YET ANOTHER ACOUSTIC REPRESENTATION OF SPEECH SOUNDS

Nobuaki MINEMATSU

Graduate School of Information Science and Technology, University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033 JAPAN
mine@gavo.t.u-tokyo.ac.jp

ABSTRACT

This paper proposes yet another representation of speech sounds. The proposed speech modeling can remove both multiplicative and linear transformational distortion from speech theoretically. It means that speech sounds are represented without being affected by any static distortion inevitably involved in production, encoding, transmission, decoding, and hearing processes, such as differences in vocal tract length, gender, age, microphone, room, line, auditory characteristics, and so on. The method acoustically models not individual phones but their entire system, where only acoustic interrelation embedded in all the kinds of phones is focused. Since the method provides us with no absolute acoustic properties of phones, it cannot recognize or synthesize even a single phone. On the contrary, the proposed method is shown to be able to be applied to pronunciation assessment effectively and reliably, where the proficiency of pronunciation is estimated without using acoustic models of the individual phones directly in the matching.

1. INTRODUCTION

In every speech application, speech sounds are modeled based upon acoustic phonetics. In speech recognition, context-dependent phone models are built using cepstrum parameters and, in speech synthesis, context-dependent (poly-)phone waveforms or models are stored. In this paradigm, the individual speech units have their own acoustic templates. But acoustic properties of a speech unit is easily affected and distorted by various factors such as microphone, room, line, speaker, and so on. If the templates are used in conditions different from those where the templates were built, some unexpected results are often seen. This is called "mismatch problem" and, as far as the author knows, all of the previous studies tried to solve it in one of the following two methods. In the first one, the templates are prepared separately for each of all the conditions possible and, if this is practically impossible, the templates are *adapted* by using a small number of acoustic observations. In the second, not the templates but the observations are modified to be *normalized*. In this case, the templates are built after normalization. Even with adaptation or normalization, however, it is known that every speech recognizer still has "sheep and goats." This fact is partly attributed to limitation in the amount of speech samples used for adaptation or normalization. But the author believes that another and essential reason for the problem is that every speech system is built on an assumption that the system has to have acoustic models of individual phones. Under this assumption, which is derived from phonetics, even after normalization, every model comes to have certain acoustic properties with regard to each of the various factors mentioned above. If we continue to use the phonetics-based models of speech, strictly speaking, we may never be able to solve the mismatch problem completely.

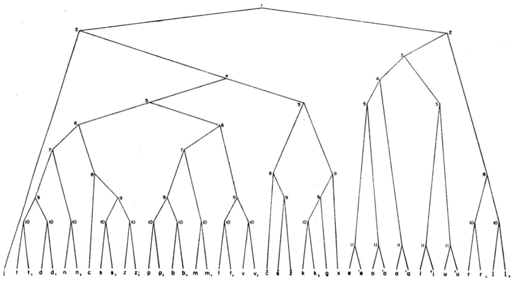
1
2
3
4
5
6
7
8
9
10
11
j t t, d d, n n, c s s, z z; p p, b b, m m, f f, v v, č š ž k k, g x e 'e o 'o a 'a i 'i u 'u r r, l l,

Fig. I-1. Branching diagram representing the morphonemes of Russian. The numbers with which each node is labelled refer to the different features, as follows: 1. vocalic vs. nonvocalic; 2. consonantal vs. nonconsonantal; 3. diffuse vs. nondiffuse; 4. compact vs. noncompact; 5. low tonality vs. high tonality; 6. strident vs. mellow; 7. nasal vs. nonnasal; 8. continuant vs. interrupted; 9. voiced vs. voiceless; 10. sharped vs. plain; 11. accented vs. unaccented. Left branches represent minus values, and right branches, plus values for the particular feature.

Fig. 1. Halle's tree diagram of Russian phonemes

In this paper, another acoustic representation of speech is proposed, which is inspired by phonology. Phonetics was born to describe phones and phonology was born to describe a language. How does phonology describe a language and its sounds? Phonology is intended to clarify a system hidden or embedded in a set of sounds of a language (phonemes) or in sequences of the phonemes. Inspired by Saussure's structuralism, Jakobson, Halle, Clements, and others have discussed the system of phonemes embedded in a language by using distinctive features[1, 2]. Figure 1 shows Halle's tree diagram proposed for Russian phonemes. One feature can classify all the phonemes into two groups. Another feature will give us four ones. The order of applying the features is determined so that a set of phonemes under every node comprise *a natural class*. Putting it another way, a set of phonemes under every node always correspond to a linguistic event specific to the phonemes of the node. In phonology, structure is extracted in a top-down way with knowledge on a specific language. But structure can be extracted in a bottom-up way where not linguistic knowledge but acoustic similarity measure between two phonemes is required. In the following sections, the bottom-up structure of phones is well discussed. It is surprising that, partly because not the individual phones but their entire structure is focused, the structure extraction can remove both multiplicative and linear transformational distortion from speech as gracefully as cepstrum smoothing can remove source information from log power spectrum. Structurally represented speech events have no dimensions for static distortion inevitably involved in production, encoding, transmission, decoding, and hearing processes. This implies that Jakobson's tree, 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.

2. SPEECH DATABASE USED IN THE ANALYSIS

The author and his co-workers designed and developed an English database read by Japanese students and General American (GA)

0-7803-8484-9/04/$20.00 ©2004 IEEE

speakers for CALL researches[3], which was used as speech samples in this work. The database is divided into two sets. One is related to segmental aspect of pronunciation and the other is to its prosodic aspect. In recording, a speaker was asked to repeat reading given words or sentences until he/she judged that the correct pronunciation was done. As for Japanese English (JE) samples, the resulting database can be said to contain only correct English utterances at least for Japanese students. A subsequent analysis of pronunciation errors showed that the JE samples still contained a large number of pronunciation errors[3]. The number of speakers is 222 (100 male and 102 female Japanese and 8 male and 12 female Americans). Pronunciation proficiency was rated by American teachers for the individual students in terms of the three aspects of pronunciation; segmental, rhythmic, and intonational aspects.

3. ANOTHER REPRESENTATION OF SPEECH

3.1. Training of speaker-dependent monophone HMMs

With the GA and JE material of the database, monophone HMMs were trained for each speaker. Here 60 sentence utterances were used for the training and the 60 sentences were a part of TIMIT phonetically rich sentences. Content of the sentences depended on speakers. Due to the rather small number of training samples, several diphthongs were not found. Then, HMMs of all the monophthongs and consonants were trained. Table 1 shows acoustic conditions of the training. Phonemic transcriptions of the training data were automatically generated by looking up PRONLEX lexicon and, consequently, pronunciation errors in the JE samples were not represented explicitly in the transcriptions.

3.2. Bottom-up clustering of phone HMMs

Bottom-up clustering of some elements is possible only with distances between any two of the elements (distance matrix). Here, square root of Bhattacharyya distance (BD) measure was adopted. BD between two elements (distributions) is formulated as follows.

$$\begin{aligned} & BD(P_u, P_v) \\ & = -\ln \int_{-\infty}^{\infty} \sqrt{P_u(x)P_v(x)}dx \\ & = \frac{1}{8}\mu_{uv}\left(\frac{\sum_u + \sum_v}{2}\right)^{-1}\mu_{uv}^T + \frac{1}{2}\ln\frac{|\frac{\Sigma_u+\Sigma_v}{2}|}{|\sum_u|^{\frac{1}{2}}|\sum_v|^{\frac{1}{2}}}, \end{aligned} \quad (1)$$

where μ_u is average vector in element u and μ_{uv} is $\mu_u - \mu_v$. Σ_u is variance and covariance matrix for u. It is assumed that element u can be modeled appropriately as Gaussian distribution. This distance measure is derived based on information theory and can be interpreted as amount of self-information of joint probability of two independent events (elements).

Table 1. Acoustic conditions for the analysis

sampling	16bit / 16kHz
window	25 ms length and 10 ms shift
parameters	FFT-based cepstrums and their derivatives
speakers	202 Japanese and 20 Americans
training data	60 sentences per speaker
HMMs	context-independent and 1-mixture monophones with diagonal matrices
topology	5 states and 3 distributions per HMM
monophones	b, d, g, p, t, k, jh, ch, s, sh, z, zh, f, th, v, dh, m, n, ng, l, r, w, y, h, iy, ih, eh, ae, aa, ah, ao, uh, uw, er, ax

The distance matrix provides us with all the information on interrelation among the elements and enables us to cluster the elements for visualization. Among some widely-used clustering algorithms, Ward's method was adopted in this work because the author considers that the clustering criterion of the method is more effective statistically than those of the others. A state-based matrix and a phone-based one give us a state-based diagram and a phone-based one, respectively. Phone-based distance is calculated, for example, as averaged state-level distance between two phones.

Figure 2 shows an example of a phone-based tree diagram of a GA speaker. While the top split of Halle's tree in Figure 1 divides phonemes into vocalic and non-vocalic ones, that of the bottom-up tree does phones into vowels, nasals, liquids, and glides and the others. Since the above four kinds of phones are well-known to have much in common with regard to their articulatory and acoustic properties, the bottom-up tree is considered more reasonable phonetically. Also at the other splits in Figure 2, as expected, phonetically good and valid clustering is seen.

3.3. Characteristics of the new representation

Since HMMs are trained purely based upon acoustic properties of phones, it is quite natural that their interrelation gives us a phonetically reasonable tree. Here, let us consider what kind of change can be see in the tree structure by linearly transforming cepstrum vector at time t, c_t, in training data.

$$c'_t = Ac_t + b, \quad (2)$$

where A and b are a constant matrix and a constant vector, respectively. With this transformation, mean vector μ' and variance and covariance matrix Σ' of c'_t are represented using μ, Σ, A, and b.

$$\mu' = E(c'_t) = A\mu + b \quad (3)$$

$$\Sigma' = E(c'_t - \mu')(c'_t - \mu')^T = A\Sigma A^T \quad (4)$$

Then, Equations 1, 3, and 4 lead to the following equality.

$$\begin{aligned} & BD(\mu'_u, \Sigma'_u, \mu'_v, \Sigma'_v) \\ & = BD(A\mu_u + b, A\Sigma_u A^T, A\mu_v + b, A\Sigma_v A^T) \\ & = BD(\mu_u, \Sigma_u, \mu_v, \Sigma_v) \end{aligned} \quad (5)$$

Bhattacharyya distance between two distributions is not changed by any constant and linear transformation. This directly indicates that any constant and linear transformation cannot change the distance matrix and its structure.

As is well-known, linear transformation of $Ax+b$ is referred to as affine transformation. In Euclidean space, affine transformation

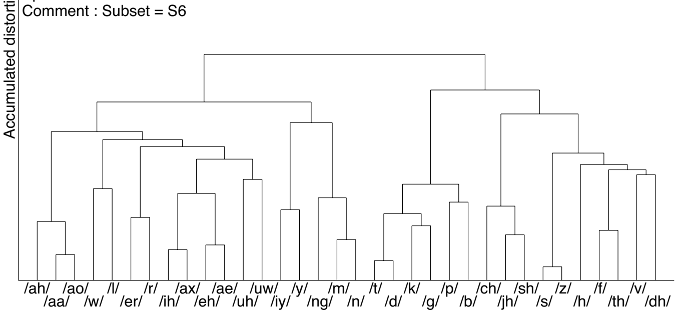

Fig. 2. An example of a phone tree diagram of a GA speaker

changes a structure into another and the change is often classified into several types. Matrix A realizes scaling, warping, or rotation of a structure and also realizes their combination. Vector b realizes shift of a structure. In Bhattacharyya space, since distance between two elements is not changed by linear transformation, matrix A and vector b correspond to rotation and shift, respectively, where structure is not changed by any constant linear transformation.

What is acoustic-phonetic interpretation of $Ax+b$? Since addition of b corresponds to that of a log spectrum pattern to an original pattern, b corresponds to multiplicative distortion caused by differences in microphone, room, line, and so on. In speaker recognition, a speaker is often modeled as a GMM, which is an average pattern of log spectrum. This means that speaker individuality can be partly modeled as b. What's A then? In [4], it is proved that any monotonously continuous frequency warping in spectrum domain can be theoretically converted into a constant A in cepstrum domain. In speech recognition, the frequency warping of spectrum is often used to represent vocal tract length differences. This directly means that A can simulate acoustic distortion caused by vocal tract length differences and it implies that distortion caused by physical growth of an individual as well as speaker differences in size cannot be seen in the structure. A part of matrix space of A corresponds to the frequency warping and other parts are expected to give us some different functions. And any kind of A, if constant, cannot change the structure.

In MLLR adaptation of acoustic models in speech recognition, multiple matrices are usually used for mixture-based bottom-up clustering of triphone HMMs[5]. Triphones are trained with a large amount of data in which different speakers read different sentences. This implies that different parts of a triphone set have different speaker individuality[6] and that this is why multiple matrices are required. At least in MLLR adaptation of triphones in HMM speech synthesis, some techniques are introduced to realize even individuality in every part of a triphone set[6] and one or a few matrices can be used for the adaptation effectively.

To sum up, speech events structurally represented by the proposed method have no dimensions for static distortion inevitably involved in production, encoding, transmission, decoding, and hearing processes. The author supposes that if acoustic matching is possible based on only the structural representation, without direct use of phone models, it may realize the most stable and robust speech application. But since the representation cannot provide us with any absolute acoustic properties of phones, it cannot recognize or synthesize even a single phone. What's possible? In the following section, the proposed method is applied to the application area requiring the most stable technology, which is education.

4. APPLICATION OF THE NEW REPRESENTATION

4.1. Quantitative description of individual students

In pronunciation training, it is often said that no two students are the same. Every student produces sounds of the target language by his/her own way and it is desirable to instruct the individual students after *knowing* what kind of states they are in. Even if teachers have good knowledge of phonetics and phonology of the target language, it is often impossible to describe students adequately because, for example, there is no science such as Japanese English phonetics or phonology. Some teachers are trying to solve this problem with tools of acoustic phonetics. But what the tools give is very noisy representation of speech in that many things irrelevant to the proficiency are on the acoustic representation. Even if they are not seen, the representation requires good knowledge of acoustics and physics. Figure 3 shows an example of a poor student in the proposed representation, where there is no noises or acoustics. What can be seen there is quantitative dependency of his/her pronunciation on their mother tongue. In the figure, 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 closer to each other because there is only one mid and low vowel in Japanese. Schwa is closely located to the above vowels. These findings are just common belief and the analysis showed that different students drew different trees. If distance measure is defined adequately between two trees, 202 students in the database can be clustered as their phones are clustered. Although the author already did the clustering to define *types* of Japanese English, due to limit of space, the types will be reported elsewhere.

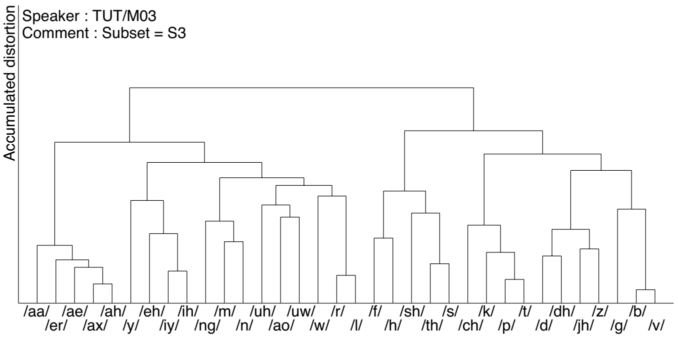

Fig. 3. An example of a phone tree diagram of a poor student

4.2. Automatic estimation of the proficiency

Most of the previous studies of pronunciation proficiency estimation did acoustic matching between native models and a student's utterances[7, 8]. In this case, since the matching is done with the individual phone models, nobody cannot guarantee that the mismatch problem will not occur. Actually, recent reports from teachers say that machine estimation of the proficiency seems not reliable or pedagogically-sound enough[9]. The author supposes that the most probable reason for that is the mismatch problem and expects that the proposed method can solve it effectively.

Before matching between a teacher's structure and a student's one, characteristics of a single structure is described. If M points are in N-dimensional Euclidean space as in Figure 4, the following equation is true, where G_P is a gravity center of $\{P_i\}$.

$$\sqrt{\frac{1}{M^2}\sum_{i<j}\overline{P_iP_j}^2} = \sqrt{\frac{1}{M}\sum_{i}\overline{P_iG_P}^2} \tag{6}$$

If Bhattacharyya distance (BD) is used for Euclid distance, the above equation is not satisfied. But square root of BD satisfies the equation approximately. Correlation of left and right quantities of the equation in JE phones over the students was 0.997. This is why square root of BD was used. Now, let us consider two structures. If M points are phones in cepstral space with their distributions, then the following equation is approximately true for JE phones.

$$\sqrt{\frac{1}{M^2}\sum_{i<j}\overline{P_iP_j}\times\overline{Q_iQ_j}} \approx \sqrt{\frac{1}{M}\sum_{i}\overline{P_iG_P}\times\overline{Q_iG_Q}} \tag{7}$$

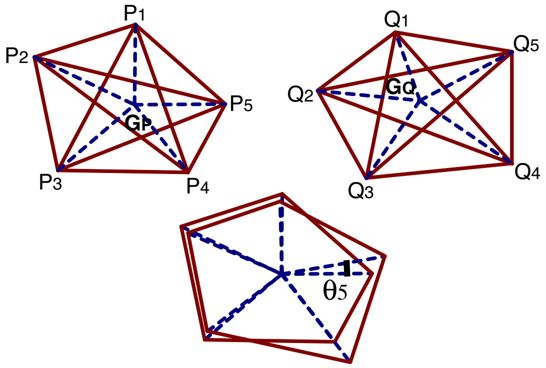

Fig. 4. Two structures and their shift & rotation for fitting

Correlation was 0.999. Equations 6 and 7 lead to another equality.

$$\sqrt{\frac{1}{M^2}\sum_{i<j}(\overline{P_iP_j}-\overline{Q_iQ_j})^2} \approx \sqrt{\frac{1}{M}\sum_{i}(\overline{P_iG_P}-\overline{Q_iG_Q})^2} \quad (8)$$

The right term is approximated to be average of cepstrum distances between two corresponding phones of the two structures after shift and rotation, where the two gravity centers are put at a position and one of the two structures is rotated so that the $\sum|\theta_i|$ (see in Figure 4) should be minimized. The left term is Euclid distance between two distance matrices by viewing a matrix as a vector. In brief, Euclid distance between two matrices, structural distortion henceforth, approximates cepstrum distance averaged over all the corresponding phones of the two structures *after full adaptation*.

Figure 5 shows the structural distortion and the positional distortion, which is average of cepstrum distances with no shift or rotation, for two cases. One is distortion between two GA speakers (GA-GA) and the other is between a GA speaker and a Japanese (GA-JE). In the left, only vowels are used and, in the right, all the phones are used. In the vowel graph, while GA-JE and GA-GA distributions are overlapped in the positional distortion, they are clearly separated in the structural distortion. This was much to be expected because the two kinds of distortion differ in whether adaptation is done. In the other graph, even the structural distortion shows less clear separation. This result implies some adequate selection of phones or phone pairs should be done before calculating the structural distortion. It is easily expected that the selection based on phone pairs will give *finer* definition of the distance measure because the number of phone pairs is much larger than that of phones. The author supposes that the selection should depend on application. For example, if automatic estimation of the pronunciation proficiency is required, the selection should be done so that correlation between $5-tr$ (Teachers' Rating, where 5 is the maximum score) and the structural distortion between students and a teacher should be maximized. Figure 6 shows the results with very high correlation between $5-tr$ and the estimated scores. Although the author already did several experiments as for robustness improvement, due to limit of space, they will be reported elsewhere.

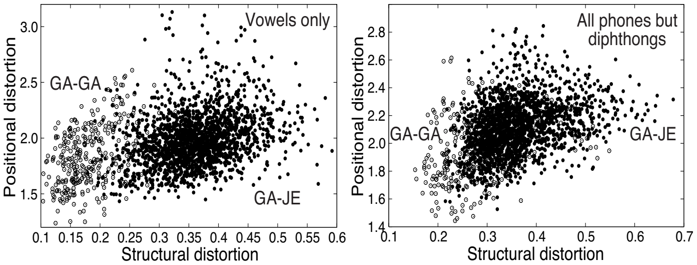

Fig. 5. Structural and positional distortion for the two cases

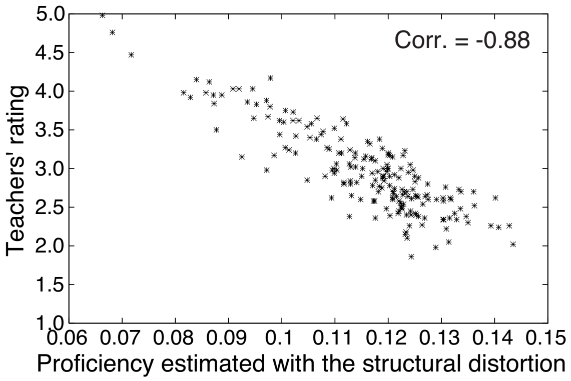

Fig. 6. Relation between teachers' rating and the estimated scores

5. CONCLUSIONS

This paper proved that the universal and essential structure of speech exists in pure acoustics of speech, where multiplicative and linear transformational distortion is removed completely and theoretically. Then, the structural representation was applied to describe individual language students quantitatively and estimate their proficiency successfully. Based upon only the new representation of speech, i.e., no individual phone models, the author already did clustering of the students to define *types* of JE and automatic generation of pedagogical instructions on *which phone should be corrected among others*. The generation was based upon *intelligibility* criterion[10] not upon *native-sounding* criterion and the results will be reported elsewhere with teachers' comments on them.

6. REFERENCES

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

[2] M. Halle, "The sound patterns of Russian," The Hague: Mouton (1959)

[3] N. Minematsu *et al.*, "English speech database read by Japanese learners for CALL system development," Proc. LREC'2002, pp.896–903 (2002)

[4] Michael Pitz *et al.*, "Vocal tract normalization as linear transformation of MFCC," Proc. EUROSPEECH, pp.1445–1448 (2003)

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

[6] J. Yamagishi, *et al.*, "A context clustering technique for average voice models," IEICE Trans. Inf. & Syst., vol.E86-D, no.3, pp.534–542 (2003)

[7] S. Witt, *et al.*, "Language learning based on non-native speech recognition," Proc. EuroSpeech, pp.633–636 (1997)

[8] B. Sevenster, *et al.*, "Evaluation and training of second-language learners' pronunciation using phoneme-based HMMs," Proc. STiLL, pp.91–94 (1998)

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

[10] N. Minematsu *et al.*, "Corpus-based analysis of production and perception of Japanese English in view of the entire phonemic system of English," Proc. ICPhS, pp.1569–1572 (2003)