Optimal Event Search Using a Structural Cost Function
— Improvement of Structure to Speech Conversion —

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

[1]Graduate School of Engineering, The University of Tokyo, Japan
[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. As developmental psychology claims, infants acquire the holistic sound patterns of words from the utterances of their parents, called word Gestalt, and they reproduce them 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 a word utterance was proposed. We already applied the word Gestalt to ASR, CALL, and also speech generation, which we call structure to speech conversion. Unlike reading machines, our framework simulates infants' vocal imitation. In this paper, a method for improving our speech generation framework based on a structural cost function is proposed and evaluated.

Index Terms: speech synthesis, the structural representation, vocal imitation, a structural cost function

1. Introduction

Most of the speech synthesizers are text-to-speech converters, which take a phoneme sequence as input and generate a 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 and length of their 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 phonemic representation underlies their speech communication but researchers of infant study deny this claim. This is because infants' 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 patterns 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.

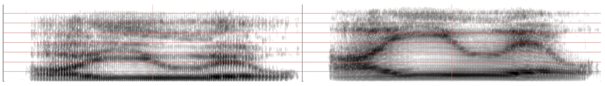


Figure 1: /aiueo/ utterances of a tall speaker and a short speaker.

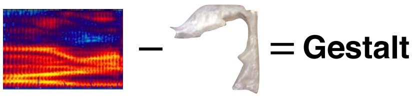


Figure 2: Speech sounds − vocal tube(size&length) = Gestalt.

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, so as to satisfy the structural constraints better, a method for improving our speech generation framework by using a structural cost function is proposed and evaluated.

2. Acoustic definition of the Gestalt

2.1. Requirements of the Gestalt

What kind of acoustical conditions should be satisfied by the word Gestalt? 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, he/she 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 derive 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

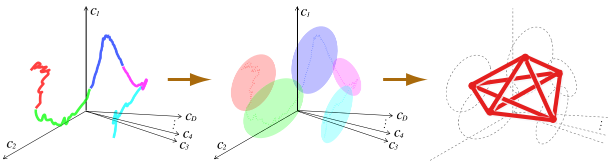


Figure 3: Invariant structuralization of an utterance.

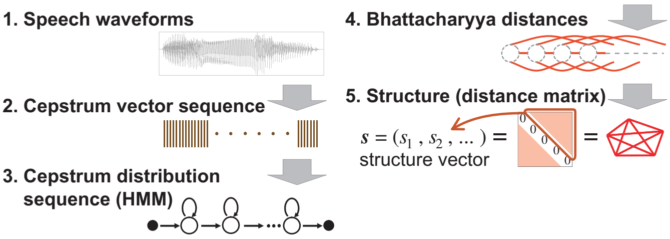

1. Speech waveforms
2. Cepstrum vector sequence
3. Cepstrum distribution sequence (HMM)
4. Bhattacharyya distances
5. Structure (distance matrix)
s = (s1 , s2 , ...) =
structure vector
=

Figure 4: Structure extraction as HMM training of an utterance.

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

One may claim that phonemic representation is a speaker-invariant or vocal-tube-invariant representation of speech. However, since infants' phonemic awareness is very immature, it is difficult for them to decompose an utterance into a phonemic sequence. From this point of view, the Gestalt of an utterance should be extracted without phonemic decomposition.

2.2. Mathematical derivation of the Gestalt

In the above section, the requirements of Gestalt were considered. It should be a speaker-invariant and holistic feature. 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. That is to say, invariance with speaker difference means mapping invariance. The distance measure of Equation 1, called f-divergence, satisfies this mathematical property. f-divergence is invariant with any kind of invertible and differential mapping [9].

$$f_{div}(p_i, p_j) = \oint p_j(x) g\left(\frac{p_i(x)}{p_j(x)}\right) dx. \qquad (1)$$

Based on this invariant feature, we introduced a transform-invariant representation of an utterance, shown in Figure 3. A sequence of cepstrum vectors is converted into a sequence of 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 Bhattacharyya distance (BD), which is a member of f-divergence family. An utterance is represented as a transform-invariant distance matrix, which can uniquely characterize a geometrical structure, i.e. a holistic pattern. We call this matrix-based representation as structural representation and believe that the structure corresponds to the Gestalt. In [5], this procedure was implemented as MAP-based HMM training for an utterance, shown in Figure 4.

Figure 3 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 are ignored completely (See Figure 2).

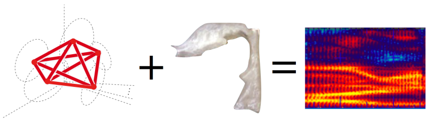


Figure 5: Structure + vocal tube(size&length) = speech sounds

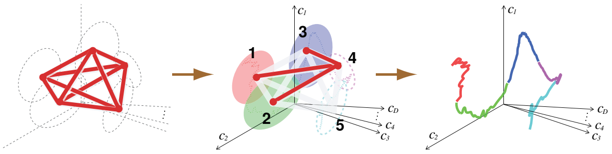


Figure 6: Search for the next target under structural constraints (a reverse process of structuralization of Figure 3).

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 3. 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 5 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. Searching a cepstrum space for target speech events

Here, conversion from a given structure to a speech sound sequence is implemented as follows. Several events of 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 events of the structure are searched for in a cepstrum space. Figure 6 shows how to search for the next target using 3 already determined events (colored ellipsoids) 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.

3.2. Geometrical solution of the problem

How do we solve this searching problem? In our previous work, a geometrical approach was adopted [8]. This section describes the previous method briefly. When the two distributions are Gaussian, i.e. $\mathcal{P}_1{=}\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{P}_2{=}\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, BD is formulated as follows,

$$BD(\mathcal{P}_1, \mathcal{P}_2) = \frac{1}{8}\boldsymbol{\mu}_{12}^t \boldsymbol{V}_{12}^{-1} \boldsymbol{\mu}_{12} + \frac{1}{2} \ln \frac{|\boldsymbol{V}_{12}|}{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}} |\boldsymbol{\Sigma}_2|^{\frac{1}{2}}}, \qquad (2)$$

where $\boldsymbol{\mu}_{12}{=}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\boldsymbol{V}_{12}{=}\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}$. In this case, BD is invariant to any common linear transform. Now let us consider a D-dimensional cepstrum space. Suppose that $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\mu}_2$ are already determined speech features and that we have to locate $\boldsymbol{\mu}_1$ in the cepstrum space using Equation 2 as structural constraint. In this case, the locus of $\boldsymbol{\mu}_1$ is found to draw a hyper-ellipsoid, ellipse in a D-dimensional space. Similarly, constraint $BD(\mathcal{P}_1, \mathcal{P}_i)$ draws another hyper-ellipsoid for $\boldsymbol{\mu}_1$.

From this fact, the intersection of multiple ellipses gives us the final solution for $\boldsymbol{\mu}_1$. In other words, solving simultaneous equations with a D-dimensional unknown vector will find a candidate for a target event. However, simultaneous equations in the quadratic form (e.g. Equation 2) with D unknowns generally have multiple solutions. For solving this ambiguity, each target was estimated by merging multiple candidates from several sets of simultaneous equations derived from the structural constraints and the initial conditions.

3.3. A structural cost function

In the previous section, we explained a primitive implementation of solving a search problem for STS. However, the above formulation has two problems. The first problem lies in simultaneous equations. Let us assume that we have to estimate a target in a D-dimensional space using m initial conditions and the structural constraints related to them. In this case, if $D > m$, the resulting simultaneous equations are ill-formed. If $D < m$, on the other hand, ${}_mC_D$ sets of simultaneous equations are possible and it takes a long computation time to solve them, especially when D is high. In addition, merging (averaging) several candidates does not guarantee an optimal solution.

The second problem is that each target was estimated independently. That is to say, when we have multiple targets, the searching in the previous section does not give us the targets that can satisfy their structural constraints fully because structural constraints between the estimated targets were ignored.

To solve these two problems, in this paper, we propose a searching method based on a structural cost function for the first problem, and stepwise updating for the second problem. Now we assume that all of covariance matrices are given, and that we have to locate a mean vector $\boldsymbol{\mu}$ of a target event from m initial conditions. We introduce a cost function $J(\boldsymbol{\mu})$ as

$$J(\boldsymbol{\mu}) = \sum_{i=1}^{m} \left(bd(\boldsymbol{\mu}, \boldsymbol{c_i}) - BD_i\right)^2, \qquad (3)$$

where BD_i is a structural constraint between the already estimated event i and the target event, and $bd(\boldsymbol{\mu}, \boldsymbol{c_i})$ represents BD between event i and $\boldsymbol{\mu}$, which is estimated so far. From Equation 2, $bd(\boldsymbol{\mu}, \boldsymbol{c_i})$ becomes

$$bd(\boldsymbol{\mu}, \boldsymbol{c_i}) = (\boldsymbol{\mu} - \boldsymbol{c_i})^t \boldsymbol{A_i} (\boldsymbol{\mu} - \boldsymbol{c_i}) + \epsilon_i, \qquad (4)$$

where ϵ_i represents the second term and $\boldsymbol{A_i}$ represents $\frac{1}{8}\boldsymbol{V}_{12}^{-1}$ in Equation 2. To acquire the optimal $\boldsymbol{\mu}$, updating equations

$$\begin{aligned} \left(\frac{\partial^2 J}{\partial \mu^2}\right) \Delta\mu &= \left.\frac{\partial J}{\partial \mu}\right|_{\mu_n} && (5) \\ \mu_{n+1} &= \mu_n - \Delta\mu, && (6) \end{aligned}$$

are used until $\Delta\mu$ becomes sufficiently small.

For the second problem, stepwise updating is used. The concept of this method is that estimated events are used as initial conditions for re-estimation. Let us assume the case of n targets and m initial conditions. As Step 1, each target is estimated independently. In Step 2, one event out of n estimated events is selected and re-estimated using the other $n-1$ estimated events as initial conditions. This step is repeated for each of the other $n-1$ events. After that, all the $n+m$ events are dealt equally. The same re-estimation process in Step 2 was repeated in 2 times.

4. Experiment

4.1. Experimental conditions

For evaluation of the proposed framework, experiments using Japanese /aiueo/ utterances were carried out. We used speech samples from 6 speakers (M1, M2 and M3 as male, and F1, F2 and F3 as female). The word Gestalt was extracted from utterances of M1 and F1, and used as structural constraints when searching for target events.

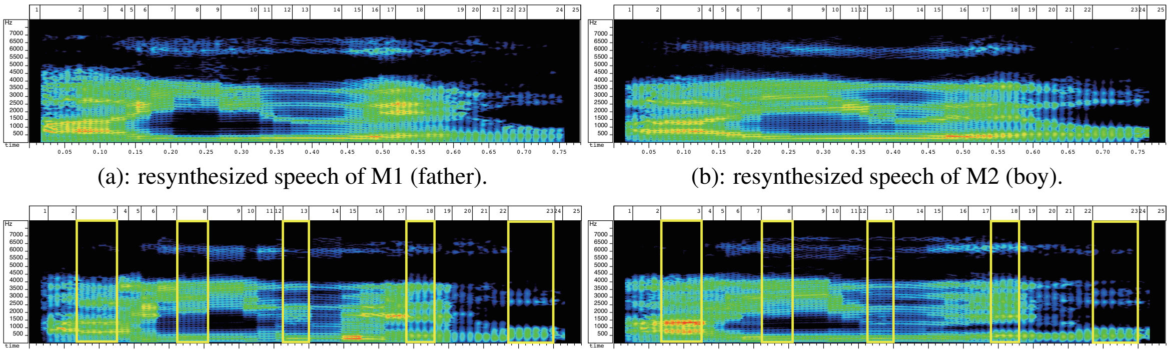
For converting a spectrum sequence to a cepstrum sequence, STRAIGHT analysis [10] was adopted and a sequence of 40 dimensional vectors was obtained. For converting a cepstrum sequence to a distribution sequence, MAP-based HMM parameter estimation was adopted since all the distributions had to be estimated 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 cepstrum stream, low dimensional sub-streams were obtained. In this experiment, The number of dimensions for each sub-stream was changed from 1 to 5. The searching problem was solved in each sub-space.

Some portions of 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 also 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.

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 vocal tubes.

4.2. Results

Figure 7 shows (a) the spectrogram of a resynthesized utterance of M1, (b) that of a resynthesized utterance of M2, and (c) and (d) are those of synthesized utterances with the M1's structure and the M2's initial conditions (the M2's imitation through the M1's Gestalt). (c) is the result from the previous method [8], and (d) is the result from the proposed one. The number of sub-streams is 40 (i.e. one-dimensional sub-streams) in (c), and 10 (i.e. four-dimensional sub-streams) in (d). In (c) and (d), the spectrum slices in five square boxes were given as initial conditions. Comparing (c) and (d) with (a) and (b) visually, we can find that spectrograms of (c) and (d) are closer to that of (b). In addition, the spectrogram of (c) includes some discontinuities, but that of (d) does not. It implies that the speaker identity is well realized in (c) and (d), and that moreover, searching based on a structural cost function effectively improves the quality of synthesized speech. We stored these four utterances in the conference CD-ROM; (a) a.wav, (b) b.wav, (c) c.wav and (d) d.wav.



(a): resynthesized speech of M1 (father).

(b): resynthesized speech of M2 (boy).

(c): Output speech synthesized with M1's structure and M2's initial conditions by using the method in [8].

(d): Output speech synthesized with M1's structure and M2's initial conditions by using the proposed method.

Figure 7: Spectrograms of resynthesized speech (a and b) and synthesized speech (c and d); (a) M1 (father), (b) M2 (boy), (c) M1's structure + M2's initial conditions (geometrical solution) and (d) M1's structure + M2's initial conditions (cost function based).

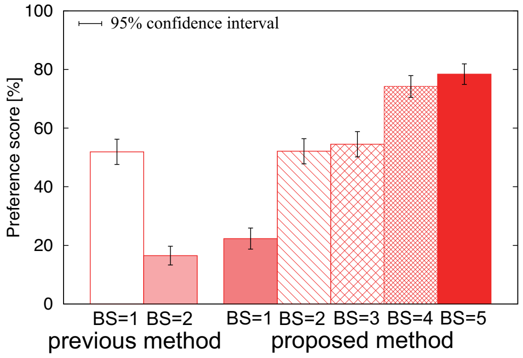


Figure 8: Results of subjective evaluation.

5. Subjective evaluation

5.1. Conditions

A listening test was carried out to evaluate naturalness of the speech samples generated by the proposed method. The test was conducted with 11 subjects to compare the utterances synthesized by the proposed method and those by our previous one [8]. All the samples for evaluation were /aiueo/ utterances. The conditions are in terms of (1) combination of 2 parents × 4 children and (2) the number of dimensions for sub-streams. In addition, for the term of (2), the samples generated by our previous method in one- and two-dimensional sub-streams were also assessed. This listening test was a paired comparison. Each subject listened to a pair of stimuli on different conditions where only the term of (2) is different between the two. After that, each subject judged which sample was more natural.

5.2. Results

Figure 8 shows preference scores of the subjective test. In Figure 8, block size (BS) means the number of dimensions for each sub-stream. From Figure 8, in the previous method [8], the higher number of dimensions degrades the quality of synthesized speech. However, in the proposed one, the quality improves when the number of dimensions is higher. Especially in the cases of BS=4 and BS=5, the preference scores of our new methods exceed those of the method [8]. In addition, computational cost of our new method is lower than that of the previous one even in the case of higher block size. This result means that it is much easier in a high dimensional space than a low dimensional space to find the optimal speech event if we derive a proper constraint, i.e. a structural cost function. On the other hand, in the previous method, degradation of the quality in the case of BS=2 is caused by the difficulty of accurate solution of simultaneous equations in a high dimensional space.

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 the sound quality of synthesized speech. One of reasons of these improvements is that a structural cost function enables us to find the optimal speech event in the high dimensional space. For more improvements of our framework, we're planning to synthesize words including consonants and to integrate the prosodic aspect into the framework.

7. 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] Y. Qiao *et al.*, "f-divergence is a generalized invariant measure between distributions," Interspeech2008, pp.1349–1352, 2008.

[10] 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.