ICSP2008 Proceedings

Experimental Study of Structure to Speech Conversion

-- An implementation of Infant-like Vocal Imitation on a Machine --

Nobuaki Minematsu, Daisuke Saito, Keikichi Hirose
The University of Tokyo
{mine,dsk_saito,hirose}@gavo.t.u-tokyo.ac.jp

Abstract

Most of the speech synthesizers have been developed as text (phoneme sequence) to speech converters and, in this framework, text input is a precondition for speech production. However, we can say that no child acquires spoken language by reading a given text out. Children are explained to acquire spoken language by imitating the utterances of their parents but they never imitate the voices of their parents. Developmental psychology claims that they extract a holistic and speaker-invariant sound pattern embedded in a given utterance, called word Gestalt, and realize the pattern acoustically using their short vocal tubes. In our previous studies, we mathematically defined this holistic and speaker-invariant pattern and used it for ASR [1,2,3,4]. Here, we experimentally implement its inverse process, i.e. Gestalt-to-utterance conversion, on a computer.

1. Introduction

Vocal imitation is found only in a very few kinds of animals. No other primates than humans perform vocal imitation [5]. This performance can be found in some species of birds and whales, but their imitation is basically the imitation of sounds [6]. For example, myna birds imitate many sounds such as cars, doors, dogs, cats as well as human voices. Hearing an adept myna bird say something, one can guess its owner [7]. Hearing the voices of an infant, however, it is impossible to guess its parents. This fact indicates that human vocal imitation is performed beyond scaling. Developmental psychology explains that children extract a holistic and speaker-invariant sound pattern embedded in a given utterance, called word Gestalt, and realize the pattern acoustically with their very short vocal tubes [8,9,10].

Word Gestalt can be considered to be a speech representation where the actual size of the vocal tract of a speaker is canceled. It is drawn conceptually in Fig. 1. We already defined word Gestalt mathematically and used it for ASR, where speaker-independent speech recognition was implemented only with several training speakers. Here in this study, Gestalt-to-utterance conversion is studied and implemented on a computer.

978-1-4244-2179-4/08/$25.00 ©2008 IEEE

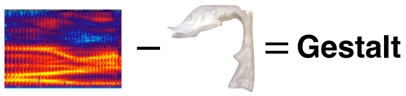


Figure 1: Utterance – vocal tract size = Gestalt

2. Acoustic definition of the Gestalt

2.1. Speaker-invariant speech pattern

In speaker conversion studies, it is often assumed that speaker difference is well modeled as space mapping, shown in Fig. 2. We already found that mapping-invariant features can be extracted from a speech stream if the mapping is invertible and differentiable everywhere. Fig. 3 shows a procedure of extracting a speaker-invariant and holistic sound pattern embedded in a given utterance. A cepstrum trajectory is converted into a sequence of distributions by merging acoustically similar frames. Then, the Bhattacharyya distances (BD) between any pair of distributions is calculated to form a BD-based distance matrix. Since a distance matrix can determine a unique shape of geometrical structure, the above BD-based matrix also has its own shape. BD is an invariant measure with any kind of invertible mapping including non-linear mappings. The BD-based structure can be robustly speaker-invariant.

Using this structural representation only, which has no absolute speech features, we showed that remarkably robust speech recognition is possible [1,2,3,4].

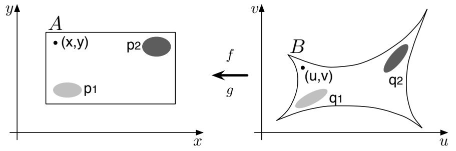


Figure 2: Mapping between space A and space B

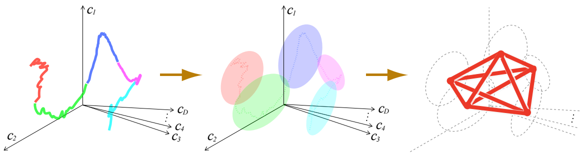


Figure 3: Speaker-invariant and holistic structure

2.2. Experimental verification of the invariance

[4] verified the invariance experimentally and showed the super robustness of the proposed framework for ASR. The task was recognizing isolated words and the words were defined as sequences of 5 Japanese vowels such as /aeoui/. It is well-known that vowel sounds are much more dependent on speakers than consonant sounds. Since Japanese has only 5 vowels, the vocabulary size was 120. Utterances of 4 male and 4 female adult speakers were used to train two recognizers, one was with word-HMMs and the other was with structures. The difference between the two kinds of acoustic models is what was modeled acoustically. The former modeled speech substances statistically and the latter modeled speech contrasts statistically. Another set of utterances of other 4 male and 4 female speakers were used as testing data. The performance of the two recognizers is shown in Tab. 1. In addition to the original test utterances of the 8 adult test speakers, we used the utterances obtained by applying spectrum warping to the original ones. By warping, we can simulate the utterances of boys and girls. In the table, the average body height (ft) is shown. It should be noted that no adaptation technique was used in the two recognizers. The table clearly shows the invariance of structures and the super-robustness of structure-based ASR.

Table 1: Comparison between HMMs and structures

height (ft)	5.30	4.97	4.64	4.31	3.98	3.64	3.31	2.98	2.65
HMMs	83.9	78.2	63.1	44.5	24.8	8.85	1.88	1.00	0.67
structures	86.4	86.8	87.2	88.1	88.0	87.9	88.7	89.2	88.9

Chance level = 1/120 = 0.86 %

3. Structure to speech conversion

3.1. A basic concept

An utterance is converted into its structure, which is a holistic and speaker-invariant representation of the utterance. If we want to convert the structure back to sounds again, we have to specify the size of the vocal tract based on who will realize this structure in an actual sound (voice) space. This process is visualized conceptually in Fig. 4. Given a structure, dynamic motions of articulatory organs may be able to be generated. In this study, however, this process is not implemented as it is due to difficulty of constructing an articulatory acoustic interface. This paper aims at realizing a structure in an actual sound space without that interface. However, it is true that, only with a structure, this process is impossible to execute because a structure contains no information at all on where in the sound space each event (distribution) of a BD-based matrix should be realized. Here, we introduce acoustic instances of a few events of the matrix as initial conditions. Then, using an entire matrix as constraint conditions, the remaining events are realized acoustically.

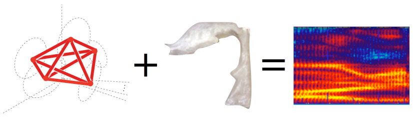


Figure 4: Structure + vocal tract size = utterance

This process is explained in Fig. 5. Suppose that four events are already realized in a voice space. The next event is searched for in this space by considering structural constraints among the new event and the four already realized events. In the case of children's vocal imitation, the structural constraints are given from their parents. About the initial conditions, children may use some speech sounds that they already produced in vocal communication with their parents.

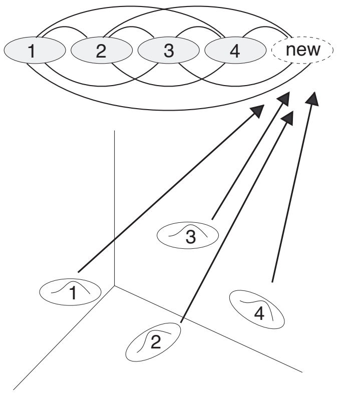


Figure 5: Search for the next under structural constraints

3.2. Space search for the new target

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,

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

where $V_{12}=\frac{\Sigma_1+\Sigma_2}{2}$. In this case, BD is invariant to any kind of common 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 the above equation 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.

1. From the distance matrix, equations of the hyper-ellipsoid are obtained.
2. Vectors of the initial conditions are substituted in 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 the target, where V is also diagonal. When BD between A and C is referred to as BDa and BD between B and C is as BDb, the structural constraint is translated into a simultaneous equation as

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

where indices x and y correspond to each dimension and ϵ represents the second term in the definition of BD. In a two-dimensional case, solution of the above equation geometrically corresponds to calculation of the intersection of two ellipses. Generally speaking, the number of intersections of two ellipses is more than one. 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 an n-dimensional space, we can say that we need at least *n+1* events as initial condition. Fig 6. shows an example in a two-dimensional space. The target event is obtained as intersections of three ellipses, the origins of which are speech events given as initial conditions.

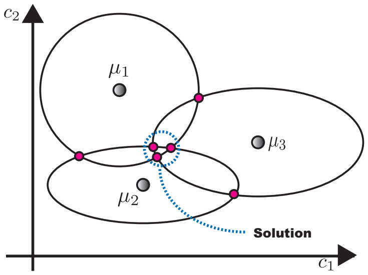


Figure 6: Solution of the search problem. In this figure, the intersection of three ellipses becomes the solution.

4. Experiments

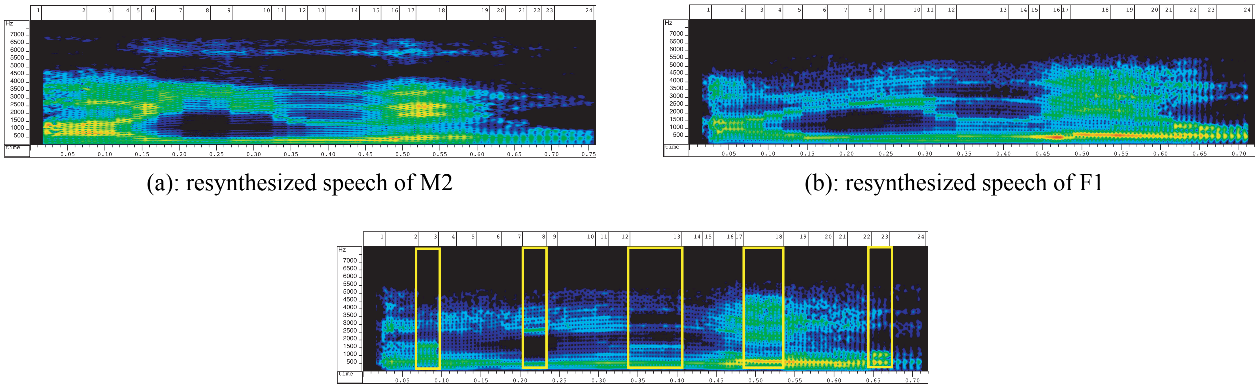
4.1. Experimental conditions

For initial assessment 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, corresponding to the structural constraints considered 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 [3] 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 dealt with in each dimension.

A part of 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 sequences of 25 diagonal Gaussians. After that, 5 mean vectors (3rd, 8th, 13th, 18th, and 23rd ones in the 25 Gaussians) were extracted and used as 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. Using the prosodic features extracted above and a sequence of the estimated distributions, utterances of M1 and F1 were synthesized acoustically. When we compare 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

Fig. 7. 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 a listening test was not done yet, when we compare (c) with (a) and (b) visually, we can find that the spectrogram of (c) is closer to that of (b). This means that speaker individuality is well realized in (c). This was verified through listening. We listened to three /aiueo/ utterances in Fig. 7. We can say that it is very easy to recognize that (c) is generated by F1 and that its linguistic content is /aiueo/. We stored these



(a): resynthesized speech of M2

(b): resynthesized speech of F1

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

Figure 7: 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

three utterances in the conference CD-ROM; (a) gestalt.wav, (b) initial.wav, and (c) proposed.wav. We believe that readers accept our judgment. Although this experiment is very small and preliminary, we can say that structure-to-speech conversion certainly works.

This paper has tried 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. As described in Section 1, it is known that the vocal imitation done by animals is the imitation of sounds. It is only humans that do not imitate the sounds. Our imitation is beyond scaling. As far as we know, all the speech synthesizers developed so far imitate the sounds (voices). It is true that, hearing a good speech synthesizer say something, one can guess its training speaker. This resembles animal-like imitation well. We can say that our synthesizer is the only one that 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] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," Proc. ICASSP, pp.889-892 (2005)

[2] Y. Qiao, et al., "Random discriminant structure analysis for continuous Japanese vowel recognition," Proc. ASRU, pp.576-581 (2007)

[3] S. Asakawa, et al., "Multi-stream parameterization for structural speech recognition," Proc. ICASSP, pp.4097-4100, (2008)

[4] S. Asakawa, "Speech recognition with super robustness to speaker variability based on discriminant analysis and speech structures," Proc. Autumn Meeting of Acoust. Soc. Jpn., 2-P-3 (2008)

[5] W. Gruhn, "The audio-vocal system in sound perception and learning of language and music," Proc. Int. Conf. on language and music as cognitive systems (2006)

[6] K. Okanoya, "Birdsong and human language: common evolutionary mechanisms," Proc. Spring Meeting of Acoust. Soc. Jpn., 1-7-15, pp.1555-1556 (2008)

[7] K. Miyamoto, Making voices and watching voices, Morikawa Pub. (1995)

[8] M. Hayakawa, "Language acquisition and matherese," In: Language, Taishukan Pub., vol.35, no.9, pp.62-67 (2006)

[9] S. E. Shaywitz, Overcoming dyslexia, Random House (2005)

[10] K. Hara, "Phonological disorders and phonological awareness in children," J. Communication Disorders, vol.20, no.2, pp.98-102 (2003)

[11] H. Kawahara et al., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," Speech Communication, vol.27, pp.187-207 (1999)