DEVELOPMENT OF A FORMANT-BASED ANALYSIS-SYNTHESIS SYSTEM AND GENERATION OF HIGH QUALITY LIQUID SOUNDS OF JAPANESE

Nobuyuki Nishizawa, Nobuaki Minematsu and Keikichi Hirose

Department of Information and Communication Engineering
Graduate School of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
nishi@gavo.t.u-tokyo.ac.jp mine@gavo.t.u-tokyo.ac.jp hirose@gavo.t.u-tokyo.ac.jp

ABSTRACT

Although flexible control of acoustic features is possible in formant-based speech synthesizers, their development requires precise estimation of parameters related to vocal tract and source. This requirement is difficult to satisfy and often results in limiting quality of the synthesized speech. The difficulty is derived from the fact that estimation of the parameters is a non-linear problem. Therefore, the completely automatic estimation of the parameters is quite difficult and some approximations or manual modifications of parameters with *a priori* knowledge are required in the development. In this study, mainly to make the estimation more efficient and/or to assist developers doing the manual modifications of parameters, a formant-based analysis-synthesis system is build. The system introduces pitch-synchronous acoustic analysis to reduce fluctuation of the estimated parameters. Experiments show that quality of synthetic speech of Japanese /r/ sounds is significantly improved by using the proposed system.

1. INTRODUCTION

Recently, speech synthesizers based on waveform concatenation, which can be built with a quite large amount of speech data, have been widely studied and used. These synthesizers can generate synthetic speech with high quality when the synthesis requires only a little modification from the original speech waveforms. However, recent speech applications, such as spoken dialogue systems, have come to request synthetic speech with various speaking styles and emotions. To synthesize these speech with the above synthesizers, a further larger speech corpus is necessary. In other words, without such a huge speech corpus, the above synthesizers can hardly generate high quality synthetic speech of various speaking styles.

On the other hand, since a formant-based synthesizer can well simulate human processes of speech generation in frequency domain, flexible control of acoustic features is possible [1][2]. This means that, once a formant-based synthesizer is built with a relatively small amount of speech data, it may be possible to generate high quality synthetic speech of various speaking styles and emotions.

However, several problems must be solved to realize the high quality formant-based synthesis. The main problem is the difficulty of precisely estimating acoustic parameters related to vocal tract and glottal source. Although the formant-based synthesis has been studied in our laboratory [3], quality of the synthetic speech is often limited by this difficulty, which is mainly derived from the fact that estimation of the target parameters is a non-linear problem. It indicates that the completely automatic estimation of the parameters is quite difficult and some approximations or manual modifications of parameters with *a priori* knowledge are required.

Inappropriate estimation of parameters easily leads to poor quality of synthetic speech. This is especially the case with sounds characterized by their transitional portions such as stops and liquids, and those characterized by zeros such as nasals. In order to cope with these problems, we have developed a formant-based analysis-synthesis system, where pole-zero analysis with glottal source waveforms is carried out based on ARX (Auto-Regressive and eXogenous input) model. Though an ARX model-based analysis-synthesis system [4] was proposed previously, unlike this system, our system has been built mainly to assist the manual modifications of parameters during acoustic analysis. In the system, to decrease unexpected perturbations of parameter values, pitch-synchronous acoustic analysis is introduced.

For evaluating the developed system, it is applied to the analysis of Japanese /r/ sounds to improve their quality in synthetic speech. Though Japanese /r/ sounds are generated similarly to vowels within a short period, their formant movement is more transitional. This characteristic usually makes it rather difficult to extract formant frequencies precisely.

In this paper, basic analysis methods in the system are presented in Section 2. In Section 3, our formant-based analysis-synthesis system is described. And application of the system to high quality synthesis of Japanese /r/ sounds and evaluation of the sounds are discussed in Section 4. Finally, Section 5 concludes the paper.

2. ACOUSTIC ANALYSIS

2.1. Modeling based on ARX model

Our formant-based synthesizer is configured with four paths of cascade connection of pole/zero filters and three source waveform generators [3] (see Figure 1). And voiced sounds are generated by pole-zero filters excited by waveforms of a glottal source model. FL (Fujisaki-Ljungqvist) model [5] is adopted as the glottal source model. This model represents a voiced source waveform as polynomials with six parameters and divides one excitation cycle into four sections. In this study, we used this model with two parameters being fixed and the remaining four parameters being controlled dynamically and adequately. This is done to make the estimation process easier.

The process of voiced sound generation in the synthesizer is modeled by ARX model, which is represented in the following equation.

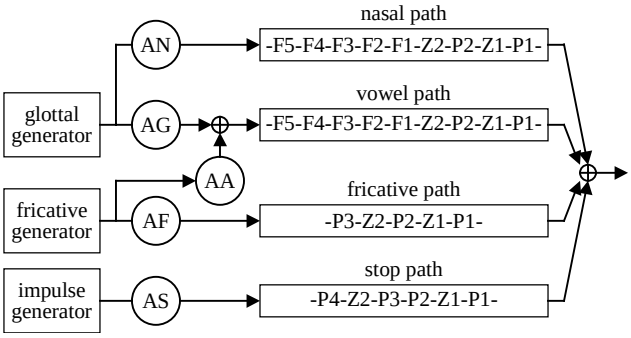


Figure 1: Configuration of our formant speech synthesizer. Symbols A*x*, F*x*, P*x*, and Z*x* denote amplifier, formant (pole) filter, pole filter, and zero filter respectively [3].

$$y(n)+\sum_{i=1}^{p} a_i y(n-i)=\sum_{j=0}^{q} b_j u(n-j)+e(n) \qquad (1)$$

And Figure 2 shows a block diagram of ARX model.

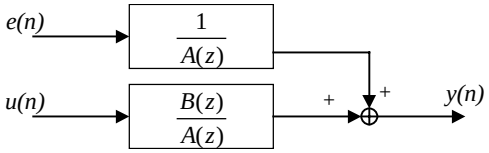


Figure 2: Block diagram of ARX model.

If input *u(n)* is known, estimation of the ARX model parameters can be done by solving linear equations. However, since it is impossible to observe source waveforms of speech directly, *u(n)* has to be treated as unknown. For this reason, it is required to estimate parameters of the glottal source waveform model (FL model), which are used to estimate ARX model parameters. This estimation is done by minimizing error terms and the minimization is a non-linear problem to solve. Therefore, appropriate values should be given as initial values for successive approximation and this should be done by using *a priori* knowledge.

2.2. Pitch-synchronous analysis

In speech waveforms, their power tends to concentrate on a short segment and the segment is often found near glottal closure. Since an interval between two consecutive glottal closures is not always constant, acoustic analysis with fixed frame rate usually gives undesired perturbations of estimated values of the target parameters.

One solution to this problem is lengthening an analysis frame to the extent where the power distribution can be considered even over the frame. In this case, however, results of the analysis do not reflect dynamic features in speech. Another one is to reduce the perturbations between a frame and its neighboring ones. For the reduction, pitch-synchronous analysis is introduced here. A reference point used to determine the position of an analysis frame is called a pitch mark. It is desirable that the center of the frame is located near a local peak of power and pitch marks should also be assigned to these peaks. In this study, the pitch marks are estimated from autocorrelation of LP residual waveform since the location of glottal closure correspond roughly to their local peak of power. The estimated pitch marks are also used to lessen computation load required for analysis for positioning the glottal source waveforms.

3. ANALYSIS-SYNTHESIS SYSTEM

The estimation method described above still have some difficulty to apply because pitch marks must be estimated with good accuracy. As told above, estimated pitch marks are used to estimate other parameters. And the estimation is done by successive approximation. Therefore, in some cases, it is required to correct the estimated pitch marks and/or to give adequate initial values of the successive approximation.

Even if pitch marks are detected correctly and the initial values are given adequately, it does not always follow that quality of the synthesized speech is improved. In most of these cases, the estimated parameters still show unexpected perturbations and those are considered to degrade the quality. To improve the quality even in those cases, manual iterative modification of parameters are inevitably required and the modification should be done by listening to speech synthesized by updated parameters, not only by inspecting parameters visually on a spectrum level.

To facilitate these operations, we developed a formant-based analysis-synthesis system, which provides several functions for developers. 1) Waveform input/output from/to disks or audio devices, 2) editing waveforms, 3) acoustic analysis described in Section 2, 4) editing pitch marks for pitch-synchronous analysis, 5) editing initial values for parameter estimation, and 6) editing the estimated parameters directly used in our formant-based synthesizer.

To developers, the system gives graphical user interface, which is consisted of several windows. Waveforms, spectral features, residual waveforms, and estimated parameters are presented in individual windows on PC. Editing of several acoustic parameters described above is possible by using a mouse. And speech synthesis can be done instantaneously by using the latest parameters. The interface supports interactive operations of analyzing and editing. For example, estimated parameters can be evaluated by a waveform/spectrum image and/or re-synthesized speech. Once appropriate parameters are obtained, analysis of the next frame can be started by using the obtained parameters as initial parameter values of the frame. And this can significantly reduce computational load for the estimation.

Figure 3 shows this system schematically. It says that the system is comprised of several modules. Since the structure of object files is extensible, new functions can be added easily to the current system.

4. FORMANT-BASED SYNTHESIS OF JAPANESE /r/ SOUND

The formant-based analysis-synthesis system was applied to making a sequence of parameters to synthesize Japanese liquid (/r/) sounds. Though /r/ sounds are generated through a process similar to vowels, formant movements of /r/ sounds are transitional. /rV/ (V is one of Japanese vowels, /a, i, u, e, o/) templates for the formant-based synthesis are newly built with our proposed system. Collection of the templates was done by using /VrV/ sounds (two vowels are the same). They were

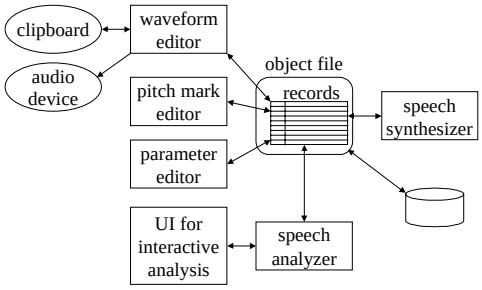


Figure 3: Block diagram of the proposed formant-based synthesis-analysis system.

spoken by a native Japanese female speaker and recorded with 12 kHz and 16 bit sampling. The use of utterances in the form of /VrV/, not /rV/, is because /r/ sound at the beginning of an utterance is quite difficult to analyze.

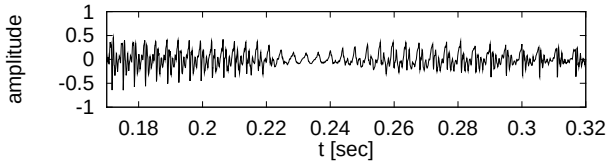
Figure 3 shows examples of applying the system to analyzing an /ara/ sound.

4.1. Evaluation experiments using /VrV/ samples

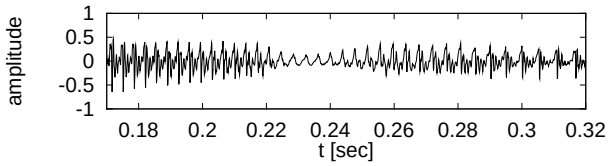
To evaluate quality of the speech synthesized newly with the /rV/ templates, listening experiments were designed and carried out. Reference speech were the /VrV/ natural speech collected above. With the reference speech, several types of synthesized speech were compared. (a) /VrV/ samples simply generated by analysis-re-synthesis, (b) /VrV/ samples generated by formant-based synthesis of /V/ and /rV/ templates using our proposed system and their concatenation, and (c) the same as (b) except that our former /rV/ templates were used. In order the comparison to be made only in terms of segmental features of the speech samples, F0 and power observed in the reference speech were used. The concatenation in (b) and (c) was based on the rules made in the development of our former formant-based synthesizer, where discontinuous formant frequency parameters were interpolated by a step response function of a critically damped or second-order linear system and the other parameters were interpolated linearly [3].

Procedure of the stimulus presentation was as follows. Pairs of the reference natural speech and a synthetic speech sample were presented in sequence to 10 native Japanese speakers individually through headphones. Here, the reference sample was always given as the first stimulus. After hearing them, they were asked to judge the naturalness of the sample in a five-degree scale (1 to 5), where the first stimulus, i.e. the reference sample, was assumed to rank at 5.

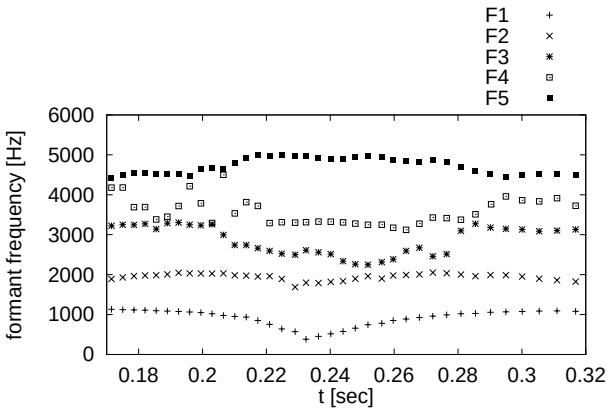
Figure 5 shows the naturalness score of each /VrV/ sample averaged over the subjects. It indicates that quality of new templates is higher than that of old templates and that naturalness degradation from re-synthesized speech (a) to formant-based synthesized speech (b) can be said to be quite small. And in several /VrV/ patterns, the averaged score is over 4.0, suggesting that sufficient quality of liquid sounds is provided by formant-based synthesizers from a practical viewpoint.



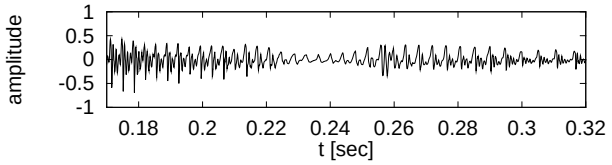
(a) natural speech waveform



(b) estimated glottal source waveform



(c) estimated formant frequencies



(d) re-synthesized speech waveform

Figure 4: Examples of applying the system to analyzing an /ara/ sound and its synthesis. (a) is original speech waveform, (b) is source waveform estimated by FL model, (c) is estimated formant frequencies, and (d) is waveform of speech synthesized with the estimated parameters.

4.2. Evaluation experiments using /V_1rV_2/ samples

/V_1rV_2/ (two vowels are different) samples generated by formant-based synthesis of /V_1/ and /rV_2/ and their concatenation were used here. In the experiments, duration, F0, and power were generated by using the rules adopted by our former formant-based synthesizer. Unlike the previous experiments, natural speech samples or analysis-re-synthesis samples were not used as stimuli here. A pair of /V_1rV_2/ samples, one was generated by new templates and the other was by old

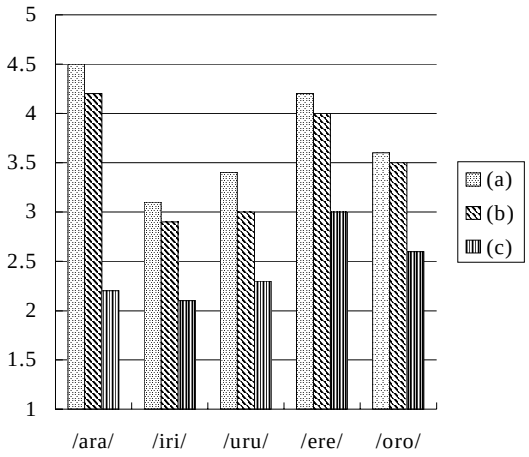


Figure 5: Averaged scores for /VrV/ samples over the subjects. (a), (b), and (c) indicate speech samples by analysis-re-synthesis, those by formant-based synthesis with new templates, and those by formant-based synthesis with old templates respectively.

templates, were arranged randomly. And twenty pairs were also presented randomly to the same subjects as those in the previous section. In this experiment, they were asked to judge which was more natural and how was the difference. And the two tasks were done simultaneously by selecting one of five answer candidates, which were from score −2 meaning that the second stimulus was quite inferior to the first one in naturalness to score +2 indicating inversely.

Figure 6 shows the preference score for each kind of stimulus averaged over the subjects. In every case, the higher score means the larger preference of samples generated with new templates. Although a few samples of new templates show lower quality in naturalness than those of old templates, the averaged preference over the kinds of samples indicates large superiority of new templates to old ones. In the previous section, several speech samples of new templates showed quite low scores, such as /iri/ and /uru/. In this experiment, speech samples using /rV/ templates of /iri/ or /uru/, namely, /*x*ri/ and /*x*ru/, also show low preference scores. This implies that quality of /V_1rV_2/ is rather dependent on /rV_2/'s quality.

5. CONCLUSION

In this paper, a new analysis-synthesis system was proposed mainly to make parameter estimation more efficient and/or to assist developers doing manual edition of the estimated parameters. In this system, to reduce perturbations of the estimated parameters, pitch-synchronous acoustic analysis was introduced and user-friendly GUI was also provided for developers. Through the GUI, they can do several operations quite easily. Evaluation experiments were carried out using speech samples including Japanese /r/ sounds. Results showed that speech templates made with the proposed system gave us higher naturalness than those with our former system. Analysis of other kinds of phonemes, such as stops and fricatives, is one of urgent future works.

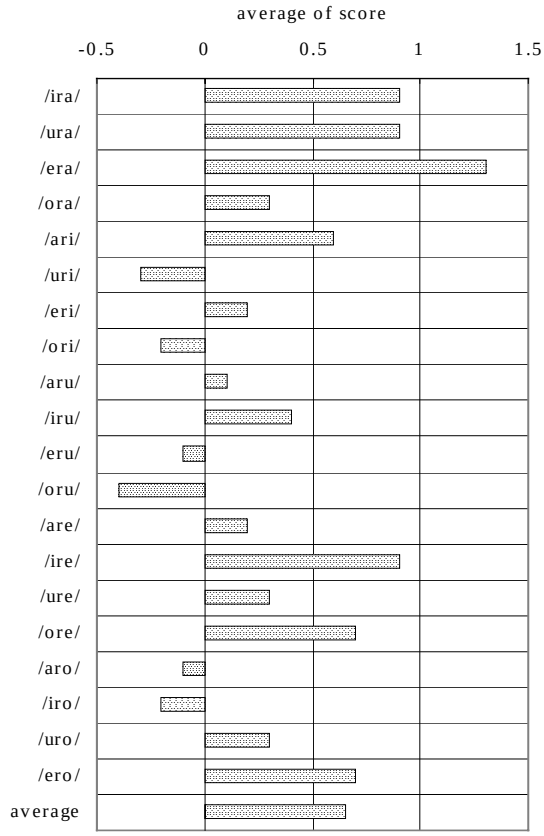


Figure 6: Preference scores for each kind of speech sample averaged over the subjects. Higher scores indicate that speech samples generated with new /V_1/ and /rV_2/ templates has better quality.

6. REFERENCES

1. Klatt, D. “Software for a cascade/parallel formant synthesizer,” J. Acoust. Soc. Am., vol. 67, no. 3, pp. 971-995, 1980.

2. Klatt D., and Klatt, L., “Analysis, synthesis, and perception of voice quality variations among female and male talkers,” J. Acoust. Soc. Am., vol. 87, no. 2, pp. 820-857, 1990.

3. Hirose, K., and Fujisaki, H., “A System for the synthesis of high-quality speech from texts on general weather conditions,” IEICE Trans. Fundamentals, vol. E76-A, no.11, pp.1971-1980, 1993

4. Zhu, W., and Kasuya, H., “A speech analysis-synthesis-editing system based on the ARX speech production model,” J. Acoust. Soc. Jpn. (E), vol. 19, no. 3, pp. 223-230, 1998.

5. Fujisaki, H., and Ljungqvist, M., “Proposal and evaluation of models for the glottal source waveform,” In. Proc. ICASSP, 31.2, pp. 1605-1608, 1986.