Improved Generation of Speech from Its Abstract and Structural Representation

Nobuaki Minematsu, Daisuke Saito, and Keikichi Hirose
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
Email: {mine,dsk_saito,hirose}@gavo.t.u-tokyo.ac.jp

***Abstract*—This paper describes an improved method for the framework of structure-to-speech conversion we proposed previously. This framework aims at building a speaking machine by simulating infants' language acquisition. Most of the speech synthesizers take a phoneme sequence as input and convert it to speech sounds, i.e. reading machines. Infants initially acquire speech communication capacity without phonemes or reading. Since their phonemic awareness is very immature, young children can hardly decompose an utterance into a sequence of phonemes but they enjoy speech communication with their parents. As developmental psychology claims, infants acquire the holistic sound patterns that underlie individual utterances, called word Gestalt. Infants reproduce this sound pattern using their very short vocal tubes, i.e. vocal imitation. In our previous studies, the word Gestalt was defined mathematically, called speech structure, and a method of extracting it from a word utterance was proposed and applied to ASR and CALL. Further, a reverse process, i.e. structure-to-speech conversion was realized. In this paper, a method of improving our speech generation framework based on a structural cost function is proposed and evaluated.**

keywords-component; speech synthesis; structural representation; invariance; vocal imitation; a structural cost function

I. Introduction

Most of the speech synthesizers are text-to-speech converters, i.e. reading machines. To enable humanoid robots to speak, text-to-speech converters are often embedded. Here, unless text or written form is provided, they have to be quiet or dumb. In contrast, human children are very chatty even long before they are able to use the written form of language.

If a speech synthesizer is build with a speech database of speaker A, then, the synthesizer will generate the voices of that speaker. Developmental psychology states that infants acquire language through imitating utterances of their parents. It is obvious, however, that they never impersonate their parents. Animal sciences tell that the vocal imitation of animals, which is found in birds, dolphins, and whales, are acoustic imitation like impersonation [1]. Even in the case of humans, however, the vocal imitation is acoustic when the performance of severely impaired autistics is observed [2], [3], who have much difficulty to acquire normal speech communication capacity.

Reviewing the findings of developmental psychology, animal sciences, and language disorders, we can say that a reading machine based on acoustic imitation is not a good option to build a *not externally but internally* human-like robot [4], [5].

What acoustic aspects of a parent's utterances, does an infant imitate? What is equivalent between a parent's utterance and an infant's imitative response? Word Gestalt [6], which is a term used in developmental psychology, represents a common sound pattern underlying both utterances but no psychologist explains it using equations. What can be said at least is that the word Gestalt has to be independent of the age, gender, size, etc of speakers. It must be a very abstract representation.

Fig. 1. Speech sounds − vocal tube (size & length) = Gestalt.

Recently, we proposed a mathematical definition of the word Gestalt [7]. Our method of extracting the Gestalt from an utterance was introduced successfully to ASR and CALL [4], [8]. In addition, we realized an inverse process, i.e. generating a speech stream from its abstract and structural representation, called structure-to-speech (STS) conversion [9]. However, formulation was insufficient for complete implementation of STS. In this paper, in order to satisfy the structural constraints better, a method of improving our generation framework is proposed by using a structural cost function.

II. Acoustic definition of the Gestalt

As we mentioned above, the Gestalt is an abstract pattern underlying an utterance, which is independent of the extra-linguistic factors such as the vocal tract length (See Figure 1).

One may claim that a phonemic representation is also a speaker-independent representation. However, since infants' phonemic awareness is very immature, it is difficult for them to decompose an utterance into phonemes [10]. We consider that the phonemic representation is not a good option if one wants to realize a human-like speaking module for humanoids.

In many studies of voice conversion, it is assumed that speaker differences are well modeled as space mapping. This indicates that invariance with speaker difference means mapping invariance. The distance measure of Equation 1, called f-divergence, satisfies this mathematical property. It is invariant with any kind of invertible and differential mapping [11].

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

Based on this invariant property, we introduced a transform-invariant representation of an utterance, shown in Figure 2. A

978-1-4244-5899-8/10/$26.00 ©2010 IEEE

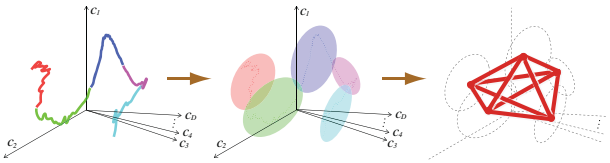


Fig. 2. Invariant structuralization of an utterance.

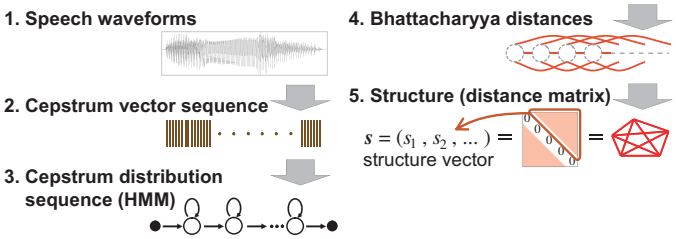


Fig. 3. Structure extraction as HMM training of an utterance.

Fig. 4. Structure + vocal tube (size & length) = speech sounds

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 the f-divergence family. An utterance is represented as a transform-invariant distance matrix, which can uniquely characterize a geometrical shape, i.e. a holistic pattern of that utterance. We call this distance matrix as speech structure and believe that this structure corresponds to the Gestalt. In [4], this procedure was implemented as MAP-based HMM training for an utterance, shown in Figure 3.

Figure 2 shows that a speech structure is estimated 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 1).

The structure is so abstract a representation of an utterance that, with it only, speech sounds cannot be recovered or determined at all, shown in Figure 2. To determine and locate the sounds of a given structure, what should be additionally needed? Looking at Figure 1, we can say that the static and default shape of the vocal tube is required for the Gestalt to be realized acoustically. Figure 4 explains this process conceptually and, in the following section, this process of structure-to-speech conversion is implemented on computers.

III. Structure to Speech Conversion

A. 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

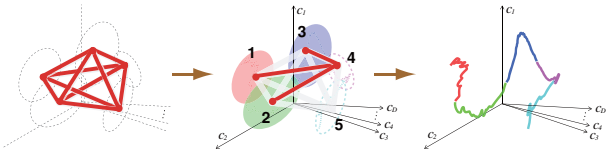


Fig. 5. Search for the next target under structural constraints.

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 5 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.

B. Geometrical solution of the problem

How do we solve this searching problem? In our previous work, a geometrical approach was adopted [9]. This section describes the previous method briefly. When 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 between them 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 $\boldsymbol{\mu}_{12}{=}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\boldsymbol{V}_{12}{=}\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}$. BD is invariant to any common linear and non-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 $BD(\mathcal{P}_1, \mathcal{P}_i)$ $(i \neq 2)$ draws an i-th 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.

C. Optimization using a structural cost function

The above formulation, which was proposed in [9], 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 each set, especially when D is high. Further, merging (averaging) several candidates does not always guarantee an optimal solution.

The second problem is that each target is estimated independently. Then, when we have multiple targets, the searching method in [9] does not give us the targets that can satisfy their structural constraints fully because the structural constraints among the estimated targets are ignored.

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

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

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

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

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

$$\left(\frac{\partial^2 J}{\partial \mu^2} \right) \Delta\mu = \left. \frac{\partial J}{\partial \mu} \right|_{\mu_n} \tag{5}$$

$$\mu_{n+1} = \mu_n - \Delta\mu, \tag{6}$$

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

For the second problem, stepwise updating is adopted. The concept of this method is that already estimated events are used as initial conditions for reestimation. 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 the n estimated events is selected and reestimated using the other $n-1$ estimated as initial conditions. This step is repeated for each of the other $n-1$ events. Finally, all the $n+m$ events are dealt equally, i.e. a target and $n+m-1$ initials. The same reestimation process in Step 2 was repeated twice.

IV. Experiment

A. Experimental conditions

To evaluate the proposed framework quantitatively, 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.

To convert a spectrum sequence to a cepstrum sequence, STRAIGHT analysis [12] 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 [4] was carried out. From a single cepstrum stream, low dimensional sub-streams were formed. 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 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.

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 are a father and a mother and target speakers are sons and daughters, who try to extract the word Gestalt underlying their parents' utterance and reproduce it acoustically using their vocal tubes.

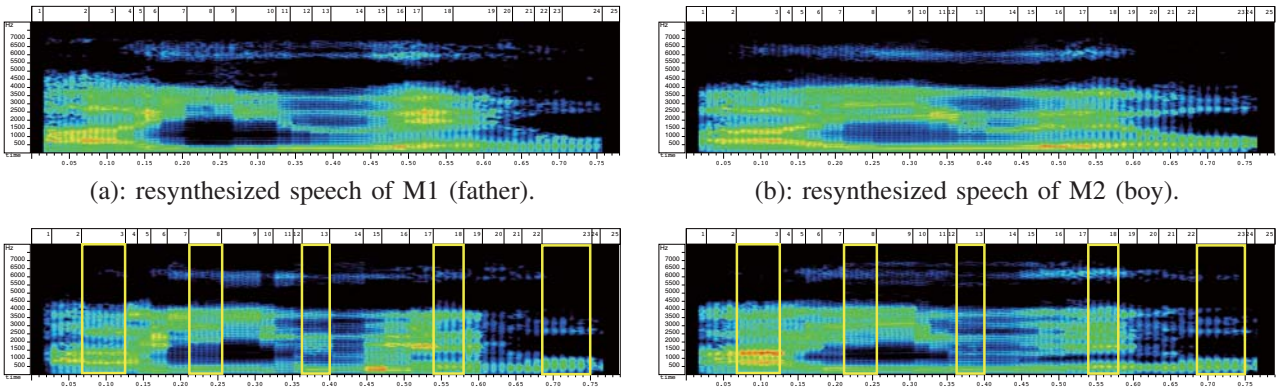
B. Results

Figure 6 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 M1's structure and M2's initial conditions (M2's imitation through M1's Gestalt). (c) is a result by the previous method [9], and (d) is a corresponding result by the proposed method. The number of sub-streams is 40 (one-dimensional sub-streams) in (c) and 10 (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 the 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 speaker identity is well realized in (c) and (d) and that a structural cost function effectively improves the quality.

V. Subjective evaluation

A. Conditions

A listening test was carried out to evaluate the naturalness of speech samples generated by the proposed method. The test was conducted with 11 subjects with normal hearing, who compared the utterances synthesized by the proposed method and those by our previous method [9]. All the samples for evaluation were /aiueo/ utterances and they were synthesized under different conditions: (1) combination of 2 parents $\times$ 4 children and (2) the number of dimensions for sub-streams. When synthesizing utterances by our previous method, the number of dimensions for sub-streams were one or two. The listening test was a paired comparison. Each subject listened to a pair of stimuli synthesized on different conditions where only the term of (2) is different. Then, he/she was asked to judge which of the two samples was more natural.



(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 [9].

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

Fig. 6. 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 solution).

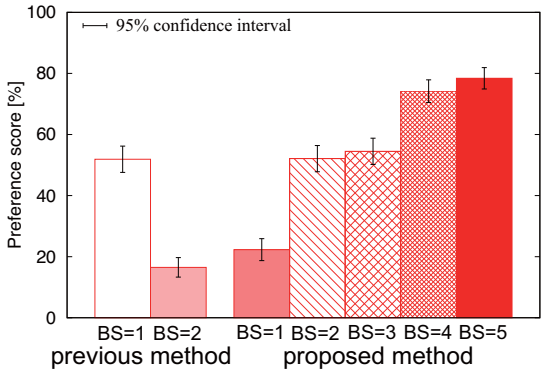


Fig. 7. Results of subjective evaluation.

B. Results

Figure 7 shows preference scores of the subjective test. In Figure 7, the block size (BS) means the number of dimensions for each sub-stream. From Figure 7, in the previous method [9], the higher number of dimensions degrades the quality of synthesized speech. In the proposed method, however, 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 [9]. In addition, computational cost of our new method is lower than that of the previous one even in the case of larger block sizes. This result means that it is easier in a high dimensional space than a low dimensional space to find the optimal speech event by using a proper constraint, i.e. a structural cost function. On the other hand, in the previous method, the quality is degraded in the case of $BS{=}2$ due to the difficulty of accurate solution of simultaneous equations in a high dimensional space.

VI. Conclusions

We have proposed a new method for the framework of structure-to-speech conversion. In this framework, the word Gestalt is extracted from an input utterance and reproduced acoustically with some initial conditions given. Here, the linguistic aspect of the input utterance retains but the extra-linguistic aspect changes. This performance is very similar to that of voice conversion, where speaker identity is changed with the linguistic content unchanged. Our method, however, has an internal and abstract representation or model of an utterance, called speech structure, that can be used directly for ASR and CALL. This is a significant difference between structure-to-speech conversion and voice conversion. This framework can also simulate infants' vocal imitation.

A method proposed in this paper has improved the sound quality of synthesized speech. One of the reasons of this improvement is that a structural cost function makes it possible to find the optimal speech event in a high dimensional space more efficiently. For further improvement of our framework, we're planning to synthesize words including consonants and to integrate the prosodic aspect into the framework.

References

[1] K. Okanoya, "Birdsongs and human language: common evolutionary mechanisms," *Proc. Spring Meet. Acoust. Soc. Jpn.*, 1-17-5, 1555–1556, 2008. (including Q&A after his presentation)

[2] T. Grandin *et al.*, *Emergence: labeled autistic,* Grand Central Publishing, 1996.

[3] L. H. Willey *et al.*, *Pretending to be normal: living with Asperger's syndrome,* Jessica Kingsley Publishers, 1999.

[4] N. Minematsu *et al.* "Speech structure and its application to robust speech processing," *Journal of New Generation Computing*, 28, 299–319, 2010.

[5] N. Minematsu, "Human speech model based on information separation," *Proc. Electronic Speech Signal Processing*, 2010.

[6] M. Hayakawa, "Language acquisition and matherese," *Language*, 35, 9, 62–67, Taishukan pub., 2006.

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

[8] M. Suzuki *et al.*, "Sub-structure-based estimation of pronunciation proficiency and classification of learners," *Proc. ASRU*, 574–579, 2009.

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

[10] M. Kato, "Phonological development and its disorders," *J. Communication Disorders*, 2, 20, 98–102, 2003.

[11] Y. Qiao *et al.*, "A study on invariance of f-divergence and its application to speech recognition," *IEEE Trans. on Signal Processing*, 58, 7, 3884–3890, 2010.

[12] 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*, 27, 187–207, 1999.