INTERSPEECH
2009 BRIGHTON

Speech Generation from Hand Gestures Based on Space Mapping

Aki Kunikoshi, Yu Qiao, Nobuaki Minematsu, Keikichi Hirose

The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{kunikoshi, qiao, mine, hirose}@gavo.t.u-tokyo.ac.jp

Abstract

Individuals with speaking disabilities, particularly people suffering from dysarthria, often use a TTS synthesizer for speech communication. Since users always have to type sound symbols and the synthesizer reads them out in a monotonous style, the use of the current synthesizers usually renders real-time operation and lively communication difficult. This is why dysarthric users often fail to control the flow of conversation. In this paper, we propose a novel speech generation framework which makes use of hand gestures as input. People usually use tongue gesture transitions for speech generation but we develop a special glove, by wearing which, speech sounds are generated from hand gesture transitions. For development, GMM-based voice conversion techniques (mapping techniques) are applied to estimate a mapping function between a space of hand gestures and another space of speech sounds. In this paper, as an initial trial, a mapping between hand gestures and Japanese vowel sounds is estimated so that topological features of the selected gestures in a feature space and those of the five Japanese vowels in a cepstrum space are equalized. Experiments show that the special glove can generate good Japanese vowel transitions with voluntary control of duration and articulation.

Index Terms: Dysarthria, speech production, hand motions, media conversion, arrangement of gestures and vowels

1. Introduction

For linguistic communication, dysarthrics have to convey their messages to others without using tongue gestures. With sign or written language, non-oral communication is possible but it requires receivers with special sign language skills or the ability of reading and writing.

We can find several technical products to support speech communication of dysarthrics, which require no special skills on the each of the receivers. Both *Say-it!* [1] and *Voice aids* [2] are portable PC based products, with which users can generate speech by touching sound symbols or word symbols. As told above, however, these products often restrict the freedom of conversation and dysarthrics are likely to lose the initiative in conversation [3]. In [4], a dysarthric engineer developed a unique speech generator by using a pen tablet. The F1-F2 plane is embedded in the tablet. The pen position controls F1 and F2 of vowel sounds and the pen pressure controls their energy. Using this machine, he demonstrated with his fingers (not his tongue) that he could generate vowel sound transitions livelily. In his talk, he said that he disliked the current TTS systems because they forced him to speak always in a monotonous style.

Another example of speech generation from body motions is [5]. With two data gloves and some additional wearing devices, body motions are transformed into parameters required to drive a formant speech synthesizer. But this system was developped for human-computer interaction, not for the handicapped.

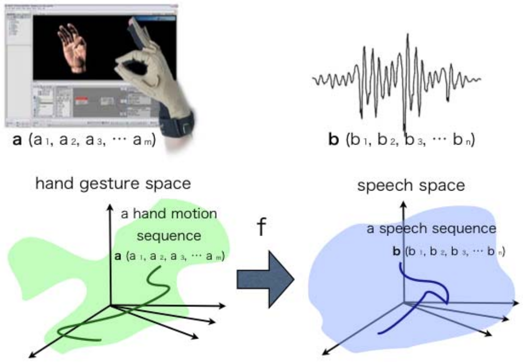


Figure 1: Media conversion based on space mapping

Recently, GMM-based speaker conversion techniques have been intensively studied, where voice spaces of two speakers are mapped to each other and the mapping function is estimated based on GMM [6, 7]. This technique was directly and successfully applied to estimate a mapping function between a space of tongue gestures and another of speech sounds [8]. In this study, GMM was used to map two spaces of different media. This result naturally lets us expect that a mapping function between hand gestures and speech can be estimated well.

People usually use tongue gesture transitions to generate a speech stream. But [4] and [5] showed that tongue gestures, which are *inherently* mapped to speech sounds, are not always required to speak. What is needed is a voluntarily movable part of the body whose gestures can be *technically* mapped to speech sounds. In [4] and [5], however, classical synthesizers were used, i.e. formant synthesizers. Partly inspired by the remarkable progress of voice conversion techniques and voice morphing techniques [9] in this decade, we develop a GMM-based hand-to-speech converter in this paper.

In the following sections, the development is described in detail. As an initial trial of hand-to-speech conversion, however, we only focus on Japanese vowel sounds. The most important issue at this point is how to design the optimal correspondence between vowels and hand gestures.

2. GMM-based media conversion

2.1. Estimation of a mapping function between two spaces

Fig.1 shows media conversion based on space mapping, where hand gesture vector $\boldsymbol{a}$ is converted into speech cepstrum vector $\boldsymbol{b}$. The mapping function f in Fig.1 can be estimated by the method proposed in [7]. For an aligned data set between a hand gesture stream and a speech stream, augment vector $\boldsymbol{z}$=[$\boldsymbol{a}$, $\boldsymbol{b}$] is formed. Then, the distribution of $\boldsymbol{z}$ is modeled by GMM, $p(\boldsymbol{z}) = \sum_{i=1}^{M} \omega_i \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes normal distribution of mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$ as below,

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^A \\ \boldsymbol{\mu}_i^B \end{bmatrix}, \; \boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{AA} & \boldsymbol{\Sigma}_i^{AB} \\ \boldsymbol{\Sigma}_i^{BA} & \boldsymbol{\Sigma}_i^{BB} \end{bmatrix}, \qquad (1)$$

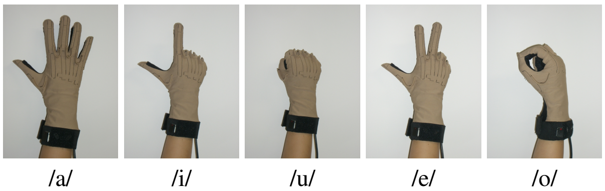


Figure 2: Gestures of the Japanese five vowels

M is the number of mixtures and ω_i is a weight for mixture i, here $\sum_{i=1}^{M} \omega_i = 1$ and $\omega_i \geq 0$.

The regression function $f(\boldsymbol{a})$ is obtained by using the above parameters and it approximates $\boldsymbol{b}$.

$$f(\boldsymbol{a}) = \sum_{i=1}^{M} p(i|\boldsymbol{a})[\boldsymbol{\mu}_i^B + \boldsymbol{\Sigma}_i^{BA}\boldsymbol{\Sigma}_i^{AA-1}(\boldsymbol{a} - \boldsymbol{\mu}_i^A)], \qquad (2)$$

where $p(i|\boldsymbol{a})$ is a posterior probability of i th GMM given $\boldsymbol{a}$.

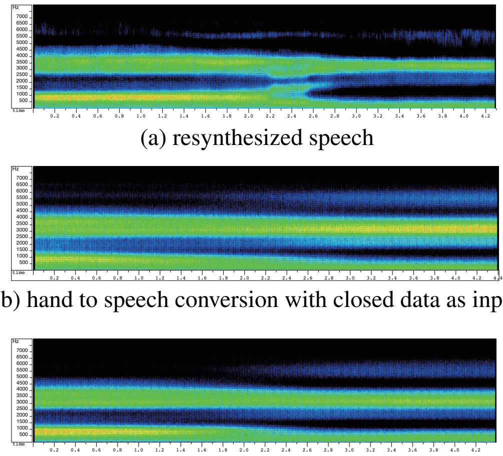
2.2. A preliminary experiment

As a preliminary experiment, hand to speech conversion was implemented for vowel transitions such as /ai/ and /oe/. The correspondence between the Japanese five vowels and hand gestures was shown in Fig.2. These gestures were determined so that a transition between any pair of vowels would not generate a third vowel. For training GMMs, a female adult recorded gesture data for the isolated vowels and ${}_5P_2$=20 transitions between every two vowels using CyberGlove made by Immersion Inc,. CyberGlove has 18 sensors and its sampling period is 10-20 ms.[1] Every gesture was recorded three times. The total number of gestures was (5+20)×3=75. In addition, a male adult speaker recorded speech for the five vowels and ${}_5P_2$=20 transitions between every two vowels. Speaking rate was adjusted to the transition rate of hand gestures. Each recording was done five times. The total number of speech samples was (5+20)×5=125. 18 dimensional cepstrum coefficients (including power) were extracted by STRAIGHT [9], where the frame length was 40 ms and the frame shift was 1 ms. Then, for every possible combination between a gesture sequence and its corresponding cepstrum sequence, after linear alignment between them, the distribution of augment vector $\boldsymbol{z}$ was estimated based on GMM, where the number of Gaussians was one. Finally, the regression function $f(\boldsymbol{a})$ was estimated. Fig.3 shows the results for /ai/. (a) indicates a resynthesized speech sample for vowel transition /ai/, (b) is a synthesized sample by using closed hand gesture data as input, and (c) shows a synthesized sample by using open hand gesture data. We used STRAIGHT for waveform generation, where F0 was fixed to be 140 Hz. According to a simple listening test of all the kinds of vowel transitions, we found that sounds of /i/, /u/, and /o/ were often confusing. In the following section, we design the correspondence between the five vowels and hand gestures so carefully that all the vowel sounds become distinct enough.

3. Design of the optimal correspondence

3.1. Variation of human hand gestures

What kind of hand gestures are possible and what kind of combination of five gestures is optimal for Japanese vowel production? In the preliminary experiment, sounds of /i/, /u/, and /o/ were often confusing and this implies that the gestures for these sounds are close to each other in the hand gesture space. In [10], 28 basic hand gestures were defined, which are shown in Fig.4. These 28 gestures were generated as follows. A hand has five fingers, each of which has two positions, high and low. Then, we have 2^5=32 combinations of five fingers, among which some are impossible to form. By deleting them, we can obtain 28 gestures. A female adult recorded gesture data for these 28 gestures twice, 2×28=56 data in total. Using these data, PCA was conducted to project 18 dimensional gesture data onto a two dimensional plane. The five gestures of the preliminary experiment, each of which had plural samples, were plotted on this plane (Fig.5). The five ovals represent regions for the five gestures and a sample trajectory of /aiueo/ is also plotted. As mentioned above, it is clear that the hand gestures of /i/, /u/, and /o/ are very close to each other. To generate distinct sounds for the individual vowels, we have to design an appropriate correspondence between vowels and gestures.



(a) resynthesized speech

(b) hand to speech conversion with closed data as input

(c) hand to speech conversion with open data as input

Figure 3: Synthesized speech for vowel transition of /ai/

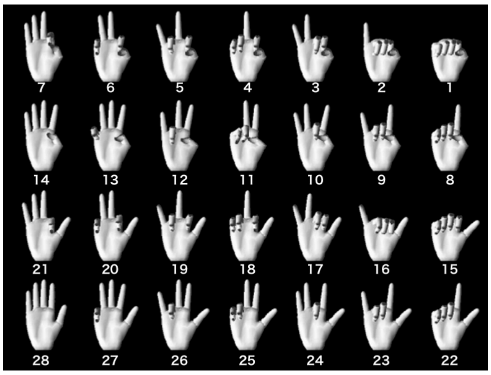


Figure 4: The 28 basic hand gestures [10]

3.2. Candidate sets of five hand gestures

We did the same PCA analysis for the 28 gestures shown in Fig.4 and the results are shown in Fig.5. Numbers in Fig.3.2 correspond to those in Fig.4. The gestures in the central purple region require special efforts to form and the remaining gestures are divided into five groups, A to E. By referring to the F1/F2 vowel chart of Japanese (See Fig.3.2), we assigned the

[1] Since the sampling period is variable, recorded data was interpolated linearly in such a way that the sampling period would be constant.

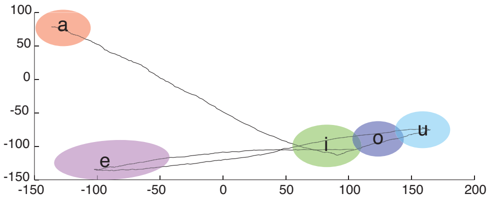
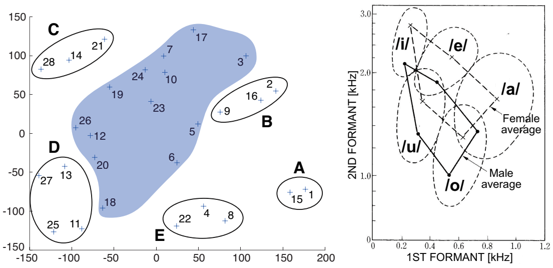


Figure 5: The five vowels in the preliminary experiment



(a) (b)

Figure 6: (a) The location of 28 gestures in PCA space, (b) the vowels chart of the five Japanese vowels

Table 1: Proposed 16 combinations of hand gestures

No.	/a/	/i/	/u/	/e/	/o/	No.	/a/	/i/	/u/	/e/	/o/
1	8	14	2	11	1	9	22	14	2	11	1
2	8	14	2	13	1	10	22	14	2	13	1
3	8	14	16	11	1	11	22	14	16	11	1
4	8	14	16	13	1	12	22	14	16	13	1
5	8	28	2	11	1	13	22	28	2	11	1
6	8	28	2	13	1	14	22	28	2	13	1
7	8	28	16	11	1	15	22	28	16	11	1
8	8	28	16	13	1	16	22	28	16	13	1

five vowels to the five region so that topological features of the five gestures in the gesture space and those of the five vowels would be equalized. For simplicity, we chose No.1 from group A and, from each of the other groups, we selected two gestures which are easier to form than the others in that group. Thus, the number of gestures we choose was nine in total. Tab.1 shows all the 16($=2^4$) combinations we selected and, out of these, we had to select the optimal one. To compare two topological patterns in different media, we used structural representation of sequence data [11, 12, 13].

3.3. Structural representation and comparison

Since speaker difference can be characterized as space mapping, mapping-invarinat features can be used as robust speech features of speech systems such as speech recognizers. [11, 14] showed that f-divergence between two distributions is invariant with any kind of invertible and differentiable transforms. In [11, 14], using Bhattacharyya distance (BD) as one of the f-divergence based distance measures, an utterance was structually represented, shown in Fig.7. A cepstrum sequence is automatically segmented and converted into a distribution sequence. Subsequently, an utterance is characterized as a total set of BDs, namely, distance matrix. Although this distance matrix is mapping invariant, by imposing some constraints, we introduce constrained invariance [15]. For example, if a distribution is assumed to be a Gaussian, the matrix is invariant only with linear transforms. In this study, a hand gesture sequence is represented as a structure (distance matrix) and a vowel sequence is also represented as another structure. Here, we assumed that the mapping function should be approximately linear. Then, we tentatively investigated whether the structural difference [11] an utterance matrix and a gesture matrix calculated with each of the 16 candidates in Tab.1 could work as evaluation function. The smaller the difference is, the better the candidate will be. Here, an utterance of /aiueo/ was used. Its distance matrix was compared to all the 16 gesture matrices of the 16 candidates. Following [15], the number of distributions was set to 25.

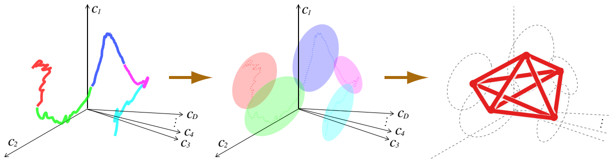


Figure 7: Structural representation of an utterance

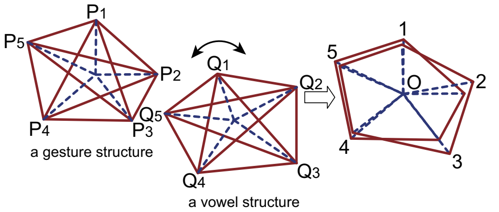


Figure 8: Structural matching between two matrices

The structural difference between two matrices is calculated as Euclidean distance between two vectors, each of which is formed by using all the elements of the upper triangle of a distance matrix. This simple measure can approximate well the minimum of total distance between the corresponding two points after shifting and rotating a structure (matrix) so that the two structures are overlapped the most optimally [11] (See Fig.8).

3.4. Results and discussions

Fig.9 shows the structural distances between an /aiueo/ utterance and a few candidates. The average distance over the 16 candidates and the distance of the hand gestures used in the preliminary experiment are also shown. Among the 16 candidates, No.5 shows the smallest distance and No.14 the largest. 10 Japanese adults participated in a listening test for five nonsense words, all of which were comprised of the Japanese five vowels such as /auoei/ and /oeiau/. The subjects were asked to transcribe the individual vowels. For each word, four versions, a re-synthesized sample, two synthesized samples with No.5 and No.14, and another synthesized one with the preliminary design were presented. The total number of non-sense word utterances was 20 and the total number of vowel sounds was 100. By randomizing the order of presentation, the 20 words were presented through headphones. The vowel-based intelligibility was 100%, 99.6%, 99.2%, and 95.2% for re-synthesized, No.5, No.14, and the preliminary design, respectively. Fig.10 shows the spectrograms of (a) re-synthesized, (b) No.5, (c) No.14, and (d) the preliminary design.

We can detect a small difference between (b) and (c) but a large one between the two and (d). The above results indicate that an adequate selection of hand gestures improves well the intelligibility and the distinctness of synthesized vowel sounds.

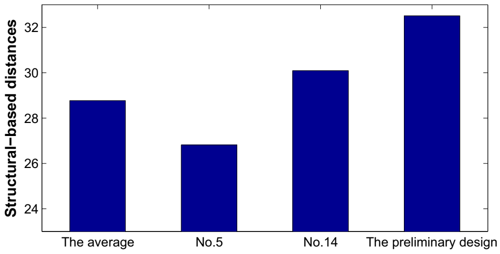
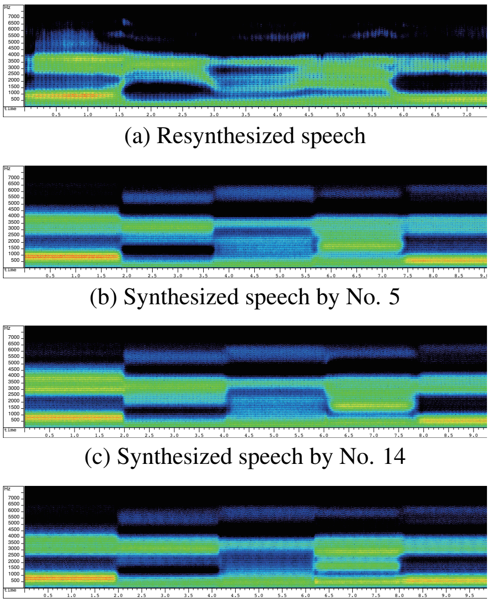


Figure 9: The structural distances for several sets of gestures



(a) Resynthesized speech

(b) Synthesized speech by No. 5

(c) Synthesized speech by No. 14

(d) Synthesized speech by preliminary design

Figure 10: Comparison between proposed designs for /aiueo/

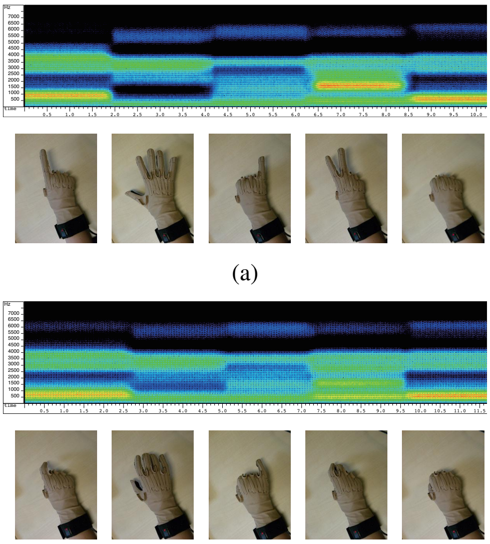
Not a small difference was found between No.5 and No.14 in Fig.9 but the difference was not perceived well auditorily and visually. In this paper, we cannot claim that the structural difference is a good enough measure when selecting a gesture set out of candidates. However, a certain measure to estimate the goodness of gestures is needed because, without that, a large number of listening tests are required to decide the optimal set of gestures. It is sometimes difficult to know in advance which parts of the body of a handicapped person are voluntarily movable. A good method to design a set of gestures automatically has to be devised in future work.

Finally, Fig.11 illustrates the spectrograms which were made by using (a) distinct (articulate) hand gestures and (b) ambiguous (inarticulate) hand gestures.[2] These speech samples were synthesized based on No. 5. By comparing (a) with (b) visually and auditorily, we can claim that our hand-to-speech generator can control the degree of articulation very easily.

4. Conclusion

We implement a speech synthesizer from hand gestures based on space mapping. By considering the topological equivalence between the structure of hand gestures in a gesture space and that of vowel sounds in the vowel space, we demonstrate how a quasi-optimal correspondence can be obtained. In the future, we will reconsider how to use the structural representation for automatic selection of the gestures and consider how to generate consonant sounds as well as prosodic features based on this framework.

[2] Readers can find wav files and movie files in the following link. http://www.gavo.t.u-tokyo.ac.jp/˜kunikoshi/index.html



(a)

(b)

Figure 11: Hand-to-speech generation in two styles (a) articulate (b) inarticulate

5. References

[1] Say-it! SAM, Words+, Inc. http://www.words-plus.com/website/products/syst/say_it_sam2.htm

[2] Voice Aids, Arcadia, Inc. http://www.arcadia.co.jp/VOCA/ (in Japanese)

[3] T. Hatakeyama, "The study for the development and the clinical application of support system for communication disorders," the symposium handout, pp12-13, 2007 (in Japanese)

[4] K. Yabu *et.al.*, " A speech synthesis device for voice disorders. – Its research approach and design concept –," IEICE Technical Report, vol.106, no.613, pp.25–30, 2007

[5] Glove Talk II http://hct.ece.ubc.ca/research/glovetalk2/index.html

[6] Y. Stylianou *et. al.*, "Continuous probablistic transform for voice conversion," IEEE Trans. Speech Audio Process., vol.6, pp.131–142, 1998

[7] A. Kain *et. al.*, "Spectral voice conversion for text-to-speech synthesis," Proc. ICASSP, vol.1, pp.285-288, 1998

[8] T. Toda *et. al.*, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," Speech Communication, vol.50, pp.215–227, 2008

[9] H. Kawahara *et.al.* "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," Speech Commun., 27, pp.187-207, 1999

[10] Ying Wu *et.al.*,"Analyzing and Capturing Articulated Hand Motion in Image Sequences," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.27, No.12, pp.1910-1922, 2005

[11] N. Minemtasu, *et.al.* , "Linear and non-linear transformation invariant representation of information and its use for acoustic modeling of speech, " Proc. Spring Meeting of Acoustic Society of Japan, 1-P-12, pp.147-148, 2007

[12] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," Proc. ICASSP, pp.889-892, 2005.

[13] N. Minematsu, *et.al.* , "Theorem of the invariant structure and its derivation of speech Gestalt," Proc. SRIV, pp.47-52, 2006

[14] Y. Qiao, *et.al.*, "Structual representation with a general form of invariant divergence", Proc. Autumn Meeting of Acoustic Society of Japan, 2-P-1, pp.105-108, 2008

[15] S. Asakawa,*et.al.*, "Automatic recognition of connected vowels only using speaker-invariant representation of speech dynamics," Proc. INTERSPEECH, pp.890-893, 2007