Improvement of Structure to Speech Conversion Using Iterative Optimization

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

[1]Graduate School of Engineering, [2]Graduate School of Information Science and Technology
The University of Tokyo, Japan
{dsk_saito,qiao,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract

This paper describes a new and improved method for the framework of structure to speech conversion we previously proposed. Most of the speech synthesizers take a phoneme sequence as input and generate speech by converting each of the phonemes into its corresponding sound. In other words, they simulate a human process of reading text out. However, infants usually acquire speech communication ability without text or phoneme sequences. Since their phonemic awareness is very immature, they can hardly decompose an utterance into a sequence of phones or phonemes. In this situation, as developmental psychology states, infants acquire the holistic sound pattern of words from the utterances of their parents, called word Gestalt, and they reproduce it with their vocal tubes. This behavior is called vocal imitation. In our previous studies, the word Gestalt was defined physically and a method of extracting it from an utterance was proposed. We already applied the word Gestalt to ASR, CALL, and also speech generation, which we call structure to speech conversion. Unlike a reading machine, our framework simulates infants' vocal imitation. In this paper, a method for improving our speech generation framework using iterative optimization is proposed and evaluated.

1. Introduction

Most of the speech synthesizers are text-to-speech converters, which take a phoneme sequence as input and generate speech stream corresponding to the sequence. To build a synthesizer, symbol-to-sound mapping is learned from a speech corpus. If a speech corpus of speaker A is used, the synthesizer learns A's voices and can read text out for him/her. A very good synthesizer may be able to deceive speaker verification systems [1].

Developmental psychology tells that infants acquire spoken language through imitating the utterances from their parents, called vocal imitation. However, they never impersonate their parents. It is impossible for infants to imitate their parents' voices due to a large difference in the shape of vocal tubes. To enable the vocal imitation in this situation, some abstract representation of utterances should exist between infants and their parents. One may claim that they communicate orally via phonemic representation but researchers of infant study deny this claim. This is because their phonemic awareness is very immature and it is difficult for them to decompose an utterance into a sequence of phonemes [2, 3]. What makes the vocal imitation possible?

Researchers answer that infants extract the holistic sound pattern from word utterances, called word Gestalt [2, 3] and they reproduce it with their short vocal tubes. Here, we can say that the Gestalt has to be speaker-invariant because, whoever speaks a specific word to infants using different voices, it seems that infants always extract the same Gestalt.

Figure 1: /aiueo/ utterances of a tall speaker and a short speaker.

Figure 2: Speech sounds − vocal tube(size&length) = Gestalt.

What is the acoustic definition of the word Gestalt? Functionally, it is a holistic and speaker-invariant pattern embedded in an utterance. Recently, the third author showed a candidate answer mathematically and verified the validity of the answer experimentally [4]. The proposed method of extracting the Gestalt from an input utterance was used successfully for ASR [5, 6] and CALL [7]. In addition, we applied the method to speech generation, which modeled infants' vocal imitation well [8]. Our speech generation framework converts the Gestalt back to speech sounds. We call it structure to speech (STS) conversion. In the previous study, however, formulation and implementation were insufficient for complete imitation of the Gestalt. In this paper, by using iterative optimization so as to satisfy the structural constraints better, a method for improving our speech generation framework is proposed and evaluated.

In the rest of the paper, we explain the details of our proposed framework. Section 2 describes what the word Gestalt is and how it is defined mathematically. In Section 3, the STS conversion framework and a proposed method to improve it using iterative optimization are shown. Section 4 reports some experimental results of the method. Section 5 reports some results of subjective evaluation for our proposed method. Finally, Section 6 concludes this paper.

2. Acoustic definition of the Gestalt

2.1. Discussions on the Gestalt from two viewpoints

Figure 1 shows two examples of /aiueo/. One is generated by a tall speaker and the other by a short one. If an infant imitates these utterances, it will generate very similar utterances because the same Gestalt is considered to exist in both the utterances of Figure 1. Then, if we try to define the acoustic definition of the Gestalt, we have to find the speech features commonly existing in both the utterances, i.e. speaker-invariant speech features.

Why are the voices of a speaker different acoustically from those of another? This is simply because the default shape (size, length, etc) of the vocal tube is different among speakers. Since speech sounds are always generated from a vocal tube, their acoustic features are inevitably influenced by the default shape of the vocal tube, which is unique to the speaker. In this sense,

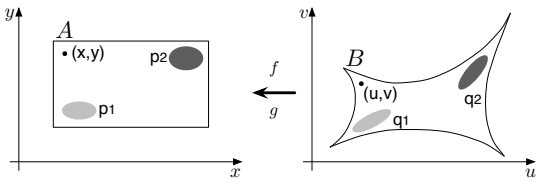


Figure 3: Linear or non-linear mapping between two spaces.

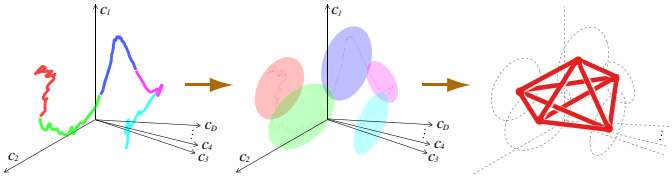


Figure 4: Invariant structuralization of an utterance.

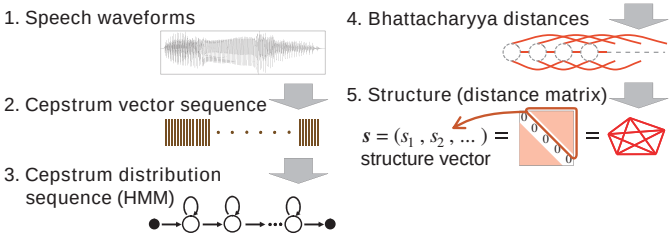


Figure 5: Feature extraction as HMM training for an utterance.

the Gestalt of an utterance is considered to be what remains after subtracting features of the default vocal tube shape from all the acoustic features of that utterance (See Figure 2).

2.2. Mathematical derivation of the Gestalt

In the above section, the Gestalt was considered from two viewpoints. In this section, it is defined mathematically. In speaker conversion studies of speech synthesis, it is often assumed that speaker differences are well modeled as space mapping. Figure 3 shows an example of invertible mapping (linear or nonlinear) between spaces A and B. In this figure the Gestalt is regarded as mapping invariant feature.

Here, in Figure 3, every event is characterized not as point but as distribution and event p_i in A is mapped to q_i in B. By considering two mapping functions of f and g, i.e. $x{=}f(u,v)$ and $y{=}g(u,v)$, we get the following;

$$q_i(u,v) = p_i(f(u,v), g(u,v))|J(u,v)|.$$

$J(u,v)$ is Jacobian. The Bhattacharyya distance (BD) is one of the well-known distance measures between two PDFs and we can prove that BD is invariant with any kind of invertible mapping functions between two spaces;

$$BD(p_1,p_2) = -\log \iint \sqrt{p_1(x,y)p_2(x,y)}dxdy$$
$$= -\log \iint \sqrt{p_1(f(u,v),g(u,v)) \cdot p_2(f(u,v),g(u,v))}|J|dudv$$
$$= -\log \iint \sqrt{p_1(f(u,v),g(u,v))|J| \cdot p_2(f(u,v),g(u,v))|J|}dudv$$
$$= -\log \iint \sqrt{q_1(u,v)q_2(u,v)}dudv = BD(q_1,q_2).$$

Based on this invariant feature, we introduced a transform-invariant representation of an utterance, shown in Figure 4. A sequence of cepstrum vectors is converted into a sequence of

Figure 6: Structure + vocal tube(size&length) = speech sounds

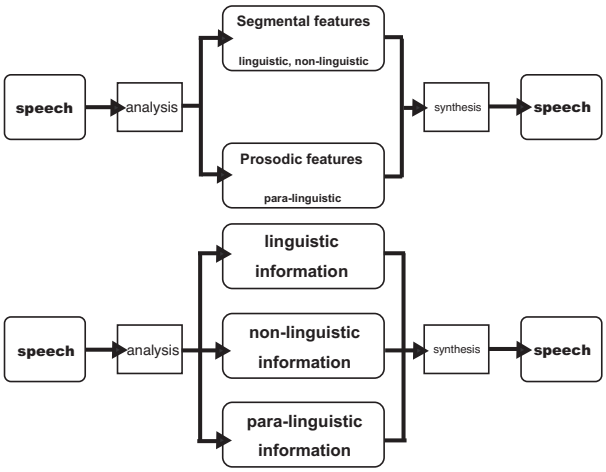


Figure 7: The conventional framework for analysis-resynthesis and the proposed one with three separate kinds of information.

distributions through merging similar frames and estimating a distribution for the merged frames. After that, every sound contrast between any two distributions, even including temporally distant ones, is calculated as BD. An utterance is represented as a transform-invariant distance matrix, which can characterize a geometrical structure uniquely. We call this matrix-based representation as structural representation and believe that the structure is the Gestalt. In [5], this procedure was implemented as MAP-based HMM training for an utterance, shown in Figure 5. Here, the number of distributions is often larger than that of phonemes existing in the utterance. We already applied this representation in ASR [5, 6] and CALL [7] successfully.

Figure 4 shows that the structural representation of an utterance is obtained by extracting speech contrasts (dynamics) only and discarding all the absolute and static features. Putting it another way, only articulatory movements are focused on and the articulatory features corresponding to the static and default shape of the vocal tube is ignored completely (See Figure 2).

The structure (the Gestalt) is so abstract a representation of an utterance that, with the structure only, speech sounds cannot be recovered or determined at all, shown in Figure 4. To determine and locate the sounds of a given structure, what should be additionally needed? Looking at Figure 2, we can say that the static and default shape of the vocal tube is required for the Gestalt to be realized acoustically. Figure 6 explains this process conceptually and, in the following section, this process of structure-to-speech conversion is implemented on computers.

3. Structure to speech conversion

3.1. Analysis-resynthesis with three kinds of information

Recently, analysis-resynthesis techniques are often utilized to modify speech. The top figure of Figure 7 shows the conventional framework of analysis-resynthesis. Speech features are divided into two kinds, segmental and prosodic. The former corresponds to the spectral envelope, which transmits linguistic as well as non-linguistic (speaker) information in speech. The

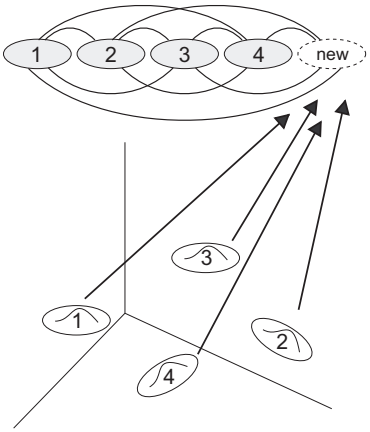


Figure 8: Search for the next target under structural constraints.

latter corresponds to fundamental frequency, power, and duration, which are said to carry para-linguistic information.

With the structural representation, we can modify the above framework into three pathways; three kinds of features for three kinds of information. The speech structure only captures spectral dynamics in an utterance and the proposed framework considers that it corresponds to linguistic information. As for non-linguistic (speaker) information, we consider that spectral bias transmits it to hearers. Using this bias feature, the structure can be located absolutely in an acoustic space, shown in Figure 4. Some readers may wonder whether words can be identified only with speech dynamics. To this question, our previous works showed that the answer is yes. With the speaker-invariant speech structures, speaker-independent speech recognition was realized successfully only with several training speakers [5, 6]. Further, the proposed structural representation was applied to dialect-based speaker classification of Chinese [9]. The speakers are classified successfully only based on their dialects (linguistic information), not based on their gender and age.

In the proposed framework, to generate speech sounds, all the three kinds of information or features have to be prepared. As told above, the default shape of the vocal tube, i.e. speaker identity, is translated acoustically as spectral bias. Then, if the center of a given structure of Figure 4 is located absolutely in an acoustic space, can we hear all the sounds from the structure subsequently? The answer is no because difference in the vocal tract length rotates a given speech structure [10]. This means that, to locate the structure completely, several points on the structure have to be determined absolutely in advance.

3.2. Searching a cepstrum space for target speech events

Here, conversion from a given structure to a speech sound sequence is implemented as follows. Several points on a given structure are fixed absolutely in advance. This step means that the default shape of the vocal tube is determined. Then, using these points as initial conditions and the structure (distance matrix) as constraint conditions, all the other points on the structure are searched for in a cepstrum space. Figure 8 shows how to search for the next target using 4 already determined events and structural constraints. In the case of infants' vocal imitation, the structural constraints are given from their parents. About the initial conditions, infants may use some speech sounds which they actually generated through vocal communication or playing with their parents.

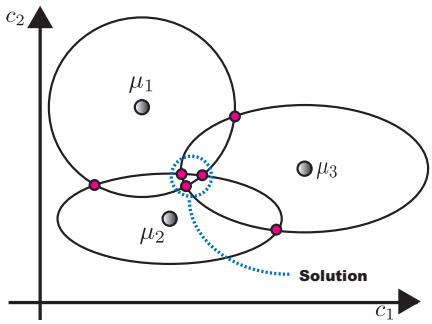


Figure 9: Solution of the search problem. In this figure, the intersection of three ellipses becomes the solution.

3.3. Solving the search problem

How do we solve this searching problem? When the two distributions are Gaussian, i.e. $p_1(x) = \mathcal{N}(\mu_1, \ \ _1)$ and $p_2(x) = \mathcal{N}(\mu_2, \ \ _2)$, BD is formulated as follows,

$$BD(p_1(x), p_2(x)) = \frac{1}{8}(\mu_1 - \mu_2)^T V_{12}^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\ln\frac{|V_{12}|}{|\ \ _1|^{\frac{1}{2}}|\ \ _2|^{\frac{1}{2}}}, \quad (1)$$

where $V_{12}=\frac{\Sigma_1+\Sigma_2}{2}$. In this case, BD is invariant to any common linear transform. Now let us consider an n-dimensional cepstrum space. Suppose that $\ _1$, $\ _2$ and μ_2 are already determined speech features and that we have to locate μ_1 in the cepstrum space using Equation 1 as structural constraint. In this case, the locus of μ_1 is found to draw a hyper-ellipsoid, ellipse in an n-dimensional space. From this fact, we take the following procedure to solve the search problem.

1. From the distance matrix, equations of hyper-ellipsoid, e.g. Equation 1, are obtained.
2. Vectors of the initial conditions are substituted to the equations obtained in 1.
3. The locus of the target event vector μ_1 is drawn by the equations obtained in 2.
4. The intersection of the loci drawn in 3 is obtained and this intersection will give us a solution.

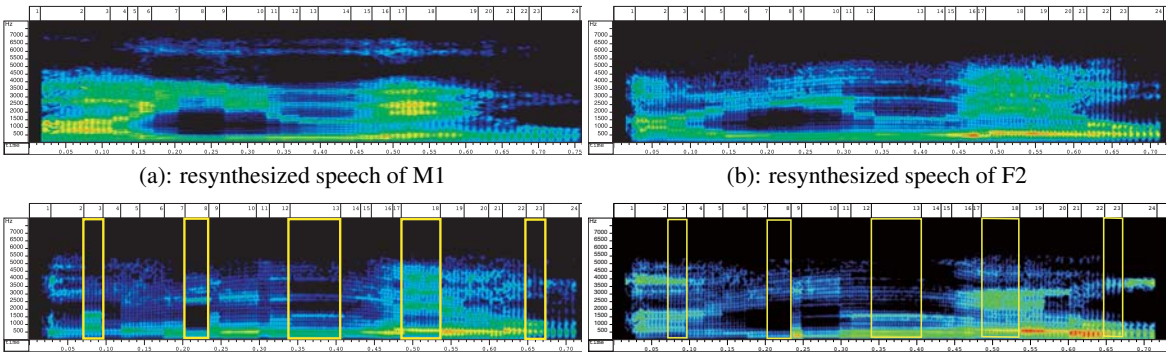
Here, we give an example of a two dimensional case. Speech events $A{=}\mathcal{N}(\boldsymbol{c_a}, \ \ _a)$ and $B{=}\mathcal{N}(\boldsymbol{c_b}, \ \ _b)$ are prepared for initial conditions, where covariance matrices of A and B are supposed to be diagonal. μ of speech event $C{=}\mathcal{N}(\mu, \ \ )$ is a target, where $\ \ $ is also diagonal and given. When BD between A and C is named as BD_a and BD between B and C is named as BD_b, the structural constraints are translated into a simultaneous equation as

$$\begin{cases} BD_a - \epsilon_a = (\boldsymbol{\mu} - \boldsymbol{c_a})^t \boldsymbol{A}(\boldsymbol{\mu} - \boldsymbol{c_a}) \\ BD_b - \epsilon_b = (\boldsymbol{\mu} - \boldsymbol{c_b})^t \boldsymbol{B}(\boldsymbol{\mu} - \boldsymbol{c_b}), \end{cases} \quad (2)$$

where $()^t$ is transpose of a vector, ϵ represents the second term in Equation 1 and $\boldsymbol{A}$ and $\boldsymbol{B}$ are

$$\boldsymbol{A} = \frac{1}{4}(\ \ _a + \ \ )^{-1}, \boldsymbol{B} = \frac{1}{4}(\ \ _b + \ \ )^{-1}. \quad (3)$$

In a two dimensional case, solving Equation 2 corresponds to obtaining the intersection of two ellipses geometrically. Generally speaking, the number of intersections of two ellipses is more than one in a two dimensional space. Hence, to determine



(a): resynthesized speech of M1

(b): resynthesized speech of F2

(c): synthesized speech with M1's structure and F2's initial conditions, without iteration

(d): synthesized speech with M1's structure and F2's initial conditions, with 10 iterations

Figure 10: Spectrograms of resynthesized speech (a and b) and synthesized speech (c and d); (a) M1 (father), (b) F2 (girl), (c) M1's structure + F2's initial conditions (no iteration) and (d) M1's structure + F2's initial conditions (10 iterations).

only one intersection for the target speech event, at least one more event is needed as initial condition. By expanding this discussion to an n-dimensional space, we can say that we need at least $n+1$ events as initial condition. Figure 9 shows a two dimensional case. The target event is obtained as intersections of three ellipses, whose origins are speech events given as initial conditions.

3.4. Iterative optimization

In the previous section, we explained a basic concept of solving a search problem for STS. However in the above formulation, each target was estimated independently. That is to say, the searching in the previous section does not give us the targets that can satisfy structural constraints fully. Now we assume that one utterance is composed of N speech events, among which, n events are considered as initial conditions and the remaining m events as targets ($m+n=N$). The number of unique elements of distance matrix for N events is ${}_NC_2 (= {}_mC_2+{}_nC_2+mn)$. The previous searching method considers only the mn elements of distance matrix as constraints. Considering the other elements of distance matrix (${}_mC_2+{}_nC_2$), we propose an iterative optimization method using already estimated events in the previous process as initial conditions again. A procedure of our new method is as follows.

1. Using the previous procedure, m targets are estimated by n initial conditions and structural constraints.
2. From the estimated m events, one event is selected. By regarding the remaining m–1 events as initial conditions, this event is re-estimated using its structural constraints to the other $m-1$ events. This process is repeated until all the m events are re-estimated.
3. From the whole N events ($N = m+n$), one event is selected and is re-estimated by the other $N-1$ initial conditions and the structural constraints. This process is repeated until all the N events are re-estimated.
4. Step 3. is repeated 10 times in this paper.

4. Experiment

4.1. Experimental conditions

For evaluation of the proposed framework, experiments using utterances composed of Japanese 5 vowels ($5! = 120$ words) were carried out. We used speech samples from 6 speakers (M1, M2 and M3 as male, and F1, F2 and F3 as female). Utterances of M1 and F1 were used to extract the word Gestalt, which was used as structural constraints when searching for targets.

For converting a spectrum sequence to a cepstrum sequence, STRAIGHT analysis [11] was adopted and a sequence of 40 dimensional vectors was obtained. For converting a cepstrum sequence to a distribution sequence, MAP-based HMM parameter estimation was adopted since all the distributions had to be estimated from a single utterance. Then, an utterance was converted into a sequence of 25 diagonal Gaussians. In addition, parameter division proposed in [5] was carried out. From a single speech stream, 20 multiple sub-streams were obtained. A structure was extracted from each two-dimensional sub-stream. The searching problem was solved in each sub-space.

The other utterances from M2, M3, F2 and F3 (henceforth target speakers) were used as initial conditions. After extracting prosodic features from these utterances with STRAIGHT, the utterances were converted into a sequence of 25 diagonal Gaussians. After that, 5 mean vectors (3rd, 8th, 13rd, 18th, and 23rd ones in the 25 Gaussians) were extracted and used as a part of initial conditions. In this experiment, all the covariance matrices of target events were given and also used as initial conditions. With these initial conditions of the target speakers and the structural constraints from M1 and F1, the remaining mean vectors were treated as targets and they were searched for. The number of iterative calculation that we proposed was changed from 0 to 10.

Finally using the prosodic features extracted above and a sequence of obtained distributions, utterances of the target speakers were synthesized. When we compare this experiment with infants' vocal imitation, M1 and F1 is a father and a mother, and target speakers are sons and daughters, who try to extract the word Gestalt in their parents' utterance and reproduce it acoustically using their short vocal tubes.

4.2. Results

Figure 10 shows (a) the spectrogram of a resynthesized utterance of M1 (father), (b) that of a resynthesized utterance of F2 (girl), and (c) that of a synthesized utterance with the girl's initial conditions (the girl's imitation through the father's Gestalt). In addition, (d) is the spectrogram of a synthesized utterance after 10 times iterative optimization. All of them are utterances of /aiueo/. In (c) and (d), the spectrum slices in five square boxes were given as initial conditions. When we compare (c) and (d)

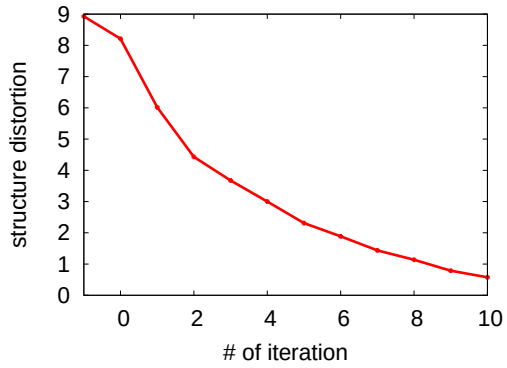


Figure 11: The number of iteration vs. structural distortion; The conditions are the same as used in Figure 10.

with (a) and (b) visually, we can find that spectrograms of (c) and (d) are closer to that of (b). In addition, comparing (d) with (c), we can find formant peaks are clearer in (d). It implies the speaker identity is well realized in (c) and (d), furthermore, iterative calculation properly improves the sound quality of synthesized speech. By preliminary listening, these were easily verified. Although the next section reports the detailed results of subjective evaluation, so far we can say that structure-to-speech conversion certainly works and that iterative optimization improves our previously proposed method.

Figure 11 shows a difference between the structure extracted from M1's utterance and the structure composed of the estimated speech events of F1 through iterative optimization. A structural difference is defined as euclidean distance between two distance matrices [4]. The graph is drawn as a function of the number of iteration (i). The experimental conditions are the same as used in Figure 10. In Figure 11, $i = -1$ means the conventional search method in [8], and $i = 0$ means the result of Step 2 in the procedure of iterative optimization. From Figure 11, we can find that the proposed iterative optimization contributes well to optimal searching constrained by the word Gestalt.

5. Subjective evaluation

5.1. Conditions

Three types of listening tests were carried out to evaluate intelligibility, naturalness and speaker identity of the speech samples generated by our method. (a) dictation test, (b) opinion score test and (c) speaker identity test were done. The purpose of (a) is to check whether our method captures linguistic information properly. (b) is for checking naturalness of the samples generated by our method. (c) is for checking whether the initial conditions can reproduce the speaker identity well.

Test (a) was carried out using 4 male subjects. Sample stimuli for evaluation were generated under different conditions. The conditions are in terms of (1) who gives the initial conditions and who gives the structural constraints (2 parents $\times$ 4 children) and (2) the number of iterative optimization (i=-1, 2 and 10). Totally, under 24(=8×3) conditions, our proposal was tested. 60 samples were presented to the subjects for each condition, namely 1,440 stimuli in total. In addition, 240 samples of resynthesized speech were also evaluated as reference. Subjects were instructed to transcribe each sample by using vowel symbols. In this test, they were told that each sample was composed of Japanese 5 vowels, i.e. word perplexity is $5! = 120$.

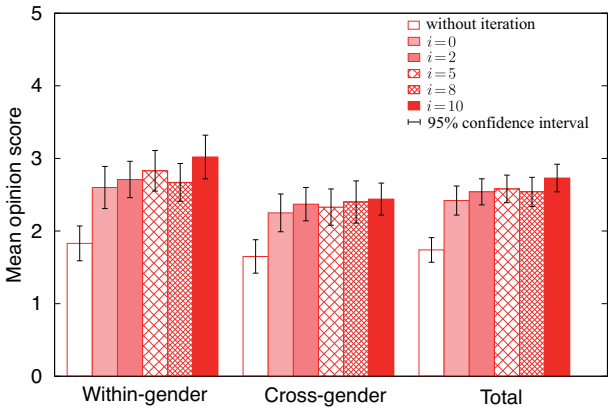


Figure 12: Results of Test (b) on naturalness. Score 5 means the most natural.

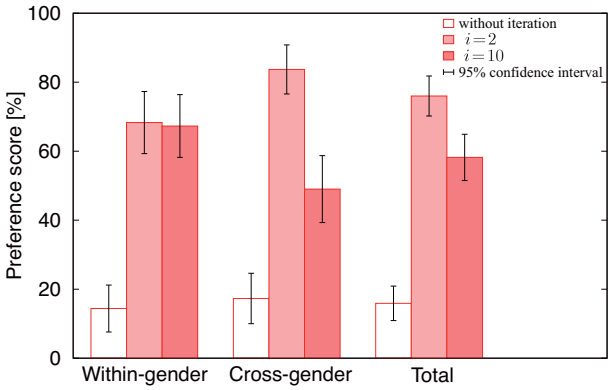


Figure 13: Results of Test (c) on speaker identity.

Test (b) was carried out with 13 subjects (10 males and 3 females). All the samples for evaluation were /aiueo/ generated under similar conditions to those in Test (a). The conditions are in terms of (1) 2 parents $\times$ 4 children and (2) the number of iteration (i =-1, 0, 2, 5, 8 and 10). i =-1 means no iteration and i =0 means Step 2 in the procedure of iteration. 1 sample was prepared for each condition, namely 48 samples in total. In addition, 20 samples of resynthesized speech were evaluated as reference. Each subject was asked to judge the naturalness of each sample by a score from 1 to 5, where 1 is the most unnatural and 5 is the most natural.

Test (c) was carried out with the same subjects of the Test (b). Samples were /aiueo/ under the same conditions as those of Test (a). Test (c) was a paired comparison using a reference stimulus. Each subject first listened to a resynthesized speech sample of the target speaker as reference, and 2 samples of different conditions where only the number of iteration is different. After that each subject judged which sample was more similar to the reference sample with respect to speaker identity.

5.2. Results

Table 1 shows the results of Test (a). In word accuracy (1), if a given word sample is perceived correctly at least by 3 subjects out of 4, the sample is counted as correct. In word accuracy (2), if 4 or 5 vowels in a given word sample are perceived correctly by more than 2 subjects out of 4, the word sample is counted as correct. We want to see whether subjects can at least extract a holistic pattern from a given sample, which may include a lo-

Table 1: Results of Test (a): M→F means a speaker who gives the structural constraint is a male and a target speaker is a female. i is the number of iteration. Accuracy rates in bold face are the highest performance.

	Word accuracy (1) [%]			Word accuracy (2) [%]			Vowel accuracy [%]		
	$i=-1$	$i=2$	$i=10$	$i=-1$	$i=2$	$i=10$	$i=-1$	$i=2$	$i=10$
M→M	**80.0**	73.3	78.3	80.8	**84.2**	81.7	88.5	90.1	**90.3**
M→F	50.8	**57.5**	43.3	59.2	**61.7**	52.5	79.2	**81.5**	75.6
F→M	64.2	64.2	64.2	66.7	**69.2**	65.8	83.8	85.0	**85.1**
F→F	61.7	**80.8**	70.0	70.0	**84.2**	73.3	83.6	**92.1**	85.8
average	64.2	**68.8**	64.0	69.2	**74.8**	68.3	83.7	**87.2**	84.2

cal error in word perception. This is why word accuracy (2) was prepared additionally to word accuracy (1). From Table 1, word accuracy (1) sometimes does not improve with iteration, e.g. the case of M→M. In word accuracy (2), however, we can say that the iterative optimization improves speech intelligibility. About 75% of the samples are perceived correctly and holistically. In addition, from a viewpoint of vowel accuracy, the iterative optimization contributes to improvement of speech intelligibility.

Figure 12 shows Mean Opinion Scores (MOS) of Test (b). The results are divided into the case that a speaker of structural constraints and a target are of the same gender (M→M and F→F in Table 1), the case that they are of different gender (M→F and F→M in Table 1), and total. In every case, we can find that iterative optimization improves naturalness of speech sounds. Final improvements after 10 iterations are about 1.2 points in the within-gender case, about 0.8 points in the cross-gender case, and totally about 1.0 points.

Figure 13 shows preference scores of Test (c). The results are divided similarly to Test (b)'s results. From Figure 13, we can find that the proposed method also makes speaker identity of speech samples closer to target speaker identity. However, in the case of cross-gender, preference score decreased in the case after 10 iterations. It means there is the optimal number of iteration for cancelling the speaker differences properly.

6. Conclusions

We have proposed a new method for the framework of structure to speech conversion. In the framework of structure to speech conversion, the word Gestalt is extracted from an input utterance and reproduced acoustically with some initial conditions given. This framework can simulate infants' vocal imitation and learning. Our proposed method in this paper has improved all of speech intelligibility, naturalness and speaker identity. One of reasons of these improvements is that discontinuity between speech events is cancelled by considering all structural constraints properly. However, in Figure 13, excessive iterative optimization sometimes causes the degradation of quality. In this case, over-fitting to structural constraints and over-smoothing may occur. For more improvements, the optimal number of iteration has to be estimated in a future work. We're also planning to integrate the prosodic aspect into the framework.

To conclude this paper, we want to discuss our STS conversion from anothor viewpoint. This paper tries to implement the process of infants' vocal imitation on machines. Infants never imitate the voices but extract the word Gestalt and reproduce it acoustically with their vocal tubes. It is known in animal sciences that the vocal imitation or vocal learning is found only in a limited kinds of animals. For example, non-human primates do not perform the vocal imitation. It is also known that the animals which do the imitation imitate the voices themselves. It is only humans that do not imitate the voices. As far as we know, all the speech synthesizers imitate the voices, i.e. animal-like imitation, and our synthesizer is the only one which performs infant-like imitation.

7. Acknowledgements

This research was partially funded by a fund for young researchers of Global COE "Secure-Life Electronics"[1].

8. References

[1] T. Masuko *et al.*, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," ICSLP2000, pp.302–305, 2000.

[2] S. E. Shaywitz, "Overcoming dyslexia,"Random House, 2005.

[3] M. Kato, "Phonological development and its disorders," J. Communication Disorders, no.2, vol.20, pp.98–102, 2003.

[4] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," ICASSP2005, pp.889–892, 2005.

[5] S. Asakawa *et al.*, "Multi-stream parameterization for structural speech recognition," ICASSP2008, pp.4097–4100, 2008.

[6] Y. Qiao *et al.*, "Random discriminant structure analysis for continous Japanese vowel recognition," ASRU2007, pp.576–581, 2007.

[7] N. Minematsu *et al.*, "Structural representation of the pronunciation and its use for CALL," SLT2006, pp.126–129, 2006.

[8] D. Saito *et al.*, "Structure to speech conversion –speech generation based on infant-like vocal imitation–," Interspeech2008, pp.1837–1840, 2008.

[9] X. Ma *et al.*, "Dialect-based speaker classification of Chinese using structural representation of pronunciation," SPECOM2009, 2009 (accepted).

[10] D. Saito *et al.*, "Directional dependency of cepstrum on vocal tract length," ICASSP2008, pp.4485–4488, 2008.

[11] H. Kawahara *et al.*, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, vol.27, pp.187–207, 1999.

[1] http://www.ee.t.u-tokyo.ac.jp/gcoe/