Gesture Design of Hand-to-Speech Conversion Based on Probabilistic Integration of Speech-to-Hand Joint Density Model and Hand Gesture Model

A. Kunikoshi, D. Saito, N. Minematsu and K. Hirose
Graduate school of the University of Tokyo
7-3-1 Hongo, Bunkyo,
Tokyo 113-8656, Japan
{kunikoshi, dsk_saito, mine, hirose}@t.u-tokyo.ac.jp

Y. Qiao
Shenzhen Institute of Advanced Technology,
1068 Xueyuan Avenue, Shenzhen University Town,
Shenzhen, P.R.China
yu.qiao@sub.siat.ac.cn

Abstract

When dysarthrics, individuals with speaking disabilities, try to communicate using speech, they often have no choice but to use speech synthesizers which require them to type word symbols or sound symbols. Input by this method often makes real-time communication troublesome and dysarthric users struggle to have smooth flowing conversations. In this study, we are developing a novel speech synthesizer where speech is generated through hand motions rather than symbol input. In recent years, statistical voice conversion techniques have been proposed based on space mapping between given parallel utterances. By applying these methods, a hand space was mapped to a vowel space and a converter from hand motions to vowel transitions was developed. It reported that the proposed method is effective enough to generate the five Japanese vowels. In this paper, we discuss the expansion of this system to consonant generation. In order to create the gestures for consonants, a Speech-to-Hand conversion system is developed using parallel data for vowels, in which consonants are not included. Thus, we are able to automatically search for candidates for consonant gestures for a Hand-to-Speech system.

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

1. Introduction

For oral 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 to read and write. We can find several technical products for dysartherics to facilitate their speech communication, which require no special skills on the part 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 mentioned above, however, these products often restrict the freedom of conversation and dysarthrics are less likely to take 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. Another example of speech generation from body motions is [5]. With two data gloves and some additional devices equipped to the user, body motions are transformed into parameters required to drive a formant speech synthesizer.

Recently, GMM-based speaker conversion techniques have been intensively studied, where the voice spaces of two speakers are mapped to each other and the mapping function is estimated based on a GMM [6, 7, 8]. This technique was directly and successfully applied to estimate a mapping function between a space of tongue gestures and other speech sounds. In this study, GMM was used to map two spaces of different media. This result naturally makes us surmise 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. However, [4] and [5] use classical synthesizers, i.e. formant synthesizers. Partly inspired by the remarkable progress of voice conversion techniques and voice morphing techniques [9] in this decade, we are in the process of developing a GMM-based hand-to-speech converter.

Section 2 describes the typical framework for the GMM-based conversion and the preliminary Hand-to-Speech system (H2S system) developed based on the framework. In Section 3, the problem of the preliminary H2S system is discussed. Based on the discussion, we try to figure out the quasi-optimal gesture design for five Japanese vowels. Then in Section 4, we discuss how to extend the framework into consonant sound generation. The experimental evaluation of the proposed method is described in Section 5. Finally, this paper is summarized in Section 6.

p3s 2011, March 14-15, 2011, Vancouver, BC, CA.

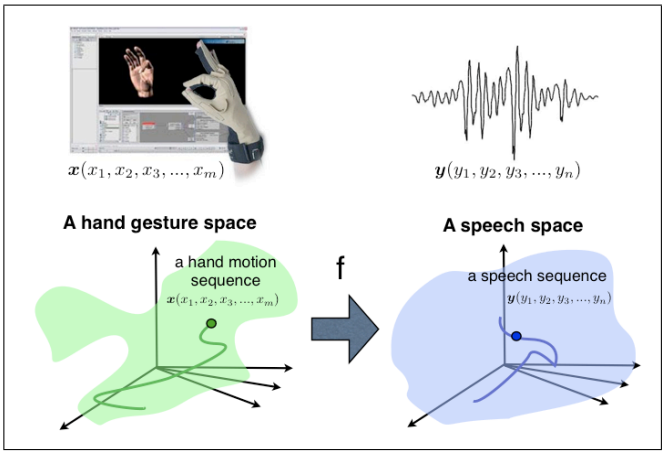


Figure 1. GMM-based media conversion

2. GMM-based media conversion

In this section, the typical framework for the GMM-based media conversion is described. Then, the preliminary H2S system is developed based on this framework.

2.1. Estimation of a mapping function between two spaces

[6, 7, 8] implemented GMM-based voice conversion between two speech spaces of a source speaker and a target speaker. Our H2S system can be considered a kind of voice conversion system in which the speech space of a source speaker is replaced by the hand gesture space. Fig.1 shows media conversion based on space mapping, where hand gesture vector $\boldsymbol{x}$ in the hand gesture space is converted into speech cepstral vector $\boldsymbol{y}$ in the speech space. The mapping function f in Fig.1 can be estimated by the method proposed in [7] as follows:

Let a vector $\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top$ be a joint feature vector consisting of the source $\boldsymbol{x}_t$ and the target $\boldsymbol{y}_t$ at time t. In GMM-based voice conversion, the vector sequence $\boldsymbol{Z} = [\boldsymbol{z}_1^\top, \boldsymbol{z}_2^\top, ..., \boldsymbol{z}_T^\top]$ is modeled by GMM $\boldsymbol{\lambda} = \{\omega_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i | i = 1, 2, ..., M\}$. The output probability $P(\boldsymbol{Z}|\boldsymbol{\lambda})$ can be computed as follows:

$$p(\boldsymbol{Z}|\boldsymbol{\lambda}) = \prod_{t=i}^{T} \sum_{i}^{M} \omega_i \mathcal{N}(\boldsymbol{z}_t | \boldsymbol{\mu}_i^{(Z)}, \boldsymbol{\Sigma}_i^{(Z)}), \quad (1)$$

$$\boldsymbol{\mu}_i^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_i^{(X)} \\ \boldsymbol{\mu}_i^{(Y)} \end{bmatrix}, \boldsymbol{\Sigma}_i^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(XX)} & \boldsymbol{\Sigma}_i^{(XY)} \\ \boldsymbol{\Sigma}_i^{(YX)} & \boldsymbol{\Sigma}_i^{(YY)} \end{bmatrix}, \quad (2)$$

where M is the number of mixtures, ω_i is the mixture weight of the i-th component, $\mu_i^{(.)}$ is the mean vector and $\Sigma_i^{(.)}$ is the covariance matrix.

The optimal sequence of the target feature vectors $\boldsymbol{Y} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, ..., \boldsymbol{y}_T^\top]^\top$ given a source feature vectore sequence $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, ..., \boldsymbol{x}_T^\top]^\top$ is described as follows:

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(x)}) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \lambda^{(z)}) P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}), \quad (3)$$

/a/ /i/ /u/ /e/ /o/

Figure 2. Gestures of the Japanese five vowels

where

$$\begin{aligned} P(m|\boldsymbol{x}_t, \lambda^{(Z)}) &= \frac{\omega_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)})}{\sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)})}, \\ P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(Z)}) &= \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(Y)}, \boldsymbol{D}_m^{Y}), \\ \boldsymbol{E}_{m,t}^{(Y)} &= \boldsymbol{\mu}_m^{(Y)} + \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(X)}), \\ \boldsymbol{D}_m^{(Y)} &= \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)-1} \boldsymbol{\Sigma}_m^{(XY)}. \end{aligned} \quad (4)$$

The regression function $f(\cdot)$ is obtained by maximizing the conditional distribution above. $f(\cdot)$ in the least squares can be written as follows:

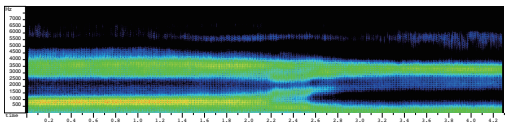
$$f(\boldsymbol{x}_t) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(Z)}) \boldsymbol{E}_{m,t}^{(Y)}. \quad (5)$$

2.2. A preliminary experiment [10]

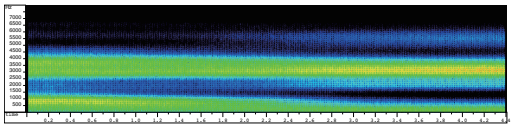
As a preliminary experiment, an H2S system was implemented for vowel transitions such as /ai/ and /oe/. The correspondence between hand gestures and the five Japanese vowels was shown in Fig.2. These gestures are designed so that a transition between any pair of vowels would not generate a third vowel. Hereafter, the correspondence between gestures and speech will be called "gesture design".

For training GMMs, a female adult recorded gesture data for the isolated vowels and ${}_5P_2$=20 transitions for each permutation of 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 cepstral coefficients were extracted by STRAIGHT [9], where the frame length was 40 ms and the frame shift was 1 ms. After every possible combination between a gesture sequence and its corresponding cepstral sequence were linearly aligned, the distribution of augment vector $\boldsymbol{z}$ was estimated based on a GMM for them, where the number of Gaussians was one. Finally, the regression function $f(\boldsymbol{x})$ was estimated.

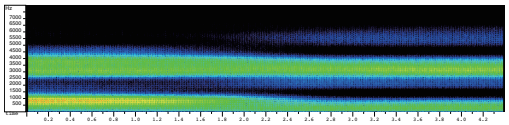
[1] Since the sampling period is variable, recorded data was interpolated linearly in such a way that the sampling period would be constant.



(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/

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. Through a simple listening test of all the kinds of vowel transitions, we found that sounds of /i/, /u/, and /o/ were often confused. In the following section, we design the correspondence between the five vowels and hand gestures so vowel sounds will have a distinct difference between them.

3. The optimal gesture design [11]

In this section, the problem found in the preliminary H2S system development is discussed. Based on the discussion, we try to figure out the quasi-optimal gesture design for five Japanese vowels so that all the vowel sounds become distinct enough.

3.1. Variation of human hand gestures

In the preliminary experiment, sounds /i/, /u/, and /o/ were often confused. This leads us to believe that the gestures for these sounds are close to each other in the hand gesture space. In [12], 28 basic hand gestures were defined, which are shown in Fig.4. These 28 gestures were generated as follows. As a hand has five fingers, each of which has two positions, high and low, we have 2^5=32 combinations for the five fingers. Among which, some are physically impossible to form. By removing those impossible-to-form gestures, 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 onto the plane (Fig.5). The

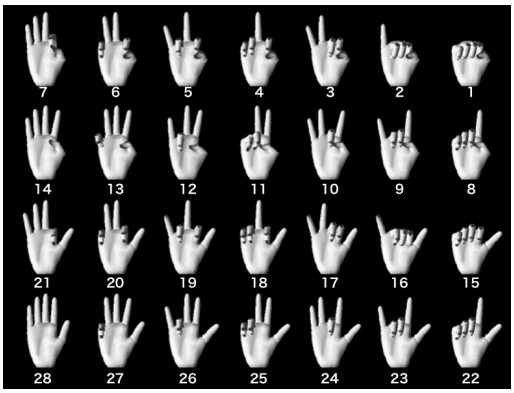


Figure 4. The 28 basic hand gestures [12]

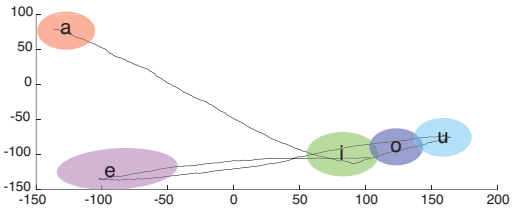
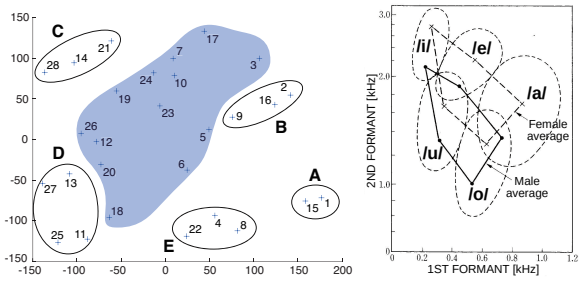


Figure 5. The five vowels in the preliminary experiment



(a) (b)

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

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.

3.2. Candidate sets of five hand gestures

We did the same PCA analysis for the 28 gestures shown in Fig.4 with the results shown in Fig.6(a). Numbers in Fig.6(a) correspond to those in Fig.4. The gestures in the central blue 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.6(b)), we designated those of the five vowels to five regions such that the topological features of the five gestures in gesture space and the five vowels would be equalized. For simplic-

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

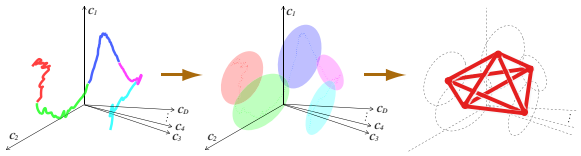


Figure 7. Structural representation of an utterance

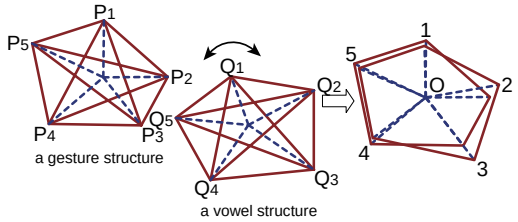


Figure 8. Structural matching between two matrices

ity, we chose No.1 from group A and, from each of the other groups, we selected two easy-to-form gestures in that group. Thus, the number of gestures we choose was nine in total. Tab.1 shows all of 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 the structural representation of sequence data [13, 14].

3.3. Structural representation and comparison

Since speaker differences can be characterized as space mapping, mapping invariant features can be used as robust speech features of speech systems such as speech recognizers. [13, 15] showed that f-divergence between two distributions is invariant with any kind of invertible and differentiable transforms.

In [13, 15], using Bhattacharyya distance (BD) as one of the f-divergence based distance measures, an utterance is structually represented as 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 introduced constrained invariance [16]. 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 represented as another structure. Here, we assumed that the mapping function should be approximately linear. Then, we tentatively investigated whether the structural difference [13] of 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, the utterance /aiueo/ was used. Its distance matrix was compared to all 16 gesture matrices for the 16 candidates. Following [16], the number of distributions was set to 25.

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 [13] (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 with 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 nonsense word utterances was 20 and the total number of vowel sounds was 100. The order of presentation was randomized, the 20 words were presented through headphones. The vowel-based intelligibility was 100%, 99.6%, 99.2%, and 95.2% for the re-synthesized sample, 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. A small difference is visible 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 the intelligibility and the distinctness of synthesized vowel sounds. A visible difference was found between No.5 and No.14 in Fig.9 but the difference was not well perceived auditorily and visually. We are unable to claim that the structural difference is sufficient to select 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 can be difficult to know in advance which parts of the body of handicapped individuals are voluntarily movable. A good method to design a set of gestures has to be devised automatically in future.

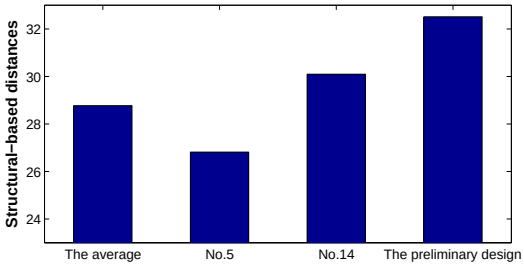
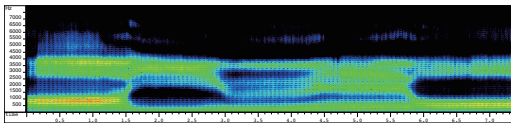
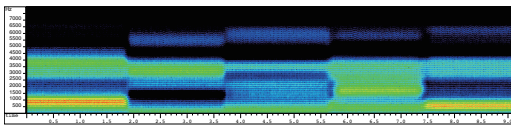


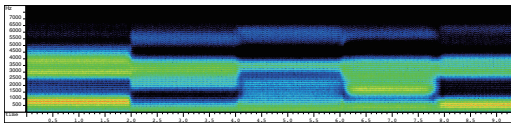
Figure 9. The structural distances for several sets of gestures



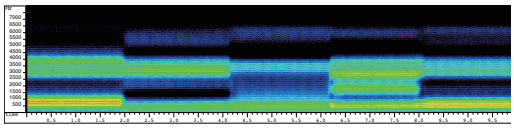
(a) Resynthesized speech



(b) Synthesized speech by No. 5



(c) Synthesized speech by No. 14



(d) Synthesized speech by preliminary design

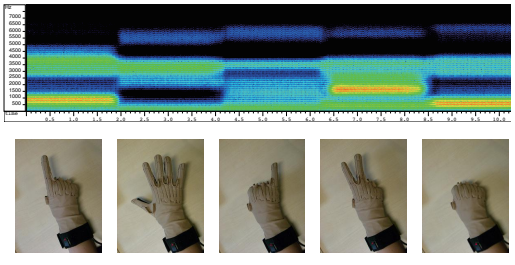
Figure 10. Comparison between proposed designs for /aiueo/

Finally, Fig.11 illustrates the spectrograms which were generated by using (a) distinct (articulate) hand gestures and (b) ambiguous (inarticulate) hand gestures. 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 well.

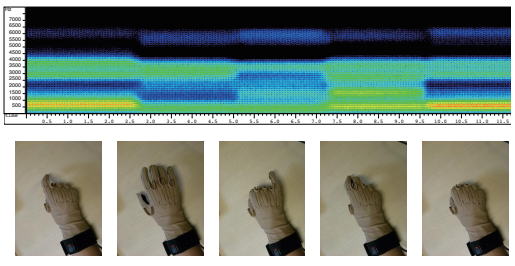
4. Consonant generation

In the previous section, we showed that the proposed method is effective enough to generate the five Japanese vowels. In this section, we will discuss how to extend this framework to consonants.

The challenge here is to figure out appropriate gestures for consonant sounds when the gesture design for vowels is given. [17] reported that inappropriate gesture designs for consonants result in a lack of transitions in synthesized speech. We have considered the reason to be: (1) the positional relation between vowels and consonants in the gesture space and that in the speech space were not equivalent, (2)



(a)



(b)

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

parallel data for transition parts from consonants to vowels did not correspond well. In order to get around those problems, we have developed a Speech-to-Hand conversion system (S2H system, the inverse system of H2S system) trained from parallel data for vowels only to infer the gestures corresponding to consonants.

In statistical machine translation studies, the maximization problem of $P(\boldsymbol{y}|\boldsymbol{x})$ for $\boldsymbol{y}$ is often solved using the Bayes rule [18]. Notation $\boldsymbol{x}$ stands for Japanese and $\boldsymbol{y}$ English, for example. In order to directly model and maximize $P(\boldsymbol{y}|\boldsymbol{x})$, a large amount of parallel data are needed. However considering $P(\boldsymbol{y}|\boldsymbol{x})$ as $P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y})$, high performance machine translation is realized with a small amount of parallel data for $P(\boldsymbol{x}|\boldsymbol{y})$ and an accurate model $P(\boldsymbol{y})$ trained with large English corpus. This framework has also been applied to voice conversion studies. [19] proposed a new technique for voice conversion using a joint density model trained by a small amount of parallel data and a target speaker model trained by a large amount of speech from target speaker.

Our aim is to figure out the appropriate gestures for consonants. In order to achieve this aim, we propose using an S2H system $P(\boldsymbol{x}|\boldsymbol{y})$, the inverse system of H2S system, and apply Bayes rule to obtain the gesture. In the other words, we would figure out the gestures $\boldsymbol{x}$ for speech $\boldsymbol{y}$ by $\underset{x}{\operatorname{argmax}}\, P(\boldsymbol{x}|\boldsymbol{y}) = \underset{x}{\operatorname{argmax}}\, P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})$. Here, $P(\boldsymbol{y}|\boldsymbol{x})$ is the joint density model for parallel data of vowels, $P(\boldsymbol{x})$ is the statistical gesture model trained by large amount of gesture data. Then the gestures should be obtainable by inputting the consonants into the S2H system. A system only using $P(\boldsymbol{y}|\boldsymbol{x})$ may derive hard-to-form gestures. However ours with the well trained $P(\boldsymbol{x})$ is anticipated to take account of the naturalness of gestures.

5. Experiments

In order to verify that S2H system is able to derive the gestures for speech, not included in the parallel data, the experiments below are carried out.

5.1. Speech-to-Hand conversion sytem

First off, an S2H system is developed using parallel data based on the gesture design for vowels and the statistical gesture model. By inputting the features for the consonants speech, the gestures for consonants are obtained.

For the gesture model, 16 gestures are chosen in Fig.6(a). Those are 15 gestures not within the difficult-to-form area and the easiest gesture in the difficult-to-form area No.7. Recorded gesture data from a female adult for the isolated gestures and the transitions of all pairs, $16 + {}_{16}P_2 = 256$ in total. The gesture model is developed using this gesture data. The number of mixtures is 64.

Next, the gesture design for the five Japanese vowels are chosen from among those 16 gestures. For simplicity, we chose No.28 for /a/. Considering the discussion of section 3, we did not use gesture designs satisfying the condition that the Euclidean distances in the gesture space between a–i and a–u are smaller than those between a-e and a–o, respectively. We get 8190 candidates for gesture designs of the five Japanese vowels. Then the joint density models for S2H system are obtained for every gesture design as follows.

The gestures for the isolated five Japanese vowels and ${}_5P_2 = 20$ transitions between every two vowels, $5+20 = 25$ gesture data in total, are extracted from the gesture data set above. In addition, a male adult speaker recorded speech for the five Japanese vowels and ${}_5P_2 = 20$ transitions between every two vowels. Each recording was done once. The total number of speech samples was $5+20 = 25$. Then, cepstrum extraction and interpolation are carried out in the same way as in section 2.2. Using these gesture sequences and cepstral vector sequences, augmented vectors are made and the joint density model is trained with them. The number of mixtures of the joint density model is 8.

By inputting the /n/, /m/, /r/, /p/, /b/ sounds, the gestures for those consonants are obtained. Thus, we get 8190 gesture designs for /a/, /i/, /u/, /e/, /o/, /n/, /m/, /r/, /p/, /b/. Some of the obtained gesture designs are shown in Fig.12, 13. From these candidates, we have to select the best one.

5.2. Hand-to-Speech conversion sytem

In order to compare those 8190 gesture designs obtained in the previous section, the following experiments are carried out. First, for every gesture design, the H2S system is developed using the method described in section 4. The joint density models $P(\boldsymbol{x}|\boldsymbol{y})$ are trained with the same training data and the same number of mixtures as for those of S2H systems in the previous section. Speech model $P(\boldsymbol{y})$ is trained using A set 50 sentences from the ATR phoneme balanced sentences recorded by the speaker who recorded the speech

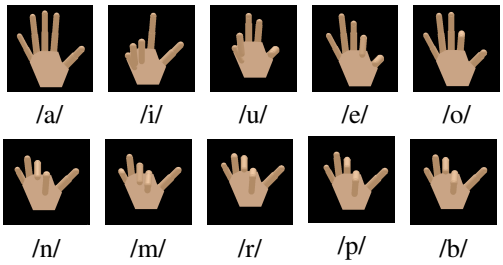


Figure 12. Gesture design for the five Japanese vowels to develop S2H system and gestures for consonants obtained through that method (sample 1)

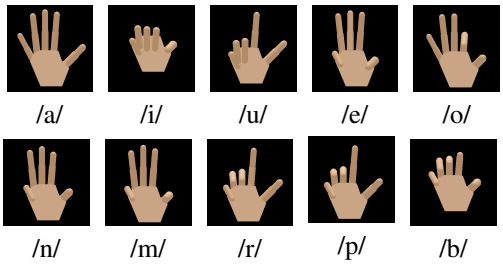


Figure 13. Gesture design for the five Japanese vowels to develop S2H system and gestures for consonants obtained through that method (sample 2)

data for the joint density model. The number of mixtures is 64, the same as that of the gesture model.

Then, gestures estimated by the S2H systems are input into those H2S systems. If those S2H and H2S systems are both ideal, input speech for S2H systems and output speech of H2S systems will be identical. However, the joint density models $P(\boldsymbol{x}|\boldsymbol{y})$ and $P(\boldsymbol{y}|\boldsymbol{x})$ do not completely describe the correspondence between the gesture space and the speech space. Therefore there will be distortion between the input speech for S2H systems and the output speech of H2S systems. We consider using this distortion as criteria to design the gestures. That is to say, we think that the smaller the cepstrum distortion between the input speech for S2H systems and the output speech of H2S systems is, the better the gesture designs will be.

Fig.14 shows the average and standard deviation of 8190 cepstral root mean square errors (RMSE). For simplicity, we only focused on the /n/ consonant. Input speech for the S2H systems are the five Japanese vowels in the training data and /na/, /ni/, /nu/, /ne/ and /no/ recorded by one speaker. $5 + 5 = 10$ samples in total. Thus, 10 mora-unit cepstral RMSEs are calculated for every gesture design. Then gesture designs are sorted for mora-unit cepstral RMSE. Therefore every gesture design has 5 ranks for each mora. The quasi-optimal design, the one where the summation of the ranks is minimal, is No. 28 for /a/, No. 22 for /i/, No. 11 for /u/, No. 7 for /e/ and No. 21 for /o/. Fig.14 also shows the cepstral RMSE of the quasi-optimal design.

Experimental evaluation shows the cepstral RMSE de-

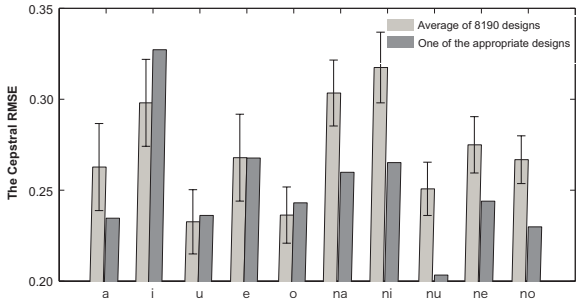
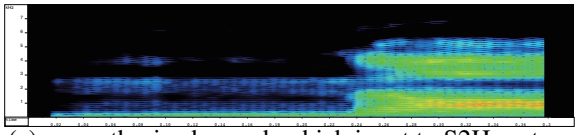
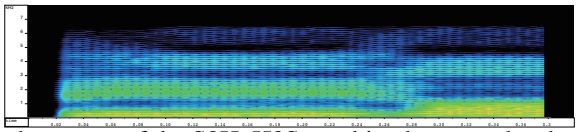


Figure 14. The cepstral RMSE between the input speech for S2H systems and the output speech of H2S systems



(a) re-synthesized speech which input to S2H system



(b) the output of the S2H–H2S combined system developed by the quasi-optimal gesture design

Figure 15. synthesized speech for /na/

pends on the mora. Of the five Japanese vowels, the smallest one is /u/ and the largest one is /i/. Overall, compared with vowels included in the training data, consonants which are not included in the training data tend to show larger cepstral RMSE. On the other hand, the cepstral RMSEs of vowels are almost the same as those of consonants in the quasi-optimal design. Fig.15 shows re-synthesized speech and the output of the quasi-optimal S2H–H2S combined system [2]. [17] reported that inappropriate gesture designs for a consonant result in the lack of transitions in synthesized speech. Then one-mora synthetic speech of /na/ are perceived as two sounds of /n/+/a/. Synthesized speech based on our proposed method are perceived as one unit in the simple listening test.

6. Conclusion

We have implemented 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 demonstrated how a quasi-optimal correspondence can be obtained. In addition, we proposed one framework to derive the gestures for consonants when the correspondence for vowels is given. In the future, we will consider how to take into account other articulatory parameters, such as the nasal cavity and the lips.

[2] Smoothing was performed before visualization

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, O. Cappe and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion, *IEEE Trans. Speech Audio Proc.*, vol.6, pp.131–142, 1998.

[7] A. Kain and M.W.Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP*, vol.1, pp.285–288, 1998.

[8] T. Toda, A. W. Black and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory, *IEEE Trans. Audio Speech Language Proc.*, vol.15, pp.2222–2235, 2007.

[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] A. Kunikoshi,*et.al.*, "Development of a speech generator from hand motions based on space mapping ", *Proc. Sprint Meeting of Acoustic Society of Japan*, 1–Q–23, pp.375–376, 2008.

[11] A. Kunikoshi *et al.*, "Speech generation from hand gestures based on space mapping," *Proc. INTERSPEECH*, pp.308–311, 2009.

[12] Y. Wu *et.al.*, "Analyzing and Capturing Articulated Hand Motion in Image Sequences," *IEEE Trans. PAMI*, vol.27, No.12, pp.1910–1922, 2005.

[13] 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.

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

[15] 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.

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

[17] A. Kunikoshi,*et.al.*, "Nasal sound generation and pitch control for the real-time hand to speech system ", *Proc. Spring Meeting of Acoustic Society of Japan*, 1–P–8, pp.419–422, 2010.

[18] P. Brown *et al.*, "A statistical approach to machine translation," Computational Linguistics Vol.16, No.2, pp.79–85, 1990.

[19] D. Saito *et al.*, "Probabilistic Integration of Joint Density Model and Speaker Model for Voice Conversion," *Proc. INTERSPEECH*, pp.1728–1731, 2010.