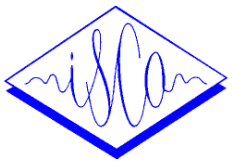

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

A DATA-DRIVEN APPROACH TO SOURCE-FORMANT TYPE TEXT-TO-SPEECH SYSTEM

Hiroki Mori, Takahiro Ohtsuka, and Hideki Kasuya

Faculty of Engineering, Utsunomiya University
7-1-2, Yoto, Utsunomiya-shi, 321-8585 Japan

ABSTRACT

A data-driven formant-type TTS system is proposed. The formant-type speech synthesizer is one of the most promising architectures to enable flexible control of various voice qualities. By applying the ARX-based speech analysis method, source and formant parameters are automatically obtained. It is shown that a TTS system can be built by using the parameters, without requiring any heuristic rules to control vocal tract characteristics.

1. INTRODUCTION

Recent progress in speech synthesis techniques has led to demands for various applications such as intelligent agent systems, entertainment software, and so on. Synthetic speech of the next generation will require not only high quality equivalent to natural voice, but also high variability to adapt its speaking style according to particular situations. The variability can be paraphrased as controllability of paralinguistic and extralinguistic information.

Waveform-concatenation-based systems, however, cannot easily achieve variability because they are highly dependent on speech corpora. In order to produce any sentence in any context, we need spoken materials from any context, but this is impossible because the number of paralinguistic or extralinguistic contexts is not finite.

One of the promising alternatives is the formant synthesizer. Although this ignores the interaction between voice source and vocal tract, it gives a reasonable approximation of the speech production system. Formants, and the shape of the glottal waveform, are sufficiently good parameters to control various aspects of voice quality[1, 2].

Nevertheless, most state-of-the-art TTS (Text-To-Speech) systems are based on waveform concatenation. The possible reasons are:

1. Formant synthesizers require a number of complex rules to produce natural sounds, especially stop and fricative consonants. As it is difficult to find such rules automatically, they have to be tuned "by art [3]."
2. Vocoded speech is thought to be far less natural than that of time-domain manipulation of waveform; it often gives a "buzzy" impression for certain types of voiced segment.

In this paper, we propose a data-driven approach for building a formant-type TTS system. To the authors' knowledge, this is the first formant-type TTS system fully based on speech corpora. To overcome the above difficulties, several key technologies are adopted.

1. Source and formant parameters are automatically obtained from speech corpora. They are simultaneously estimated with the speech analysis method based on the ARX (Auto-Regressive with eXogenous input) model [4, 5]. Due to an assumption of the input to the ARX speech production model, voice source and vocal tract parameters are effectively differentiated with the method, and no special rules are needed to synthesize with the formants, even for consonants.
2. By incorporating aperiodic components into the voicing source, fairly natural speech without "buzziness" can be produced, even for a breathy or soft voice[5].

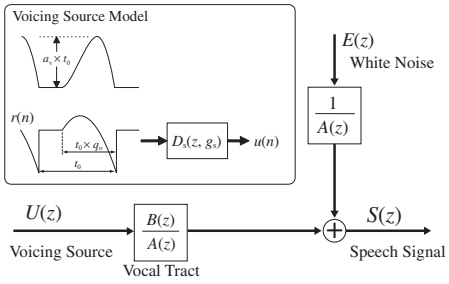

Fig. 1. ARX speech production model.

The contents of this paper are as follows. Section 2 gives a brief introduction to the ARX speech production model, Section 3 describes the corpus analysis and database preparation, and Section 4 explains the system configuration and individual modules.

2. ARX SPEECH PRODUCTION MODEL

The ARX speech production model is shown in Fig. 1 and is represented by the linear difference equation,

$$s(n) + \sum_{k=1}^{p} a_k s(n-k) = \sum_{k=0}^{q} b_k u(n-k) + e(n), \qquad (1)$$

where the input $u(n)$ denotes a periodic voicing source signal and the output $s(n)$ a speech signal. A part of the glottal noise component is simulated by the white noise $e(n)$. In the equation, a_i and b_i are vocal tract filter coefficients, and p and q are ARX model orders. We define $A(z)$ and $B(z)$ as

$$A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p} \qquad (2)$$

$$B(z) = b_0 + b_1 z^{-1} + \cdots + b_q z^{-q}, \qquad (3)$$

then the z-transform of Equation (1) can be written as

$$S(z) = \frac{B(z)}{A(z)} U(z) + \frac{1}{A(z)} E(z). \qquad (4)$$

The vocal tract transfer function is then given by $B(z)/A(z)$.

We employ the Rosenberg-Klatt (RK) model [1] to represent a differentiated glottal flow signal $u(n)$, including radiation characteristics. The rudimentary RK waveform $u(n)$ is given by

$$r(n) = r_{\rm c}(nT_{\rm s}) \tag{5}$$

$$r_{\rm c}(t) = \begin{cases} 2at - 3bt^2 & 0 \le t < q_{\rm o}t_0, \\ 0 & \text{elsewhere}, \end{cases} \tag{6}$$

$$a = \frac{27a_{\rm v}}{4q_{\rm o}^2 t_0}, \quad b = \frac{27a_{\rm v}}{4q_{\rm o}^3 t_0^2}, \tag{7}$$

where $T_{\rm s}$ is a sampling period, $a_{\rm v}$ is an amplitude parameter, t_0 is a pitch period, and $q_{\rm o}$ is an open quotient of the glottal open phase of the pitch period. The differentiated glottal flow waveform $u(n)$ is generated by smoothing $r(n)$ through the use of a low-pass filter, where the tilt of the spectral envelope is adjusted via a spectral tilt parameter, $g_{\rm s}$.

3. DATABASE DEVELOPMENT

3.1. Speech corpus

A set of 503 phonetically-balanced sentences and 468 sentences from a set of prose was digitally recorded by a female Japanese speaker. The overall duration of the recorded speech is about 160 minutes. It is downsampled to 11,025 Hz using a digital filter, then phoneme and voiced/unvoiced labeling is performed by hand. Moreover, prosodic labels including accent type and break indices are given using the J_ToBI [6] framework. Segment database construction, f_0 vector statistics and CART for duration control as explained below are all based on these data.

3.2. ARX analysis

Source and formant parameters are estimated for the recorded data with the ARX-based speech analysis method [4]. For each voiced frame, source parameters (t_0, $a_{\rm v}$, cutoff frequency for phase randomization [5] $f_{\rm a}$, $q_{\rm o}$, etc.) and formant parameters (frequency and intensity) are extracted, under the conditions of $p = 14$, $q = 0$, $g_{\rm s} = 15$ dB/3000 Hz, frame length of 35 ms, and frame shift of 5 ms. For unvoiced frames, noise source amplitude and formant parameters are estimated under the assumption of $u(n) = 0$ (i.e. AR model).

The analyzed parameters except t_0 ($= 1/f_0$) and $a_{\rm v}$ are stored and indexed so as to be extracted against a given phone sequence of arbitrary length[7] (ex. CV, VCV, CVCVCV...). The total number of stored parameter segments is about 47000 morae.

3.3. Global f_0 shape and microprosody separation

The fundamental frequency (f_0) contour is modeled as a superposition of global shape and micro structure. The global component reflects pitch accent and intonation, whereas the microprosodic component is mainly caused by larynx-tract interaction in producing consonants.

In this paper, the global component of the f_0 contour is approximated by a piecewise linear function in the logarithmic domain. Each line segment interpolates the contour by connecting f_0 sample points at the vowel center of adjacent morae. The global pattern, which is represented by values sampled at the vowel center, is stored for f_0 contour generation (Section 4.2). By subtracting the global component from the original f_0 contour of the speech corpus in the logarithmic domain, the microprosodic component is obtained.

$$\Delta f_0(n) = f_0(n) - \left\{ f_0(n_{\rm V1}) + \frac{f_0(n_{\rm V2}) - f_0(n_{\rm V1})}{n_{\rm V2} - n_{\rm V1}}(n - n_{\rm V1}) \right\}, \tag{8}$$

where $f_0(n)$ denotes the $\log f_0$ value at the n-th frame, $n_{\rm V1}$ and $n_{\rm V2}$ denote the sampled frame number of adjacent vowels, V1 and V2, respectively. The microprosodic feature is a segment-related parameter, so it is stored in the database together with formant parameters and various source parameters. The same procedure is performed for the micro structure of the $a_{\rm v}$ contour.

Figure 2 shows the concept of separation of global shape and microprosody. The model clearly does not fit well for f_0 near the boundary of /fu/ and /ri/ for this case. This is caused by the model's incompleteness, which is expected to be improved by, for example, introducing multiple samples per mora[8]. Nevertheless, f_0 and $a_{\rm v}$ fluctuations at consonants can be captured for most cases.

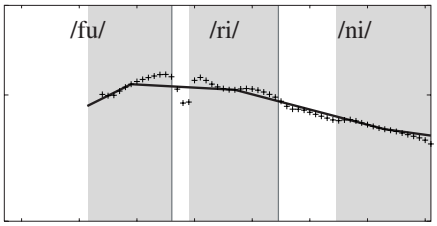

Fig. 2. The original f_0 contour ('+') and its estimated global pattern (solid line) for the utterance /furini/. Vowel regions are filled with gray.

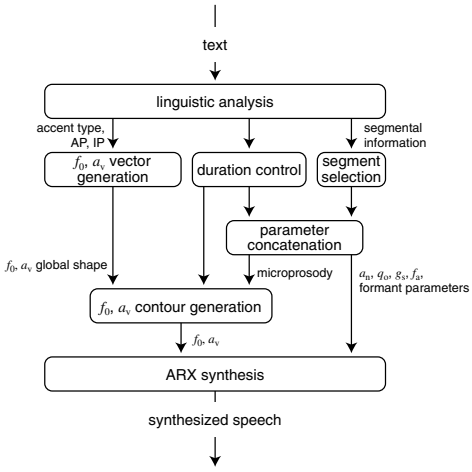

Fig. 3. Block diagram of the system.

4. SYSTEM DETAILS

An outline of the TTS system is shown in Fig. 3. Each module is standalone and communicates with the other modules through the UNIX pipeline.

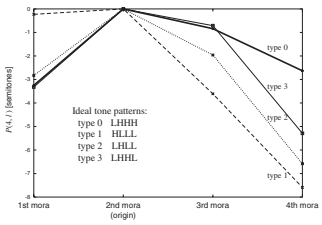

Fig. 4. Standard f_0 vectors for accent type 0, 1, 2 and 3 AP of 4-mora.

4.1. Linguistic analysis

This module consists of morphological analysis [9], accentual phrase (AP) identification that follows compound word analysis, accent type identification including accent movement, processing for vowel devoicing, and intonation phrase (IP) identification.

4.2. f_0 and a_v vector generation

The global component of the f_0/a_v contour is modeled as a sequence of sample points at the vowel center, as described in Section 3.3. The global f_0 pattern of a given AP is predicted according to the following formula:

$$\boldsymbol{P}(k,l,n) = \boldsymbol{P}(k,l) + \Delta\boldsymbol{P}(k,l) + \boldsymbol{C}(n) + \boldsymbol{t}, \qquad (9)$$

where $\boldsymbol{P}(k,l,n)$ denotes a vector whose elements are predicted $\log f_0$ values of sample points for (k-mora, type l) AP at the n-th position in IP. Standard f_0 vector, $\boldsymbol{P}(k,l)$, is determined beforehand by averaging global patterns (Section 3.3) for all (k-mora, type l) APs. $\Delta\boldsymbol{P}(k,l)$ represents tonal coarticulation and boundary tone, and is separately modeled by linear/nonlinear regression. $\boldsymbol{C}(n)$ represents downstep [10]. $\boldsymbol{t}$ determines a speaker-dependent baseline.

Figure 4 shows an example of standard f_0 vectors for 4-mora AP. The value for each mora is normalized to that for second mora when averaging.

For predicting a_v vectors, we make use of a correlation between f_0 and a_v. An a_v vector is derived from the predicted f_0 vector by applying a monotonic function.

4.3. Duration control

The process for segment duration control is divided into two steps. In the first step, the duration for each mora is predicted using CART. Controlling by mora seems to be reasonable because Japanese is a mora-timed language. A tree is built for a subset of the speech corpus described in Section 3 (15,971 morae from 503 sentences) using the segmental and prosody-related features shown in Table 1. 10% cross-validation testing reveals that the CART predicts mora duration with a precision of 18.63 ms in RMS error.

In the second step, actual vowel duration is estimated for CV mora. To take into account the phoneme-dependent and nonuniform behavior of vowel duration in lengthening/shrinking, multiple regression classes are introduced.

4.4. Segment selection

The unit for parameter concatenation is mora, that is, CV or V (including /N/). The parameter segment is selected from the nearest phonemic context in the database by the following procedure:

Table 1. CART features for duration control.

features	#	features	#
phonemic class of the mora	19	# of morae in the intonation phrase	1
phonemic class of preceding mora	8	position in the intonation phrase	1
phonemic class of following mora	19	# of morae in the breath group	1
vowel devoicing of the mora	1	position in the breath group	1
vowel devoicing of preceding mora	1	# of morae in the sentence	1
vowel devoicing of following mora	1	sentence final	1
# of morae in the accentual phrase	1	accented mora	1
position in the accentual phrase	1	content/functional	1

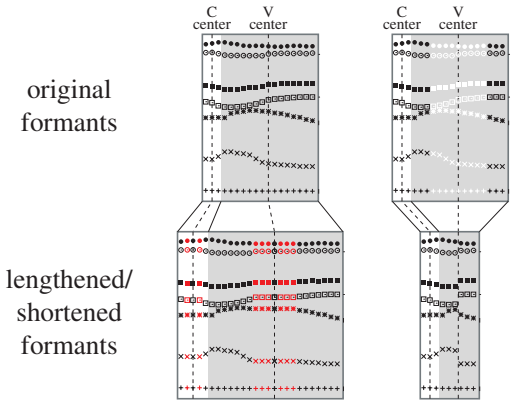

Fig. 5. Parameter interpolation/decimation. Formant frequencies are indicated by dots.

1. Search for the segment that matches the given context, starting with phoneme 4-gram. Relax the condition until the number of segments reaches a threshold.
2. Calculate the contextual distance for the preceding two morae and the following two morae as:

$$d_{\text{context}} = \boldsymbol{w}_{\text{pos}}^t \begin{pmatrix} d^{\text{C}}(-2) & d^{\text{V}-}(-2) \\ d^{\text{C}}(-1) & d^{\text{V}-}(-1) \\ d^{\text{C}}(1) & d^{\text{V}+}(1) \\ d^{\text{C}}(2) & d^{\text{V}+}(2) \end{pmatrix} \boldsymbol{w}_{\text{CV}}, \qquad (10)$$

   where $\boldsymbol{w}_{\text{pos}}$ is a weighting vector for position, $\boldsymbol{w}_{\text{CV}}$ is a weighting vector for C/V, $d^{\text{C}}(-2)$, $d^{\text{V}-}(-1)$ and $d^{\text{V}+}(1)$ represents the consonantal distance of the preceding-preceding mora, vowel distance of the preceding mora, and vowel distance of the following mora, respectively, etc.
3. Find the optimal segment sequence by using the Viterbi algorithm, according to the contextual distance and inter-segment formant distance.

4.5. Parameter concatenation

This process involves parameter interpolation or decimation. When the length of the selected segment is shorter than the specified duration, parameters at the phoneme center are copied to satisfy the length. When the selected segment is longer, it is shrunken from the phoneme center. Fig. 5 explains this concept. This module also considers some special phenomena of Japanese, such as long vowels, "sokuon (choked sound)" and vowel devoicing.

Unlike a waveform-based synthesizer, simple concatenation of formant parameters does not cause serious degradation of quality.

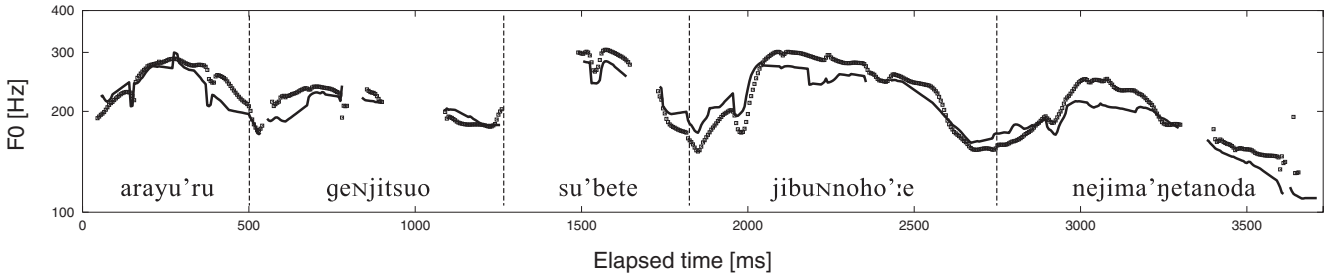

Fig. 6. Generated f_0 contour (solid line) and the natural one from the corpus (dot). IP boundaries are given. Segment duration is adjusted to that of natural speech.

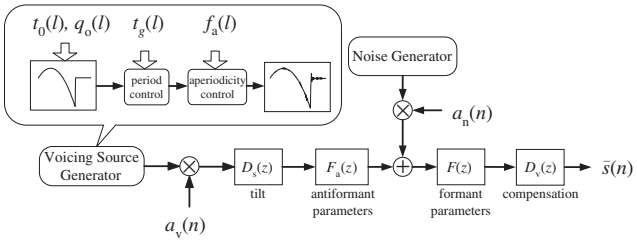

Fig. 7. Block diagram of ARX-based synthesizer.

4.6. f_0 and a_v contour generation

By superposing the microprosodic f_0/a_v contour (adjusted in Section 4.5) onto the global pattern (described in Section 4.2), the final f_0/a_v contour for synthesis is obtained. Figure 6 shows an example of the generated f_0 contour for a 5-AP sentence, as well as that of the natural utterance. The generated contour is slightly flatter, but the overall shape seems to be fairly good. It should also be noted that perturbations caused by consonants are reproduced (for example, /b/ at 1530 ms).

4.7. ARX synthesis

Now we have all source and formant parameters required to synthesize speech. The construction of the ARX-based synthesizer is shown in Fig. 7. One of the important features of the module is the aperiodicity control in voicing source generation. By randomizing the phase of the RK waveform at higher frequency than f_a, the aperiodic nature of the voicing source is simulated. The f_a (cutoff frequency) contour is estimated [5] and stored in the database. It is revealed that this feature contributes to the naturalness of synthesized speech.

5. CONCLUSION

A data-driven source-formant type TTS system is developed based on the ARX-based speech analysis/synthesis method. There remain several features to be improved, for example, AP identification considering dephrasing, better prosodic trajectory estimation, segment selection which avoids spectral mismatching. Nevertheless, as the first corpus-based formant-type TTS system, it produces sufficiently intelligible and fairly natural sounds. The system can be evaluated via an interactive demonstration page, http://tts.klab.jp/.

Acknowledgement

The authors thank Mr. Kenji Matsui at Advanced Technology Research Laboratories, Matsushita Electric Industrial Co., Ltd. for providing speech materials, and Mr. Takashi Ohtake, Mr. Kimihiro Sasaki and Mr. Kenji Watanabe for their assistance.

6. REFERENCES

[1] D. H. Klatt and L. C. Klatt, "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.

[2] H. Kasuya, K. Maekawa, and S. Kiritani, "Joint estimation of voice source and vocal tract parameters as applied to the study of voice source dynamics," in *Proc. ICPhS 99*, 1999, pp. 2505–2512.

[3] N. H. Pinto, D. G. Childers, and A. L. Lalwani, "Formant speech synthesis: improving production quality," *IEEE Trans. ASSP*, vol. 37, no. 12, pp. 1870–1887, 1989.

[4] T. Ohtsuka and H. Kasuya, "An improved speech analysis-synthesis algorithm based on the autoregressive with exogeneous input speech production model," in *Proc. ICSLP 2000*, 2000, vol. II, pp. 787–790.

[5] T. Ohtsuka and H. Kasuya, "Aperiodicity control in ARX-based speech analysis-synthesis method," in *Proc. Eurospeech 2001*, 2001, vol. 3, pp. 2267–2270.

[6] J. J. Venditti, "Japanese ToBI labelling guidelines," in *Working Papers in Linguistics*, K. Ainsworth-Darnell and M. D'Imperio, Eds., vol. 50, pp. 127–162. Ohio State University, 1997.

[7] M. Nagao and S. Mori, "A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese," in *Proc. 15th ICCL*, 1994, pp. 611–615.

[8] B. Holm and G. Bailly, "Generating prosody by superposing multi-parametric overlapping contours," in *Proc. ICSLP 2000*, 2000, pp. 203–206.

[9] A. Masayuki and Y. Matsumoto, "Extended models and tools for high-performance part-of-speech tagger," in *Proc. COLING 2000*, 2000, pp. 203–206.

[10] J. Pierrehumbert and M. Beckman, *Japanese tone structure*, chapter 3, MIT Press, 1988.