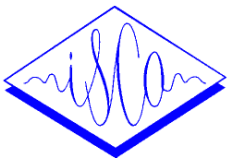

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

# A LOW BIT RATE SPEECH CODING METHOD USING A FORMANT-ARTICULATORY PARAMETER NOMOGRAM

*Hiroshi OHMURA, Akira SASOU, Kazuyo TANAKA*
Electrotechnical Laboratory,
1-1-4 Umezono, Tsukuba, 305-8568, JAPAN
*{ohmura; sasou; ktanaka}@etl.go.jp*

## ABSTRACT

In this paper, we propose a new method for low bit rate speech coding using a nomogram that is a pair of codebooks representing the functional relationship between formant frequencies and articulatory parameters. Significant features of our approach are 1) using the codebooks derived theoretically from the computation using a stylized vocal tract model and 2) independent coding by separating frequency information from the amplitude in a speech segment. From these features, the method is also characterized by little dependency upon speech databases and/or languages in the acoustic domain, so that it has a potential to construct a more flexible rule-based speech synthesis system. We have conducted articulatory encode-decode experiments with the bit rate range from 3.2kbps to 1.6kbps using speech samples in ASJ and TIMIT speech databases and confirmed that good quality speech synthesis is achieved with improvements on the bit allocation scheme and a frame sampling method.

## 1. INTRODUCTION

The basic idea in developing this speech coding system is that formant representation of speech sounds provides a natural and convenient means for organizing and processing their features. We expected that a formant frequency vectors can be easily converted to a small number of articulatory parameter vector which bring compact and informative form of sound characteristics to many speech application. To obtain these advantages in speech processing, we developed a formant based speech analysis and synthesis system [1]. In this system, formants were automatically determined and temporal, intensity, and spectral properties of speech were described by time sequences of frequency and its intensity parameters.

Next, in order to extract articulatory information from real speech samples, we introduced an articulatory encoding and decoding stage in the system using a formant-articulatory parameter nomogram. From the results of the articulatory coding experiments for 126 sentences uttered by 126 speakers included in TIMIT and ASJ databases, we found that some segment boundaries are clearly observed in the articulatory parameter domain whereas the formant patterns are smooth in the corresponding section. We also applied directly the articulatory coding method to a speech coder at 7.2kbps and confirmed that good quality speech synthesis was achieved [2].

The focus of this paper is to get compact representation of acoustic characteristics of a segment speech by independent coding that separates formant frequency information from the amplitude in a speech segment. The following sections describe the analytic investigation of bit allocation scheme, and articulatory encode-decode experiment results.

## 2. METHOD

### 2.1 Extended Three Parameter Description of the Vocal Tract Model

Acoustic characteristics for voiced sounds can be sufficiently described by the first three or four formnat frequencies. This means that a vocal tract model related with formant allocation information will become a stylized and simple model described by a small number of articulatory parameter [3].

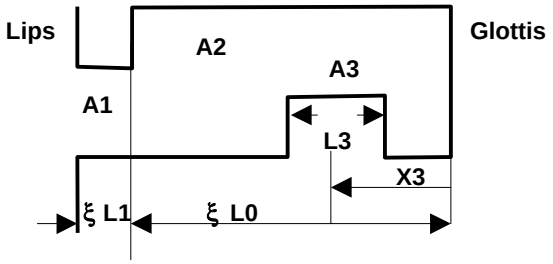


**Figure 1:** A vocal tract model described by four variables, *A1*, *X3*, *A3*, and ξ. Parameter *L0*, *L3*, *A2* are keep in constant and *L1* is a function of *A1*.

From this perspective, the three parameters model of FANT is the best suited one [4] and this model can be extended easily to have a property of speaker adaptation by expanding and contracting its vocal tract length. The current vocal tract model parameterized by four variables shown in Figure 1.

The first three variables, *A1*, *X3*, *A3* are a cross-sectional area function at the lips, a distance between the constriction center and the glottis, and a cross-sectional area function at the constriction respectively. The fourth variable ξ is a multiplier of vocal tract length that is calculated to minimize a matching error between formants generated from the model and formants extracted from input speech.

Parameter *L0* is the vocal tract length, *L0*=15cm. Parameter *L3* is the length of the constriction, *L3*=5.2cm. Parameter *A2* is a cross-sectional area function of the front and back cavities, *A2*=8 square centimeter. After a preliminary experiment using speech samples in TIMIT and ASJ databases [5], active ranges of the variables were given as follows.

$$\begin{aligned} 0.2 \le A1 \le 8.0 \\ 0.2 \le A3 \le 5.0 \\ 2.5 \le X3 \le 12.5 \\ 0.6 \le \xi \le 1.4 \end{aligned} \quad (1)$$

$$\log \xi_p = -\Big\{ \sum_{j=1}^{N} w_j \log \left( F_{i,j} / G_{p,j} \right) \Big\} / \sum_{j=1}^{N} w_j \quad (2)$$

$$E_i = \min_{(P)} [ \sum_{j=1}^{N} w_j \{ \log(F_{i,j} / G_{P,j}) \}^2 - (\log \xi_p)^2 \sum_{j=1}^{N} w_j \ ] \quad (3)$$

## 2.2 The Nomogram of Formant Frequency and Articulatory Parameter Vectors

The nomogram is a codebook representing the functional relationship between an articulatory parameter vector ***P*** of {*A1, X3, A3,* $\xi$=1.0} and a formant frequency vector ***G*** of {*G1, G2,Gn*} given by vocal tract computation. The area function of the model described by ***P*** is smoothed and sampled at an equal spatial rate for constructing a cross sectional area series.

The formant vector ***G*** is a polynomial root set of the vocal tract transfer function. The number of entries in the codebook is given by Nx3*Na1*Na3 where Nx3, Na1, and Na3 are the number of quantized samples for each articulatory parameter. Figure 2 is an example of the nomogram.

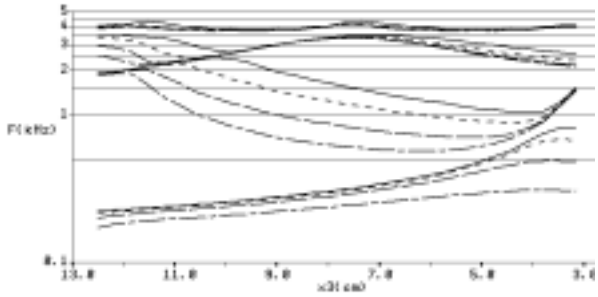

**Figure 2:** An example pattern of the formant-articulatory parameter nomogram in which A3 is a constant value and A1 is changed by four different values. The horizontal axis is *x3* coordinate.

## 2.3 Mapping Algorithm

Parameter $\xi$ causes a linear movement in the logarithmic frequency domain. The error function between formant vectors of ***G*** and ***F*** is defined in equation (3) for a given articulatory parameter vector ***P***. Parameter $\xi$ is calculated by equation (2) in advance

Where *Ei* is an error function at *i*-th frame of a running analysis; *Fi,j* is *j*-th formant extracted at *i*-th frame; and *Gp,j* is *j*-th formant produced by vocal tract computation with a given parameter set ***P***={*A1, X3, A3,* $\xi$=1 }. Symbol ***w*** is a weighting function to emphasize lower formant frequencies. The best-matched parameter set ***P*** for an input formant vector ***F***={*Fi1, Fi2,…, Fin*} is obtained by minimizing the error function among combinations of four numerical values correspond to each parameters in the given ranges.

## 2.4 Coding System

Figure 3 shows the block diagram of the articulatory encoding and decoding system. Speech parameters, e. g., fundamental frequency *F0*, buzz/hiss indicator *BH*, formant intensity vector ***I,*** and formant frequency vector ***F*** are extracted from input speech at the first stage.

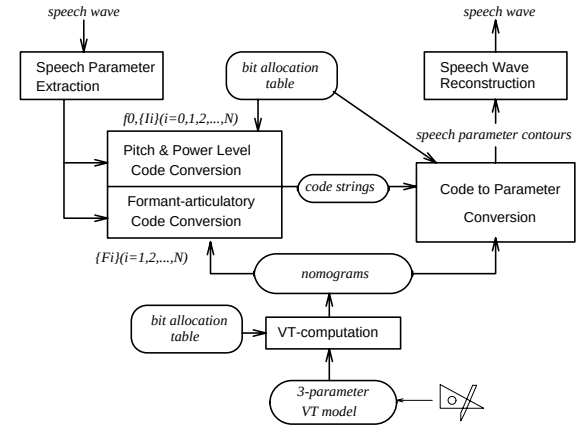


**Figure 3:** Low bit rate speech coding system.

The second is the mapping stage between formant frequencies and articulatory parameters using the nomograms, in which a parameter vector ***P*** minimizing equation (3) is determined frame by frame. Extracted code strings are fed into the code to parameter conversion stage and a speech waveform is reconstructed by the acoustic parameter contours.

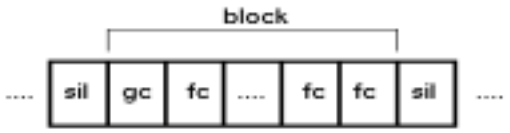


**Figure 4:** Code string format where *sil* = silence code, *gc* = group code, and *fc* = frame code.

## 2.5 Format of a Code String

We have conducted articulatory encode-decode experiments to measure formant distortion characteristics for several bit allocation schemes of ***P***. Speech samples for the experiments are 126 sentences uttered by 126 speakers included in TIMIT and ASJ databases.

There are several acoustic characteristics with regard to speaker's individuality such as average pitch level (Av-*F0*) or average vocal tract length (Av-ξ). These characteristics have no need to be reported with frame by frame. We introduced a group code for these two parameters that affects the group of consecutive frames for presenting these characteristics.

Figure 4 shows a code string format. One group code and one consecutive frame code set construct one block code. Table 1 shows contents for each code of the format.

| Code | Contents |
|---|---|
| *sil* | b/h/sil |
| *gc* | b/h/sil, Av-ξ, Av-*F0* |
| *fc* | b/h/sil, ***P***, ***I***, *F0* |

**Table 1:** Contents of each code in the format. Parameter b/h/sil is a buzz/hiss/silence index,. Av-ξ is an average of vocal tract length multiplier over a block. Av-*F0* is an average of fundamental frequency. Vector ***P*** is an articulatory vector,{*A1, X3, A3,* ξ}. Vector ***I*** is a formant intensity vector, and *F0* is fundamental frequency for each frame.

The Silence code and the group code have equal bit length to the frame code. Therefor, these two codes include several dummy bits.

# 3. EXPERIMENTS

## 3.1 Formant Frequency Distortion

Average formant distortion function is defined by the following equation. In the equation (4), *Gij* is a reconstructed j-th formant frequency at i-th frame.

$$e_j = \sqrt{\sum_{j=1}^{N} w_j \{10\log(F_{ij}/G_{ij})\}^2 / \sum_{j=1}^{N} w_j} \qquad (4)$$

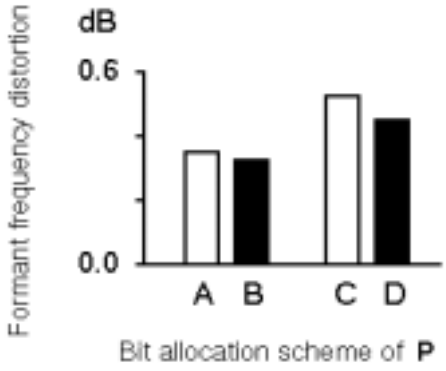


**Figure 5:** Average formant frequency distortion for four bit allocation schemes of ***P***.

Figure 5 shows average formant distortions for the bit allocation schemes in Table 2.

| Scheme | *A1* | *X3* | *A3* | ξ | *Av-ξ* |
|---|---|---|---|---|---|
| A | 4 | 6 | 4 | 4 | 0 |
| B | 4 | 6 | 4 | 4 | 3 |
| C | 2 | 5 | 2 | 3 | 0 |
| D | 2 | 5 | 2 | 3 | 3 |
| | | | | | Bit |

**Table 2:** Bit allocation schemes for formant distortion measurements.

The first scheme A has the same bit allocation pattern used in the 7.2kbps coder [2] and it's bit length is 18. The scheme C is tried one in this paper and the bit length is 12. For improvement of formant distortion, the scheme B and D are added 3 bits for Av-ξ to A and C respectively. The change of the distortion for the scheme D ranged from 0.32 to 0.62dB with 126 speech samples and the average distortion is 0.45dB.

The supplement of 3 bits to Av-ξ had a large effect on the short bit scheme rather than the other for reducing format distortion.

## 3.2 Formant Intensity Distortion

The largest dynamic range of formant frequency is generally for the first formant and it's range is 15dB or below. However, intensity dynamics for the first 4 formants have ranges above 40dB. It will be hard to realize the same distortion level as in the formant frequency domain with a low bit length.

From this thought, we allocate more bits to the first 4 formant intensities for reducing spectral envelope distortion rather than keeping an absolute speech power level. Table 3 shows the intensity distortion levels for the two bit allocation schemes. The bits were allocated to the first 4 formant intensities and the distortion level was an averaged over the formants. Scheme A is that of the 7.2kps-coder [2]. Scheme B is tried one in this paper and the distortion level was about 2.7 times of the formant frequency distortion in D (Figure 5).

| Scheme | Bit length | Intensity distortion(dB) |
|---|---|---|
| A | 18 | 0.30 |
| B | 10 | 1.23 |

**Table 3:** Formant intensity distortion for the bit allocation schemes of ***I***.

## 3.3 Low Bit Rate Coding

We have conducted articulatory encode-decode experiments with the bit rate of 3.2kbps (the frame interval 10 ms) and 1.6kbps (the frame interval 20ms) for speech samples uttered by 4 speakers (two males and two females) in ASJ and TIMIT speech databases.

Table 4 shows the bit allocation form for the experiment. Formant information described by the first 4 parameters in the table spends 12 bits per frame. Sound source parameters of *P0, F0, B/H/sil*-index, and formant intensity vector ***I***={*I1, I2, I3, I4*} spend 20 bits. Parameter Av-$\xi$ and parameter Av-*F0* in the group code (Figure 4) were assigned 3 bits on each parameter.

| Parameter | Bits/frame |
|---|---|
| *X3* | 5 |
| *A1* | 2 |
| *A3* | 2 |
| $\xi$ | 3 |
| *P0, F0, B/H/sil* | 3,5, 2 |
| ***I*** | 10 |
| | **Total: 32bits** |

**Table 4:** Bit allocation form of the frame code (Figure 4) for the low bit rate coder.

We have confirmed that good quality speech synthesis is achieved in both conditions of 3.2kbps and 1.6kbps. However, we presumed the case of 1.6kbps that unequal rate of frame sampling process will be better to preserve articulatory intelligibility of a high speaking rate utterance. Wave files, 00644_01.WAV and 00644_02.WAV, are reconstructed speech samples for English and Japanese speakers respectively. The contents are formed by a sequence of male-voice-3.2kbps and male-voice-1.6kbps for both of English and Japanese materials.

# 4. Discussion

We have presented a new low bit rate coder using a formant-articulatory parameter nomogram. From the experiments, we confirmed that articulatory parameter representation of speech sounds has the advantage of providing a compact and universal code set for multi-language synthesis application. The coding system is still under development. It will be needed to introduce a method of unequal rate of frame sampling and articulatory interpolating/smoothing algorithms for improvement of quality of synthesized speech.

# 5. ACKNOWLEDGMENT

We wish to thank Dr. Nobuyuki Ohtsu, Director of the Machine Understanding Division and all the members of the Speech Signal Processing Laboratory for the usual discussion and support.

# 6. REFERENCES

[1] H. Ohmura, K. Tanaka, "Evaluation of a Speech Synthesis Method for Nonlinear Modeling of Vocal Folds Vibration Effect", Proc. ICASSP97, pages 935-938, 1997.

[2] H. Ohmura, K. Tanaka, "Segmental feature extraction and coding for speech synthesis," Proc. EUROSPEECH'99, Vol. 3, pp1471-1474 (Sep. 1999).

[3] M. Bavegard, G. Fant, "Parameterized VT Area Function Inversion", Proc. ICSLP96, Pages 961-964, 1996.

[4] G. Fant, Acoustic Theory of Speech Production, page 74, Mouton, 1960.

[5] T. Kobayashi, S. Itahashi, S. Hayamizu, T. Takezawa, "ASJ Continuous Speech Corpus for Research", J. A. S. J. Vol. 48, pages 888-893, 1992.