Holistic and Prosodic Representation of the Segmental Aspect of Speech

N. Minematsu[†], *T. Nishimura*[‡], *D. Saito*[†], *S. Asakawa*[†], *Y. Qiao*[†]

[†] Graduate School of Engineering, The University of Tokyo
[‡] Graduate School of Medicine, The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan
{mine,dsk_saito,asakawa,qiao}@gavo.t.u-tokyo.ac.jp, nt-tazuko@ams.odn.ne.jp

Abstract

Speech communication has several steps of encoding, transmission, and decoding. In each step, various acoustic distortions are inevitably induced by non-linguistic factors such as differences of age, gender, microphone, line, room, auditory characteristics of a hearer's ears, etc. In spite of this large variability, humans can perform very precise speech processing. Recently, the first author proposed a novel representation of speech [1, 2], which is invariant with these factors at all. Only the dynamic motions in speech are focused on and the static features in speech are completely discarded. The high validity of this new representation for speech recognition was already verified experimentally [3, 4, 5]. In this paper, we show that the new representation of the segmental aspect of speech can be interpreted as a kind of holistic and prosodic feature because the representation captures speech as music, i.e. *timbre*-based melody.

1. Introduction

Many speech sounds are produced as standing waves in a vocal tube and their acoustic characteristics mainly depend on the shape of the tube. No two speakers have the same tube in general and speech acoustics come to have speaker variability. Different shapes cause different resonance, which causes different timbre[1]. In the same way, different vowels are produced in a vocal tube by changing its shape. Acoustically speaking, both speaker difference and vowel difference are caused by the same reason. Further, the timbre of speech can be easily changed also by other factors such as microphone, room, line, etc.

Despite this large acoustic variability, humans can perform accurate speech perception easily. How is this done? Even after the long history of speech science, this still remains one of the unanswered questions, i.e. the variability of speech acoustics and the invariance of speech perception [6]. Speech engineering has tried to answer it by collecting a large number of samples of the individual linguistic categories, e.g. phonemes, and modeling them statistically. For example, IBM announced that they had collected speech samples from 350 thousand speakers to build a speech recognizer [7]. As far as we know, however, no child needs such an enormous number of samples to be able to understand speech. A major part of the speech it hears are from its mother and father. After it begins to talk, as *speech chain* implies, about a half of the speech it hears is its own speech.

Developmental psychology explains that infants acquire spoken language through imitating utterances of their parents. Here, we can say that infants never imitate the voices of their parents, which is a clear difference from the vocal imitation of myna birds. They imitate many sounds of cars, doors, animals, etc and they also imitate human voices. Hearing an adept myna bird say something, one can guess its keeper [8]. Hearing a very good child say something, however, it is impossible to guess its keeper. What in the voices is imitated by infants? Due to poor phonological awareness, it is difficult for them to decode an input utterance into a string of phonemes [9, 10, 11] and therefore, it is also difficult to convert the individual phonemes into sounds. In this situation, what is imitated by infants? Developmental psychology claims that they extract the holistic sound pattern of an input word, called *word Gestalt* [10, 11], and reproduce it with their mouths. Then, what is the acoustic definition of that Gestalt? It must be speaker-invariant because infants can extract the same Gestalt whoever speaks that word to them using different voices. As far as we surveyed, no researcher had yet succeeded in deriving its acoustic definition [12].

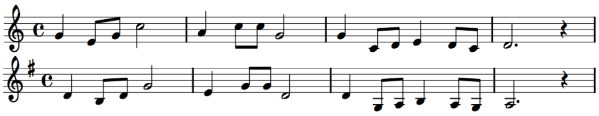

Figure 1: A musical piece and its transposed version

No child learns speech sounds as they are but myna birds try to learn them as they are. We can say that every speech synthesizer learns speech sounds as they are and therefore, by hearing an output speech sample of the synthesizer, one can guess the original speaker easily. In this sense, every speech synthesizer is a myna bird simulator, not a human simulator. We can also say that every speech synthesizer convert phoneme sequences into sounds but, as noted above, no child acquires spoken language by reading phoneme sequences into sounds. They acquire word Gestalt, which does not convey speaker information, and try to reproduce that Gestalt with their *short* vocal tubes, that is the vocal imitation of infants. Then, we hear their *sweet* voices.

In the present paper, at first, we describe how we derived a speaker-invariant representation of speech. The derivation was carried out by considering intrinsic similarity between speech and music. After summarizing some experimental results, we discuss that the representation can be interpreted as a kind of holistic and prosodic feature of speech although, only with the new representation, spoken words are correctly identified.

2. Speaker-invariant representation

2.1. Key-invariant representation of music

Figure 1 shows a musical piece and its transposed version. The upper melody is C-major and the other is G-major. Hearing these musical pieces, it is usually easy to recognize the equivalence between the two although they are acoustically different. People with strong absolute pitch (AP) show some difficulty in perceiving the equivalence [13]. When hearing these pieces, they are automatically converted into pitch name sequences. For

[1] In musicology, timbre means the spectral envelope of a sound.

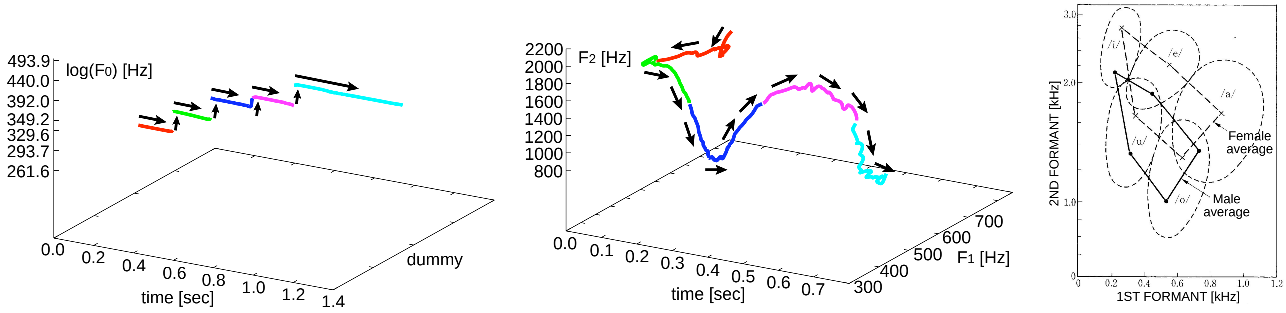

Figure 2: Dynamic changes of pitch in CDEFG and those of timbre in /aiueo/ with the Japanese vowel chart

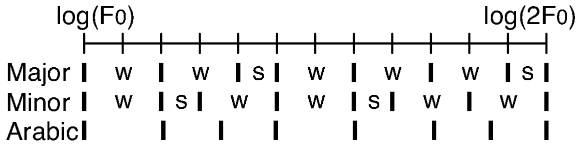

Figure 3: Three musical scales of Major, Minor, and Arabic

example, the upper melody is converted into "GEGC ACCG GCDEDC D". People with very strong AP have to consciously transform this symbol sequence into the sequence of the second melody to check the equivalence. This is considered to be a reason why people with AP takes a longer time to perceive the equivalence between a melody and its transposed version [13].

We can find people who cannot transcribe a melody as a sequence of sound symbols of pitch names or syllable names. For them, however, it is easy to perceive the equivalence between the two musical pieces in Figure 1. It is evident enough that the equivalence perception does require not sound identification but melody contour comparison. A melody contour is defined as a sequence of *local* pitch movements. If $\Delta F0_t$ is defined as $\Delta F0_t{=}F0_t{-}F0_{t-1}$, a sequence of $\Delta F0_t$ represents the melody contour. In Western music, an octave is divided into 12 semitone intervals and a musical scale is composed of 8 tones, which have 5 whole-tone intervals (Ws) and 2 semi-tone intervals (Ss). It should be noted that the tones' relative arrangement is invariant with key. Figure 3 shows two well-known keys, Major and Minor, and Arabic musical scale. If C is used as Tonic sound (the first sound) in major key, the scale is called C-major. For any major key, the tonal arrangement is the same, which means that the melody contour or $\Delta F0_t$ sequence is key-invariant.

2.2. Speaker-invariant representation of speech

In music, absolute acoustic properties of individual tones are key-dependent but their melody contour is key-independent. In speech, those of individual sounds are speaker-dependent. Is their timbre contour speaker-independent?

Figure 2 shows a melody (pitch) contour of CDEFG and a timbre contour of /aiueo/. Both contour patterns are visualized in a phase space. Here, pitch is a one-dimensional feature of F_0 and timbre is tentatively shown as a two-dimensional feature of F_1 and F_2. Transposition of music translates the pitch contour but the shape of the contour is not altered. The Japanese F_1/F_2-based vowel chart is also shown in Figure 2. It is seen that the male vowel system is translated to fit to the female vowel system in which the vowel arrangement is not changed. If this vowel system invariance is always satisfied independently of any kind of the non-linguistic factors, then, the timbre contour can be considered as the acoustic definition of word Gestalt. As explained shortly, however, this simple derivation does not provide a good answer to "what is the acoustic definition?"

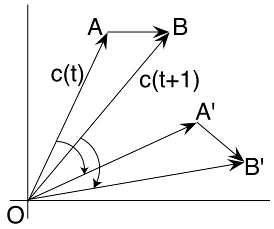

Figure 4: Rotation of two cepstrum vectors and their Δ vector

Using cepstrum c, *local* timbre movements are represented as Δcepstrum, which is derived by the following equation,

$$\Delta c_t = \frac{\sum_{\tau=1}^{T} \tau(c_{t+\tau} - c_{t-\tau})}{2\sum_{\tau=1}^{T}\tau^2}.$$

If we use 1 as T, the shortest window length of 2, Δc_t is calculated as $\frac{1}{2}(c_{t+1}{-}c_{t-1})$. Although both ΔF_0 and Δc are the velocity components of observations, a sequence of ΔF_0 is key-independent but that of Δc is strongly speaker-dependent.

2.3. Directional dependence of cepstrum on speakers

Difference of the vocal tract length changes formant frequencies. If it becomes shorter or longer, they will become higher or lower, respectively. This change is often modeled as frequency warping of a spectrum envelope in the spectral domain and as multiplication of matrix A on c in the cepstral domain [14].

$$A = \begin{pmatrix} 1{-}\alpha^2 & 2\alpha{-}2\alpha^3 & \cdots & \cdots \\ -\alpha{+}\alpha^3 & 1{-}4\alpha^2{+}3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \quad (|\alpha| < 1.0)$$

Using $c'{=}Ac$, it is possible to convert a speech sample of a male adult into that of a boy. The element a_{ij} is described as

$$a_{ij} = \frac{1}{(j-1)!}\sum_{m=\max(0,j-i)}^{j}\binom{j}{m}$$
$$\times\frac{(m+i-1)!}{(m+i-j)!}(-1)^{(m+i-j)}\alpha^{(2m+i-j)},$$

where

$$\binom{j}{m} = \begin{cases} {}_jC_m & (j \geq m) \\ 0 & (j < m). \end{cases}$$

We carried out a geometrical analysis of this matrix and found that A has a very strong function of rotating cepstrum vectors although A is not completely a rotation matrix [15]. This rotation is dependent on vocal tract length difference and reasonably

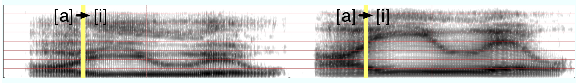

Figure 5: Original (left) and warped (right) speech samples

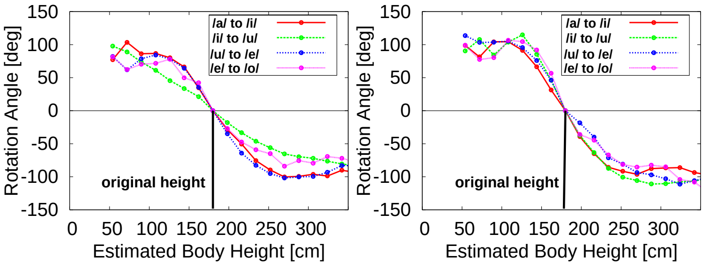

Figure 6: Rotation of cepstrum vectors and their Δs

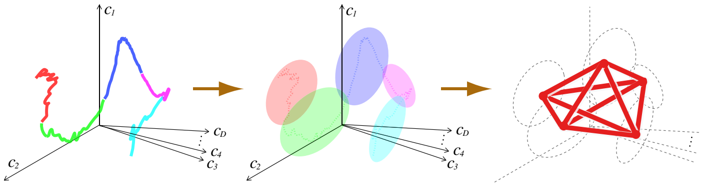

Figure 7: Distribution-based structuralization

independent of speakers and phonemes. As shown in Figure 4, if two consecutive cepstrum vectors are rotated similarly, then, their Δ vector is also rotated in the same way.

Figure 5 shows two speech samples of /aiueo/; an original one (male adult) and its warped version (boy) using A. The formant frequencies are clearly shifted higher. Figure 6 shows some results of analyzing the relation between rotation angles and the degree of body height change through warping. As in Figure 5, an /aiuoe/ utterance of a male adult was warped into that of speakers of different heights. The original height was 167 cm and the height was changed into 50 cm to 350 cm. From these utterances, four fixed points were detected, i.e. the central positions of transition of /a/ to /i/, /i/ to /u/, /u/ to /e/, and /e/ to /o/. In Figure 5, the position of /a/ to /i/ transition is shown. We can say from Figure 6 that the rotation of cepstrum vectors and that of their Δs are very similar and that the rotation is vowel-independent. Similar analysis was done with Δ^2cepstrum and other male and female speakers. Then, it was shown that the rotation of Δ^2 was also very similar and that the rotation was very gender- or speaker-independent. These results claim that a sequence of Δc is very size- or age-dependent within a speaker.

2.4. Robust and structural invariance in speech

If matrix A is completely a rotation matrix, speaker-invariant features can be obtained as follows. A speech stream is converted into a sequence of N cepstrum vectors. If every distance is calculated between any pair of the N cepstrums, which provides an $N\times N$ distance matrix, the matrix is invariant. In the cepstral domain, difference of microphones or lines is represented as addition of another static vector b, $c'=c+b$. And it is very clear that the matrix is also invariant with any kind of b. It seems that the distance matrix can be a good candidate to the acoustic definition of word Gestalt but we have to note that matrix A is not completely a rotation matrix. So, the distance matrix is easily modified by difference of speakers.

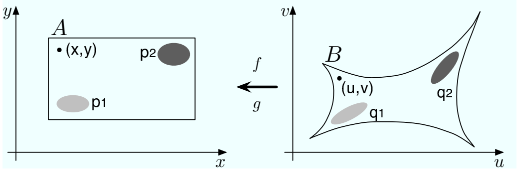

Figure 8: Linear and non-linear transform between two spaces

Is there a good method to make the distance matrix invariant? The answer is to calculate the matrix as *distribution-based* matrix. Figure 7 shows a timbre contour in a cepstrum space. In the figure, it is converted into a sequence of distributions, from which a distance matrix is extracted. It should be noted that distance is calculated also from *temporally distant* events. We can guarantee mathematically that this distance matrix is invariant with any kind of linear or non-linear transform function [16].

In Figure 8, there are two spaces, one of which is mapped into the other by a linear or non-linear transform. Point (x,y) in space A is mapped uniquely on (u,v) in space B, where $x=f(u,v)$ and $y=g(u,v)$. Using f and g, any integral operation in space A can be rewritten as its counterpart in B.

$$\begin{aligned}\iint \phi(x,y)dxdy &= \iint \phi(f(u,v),g(u,v))|J(u,v)|dudv \\ &= \iint \psi(u,v)dudv,\end{aligned}$$

where $J(u,v)$ is Jacobian. Then, acoustic event p_i in A, which is characterized as distribution, is mapped onto q_i in space B.

$$q_i(u,v) = p_i(f(u,v),g(u,v))|J(u,v)|,$$

We can show that Bhattacharyya distance between two distributions is invariant with any kind of linear or non-linear transform.

$$\begin{aligned}BD(p_1,p_2) &= -\log \oiint \sqrt{p_1(x,y)p_2(x,y)}dxdy \\ &= -\log \oiint \sqrt{p_1(f(u,v),g(u,v))|J|\cdot p_2(f(u,v),g(u,v))|J|}dudv \\ &= -\log \oiint \sqrt{q_1(u,v)q_2(u,v)}dudv = BD(q_1,q_2)\end{aligned}$$

The distribution-based distance matrix is invariant robustly. The shape of a triangle is determined uniquely if the length of all the segments is given. Similarly, the shape of an n point geometrical structure is determined uniquely if the length of all the ${}_n\mathrm{C}_2$ segments, including the diagonal ones, is given. This is why we call matrix-based representation as structural representation.

The mathematically same framework is used in quantum chemistry. Structural analysis of molecules is carried out using distribution-based distance matrices. Here, a distribution means an electron cloud but "$-\log$" operation is not done to calculate distance between two clouds. $\oint\sqrt{p_1(x)p_2(x)}dx$ is called overlap integral and the matrix is called overlap matrix, both of which are very fundamental measures of quantum chemistry. Many of chemical functions of matter are determined based on the structural and morphological features of its molecule.

2.5. Some experimental results

An utterance is converted into a distance matrix (a structure) and another utterance is done into another matrix, both of which are invariant with any kind of static non-linguistic factors. We already proposed a distance measure between two matrices [17] and, using it, structural speech recognition was examined experimentally [3, 4, 5], where static and absolute speech features

Table 1: Recognition rates of HMMs and the structural models

	HMM	Proposed
#speakers	4,130	8
word-based	97.4	98.3
vowel-based	98.8	99.3

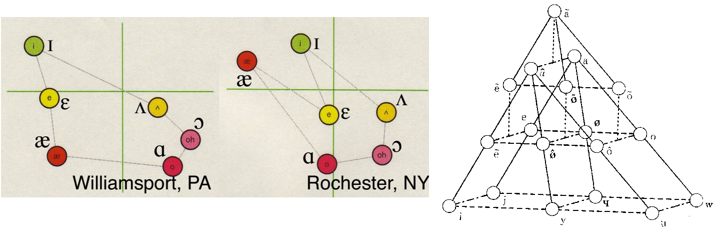

Figure 9: Two accented pronunciations of American English [19] and Jakobson's geometrical structure of the vowels [20]

were completely discarded. The task was isolated word recognition, where word was defined artificially as vowel sequence of $V_1V_2V_3V_4V_5$ ($V_i \neq V_j$) like /aeiou/. Since Japanese has only 5 vowels, the vocabulary size is 120. Since speaker differences are removed well, the matrix-based acoustic models were built by using only 4 male and 4 female speakers for training. Test utterances of $V_1V_2V_3V_4V_5$ were given by other 4 male and 4 female speakers. The total number of test utterances was 4,800.

The performances are shown in Table 1. For comparison, an isolated word recognizer was built using tied-state triphone HMMs, trained by 4,130 speakers using MFCC and its Δ [18]. The proposed framework showed almost the same performance. Strictly speaking, however, we have to note that the direct comparison is not fair because the proposed method was examined in a task-closed experiment and the HMMs were done in a task-open one. But we can say that the holistic and structural representation, which does not have any static and absolute speech features, has a very good function of identifying spoken words. Detailed descriptions on the experiments are found in [3, 4, 5].

3. Discussions and conclusions

In Figure 3, Arabic scale is shown. If a western music is performed with this scale, it will take on a very different color. This means that the sound arrangement pattern can easily change the color of music. This is the case with vowels. If the vowel arrangement pattern is changed, it will indicate a regionally accented pronunciation. Figure 9 shows two examples of the accented pronunciation of American English. The vowel arrangement pattern can easily change the color of pronunciation.

Suppose that the parents of identical twins get divorced immediately after birth. A twin is taken in by the mother and the other is by the father. What kind of pronunciation do they acquire ten years later? Do the twins produce mother-sounding and father-sounding voices? No way! They are not myna birds! But there is an exceptional case that the twins' pronunciations are very different. The case is that the parents are speakers of different regional accents. Timbre difference based on speakers does not affect the pronunciation but that based on regional accents affects it. Why? The simplest explanation is that infants don't learn the sounds as they are but infants learn the sound system embedded in spoken language. The proposed structural representation extracts the embedded invariant system in an utterance and we consider that this is the answer to the question.

Phonetics discusses the absolute values of language sounds. Phonology does their relative values and often focuses on *contrasts* in speech. Figure 9 shows Jakobson's structure of the French vowels. He was inspired by Saussure's claim that language is a system of conceptual differences and phonic differences. We consider that the proposed structural representation is a physical implementation of structural phonology and, in the representation, distant contrasts as well as local ones are considered. In a sense, we can say that phonetics looks at speech atoms and phonology looks at speech molecules.

The new representation was obtained by making a timbre contour of Figure 2 invariant. The pitch contour as in Figure 2 is one of the prosodic features. We consider that the proposed holistic and structural representation based on the timbre contour is yet another prosodic feature. This is because the representation is very supra-segmental and cannot identify any isolated sound or segment although it can identify a word.

In studies of speech recognition and speech perception [21], speech features are often divided into two kinds, static and dynamic. In this study, another criterion is given, which divides the features into local (atomic) and holistic (molecular or morphological). This division surely corresponds to phonetics and phonology and, as discussed in Section 2, we consider that this division is more valid linguistically and psychologically. The effective integration of both features is left as future work.

4. References

[1] N. Minematsu *et al.*, "Theorem of the invariant structure and its derivation of speech Gestalt," *Proc. SRIV*, 47-52, 2006.
[2] N. Minematsu, "Are learners myna birds to the averaged distributions of native speaker? – a note of warning from a serious speech engineer –," *CD-ROM of SLaTE*, 2007.
[3] S. Asakawa, *et al.*, "Automatic recognition of connected vowels only using speaker-invariant representation of speech dynamics," *Proc. InterSpeech*, 890-893, 2007.
[4] Y. Qiao *et al.*, "Random discriminant structure analysis for continous Japanese vowel recognition," *Proc. ASRU*, 576-581, 2007.
[5] S. Asakawa *et al.*, "Mult-stream parameterization for structural speech recognition," *Proc. ICASSP*, 2008 (to appear.)
[6] K. Johnson *et al.*, *Talker variability in speech processing,* Academic Press, 1997.
[7] http://tepia.or.jp/archive/12th/pdf/viavoice.pdf
[8] K. Miyamoto, *Making voices and watching voices,* Morikawa Pub., 1995.
[9] P. W. Jusczyk, *The discovery of spoken language,* The MIT Press, 2000.
[10] S. E. Shaywitz, *Overcoming dyslexia,* Random House, 2005.
[11] M. Kato, "Phonological development and its disorders," *J. Communication Disorders*, 20, 2, 84-85, 2003.
[12] N. Minematsu *et al.*, "Universal and invariant representation of speech," *CD-ROM of Int. Conf. Infant Study* (2006)
http://www.gavo.t.u-tokyo.ac.jp/~mine/paper/PDF/2006/ICIS_t2006-6.pdf
[13] K. Miyazaki, "How well do we understand absolute pitch?," *J. Acoust. Soc. Jpn.*, 60, 11, 682-688, 2004.
[14] M. Pitz *et al.*, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech and Audio Processing,* 13, 5, 930-944, 2005.
[15] D. Saito *et al.*, "Directional dependency of cepstrum on vocal tract length," *Proc. ICASSP*, 2008 (to appear.)
[16] N. Minematsu, *et al.*, "Linear and non-linear transformation invariant representation of information and its use for acoustic modeling of speech," *Proc. Spring Meeting Acoust. Soc. Jpn.*, 147-148, 2007.
[17] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," *Proc. ICASSP*, 889-892, 2005.
[18] T. Kawahara *et al.*, "Recent progress of open-source LVCSR engine Julius and Japanese model repository," *Proc. ICSLP*, 3069-3072, 2004.
[19] W. Labov *et al.*, *Atlas of North American English*, Walter De Gruyter, 2001.
[20] R. Jakobson *et al.*, *Notes on the French phonemic pattern*, Hunter, 1949.
[21] J. Ryalls, *A basic introduction to speech perception*, Singular, 1996.