MATHEMATICAL EVIDENCE OF THE ACOUSTIC UNIVERSAL STRUCTURE IN SPEECH

Nobuaki MINEMATSU

Graduate School of Information Science and Technology, University of Tokyo

mine@gavo.t.u-tokyo.ac.jp

ABSTRACT

This paper mathematically shows that there exists the acoustic universal structure in speech, which can be interpreted as physical implementation of structural phonology. The structure has completely no dimensions of multiplicative and linear transformational distortions, which are inevitably involved in speech communication as differences of vocal tract shape, gender, age, microphone, room, line, hearing characteristics, and so on. A speech event, such as a phone, is probabilistically modeled as a distribution of parameters calculated by linear transformation of log spectrum, e.g., cepstrums. A set of the events, such as a word, is relatively captured as structure composed of the distributions. An n-point structure is uniquely determined by fixing lengths of its ${}_nC_2$ diagonal lines, namely, the distance matrix among the n points. Distance between two distributions is calculated as Bhattacharyya distance. The resulting structure has very interesting characteristics. Multiplicative and linear transformational distortions are geometrically interpreted as shift and rotation of the structure, respectively. This fact implies that there always exists a distortion-free communication channel between a speaker and a listener.

1. INTRODUCTION

Speech communication has several steps of production, encoding, transmission, decoding, and hearing. In every step, multiplicative or linear transformational distortions are inevitably caused. With the distortions, however, humans can extract linguistic information from speech so easily as if the distortions cannot disturb the communication. One may hypothesize that the linguistic information in speech is acoustically represented in brain where no dimensions of the above distortions exist, namely, 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 various factors and this causes the "mismatch problem". As far as the author knows, 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 author believes 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 every 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 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 have no dimensions of the inevitable multiplicative and linear transformational distortions. The abstraction is not mental but physical.

2. INEVITABLE ACOUSTIC DISTORTIONS IN SPEECH

What kind of distortions are involved in speech communication and which ones are inevitable? The author considers 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 represented rather well by the average pattern of log-spectrum of the individual. 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 certain human and recorded by a certain acoustic device. If a speech event is represented by cepstrum vector c, the multiplicative distortion is addition of vector b and the resulting cepstrum is shown 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],

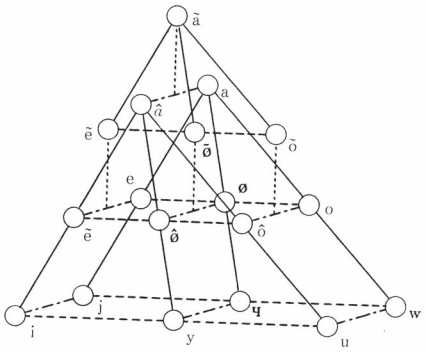

Fig. 1. *Jakobson's geometrical structure of some French phonemes*

any monotonous frequency warping of the log spectrum can be mathematically converted into multiplication of matrix A in cepstrum domain. The resulting cepstrum is shown 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 or b. 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 author believes that this is the most essential reason why every speech system has "sheep and goats" inevitably.

3. PHYSICAL IMPLEMENTATION OF STRUCTURAL PHONOLOGY

3.1. 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 unable to provide the representation by itself. In this work, another speech science is focused on, which is phonology. In phonology, the inevitable distortions are mentally ignored in researchers' brain and speech sounds are represented as abstract entities named phonemes. Phonology is intended to clarify a system or structure hidden or embedded in a set of the phonemes of a language or in sequences of phoneme instances. Here, structural phonology, i.e., structure in the phoneme set, is focused on.

Inspired by Saussure's claim on the language; *Language is a system of conceptual differences and phonic differences*[2], Jakobson, Halle, and others have discussed a system of the phonemes embedded in a language by using distinctive features[3]. Figure 1 shows Jakobson's geometrical structure proposed for some French phonemes and he clams that this structure is invariant among native speakers of French. In phonology, structure is extracted from the sounds of a language in a top-down way based on researchers' knowledge on the language. Then, two researchers may show two different structures because their knowledges are different.

Viewing n elements as structure means that the elements are observed only relatively. Then, extracting the structure can be regarded as a process of ignoring some information in 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 representation the author pursues.

3.2. Sufficient conditions for the physical implementation

Structure can be extracted in a bottom-up way where not knowledge but all the distances between two elements are required. Geometrically speaking, an n-point structure is uniquely determined by obtaining its distance matrix of the n points, equivalent to lengths of its ${}_nC_2$ diagonal lines. With the matrix, a tree diagram can be drawn by a clustering algorithm and the diagram is just a method to visualize the structure. Phonology claims that the phonological structure is universal with regard to speakers, ages, genders, microphones, rooms, lines, listeners, 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 cepstrum space. If n phonemes are found in the space, an n-point structure is defined. Phonology claims that the n-point structure should not be distorted by affine transformation of $c' = Ac + b$ because the transformation represents the 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 author wonders 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. In the above, phoneme is regarded as point in cepstrum space, which represents a single spectrum slice. In this paradigm, the above discussion implies that structural phonology is just an illusion physically and mathematically.

Every speech researcher knows that repetitions of a single pitch waveform, even extracted from *natural* speech, sound like a buzzer. Acoustic perturbations are inevitably observed in speech and a spectrum slice cannot represent this essential characteristics of speech. Then, let phoneme x be represented as distribution $d_x(c)$ in cepstrum space. Since an n-point structure can be determined uniquely by fixing lengths of its ${}_nC_2$ diagonal lines, a necessary and sufficient condition for the implementation is that distance between any two distributions should not be changed by any of a single affine transformation. It there any distribution-to-distribution distance measure that satisfies the above condition?

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

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

where $0.0 \leq \int_{-\infty}^{\infty} \sqrt{d_x(c) d_y(c)} dc \leq 1.0$. This distance measure is derived based on information theory and can be interpreted as amount of self-information of joint probability of the two independent distributions $d_x(c)$ and $d_y(c)$. If the two distributions follow Gaussians, the following is obtained.

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

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 $d_x(c)$ and $d_y(c)$ is not changed.

$$\begin{aligned} &BD(A\mu_x + b, A\Sigma_x A^T, A\mu_y + b, A\Sigma_y A^T) \\ &= BD(\mu_x, \Sigma_x, \mu_y, \Sigma_y) \end{aligned} \qquad (3)$$

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.

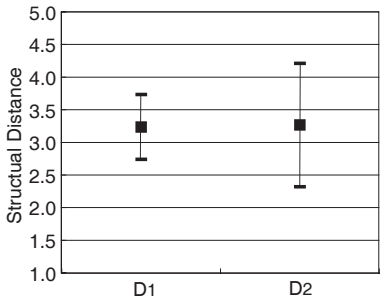

Fig. 2. *Cancellation of the inevitable distortions from speech*

Now, it has been shown 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 called *the acoustic universal structure in speech*.

3.4. Cancellation of the inevitable distortions from speech

A simple experiment was done 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, twelve 5-vowel structures were obtained, three structures for each speaker. Each vowel is represented as a single Gaussian. Distance between two 5-vowel structures, P and Q, is defined as

$$D = \sqrt{\frac{1}{M^2}\sum_{i<j}(\overline{P_iP_j} - \overline{Q_iQ_j})^2}. \quad (4)$$

i is vowel index and M is the number of vowels. $\overline{P_iP_j}$ is distance between vowels i and j in structure P. D is approximately equal to average distance between corresponding vowels of P and Q after full adaptation with regard to A and b[4]. If the inevitable distortions are canceled well by extracting the structure, intra-speaker structural distance, D_1, and inter-speaker structural distance, D_2, should be the same. Figure 2 shows D_1 and D_2 with almost no differences ($p = 90[\%]$). 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.

4. STRUCTURALIZATION OF SPEECH — FROM LANGUAGES TO INDIVIDUALS —

The above sections showed that structural phonology, structuralization of speech sounds of a language, can be implemented on physics and that the 5-vowel acoustic structure is invariant and universal among Japanese natives.

Native speakers of the same local accent provide the same structure. How about non-native speakers? It is often said that no two language students are the same because they have their own forms of the target language. Strictly speaking, even in the case of native speakers, no two native speakers may not be the same because everybody may have his/her own habit of pronunciation. It would be better to understand that the structural (phonological) difference between two native speakers is much smaller than that between two non-native speakers.

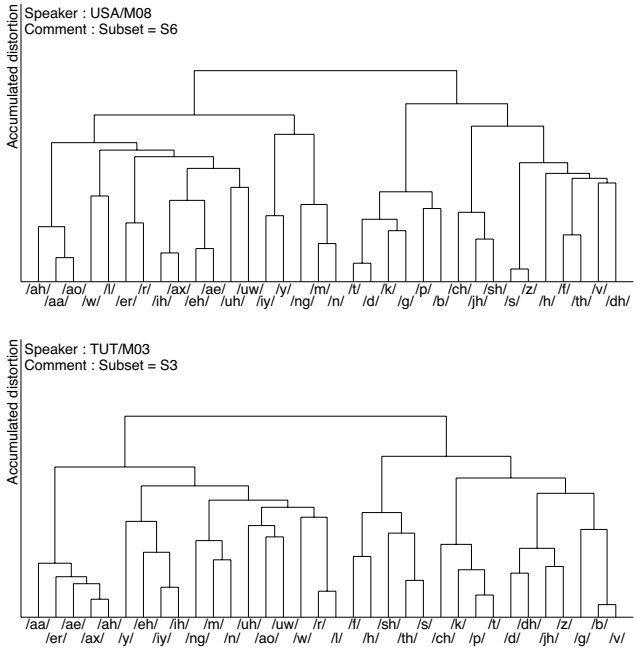

Fig. 3. *Examples of the acoustic universal structure extracted from an American (above) and a Japanese (below)*

If the acoustic universal structure is extracted from utterances of a non-native individual, what does the structure represent? As is shown in the previous sections theoretically and experimentally, non-linguistic information such as vocal tract shape, age, gender, microphone, room, line, and so on are completely unseen. Then, the extracted structure is easily expected to have phonological distortions and they are considered pure interference of the individual's mother tongue on his/her pronunciation of the target language. Figure 3 shows two phonological structures of an American and a Japanese, both speaking English. Clearly shown in the figures, the American tree is much more reasonable in view of phonetics. In the Japanese tree, the well-known Japanese habits of English pronunciation are clearly seen. Confusions of /r/&/l/, /s/&/th/, /z/&/dh/, /f/&/h/, /iy/&/ih/, /v/&/b/, and so on are found. Mid and low vowels of English are located very close to each other because there is the only one mid and low vowel in Japanese. Schwa is close to the above vowels because Japanese often produce the mid and low Japanese vowel for schwa.

Many CALL (Computer Aided Language Learning) systems were built so far but all the systems were built on acoustic phonetics. As is discussed in Section 1, speech representation provided by acoustic phonetics inevitably includes non-linguistic information, which is irrelevant to pronunciation assessment. Acoustic phonetics can give only *noisy* representation and researchers tried to solve this problem by collecting a large amount of data to build speaker-independent models. But the models require adaptation or normalization techniques because the speaker-independent models cannot be really speaker-independent. Recently, some reports indicate the unreliability of CALL systems[5]. What the author did is completely different from what the others had done, which is deletion of the dimensions of the non-linguistic information from speech acoustics. Although some para-linguistic information such as speaking rate and style is thought to modify the acoustic universal structure, if this effect can be ignored, the author believes

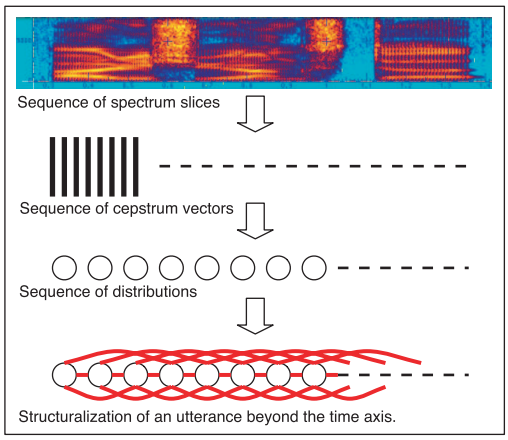

Fig. 4. *Utterance-level structuralization of speech*

that the structure is the only methodology to describe purely and exclusively the pronunciations of individual students. Further, it is very valid to consider that the structure is the student's phonological awareness of the target language, which is estimated from his/her utterances. The author already applied the acoustic universal structure to CALL researches[6, 7].

5. STRUCTURALIZATION OF SPEECH — FROM INDIVIDUALS TO UTTERANCES —

In the previous sections, the acoustic universal structure was discussed with linguistic units such as phonemes or phones. But the universal structure requires not linguistic but acoustic events modeled as distributions. As a word HMM can be trained from a single word utterance, the structuralization of speech is possible only with a single utterance. Figure 4 shows a method of the utterance-level structuralization. Acoustically similar consecutive frames are merged to compose a distribution. Then, the structuralization is done beyond the time axis by considering temporally-distant distributions. The resulting structure cannot be mathematically distorted by the inevitable distortion sources between a speaker and a listener. It seems that the utterance-level structuralization process ignores the temporal order of acoustic events. The author can claim that the temporal order is considered well when two structures, composed of the same number of distributions, are matched. As is described in Section 3.4, D is approximately equal to average distance between the *corresponding* phones of P and Q after full adaptation with regard to A and b[4]. This matching may imply mathematical possibility of improving speech recognition only with relative acoustic properties of speech, that is the structure.

Do human listeners use the utterance-level acoustic universal structure when they extract linguistic information from speech acoustics? The author already did a perceptual experiment to try to answer this question. Although the experiment is described in another paper[8] due to limit of space, the results indicate that easy and rough listening exploits the structure although intensive and analytic listening does not use the structure.

The utterance-level acoustic universal structure can be viewed as speech modeling only with speech dynamics. But the universal structure differs from the other dynamics modeling methods in two points. The universal structure considers not only the dynamics between two consecutive speech events but also the dynamics between distant speech events. The universal structure does not consider the direction of speech dynamics, but only the magnitude of speech dynamics. Acoustic change caused by vocal tract length difference is mathematically interpreted as rotation of the structure. This expects that Δ cepstrum *vectors* mathematically reduces the robustness of acoustic models with regard to vocal tract length differences. Additional parameters are supposed to increase data dependency and, in the case of Δ cepstrums, the dependency is mathematically shown to be on vocal tract length differences.

Mel warping is often modeled as frequency warping. This operation is meaningless in the acoustic universal structure because it is just rotation. Two listeners have different hearing characteristics as two speakers have different vocal tract shapes. The author considers that what is important is not to warp a frequency axis to the average hearing characteristics, but to find out speech parameterization which is invariant with any frequency warping functions, i.e., with any speakers and any listeners.

6. CONCLUSIONS

This paper proposes a new method of observing speech acoustics. The method is derived from structural phonology and realized by implementing structural phonology on physics. Speech events are modeled probabilistically as distributions, distance between any two of the events is calculated based upon information theory, and the events are relatively captured as structure. The resulting structure is invariant and universal with regard to the non-linguistic information inevitably involved in speech. 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 current paper may imply possibility of yet another speech engineering based on structural phonology implemented on physics. It may claim that speech cannot be distorted mathematically if it is produced by a human speaker and that distorted speech can be obtained only by adding para-linguistic information on speech[9].

7. REFERENCES

[1] M. Pitz *et al.*, "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 *et al.*, "Preliminaries to speech analysis: the distinctive features and their correlates," MIT Press, Cambridge (1952)

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

[5] A. Neri *et al.*, "Automatic speech recognition for second language learning: how and why it actually works," Proc. ICPhS, pp.1157–1160 (2003)

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

[7] N. Minematsu, "Pronunciation assessment based upon the compatibility between a learner's pronunciation structure and the target language's lexical structure," Proc. ICSLP, pp.1317–1320 (2004)

[8] N. Minematsu *et al.*, "Yet another acoustic representation of speech based on physical implementation of structural phonology," Technical report of IEICE, SP2004-28, pp.53–58 (2004, in Japanese)

[9] N. Minematsu *et al.*, "The acoustic universal structure in speech and its correlation to para-linguistic information in speech," Proc. Int. Workshop on Man-Machine Symbiotic Systems, pp.69–79 (2004)