)

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

[9] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," Bulletin of Calcutta Maths Society, vol.35, pp.99–110 (1943)

[10] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," IEEE transaction on communication technology, vol.15, no.1, pp.52–60 (1967)

[11] N. Minematsu *et al.*, "Development of English speech database read by Japanese to support CALL research," Proc. ICA, pp.557–560 (2004)

[12] N. Minematsu *et al.*, "Acoustic modeling of sentence stress using differential features between syllables for English rhythm learning system development," Proc. ICSLP, pp.745-748 (2002)

[13] J. Clark and C. Yallop, An introduction of phonetics and phonology, 2nd edition, Blackwell Publishers Inc. (1995)

[14] P. Ladefoged, A course in phonetics, 4th edition, Heinle & Heinle (2001)

[15] http://www.r-project.org