Formant-based articulatory normalization and its application to vowel restoration

Yuichi Ueda*, Kosuke Tominaga and Tadashi Sakata

*Corresponding author's address: Graduate School of Science and Technology, Kumamoto University, 2-39-1 Kurokami, Chuoh-ku, Kumamoto, 860-8555, Kumamoto, Japan, ueda@cs.kumamoto-u.ac.jp

Visual feedback of spontaneous speech is effective for articulatory training in deaf children and speech rehabilitation in patients with dysarthria. Particularly, visual representations of vowel formant frequencies have been used directly and indirectly for such purposes because these acoustical parameters reflect the articulatory behavior. However, since not only the shape but also the size of the vocal tract affects the formant frequencies, minimization of the effects due to differences in size of the vocal tract is required. In such speaker normalization, we define a color space consisting of three ratios of formant frequencies and apply it to the color visualization of vowels. In this paper, we propose normalized articulation space as an expansion of color space, where we assume that the neutral vowel of any speaker is mapped onto a unique point. In addition, since the proposed articulatory space is regarded as a speaker-independent representation of the vocal tract shape, we also propose a method for converting the modified articulation plane into formant space for the purpose of correcting degraded speech.

©2013 Acoustical Society of America [DOI: 10.1121/1.4800620]
Received 21 Jan 2013; published 2 Jun 2013
INTRODUCTION

In speech learning in deaf children and speech rehabilitation in patients with dysarthria, providing visual feedback of speech features is advantageous because it makes training more efficient, interesting, and easy to conduct. Particularly, formant frequencies have often been used as visual “formant graphs” in vowel articulation training. However, since those frequencies reflect not only the shape but also the size of the speaker’s vocal tract, it is difficult for users to derive articulatory states from formant information alone. Previously, we developed a speech visualization system [1][2] where the vowel formant frequencies were converted into the three primary colors and consonantal features were overlapped onto resulting color image with some texture. This tool allows users to observe articulatory states as visual images in real time. Therefore, visual representation is intuitive and easy for children to become familiar with.

However, since directly associating articulatory positions with corresponding vowel colors is difficult, it was considered that some articulatory information, such as the positions of the tongue or the jaw, should be added to the color image. Thus, we reconsidered formant-converted color signals and defined a pair of new variables (denoted as \( h \) and \( v \)), which served the purpose of speaker normalization and description of the articulatory position. This allowed us to evaluate the articulatory position of vowels in a speaker-independent \( h-v \) plane, as well as to modify the articulatory positions in the \( h-v \) plane, that is, correcting formant frequencies degraded by abnormal articulation. To do so, we needed to establish how to map the \( h-v \) plane onto formant space by considering speaker individuality.

In this background, we designed a prototype of an experimental system performing speech analysis, modification, and synthesis in real time, as shown in Figure 1. In this paper, we describe formant-based speaker normalization and its articulatory representation following the block diagram in Fig. 1. We also discuss the restoration of formants from the modified articulatory position for the purpose of restoring vowels.

FIGURE 1. Block diagram of the experimental system for vowel restoration based on the articulatory modification. The shaded area denotes the parts discussed in this paper.

FORMANT SPACE TO ARTICULATORY POSITION

Formant Distributions of Five Japanese Vowels

Speech materials

Speech materials were five Japanese vowels uttered multiple times in isolation. Speakers were divided into three groups: Groups M and F, consisting of 75 adult males and 75 adult females, respectively, and Group C, consisting of 80 children (40 boys and 40 girls) aged 7 to 11 years. Utterances in Groups M and F were selected from the ATR Japanese speech database, and those in Group C were recorded at a local elementary school. We randomly selected two utterances from each speaker and prepared 2300 speech files (130 speakers \( \times \) 5 vowels \( \times \) 2 times). However, utterances with lower SNR and/or with vague phonemes were eliminated, as a result of which the speech materials to be investigated consisted of 2252 files, where each set for Groups M, F, and C included 736, 716, and 800 files, respectively.

Formant frequencies and their distributions

Using an inverse filter control-based formant tracker [3], the four lowest formant frequencies (\( F_1 \) to \( F_4 \)) were extracted from each speech file by frame (10 ms), and then averaged over 5 frames in the central segment of
each formant’s trajectory. Consequently, we obtained a formant data set of 2252 data \((F_1,F_2,F_3)\). Figure 2(a) shows the distribution of the formant set in \(F_1-F_2-F_3\) space. It is well known that the radial spread of each vowel group is mainly due to differences in vocal tract length between speakers. Particularly, such speaker-dependence appears strongly in the distributions in the \(F_1-F_2\) plane shown in Fig. 2(b) and (c), which generally reflects the relative positions of tongue and jaw.

**FIGURE 2.** Formant distributions of five Japanese vowels uttered by three groups of Japanese speakers: M (male), F (female), and C (child). The ellipses indicate \(2\sigma\)-distributions in principal component analysis using formant data in each group.

### Color Representation Using Normalized Formant Frequencies

**Formant conversion into color space**

We proposed a color visualization method [1] of vowels which was used in an articulation training tool for deaf children. In this visualization method, a formant point \((F_1,F_2,F_3)\) is converted into a color \((R,G,B)\) by using the three primary colors in accordance with the following equation:

\[
(R,G,B) = (5F_1/F_3, 3F_1/5F_2, F_2/3F_3),
\]

where the coefficients \(5\), \(3/5\), and \(1/3\) were chosen so that a neutral vowel with a formant ratio \(F_1:F_2:F_3 = 1:3:5\) would be colorless. Figure 3(a) shows the color distributions of all formant points in Fig. 2(a). The differences between the three speaker groups in RGB space are reduced in comparison with Fig. 2(a). Fig. 3(b) shows the color display of the speech visualization system [2], where individual vowels are perceived as particular colors.

**FIGURE 3.** Color distribution corresponding to five Japanese vowels for the same data as in Fig. 2. Ellipses in (a) indicate \(2\sigma\)-distributions of the color data in the three groups (M, F, and C) for each vowel. Color patterns in (b) flow from right to left on the display in real time.
Speaker normalization in color representation

It can be seen from Eq. (1) that formant points with the same formant ratios are mapped onto a particular color point that is located on the curved surface $RGB=1$ in Fig. 3(a). According to the acoustic tube model, the cyclic formant ratios in Eq. (1) have an effect of normalizing differences in vocal tract length. Using such formant ratios, Sussman proposed a psycho-auditory model of speaker normalization in vowel perception [4]. Ladefoged used those ratios to characterize vocal tract shapes in vowel articulation [5]. On the other hand, assimilation in single-color areas and contrast effect between adjacent colors can be expected in the vowel color image shown in Fig. 3(b). Therefore, the color representation of vowels also has a visual advantage in terms of the normalization effect [1][2].

Formant-based Articulatory Normalization and Visualization

Articulatory distribution

Based on the relation between the $F_1$-$F_2$ plane and the vowel articulatory positions, we defined a pair of variables $(h$ and $v)$ representing the normalized articulatory position as follows:

$$ (h,v) = (G^{-1}, R) = (5F_2 / 3F_3, 5F_1 / F_3), $$

(2)

where $h$ and $v$ represent the relative horizontal and vertical positions of the tongue body, respectively. As seen from Eqs. (1) and (2), since the variables $h$ and $v$ are derived from RGB space, the performance of speaker normalization in the $h$-$v$ plane is the same as that in RGB space. Figure 4(a) shows the distributions of articulatory positions as well as the $2\sigma$ ellipses and mean positions for each group on the $h$-$v$ plane. From these figures, we can see that the conventional formant graph (the $F_1$-$F_2$ plane) is normalized over the speaker groups.

Application of the normalized articulatory plane

Fig.4(b) shows the developed vowel articulation training tool using the normalized and colored formant graph [6]. Users produce vowel using real-time visual feedback which indicates the articulatory position in the $h$-$v$ plane.

![Vowel distributions in the normalized articulatory plane](image1)

![Vowel articulation training tool using the $h$-$v$ plane](image2)

**FIGURE 4.** Articulatory distributions of the five Japanese vowels (a) and its application tool (b). Ellipses in (a) indicate $2\sigma$-distributions in the $h$-$v$ plane for each vowel and group. Ellipses in (b) correspond to M, F and C and are colored at the corresponding points. The character size of the Japanese hiragana is inversely proportional to the distance from each vowel center.

ARTICULATORY POSITION TO FORMANT SPACE

Mapping Articulatory Position into Formant Frequencies

In this section, we consider an inverse mapping method from the $h$-$v$ plane to $F_1$-$F_2$-$F_3$ space in the formant-based speech modification as shown in Fig. 1.

**Principle of formant restoration**

Since the mapping from the $h$-$v$ plane to formant space is of the one-to-many variety, we must select the most reliable estimates $(F'_1,F'_2,F'_3)$, which are derived from the original $(F_1,F_2,F_3)$ as follows:
Here, $k$ in $c=500k$ is a scale factor that is taken as a function of articulatory point $(h,v)$ as follows:

$$k = g(h,v).$$

For instance, when a male with a vocal tract length of 17.5 cm utters a neutral vowel, the restored formant frequencies are $(F'_1,F'_2,F'_3) = (5000,1500,2500)$ because $k=g(1,1)=1.0$. Consequently, a particular position $(h,v)$ represents a speaker-independent articulatory state, and the scale factor $g(h,v)$ defines a speaker-dependent function.

### Restoration errors in formant mapping

Since $g(h,v)$ is a speaker-dependent function, we need to estimate it using multiple formant points of various vowel sounds uttered by the target speaker. Given the function $g(h,v)$ and the original formant point, we can calculate the restored point $(F'_1,F'_2,F'_3)$ from Eq. (3), after which we can evaluate the restoration error rates $e_n$:

$$e_n = |F' - F| / |F| = |F' - 1|; \quad n = 1, 2, 3$$

From the following relations (Eq. (6)), the error rates $e_1$, $e_2$, and $e_3$ are equal to each other and determine the accuracy of the restoration function $g(h,v)$:

$$F'_1 / F_1 = F'_2 / F_2 = 5x / F_3 = 2500k / F_3 = F'_3 / F_3 = k / K,$$

where $K$ is defined as $F_3/2500$. For the function $g(h,v)$, we assumed a polynomial expression with variables $h$ and $v$ and evaluated the relation between the order and the errors in Eq. (5). As a result of evaluating connected vowels uttered by a male and a female, the cubic polynomial was optimized in minimizing the errors.

### Example of formant restoration using a restoring function

Figure 5 presents an example of the results of closed evaluation using optionally connected vowels uttered by a male speaker, where (a) shows the extracted and restored formant trajectories overlapped on the sound spectrogram, (b) shows a scatter diagram of all articulatory points in the $h$-$v$ plane, where the colored contours illustrate the values of the estimated $g(h,v)$, and (c) shows the restoration accuracy in the formant mapping. Although this result indicates the closed evaluation using a male voice, the restoration error rates of formant restoration are within about 10%.

### Restoration and modification in formant mapping

Using two types of restoring functions for male (M) and female (F) speakers, we carried out closed/open evaluations for formant restoration of connected vowels (/aeiou/). Figure 6 presents the results, where (a) shows the extracted (original) and restored formant trajectories, which are plotted in (c) as M to M (closed), (b) shows the articulatory trajectories in the $h$-$v$ plane, and (d) shows the modified formant trajectories which were obtained by modifying the original trajectories in $h$-$v$ space for male speakers and the formant trajectories restored using the restoring function $g(h,v)$ for female speakers. Particularly, in the case of (d), M to F means an example of speaker conversion in the open test of formant restoration.

![Image](image_url)

**FIGURE 5.** Example of speech material to be used in generating the restoring function (connected vowels /aoiueo.../ uttered by a male) and individual parameters. Contours and colors in (b) indicate the estimated function $g(h,v)$. 

---

### Mathematical Formulas

1. **Formant Trajectories Calculation**

   

   $$(F'_1,F'_2) = (vF'_3/5, 3hF'_3/5)$$

   

   $$(F'_1,F'_2) = (vF'_3/5, 3hF'_3/5, F'_3) = (v; 3h; 5c) = (500v/k, 1500h/k, 2500k)$$

   

   Here, $k$ in $c=500k$ is a scale factor that is taken as a function of articulatory point $(h,v)$ as follows:

   $$k = g(h,v).$$

2. **Error Rate Calculation**

   

   $$e_n = |F' - F| / |F| = |F' - 1|; \quad n = 1, 2, 3$$

3. **Restoration Accuracy**

   

   $$F'_1 / F_1 = F'_2 / F_2 = 5x / F_3 = 2500k / F_3 = F'_3 / F_3 = k / K,$$
Articulatory Modification and Its Application

As a preliminary simulation of the articulatory modification and formant restoration, we applied the following two kinds of simple transformations to the original articulatory position \((h, v)\):

\[
(h', v') = (r(h - 1), r(v - 1))
\]

\[
(h', v') = (h' \cos \theta + v' \sin \theta + 1, h' \sin \theta - v' \cos \theta + 1)
\]

where \(r\) is a multiplier for controlling the articulatory degree and \(\theta\) is a rotation angle for modifying the phoneme. In the above transformations, we assign a neutral point \((h, v) = (1, 1)\) as the origin of the transformation, which denotes the unique articulation of the ideal neutral vowel. In Fig. 6 (b), the transformed trajectories in the \(h-v\) plane are plotted for the case of \(r = 1.05\) and \(\theta = 30^\circ\), and the modified formant trajectories are shown in Fig. 6(e).

We intend to use the above articulatory modification and formant restoration method for normalization of degraded speech in dysarthric patients. In such application, the complexity of transformations for modifying the articulations is expected to increase in a nonlinear manner since they must be determined individually. In future work, we plan to investigate how to determine such transformations for individual patients.

**FIGURE 6.** Results for formant frequencies of male voice: (a) Original spectrogram and extracted formants; (b) Trajectories of the original and modified data in the \(h-v\) plane; (c) Restoration of original formants: \(M \to M\); (d) Conversion to female formants: \(M \to F\); (e) Modification of original formants: \(M \to M'\).

**CONCLUSION**

In this paper, we proposed a formant-based articulatory modification method for restoring speech or converting between different types of speaker formants. In this method, the formant space is divided into a speaker-independent articulatory plane and a speaker-dependent parametric space. Although the formant restoration accuracy depends on estimating one of the restoring functions and the method of generating the modification rules, we can apply the modified formant frequencies in a formant-based speech synthesizer, as shown in Fig. 1. In future work, we need to determine and select the modification rules for individuals.

**REFERENCES**