Generative model of spectra for a word using Fujisaki model and genetic algorithm

Tomio Takara*, Akichika Higa*, Ryoichi Eto* and Go Ishikawa*

University of the Ryukyus,
1 Aza-Senbaru, Nishihara-cho, Okinawa, 903–0213 Japan

(Received 1 June 2017, Accepted for publication 20 October 2017)

Keywords: Fujisaki model, Spectrum, Genetic algorithm, Word acquisition, Imitation

PACS number: 43.72.+q [doi:10.1250/ast.39.147]

1. Introduction

Human infants learn articulatory behaviors such as the tongue hump position and the degree of constriction, which are human internal information, by controlling their speech organs and imitating other people’s speech [1]. This is the first step of human communication because it means that not only the other people’s internal information but their intention is known.

We propose a new model of this speech imitation process of infants [2]. We regard the vowel space parameters [3] as the tongue hump position and the degree of constriction of vowels. We represent the time-varying pattern of the vowel space parameters using Fujisaki’s generative model of speech [4]. We model a trial and error process of the imitation using the genetic algorithm (GA) [5].

In our model, “command” in the Fujisaki model is detected from the spectral sequence using the GA. In other words, the original phonemic information is extracted from an ambiguous spectrum by representing the coarticulation in the Fujisaki model. Our model also expresses the hypothesis that human infants acquire the normalization skill of coarticulation through the process of imitating spoken words.

2. Speech analysis and synthesis using vowel space parameters

We proposed vowel space parameters for speech synthesis. The vowel space parameters are the principal components obtained by principal analysis of the log spectra of isolated vowels. Figure 1 shows the distribution of the vowel space parameters of isolated vowels spoken by a male speaker. From this figure, we know that the first and second vowel space parameters can be regarded as the tongue hump position and the degree of constriction of vowels, respectively, because of their position.

The log spectrum can be approximately reconstructed by a linear combination of a few principal vectors of vowel space parameters. We can synthesize a spoken word by reconstructing the time sequence of the log spectra from the vowel space parameters.

3. Fujisaki model

The Fujisaki model is an effective model for approximating the contour of the fundamental frequency precisely for the source model of speech synthesis. Saito et al. proposed an expression representing the coarticulation of formant frequencies using the Fujisaki model [6]. Akagi et al. introduced an equation for a target model of speech perception and showed that the resolution of this equation is the same as the expression of the Fujisaki model [7].

Following these studies, we approximate the time pattern of the vowel space parameters by the Fujisaki model. Figure 2 shows an example of a command and the resulting contour of the Fujisaki model.

4. Generation of Fujisaki’s contour using genetic algorithm

We executed the GA in which Fujisaki’s parameters are adopted as genes. The fitness was defined as the similarity between the generated contour of the Fujisaki model and the target time pattern of the vowel space parameter. For the source information, such as voiced/unvoiced information, the fundamental frequency, and power, we used the analyzed parameters themselves. In these processes, speech spectra are imitated and a spoken word is synthesized.

Figure 3 shows the genes of our algorithm. They are real numbers representing the value and starting time for each phoneme of a word. For the value, as the crossover, random numbers between the values of two genes are adopted as new genes. As the mutation, a random number between −20% and +20% values of the former one is adopted. For the starting time, as the crossover, the average value of two genes is adopted as a new gene. The mutation, a random number between 80% of the former value and the next initial starting time is adopted. We set the number of phonemes manually, and the initial starting times of the phonemes are set at equal intervals. The GA is executed for each vowel space parameter.

Figure 4 shows the average fitness (difference) of 100 trials to the original vowel space parameter of the best gene at each generation and the best and the fifth best genes at the initial generation. They are also expressed as spectral errors in dB from the original spectrum. The spectral error of the seven dimensional vowel space parameters is also shown. In this evaluation, the word was /hinokuruma/, whose number of phonemes was 10. The number of individuals per generation was 10. The mutation rate was 0.6. We adopted the elite strategy.

Figure 5 shows the target vowel space parameter and the generated Fujisaki model contour of the first component. The word was /hinokuruma/ uttered by a male. The contour had

*e-mail: takara@ie.u-ryukyu.ac.jp
the best fitness among 40 individuals at the third generation, whose commands were the same as in Fig. 2. The value of the command for each vowel part in Fig. 2 ([-13, 8, 3, 7, 9] for /i, o, u, a/, respectively) is reasonably similar to the value on the horizontal first axis at the corresponding vowels in Fig. 1 ([−10, 5, 4, 4, 8]). Therefore, the command is detected correctly from the vowel space parameters of the corresponding vowels even though the parameters are ambiguous because of the coarticulation effect.

5. Listening test

In order to evaluate the quality of the imitated speech, we performed a listening test using isolated vowels synthesized from three-dimensional vowel space parameters. The number of individuals for each generation was 10. The number of phonemes was set to three. Figure 6 shows the spectral errors of the test data. Each data is presented ten times randomly. The listeners were asked “what Japanese vowel does it sound like?” The listeners were five Japanese university students. For comparison, the original three-dimensional vowel space parameters were also used. The result of the listening test is shown in Fig. 7. A high rate of correct answers of over 80% was attained at an early generation.

We also evaluated the speech quality of the synthesized word for each generation. The dimension of the vowel space parameters was set to seven. The numbers of phonemes were
set for each word. For comparison, the original seven-dimensional vowel space parameter and the analysis-synthesis speech were used. The spectral errors of the word /hino-kuruma/ are shown in Fig. 4, which show the physical qualities for each synthesized speech. Similarly, the words /aozora/ and /mikazuki/ were synthesized. The listening test was performed for each word. The listeners were asked to evaluate the spoken words in terms of their clarity as Japanese words on the scale of very good (4), good (3), fair (2), and bad (1). The other experimental conditions were the same as above. Figure 8 shows the results of the evaluation tests on the quality for each synthesized speech. Nearly ‘fair’ responses were obtained for the synthesized speech by the GA. However, the quality was sufficient to correctly recognize the words. We confirmed that the quality of the manually fitted synthesized speech was almost the same as that of the speech synthesized using the original vowel space parameters. This shows the possibility of higher performance of the proposed generative model by tuning the GA.

Therefore, we confirmed similar performance of the proposed generative model to that of a human infant who can pronounce speech with reasonable quality after a few trials.

6. Conclusion

In this study, we proposed a new generative model of spectral sequences, using the Fujisaki model. We obtained synthesized speech that can be understood as spoken words at an early generation of the GA. As a result, a listening test on isolated vowels attained a correct rate of over 80% at an early generation. The speech quality of the words was reasonable for a spectrum generated from only seven-dimensional vowel space parameters. Consequently, it was shown that the proposed model can be used to model the acquisition of the normalization skill of coarticulation by human infants.

Future works are to imitate adult speech and generate infant’s speech, to automatically detect the number of phonemes, and to apply the GA to the source of the speech generation model.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP25330201 and JP24300278.

References