A Phonological Control Method on A Speech Compensation System for Dysarthria Using A Standardized Space

Yukinori Hetsugi, Tadashi Sakata, Yuichi Ueda

Abstract: We have developed a speech compensation system for dysarthria. The system aims at improving the phonological properties of vowels without losing speaker individuality. We propose a method for phonological control of vowels using a standardized space to control vowels in the normalized articulation space, normalized for speaker individuality. The method maps an original dysarthric speaker’s normalized articulation space to a standardized space, then from the standardized space to the target speaker’s normalized articulation space assuming normality to improve the phonological properties of vowels. We confirm phonological control of vowels by performing a processing simulation, comparison different target speakers and a processing simulation using a dummy original speaker as a dysarthria.

Keywords: Vowel Formant, Articulation Space, Phonological Control, Vowel Normalization, Dysarthria

I. INTRODUCTION

Speech is one of the important means of communication in daily life, and deterioration of the phonological features of speech due to dysfunction of the speech organs can make communication difficult. To assist people with speech impairments, we have been developing a speech compensation system. In a previous study, we proposed a method for normalizing degraded formant frequencies in an articulatory space called the “hv-plane,” which normalizes speaker differences in vocal tract length [1]. We also proposed a function for inverse mapping (formant restoration) from the hv-plane to a unique coordinate point in the actual formant space [2]. If the speaker from which the formant restoration function is created and the speaker to be restored are the same, formant restoration can be performed with an average restoration error rate of about 5%, regardless of speaker or context [3]. The hv-plane can normalize speaker individualities from the actual formant space and can separate phonological features, suggesting that the formant restoration function includes speaker individualities such as vocal tract length and shape [4].

In this paper, we propose a vowel phonological control method for speech compensation systems. Dysarthric speakers have different vowel hv distributions depending on the type and severity of the disorder. Various vowel hv distributions for each dysarthric speaker are standardized using a standardized space, and vowel phonological control is performed to obtain a normal vowel hv distribution.

II. SPEECH COMPENSATION SYSTEM

A. System overview

A speech compensation system improves the phonological features of vowels in impaired speech without losing the individuality features of speakers [5]. As Figure 1 shows, this system consists of three blocks: an analysis block, a normalization block, and a synthesis block.

An analysis block analyzes the acoustic features of input dysarthric speech. It estimates formant frequencies \((F_1,F_2,F_3)\) by an inverse filter control (IFC) method [6], estimates the fundamental frequency \((F_0)\) by autocorrelation function peak picking of a vocal fold source differential wave obtained by passing input speech through an inverse filter composed of estimated formant frequencies, and extracts the root mean square (RMS) from input speech.

A normalization block normalizes degraded vowel formant frequencies using features extracted by the analysis block. It first performs mapping from an actual formant space to a normalized articulation space that normalizes differences in vocal tract length for each speaker. This normalized articulation space is the hv-
plane, and we call mapping to the $hv$-plane “$hv$ conversion.” This block next performs normalization processing on the $hv$-plane, then finally performs inverse mapping from the $hv$-plane to the actual formant space (formant restoration). In this paper, we propose a vowel phonological control method using a standardized space as a normalization process.

A synthesis block performs speech synthesis by formant analysis and synthesis using the formant frequencies normalized by the normalization block.

**B. $hv$ conversion**

To normalize differences in speaker vocal tract lengths from the actual formant space consisting of the first to third formant frequencies, variables $h$ and $v$ are defined as

$$h = \frac{5F_2}{3F_3}, \quad v = \frac{5F_1}{F_3},$$

where $h$ (horizontal) is a parameter corresponding to the front back position of the tongue and $v$ (vertical) corresponds to opening of the chin. $(h,v)=(1,1)$ indicates the articulation position of the neutral vowel $(F_1:F_2:F_3=1:3:5)$ in the actual formant space. Vowel colorization has been proposed as a method for normalizing speaker individuality in a formant space [7]. In this method, the first to third formant cyclic ratios correspond to the primary color signals RGB, with vowels colorized by mapping from an actual formant space to RGB space as

$$(R,G,B) = \left(\frac{5F_1}{F_3}, \frac{3F_3}{5F_2}, \frac{F_2}{3F_1}\right),$$

thereby rendering neutral vowels colorless.

In the speech compensation system, the $hv$-plane is used as a space in which the articulation state can be expressed in addition to the speaker normalization ability of the RGB space.

**C. Formant restoration**

Articulation points $(h,v)$ in the $hv$-plane have a one-to-many relationship with sample points in the actual formant space $(F_1,F_2,F_3)$ and cannot be uniquely inverse-mapped to the actual formant space. If an arbitrary articulation point in the $hv$-plane is $(h',v')$ and the formant space point obtained by inverse mapping is $(F'_1,F'_2,F'_3)$, it can be expressed as a modification of (2) as

$$F'_1 = \frac{v'F_3}{5}, \quad F'_2 = \frac{3h'F_3}{5}.$$  

When (3) is applied to the relation of the formant ratio,
By providing a constant (c), it is possible to inversely map from the hv-plane to a unique point in the actual formant space, as

\[
(F'_1, F'_2, F'_3) = (cv', 3ch', 5c).
\]

(5)

If \((h',v') = (1,1)\) and \(c=500\) in (5), the formant frequency \((F'_1,F'_2,F'_3) = (500,1500,2500)\) is the neutral vowel of an adult male with a vocal tract length of 17.5 cm. For the articulation shape corresponding to articulation point \((h',v')\), the formant ratio of each speaker is uniquely determined by (4), and the absolute value of the formant is determined by (5) with a constant \(c\) depending on the vocal tract length. In (5), the constant \(c\) gives the value of F03. If we define a function \(g(h',v')\) with \((h',v')\) as a variable and set \(c=500g(h',v')\), we can express (5) as

\[
(F'_1, F'_2, F'_3) = (500gv', 1500gh', 2500g).
\]

(6)

(6) can uniquely recover the formant frequencies \((F'_1,F'_2,F'_3)\) from any articulation point \((h',v')\). The function \(g(h',v')\) is called the formant restoration function, and inverse mapping from the hv-plane to the formant space is called formant restoration. \(g(h,v)\) is defined for each speaker as a three-dimensional surface on the hv-plane as in (7), and coefficients \(a_0, a_1, ..., a_9\) for each speaker are estimated by multiple regression analysis.

\[
g(h,v) = \frac{F_3}{2500} = a_0 + a_1h + a_2v + a_3h^2 + a_4v^2 + a_5hv + a_6h^3 + a_7v^3 + a_8hv^2 + a_9hv^2
\]

(7)

III. VOWEL PHONOLOGICAL CONTROL METHOD THROUGH A STANDARDIZED SPACE

A. Method overview

To improve the phonological properties of vowels, in this method the hv-plane of the impaired speech (the original speaker’s hv-plane) is mapped to the standardized space and the hv-plane of the target normal speech (the target speaker’s hv-plane). Dysarthric speakers have unique vowel hv distributions depending on the type and severity of the disorder. Through the standardized space, it is possible to normalize differences in the vowel hv distribution that vary depending on the dysarthric speaker. Therefore, even when the target dysarthric speaker changes (when the original speaker’s hv-plane changes), the same normalization processing can be performed.

B. Mapping from the original speaker’s hv-plane to the standardized space

From the hv distribution of dysarthric speech, five basic points \((h_0,v_0)\) corresponding to the gravity center of the five Japanese vowels “a, i, u, e, o” and a gravity center \((h_5,v_5)\) of the five basic points are calculated. The declination of \((h_5,v_5)\) when \((h_0,v_0)\) is the origin point is calculated as

\[
\theta_k = tan^{-1} \left( \frac{v_k - v_o}{h_k - h_o} \right) \quad k = a,i,u,e,o.
\]

(8)

As Figure 2(a) shows, \((h_k,v_k)\) is rotated using (9) so that the declination \(\theta_k\) of \((h_0,v_0)\) is 0 rad with origin \((h_5,v_5)\). The rotation coordinates \((h_0,v_0)\) of the point are

\[
h_{tk} = (h_k - h_o)cos(-\theta_a) - (v_k - v_o)sin(-\theta_a)
\]

\[
v_{tk} = (h_k - h_o)sin(-\theta_a) - (v_k - v_o)cos(-\theta_a)
\]

(9)

The declination \(\theta_k\) of \((h_k,v_k)\) is calculated as

\[
\theta_{tk} = tan^{-1} \left( \frac{v_{tk}}{h_{tk}} \right) \quad k = a,i,u,e,o.
\]

(10)

From the above, \(\theta_a\) is 0 rad.

Consider the case where an arbitrary point \(X(h_a,v_a)\) is input. Calculate the rotation coordinate \(X(h_{ta},v_{ta})\) and declination \(\theta_{ta}\) by setting \(k = x\) in (9) and (10). As Table I shows, segments seg1 to seg5 are set from \(\theta_a\) to \(\theta_o\) according to the argument range. seg1 to seg5 determine which segment \(\theta_a\) belongs to, and set the two

| segm | \(\theta_a\) range | \((h_1,v_1)\) | \((h_5,v_5)\) |
|------|-----------------|--------------|--------------|
| m=1  | \(0 < \theta_a < \theta_0\) | \((h_0,v_0)\) | \((h_0,v_0)\) |
| m=2  | \(0_0 < \theta_a < \theta_0\) | \((h_0,v_0)\) | \((h_0,v_0)\) |
| m=3  | \(0_0 < \theta_a < \theta_0\) | \((h_0,v_0)\) | \((h_0,v_0)\) |
| m=4  | \(0_0 < \theta_a < \theta_0\) | \((h_0,v_0)\) | \((h_0,v_0)\) |
| m=5  | \(0_0 < \theta_a < 2\pi\) | \((h_0,v_0)\) | \((h_0,v_0)\) |
Rotation of original speaker’s $hv$-plane

Calculation of radial ($r_n$) and declination ($\theta_n$)

Fig. 2. The standardized space

Basic points ($h_1,v_1$) and ($h_2,v_2$) corresponding to the segment to which they belong. Point $P(h_p,v_p)$ lies on a straight line passing through basic points ($h_1,v_1$) and ($h_2,v_2$) and intersecting a straight line passing through the origin point and point $X$. Point $P$ transitions linearly between two basic points, and its position can be represented as a ratio $\beta$ from 0 to 1, which is based on ($h_1,v_1$) as

$$\beta = \frac{h_p - h_1}{h_2 - h_1}. \quad (11)$$

As Figure 2(b) shows, the radials of points $X_t$ and $P$ are $r_{tx}$ and $r_p$, and radials $r_n$ and declination $\theta_n$ mapped to the standardized space can be calculated as

$$r_n = \frac{r_{tx}}{r_p}, \quad \theta_n = c_1 + c_2\theta_{tx}. \quad (12)$$

$\theta_n$ in (12) is an interpolation function for standardizing the argument. If the declinations of basic points ($h_1,v_1$) and ($h_2,v_2$) are $\theta_m$ and $\theta_{m+1}$, coefficients $c_1$ and $c_2$ in (12) are

$$c_1 = \theta_m - c_2 \cdot \frac{2\pi m}{5}, \quad c_2 = \frac{5(\theta_{m+1} - \theta_m)}{2\pi}. \quad (13)$$

Assuming that a point in the standardized space represented by radius $r_n$ and declination $\theta_n$ is $X'(h'_x,v'_x)$, the coordinates are

$$h'_x = r_n \cos(\theta_n), \quad v'_x = r_n \sin(\theta_n). \quad (14)$$

C. Mapping from standardized space to target speaker’s $hv$-plane

By mapping from the original speaker’s $hv$-plane to the standardized space, the declination and radius are standardized when the gravity centers of the basic points for the five vowels differ for each speaker. Next, we map from the standardized space to the target speaker’s $hv$-plane with normal phonological properties. For the point $X'(h'_x,v'_x)$ mapped from the original speaker’s $hv$-plane to the standardized space, the declination $\Delta_s$ to the corresponding basic point ($h_1,v_1$) is obtained. The ratio $\alpha$ of $\Delta_s$ to declination of two basic points ($h_1,v_1$) and ($h_2,v_2$) is calculated as

$$\alpha = \frac{5\Delta_s}{2\pi}. \quad (15)$$

In the target speaker’s $hv$-plane, using two basic points ($h_{t1},v_{t1}$), ($h_{t2},v_{t2}$) and ratio $\alpha$ corresponding to the segment of point $X'$ in the standardized space, a point ($h_s,v_s$) on a straight line connecting the two basic points is calculated as

$$h_s = h_{t1} + \alpha(h_{t2} - h_{t1}), \quad v_s = v_{t1} + \alpha(v_{t2} - v_{t1}). \quad (16)$$

($h_s,v_s$) in (16) corresponds to the declination in the target speaker’s $hv$-plane. Using radius $r_n$ of point $X'$ in
the standardized space, \((h', v')\) of the target speaker’s \(hv\)-plane corresponding to point \(X'\) are obtained as
\[
\begin{align*}
  h' &= r_n \times h_s, \\
  v' &= r_n \times v_s.
\end{align*}
\]  

(17)

D. Processing under proposed method

Figure 3 shows an example of processing using the vowel phonological control method through the normalized space. Figure 3(a), (b) show the original speaker’s formant frequencies and \(hv\)-plane, (c) shows the standardized space, (d) and (e) show the target speaker’s \(hv\)-plane and restored formant frequencies.

An appropriate male speaker is set as the original speaker, and a female speaker F5 of the phoneme balance words (ATR 216 words \([8]\)) is set as the target speaker F5-ATR. The squares in Figure 3 represent the gravity center position of each vowel. For target speaker \(hv\)-planes, an area limitation process is performed to prevent mapping to unusual points such as extreme articulation states in the restoration to the actual formant space. Namely, if mapping is performed outside the bold line, we remap to the point on the bold line at the intersection when the mapping point and the origin are connected by a straight line. The broken line connecting the dots in Figure 3 is an example of the
normalization process in which the \( tv \) transition obtained by \( tv \) transformation of the formant trajectory of the continuous Japanese vowels /ieuau/ is input to the original speaker’s \( tv \)-plane. In the original speaker’s \( tv \)-plane, the standardized space, and the target speaker’s \( tv \)-plane, each vowel centroid point has a corresponding relation. Therefore, \( tv \) transition on the original speaker’s \( tv \)-plane also continuously transitions in other spaces.

Figure 4 compares \( tv \) trajectories on the target speaker’s \( tv \)-plane when various speakers are set as the target speakers. The same input and original speaker’s \( tv \)-plane as in Figure 3 were used. Figure 4(a), (b) show the \( tv \) transition results when a female speaker F3 of ATR 216 words is set as the target speaker. Figure 4(c) and (d) shows the results when a male speaker M6 is set. For each speaker, the gravity center of the five vowels represented by squares has a slightly different distribution and the restricted area represented by the bold line has a unique range. It can be seen that in each target speaker’s \( tv \)-plane, the \( tv \) transition of the original speaker’s \( tv \)-plane can be converted according to the positional relationship between the five vowel centroids.

Figure 5 shows a processing simulation using a dummy speaker as a original speaker. A dummy speaker imitates a dysarthria. Figure 5 confirm the effect of normalization even in speakers with disabilities. These results suggest that when mapping from the original speaker’s \( tv \)-plane to the target speaker’s \( tv \)-plane through the standardized space, the vowel phonology can be controlled by setting the target speaker.

IV. CONCLUSION

We proposed a vowel phonological control method for a speech compensation system for dysarthria using a standardized space. In this method, different \( tv \) distributions for each dysarthria can be normalized by mapping from the original speaker’s \( tv \)-plane to the standardized space, so that even if the original speaker changes, as long as we can obtain the gravity center of the speaker’s five vowels, it is possible to map from the original speaker’s \( tv \)-plane to the target speaker’s \( tv \)-plane in the same processing. Using the proposed method, we set the original speaker and the target speaker and performed a processing simulation, thereby confirming correspondence of each space from the \( tv \) transition of the original speaker’s \( tv \)-plane, the standardized space, and the target speaker’s \( tv \)-plane.

The results suggest that vowel phonology can be controlled by setting an arbitrary target speaker. Topics for future study include implementation of speech synthesis using the restored formant frequency, as well as verification and improvement of phonological improvement effects through listening experiments.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 17K01568 and a donation by Research & Consulting of Regional Science Co., Ltd.

REFERENCES

[1] W. Shibata, K. Tominaga, T. Sakata, and Y. Ueda, “A Study on Articulation Deformation Method for Speech Compensation System,” in Record of Joint Conference of Electrical and Electronics Engineers in Kyushu (CD-ROM), Sep 2011, pp. 331–332.
[2] Y. Ono, T. Sakata, and Y. Ueda, “A Formant Restoration from an Articulation Space for Normalization of the Dysarthria Speech Sounds,” in Record of Joint Conference of Electrical and Electronics Engineers in Kyushu (CD-ROM), Sep 2013, pp. 1–2.
[3] Y. Hetsugi, Y. Ono, T. Sakata, and Y. Ueda, “A Study on Speaker Dependence in Formant Restoration from Normalized Articulation Space,” in Reports of the 11th Meeting for Students Kyushu Chapter of the Acoustical Society of Japan. The Acoustical Society of Japan, 2015, pp. 19–22.
[4] Y. Hetsugi, T. Sakata, and Y. Ueda, “Separation of Individual and Phonemic Features in a Vowel Formant Space,” in Reports of the 2018 Autumn Meeting the Acoustical Society of Japan(CD-ROM). The Acoustical Society of Japan, Sep 2018, pp. 1339–1340.
[5] Y. Ueda, K. Tominaga, and T. Sakata, “Formant-based Articulatory Normalization and Its Application to Vowel Restoration,” in Proceedings of Meetings on Acoustics, vol. 19. International Conference of Acoustics, May 2013, p. 060282.
[6] A. Watanabe, “Formant Estimation Method Using Inverse-Filter Control,” IEEE Transactions on Speech and Audio Processing, vol. 9, pp. 317–326, May 2001.
[7] A. Watanabe, “Color Display System for Connected Speech to be Used for the Hearing Impaired,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 33, pp. 164–173, Feb 1985.
[8] K. Takeda, Y. Sagisaka, S. Katagiri, and H. Kuwabara, “A Japanese Speech Database for Various Kinds of Research Purposes.” The Journal of the Acoustical Society of Japan, vol. 44, no. 10, pp. 747–754, 1988.