Gesture Design of Hand-to-Speech Converter Derived from Speech-to-Hand Converter Based on Probabilistic Integration Model

Aki Kunikoshi[1], *Yu Qiao*[2], *Daisuke Saito*[1], *Nobuaki Minematsu*[1], *Keikichi Hirose*[1]

[1]The University of Tokyo, Japan, [2]Shenzhen Institutes of Advanced Technology, China

{kunikoshi, dsk_saito, mine, hirose}@gavo.t.u-tokyo.ac.jp, 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 firstly developed using parallel data for vowels, in which consonants are not included. Then, we are able to automatically search for candidates for consonant gestures for a Hand-to-Speech system.

Index Terms: 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 dysarthrics to facilitate their speech communication, which require no special skills on the part of the receivers. *Say-it!* [1] is a portable PC based product, 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 [2]. In [3], 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 [4]. With two data gloves and some additional devices equipped to the user, body motions are transformed into parameters for 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 [5, 6]. 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 [3] and [4] 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, [3] and [4] use classical synthesizers, i.e. formant synthesizers. Partly inspired by the remarkable progress of voice conversion techniques and voice morphing techniques [7] in this decade, we are in the process of developing a GMM-based hand-to-speech converter. In this paper, we focus attention on the design of the system.

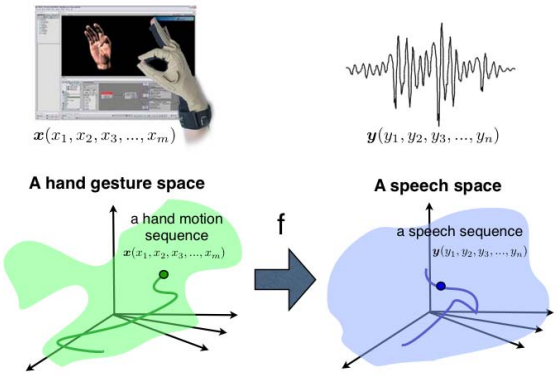


Figure 1: GMM-based media conversion

Section 2 describes the typical framework for the GMM-based conversion. In Section 3, the preliminary Hand-to-Speech system (H2S system) was developed based on the framework. 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.

2. GMM-based media conversion

[5, 6] 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 [6] 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, $\mu_i^{(\cdot)}$ is the mean vector and $\Sigma_i^{(\cdot)}$ 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:

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

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:

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

3. Japanese vowel sound generation [8]

3.1. Gesture design

H2S system was implemented previously for the five Japanese vowels. The framework of Fig.1 needs parallel data between two spaces in order to learn a GMM. For voice conversion, it is not very difficult to get correspondence between two data sequences, for example, by DTW. On the other hand, how to design the optimal correspondence between vowels and hand gestures is one of the most important issues for our framework.

In [9], 28 basic hand gestures were defined, which are shown in Fig.2. 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. [8] showed that the quasi-optimal correspondence can be obtained by considering the topological equivalence between topological features (structure) of hand gestures in the gesture space and those of vowel sounds in the vowel space. Using a CyberGlove made by Immersion Inc., a female adult recorded

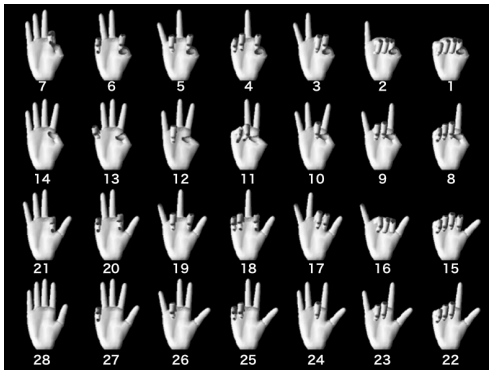
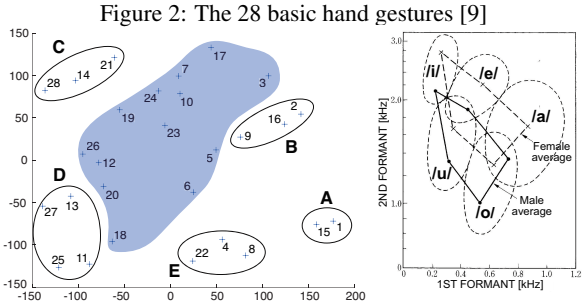


Figure 2: The 28 basic hand gestures [9]



(a) (b)

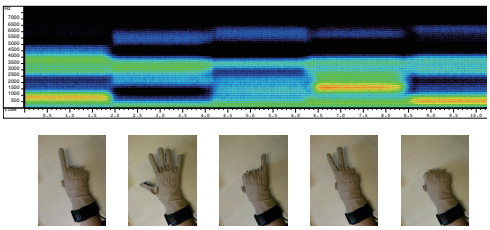
Figure 3: (a) The location of 28 gestures in PCA (b) the vowel chart of the five Japanese vowels

gesture data for these 28 gestures twice, 2×28=56 data in total. CyberGlove has 18 sensors and its sampling period is 10-20 ms.[1] PCA was conducted using these data to project 18-dimensional gesture data onto a two-dimensional plane, shown in Fig.3 (a). 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 shown in Fig.3 (b), we designed No.8 for /a/, No.28 for /i/, No.2 for /u/, No.11 for /e/ and No.1 for /o/ so that topological features of the five gestures in the gesture space and those of the five vowels would be equalized.

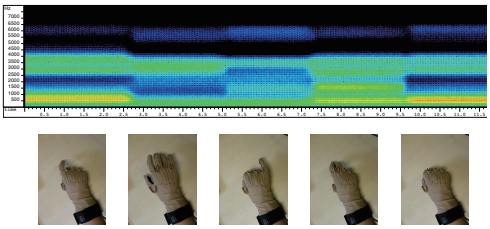
3.2. Estimation of a mapping function

For training GMMs, a female adult recorded gesture data for the isolated vowels and ${}_5P_2$=20 transitions for every permutation of two vowels using CyberGlove. Each 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 isolated vowels and ${}_5P_2$=20 transitions for each permutation of two vowels. Each recording was performed five times. The total number of speech samples was (5+20)×5=125. 18 dimensional cepstrum coefficients were extracted by STRAIGHT [7], 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 was 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. Fig.4 illustrates the spectrograms. We used STRAIGHT for waveform generation, where F0 was fixed to be 140 Hz. By comparing (a) with (b) visually and auditorily, we can claim that our hand-to-speech generator can control the degree of articulation well.

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



(a) Synthesized speech by using articulate hand gestures



(b) Synthesized speech by using inarticulate hand gestures

Figure 4: Synthesized speech for /aiueo/

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. [10] reported that inappropriate gesture designs for consonants result in a lack of smoothness in transitional segments of 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) 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. As explained later, this S2H system is used for H2S.

In statistical machine translation studies, the maximization problem of $P(\boldsymbol{y}|\boldsymbol{x})$ for $\boldsymbol{y}$ is often solved using the Bayes rule [11]. 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 is 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 a large English corpus. This framework has also been applied to voice conversion. [12] 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 the 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 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 the naturalness of gestures into account.

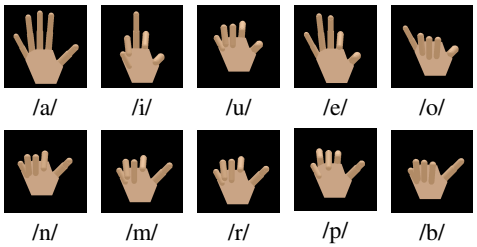


Figure 5: An example of the gesture design for the five Japanese vowels to develop S2H system and gestures for consonants obtained through that method

5. Experiments

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.3(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 18 sensors on the CyberGlove measure the angles of 18 joints in the hand. These 18 data are mutually correlated, because it is difficult to move joints completely independently. On the other hand, each dimension of speech feature vectors, cepstral coefficients, can be considered independent from each other. Considering the equivalence of two spaces, we performed PCA to every gesture data so that each dimension of gesture feature vectors will be as orthogonal as speech feature vectors.

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 3. Using these gesture sequences and cepstral vector sequences, augmented vectors are made and the joint density model is trained with them. The optimal numbers of mixtures, 64 and 8, are chosen for the gesture and joint density models, respectively.

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/. One of the obtained gesture designs is shown in Fig.5.

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 set A: 50 sentences from the ATR phoneme balanced sentences, recorded by the speaker who recorded the speech 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.6 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, 10 samples in total. Thus, 10 mora-unit cepstral RMSEs are calculated for every gesture design [2]. Then gesture designs are sorted for mora-unit cepstral RMSE. Therefore every gesture design has 5 ranks for each mora, /n/ + a vowel. The quasi-optimal design, the one where the summation of the ranks is minimal, is No.28 for /a/, No.7 for /i/, No.1 for /u/, No.21 for /e/ and No.15 for /o/. Fig.6 also shows the cepstral RMSE of the quasi-optimal design. It shows that the cepstral RMSE depends 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. Fig.7 shows re-synthesized speech and the output of the S2H–H2S combined system.

5.3. The subjective evaluation

The effectiveness of our approach was evaluated subjectively through an AB preference test, in which 15 Japanese native speakers selected the preferred one from a pair of synthesized speech in term of naturalness. A and B were output speech from the H2S systems which were trained by the conventional design [10] and the proposed design, respectively. For both designs, the gesture design for vowels was the quasi-optimal one in the previous section and the same training data for vowels as for those of S2H system was used. In the conventional design, the gesture for /n/ was chosen as No.8, without considering the topological equivalence between the gesture space and the speech space. Then gesture data for /na/, /ni/, /nu/, /ne/, /no/ were recorded once and added to the training data. In the proposed design, 10 sets of the gesture data for /na/, /ni/, /nu/, /ne/, /no/, which were obtained by S2H–H2S combined system, are added to the training data. The total frame numbers for /n/ of both designs were approximately the same. The mixture number of GMM was 64 for both designs. Preference test was conducted separately for each case of /a/, /i/, /u/, /e/, /o/, /na/, /ni/, /nu/, /ne/ and /no/. Preference of the proposed design was 48%, no preference was 27% and the preference of the conventional design was 24%. This result shows that the proposed method would be effective to derive appropriate gestures for consonants.

[2] Mora is a unit of speech production and speech rhythm of Japanese, which is usually composed of CV or V.

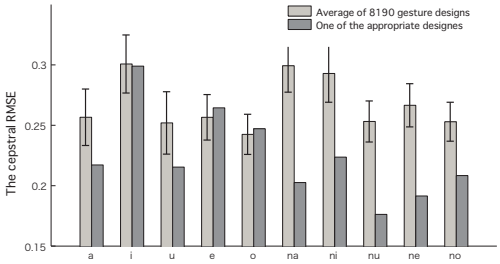
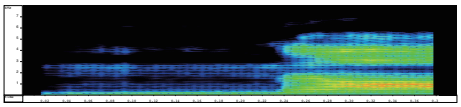
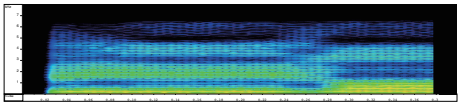


Figure 6: The cepstral RMSE between the input speech for S2H systems and the output speech of H2S systems



(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 7: synthesized speech for /na/

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 a framework to derive the gestures for consonants when only 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.

7. References

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

[2] 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).

[3] 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 (in Japanese).

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

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

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

[7] 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.

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

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

[10] 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 (in Japanese).

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

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