Structure to Speech Conversion
– Speech Generation Based on Infant-like Vocal Imitation –

Daisuke Saito[1], *Satoshi Asakawa*[2], *Nobuaki Minematsu*[1], *Keikichi Hirose*[3]

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

Abstract

This paper proposes a new framework of speech generation by imitating "infants' vocal imitation". Most of the speech synthesizers take a phoneme sequence as input and generate speech by converting each of the phonemes into a sound sequentially. In other words, they simulate a human process of reading text out. However, infants usually acquire speech generation ability without text or phoneme sequences. Since their phonemic awareness is very immature, they can hardly decompose a word utterance into a sequence of phones. 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 and used successfully for ASR and CALL. In this paper, a method of converting the word Gestalt back to speech is proposed and evaluated. Unlike a reading machine, our proposal simulates infants' vocal imitation.

Index Terms: speech synthesis, vocal imitation, word Gestalt, invariant structure, Bhattacharyya distance, searching problem

1. Introduction

Most of the speech synthesizers are text-to-speech converters, which take a phoneme sequence as input and generate speech sounds 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 imitate the voices of their parents. It is impossible for infants to create their parents' voices due to a 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 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.

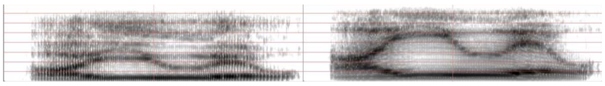


Figure 1: /aiueo/ utterances of a tall speaker and a short speaker

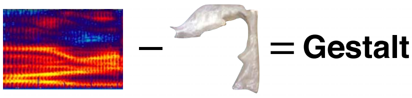


Figure 2: Speech sounds − vocal tube(size&length) = Gestalt

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 this paper, a new method of converting the Gestalt back to speech sounds is proposed. Two processes of conversion from an utterance to its Gestalt and that from the Gestalt to its acoustic version are implemented. We consider that they simulate infants' vocal imitation well.

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 small 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, 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. Here, 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

Accepted after peer review of full paper

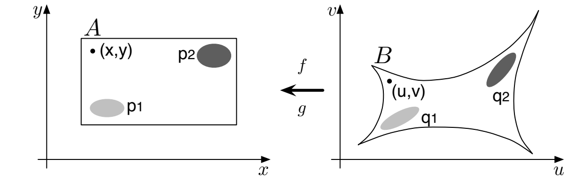


Figure 3: Linear or non-linear mapping between two spaces

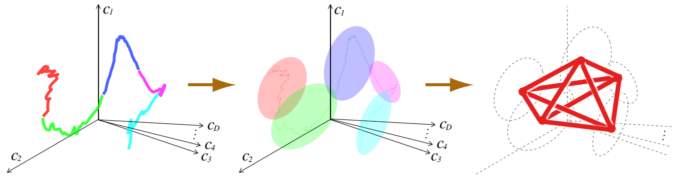


Figure 4: Invariant structuralization of an utterance

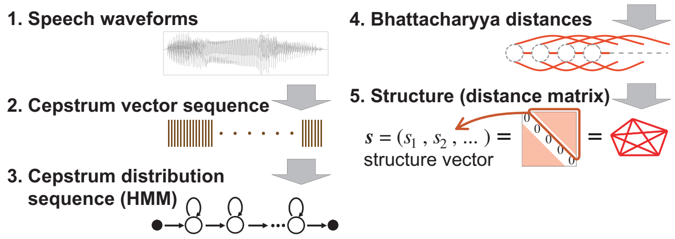

1. Speech waveforms
2. Cepstrum vector sequence
3. Cepstrum distribution sequence (HMM)
4. Bhattacharyya distances
5. Structure (distance matrix)
s = (s1 , s2 , ...) =
structure vector
0
0
0
0
=

Figure 5: Feature extraction as HMM training for an utterance

an example of invertible mapping (linear or nonlinear) between spaces A and B. In both spaces, 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;

$$BD(p_1, p_2) = -\log \oiint \sqrt{p_1(x,y)p_2(x,y)}dxdy$$

$$= -\log \oiint \sqrt{p_1(f(u,v),g(u,v)) \cdot p_2(f(u,v),g(u,v))}|J|dudv$$

$$= -\log \oiint \sqrt{p_1(f(u,v),g(u,v))|J| \cdot p_2(f(u,v),g(u,v))|J|}dudv$$

$$= -\log \oiint \sqrt{q_1(u,v)q_2(u,v)}dudv = BD(q_1,q_2).$$

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 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 larger than the number of phonemes existing in the utterance. We already applied this representation in ASR [5] and CALL [7] successfully.

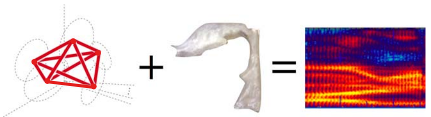


Figure 6: Structure + vocal tube(size&length) = speech sounds

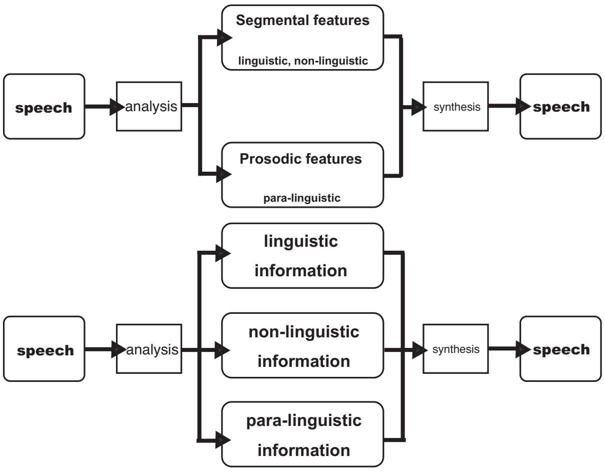


Figure 7: The conventional framework for analysis-resynthesis and the proposed one with three separate kinds of information

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 considered? 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. STRAIGHT [8] realizes very high quality in its resynthesized 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 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

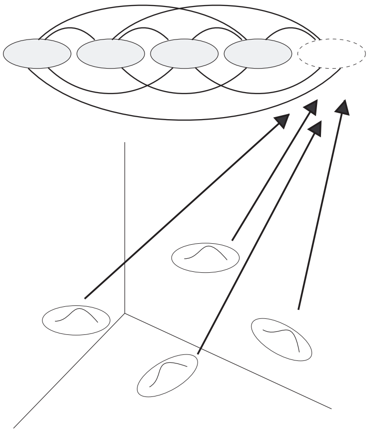


Figure 8: Search for the next target under structural constraints

with speech dynamics. To this question, our previous works showed an answer and it is possible. With the speaker-invariant speech structures, speaker-independent speech recognition was realized successfully only with several training speakers [5, 6].

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 a difference in the vocal tract length rotates a given speech structure [9]. 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 some of the 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 generate through vocal communications or playing with their parents.

3.3. Solving the search problem

How do we solve this searching problem? When the two distributions are Gaussian, i.e. $p_1(x) = \mathcal{N}(\mu_1, \Sigma_1)$ and $p_2(x) = \mathcal{N}(\mu_2, \Sigma_2)$, BD is formulated as follows,

$$BD(p_1(x), p_2(x)) = \frac{1}{8}(\mu_1 - \mu_2)^T V_{12}^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\ln\frac{|V_{12}|}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}}, \quad (1)$$

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 Σ_1, Σ_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, ellipsis in an n-dimensional space. From this fact, we take the following procedure to solve the search problem.

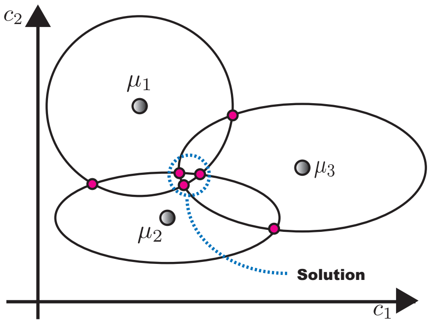


Figure 9: Solution of the search problem. In this figure, the intersection of three ellipses becomes the solution.

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 $A{=}\mathcal{N}(a, V_a)$ and $B{=}\mathcal{N}(b, V_b)$ are prepared for initial conditions, where covariance matrices of A and B are supposed to be diagonal. Speech event $C{=}\mathcal{N}(\mu, V)$ is a target, where V is also diagonal. When BD between A and C is named as BD_a and BD between B and C is named as BD_b, the structural constraint is translated into a simultaneous equation as

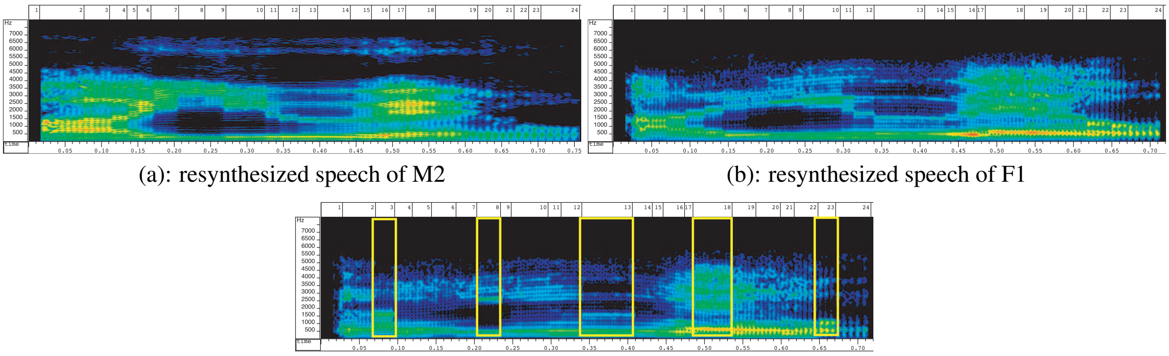
$$\begin{cases} BD_a - \epsilon_a = \displaystyle\sum_{d\in\{x,y\}} \frac{1}{4(V_d+V_{ad})}(c_d - a_d)^2 \\ BD_b - \epsilon_b = \displaystyle\sum_{d\in\{x,y\}} \frac{1}{4(V_d+V_{bd})}(c_d - b_d)^2, \end{cases} \quad (2)$$

where indices x and y correspond to each dimension and ϵ represents the second term in Equation 1. 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 only one intersection for the target speech event, at least one more event is needed as initial condition. By expanding this discussion to a n-dimensional space, we can say that we need at least $n{+}1$ events as initial condition. Figure 9 shows an two dimensional case. The target event is obtained as intersections of three ellipses, whose origins are speech events given as initial conditions.

4. Experiment

4.1. Experimental conditions

For initial evaluation of the proposed framework, experiments using /aiueo/ utterances were carried out. We used speech samples from 3 speakers (M1 and M2 as male and F1 as female). An utterance of M2 was 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,



(a): resynthesized speech of M2

(b): resynthesized speech of F1

(c): synthesized speech with M2's structure and F1's initial conditions

Figure 10: Spectrograms of resynthesized speech (a and b) and synthesized speech (c)
(a) M2 (father), (b) F1 (girl), and (c) M2's structure + F1's initial conditions

STRAIGHT analysis [8] 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 and a structure was extracted from each dimension (From a single speech stream, 40 multiple sub-streams were obtained). It means that the searching problem was solved in each dimension.

The other two utterances from M1 and F1 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 M1 and F1 were also used as initial conditions. With these initial conditions of M1 and F1 and the structural constraints from M2, 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 M1 and F1 were synthesized. When we consider this experiment and infants' vocal imitation, M2 is a father and M1 and F1 are a boy and a girl, who try to extract the word Gestalt in their father's utterance and reproduce it acoustically using their short vocal tubes.

4.2. Results and discussions

Figure 10 shows (a) the spectrogram of a resynthesized utterance of M2 (father), (b) that of a resynthesized utterance of F1 (girl), and (c) that of a synthesized utterance with the girl's initial conditions (the girl's imitation through the father's Gestalt). In (c), the spectrum slices in five square boxes were given as initial conditions. Although an objective listening test was not done yet, when we compare (c) with (a) and (b) visually, we can find that spectrogram of (c) is closer to that of (b). This means that the speaker individuality is well realized in (c). This was verified through listening. We listened to three /aiueo/ utterances in Figure 10. We can say that it is very easy to recognize that utterance of (c) is generated by F1 and that its linguistic content is /aiueo/. We stored these three utterances in the conference CD-ROM; (a) gestalt.wav, (b) initial.wav, and (c) proposed.wav. We believe that readers accept our judgment well. Although this experiment is very small and rather preliminary, we can say that structure-to-speech conversion certainly works.

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, the primates other than humans 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.

5. Conclusions

We have proposed a new framework of speech generation based on the structural representation of speech. The proposed framework extracts the word Gestalt from an input utterance and reproduce it acoustically with some initial conditions given. This framework can simulate infants' vocal imitation and learning. As a future work, we're planning to integrate the prosodic aspect into the framework and to examine whether this framework can generate speech sounds of a variety of speaker individuality.

6. References

[1] T. Masuko *et al.*, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," *ICSLP2000*, vol. 2, pp. 302–305, 2000.

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

[3] M. Kato, "Phonological development and its disorders," *J. Communication Disorders*, vol. 20, no. 2, 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," *ICASSP 2008*, 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] 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.

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