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 sounds 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 timbre⁴. 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

---

⁴In musicology, timbre means the spectral envelope of a sound.
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 \( \Delta F_0 \) is defined as \( \Delta F_0 = F_{0k} - F_{0k-1} \), a sequence of \( \Delta F_0 \) 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 \( \Delta F_0 \) 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 tone contour speaker-independent?

Figure 2 shows a melody (pitch) contour of CDEF and a timbre contour of /aiueo/ with the Japanese vowel chart.

![Figure 2: Dynamic changes of pitch in CDEFG and those of timbre in /aiueo/ with the Japanese vowel chart](image)

<table>
<thead>
<tr>
<th>Male average</th>
<th>Female average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fo</td>
<td>2Nd Formant</td>
</tr>
<tr>
<td>1400</td>
<td>800</td>
</tr>
<tr>
<td>1000</td>
<td>2Nd Formant</td>
</tr>
<tr>
<td>1600</td>
<td>800</td>
</tr>
<tr>
<td>1800</td>
<td>2Nd Formant</td>
</tr>
</tbody>
</table>

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

![Figure 3: Three musical scales of Major, Minor, and Arabic](image)

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...
independent of speakers and phonemes. As shown in Figure 4, if two consecutive cepstrum vectors are rotated similarly, then, their \( \Delta c \) 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 /aiueo/ 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 \( \Delta s \) are very similar and that the rotation is vowel-independent. Similar analysis was done with \( \Delta c \) of other male and female speakers. Then, it was shown that the rotation of \( \Delta c \) was also very similar and that the rotation was gender- or speaker-independent. These results claim that a sequence of \( \Delta c \) 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 \( \hat{c} = c + b \). 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.

Is there a good method to make the distance matrix invar-
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 /aiou/. 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 $\Delta$ [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 phonetic 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 (static) 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


---

Table 1: Recognition rates of HMMs and the structural models

<table>
<thead>
<tr>
<th></th>
<th>HMM</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>#speakers</td>
<td>4,130</td>
<td>8</td>
</tr>
<tr>
<td>word-based</td>
<td>97.4</td>
<td>98.3</td>
</tr>
<tr>
<td>vowel-based</td>
<td>98.8</td>
<td>99.3</td>
</tr>
</tbody>
</table>