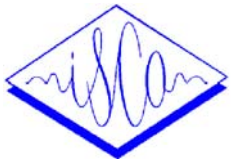

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SEGMENTAL FEATURS EXTRACTION AND CODING FOR SPEECH SYNTHESIS

H. Ohmura, K. Tanaka
Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Ibaraki 305-0045, JAPAN
ohmura@etl.go.jp, ktanaka@etl.go.jp

ABSTRACT

This paper describes a segmental feature extraction and speech coding method in an acoustic-articulatory domain using nomograms that represent a mapping between formant frequencies and articulatory parameters. The vocal tract model is a modified Fant model, in which we newly introduced a parameter for successively adjusting vocal tract lengths. We investigated first the relationship between formant contours and those of articulatory parameters and found the effectiveness of the articulatory domain for organizing acoustic-phonetic features with little dependency upon languages. Next, we applied the method to the low bit rate coder and confirmed that good quality speech synthesis was achieved in the condition of 18 bit used for articulatory code words.

10.21437/Eurospeech.1999-379

1. INTRODUCTION

Formant representation of acoustic characteristics of speech sound will provide a natural and convenient means for organizing and processing their features. In addition, it will be expected that formant contours will be easily converted to articularoty parameter sequences which brings compact and informative form of sound characteristics to many application of speech. To obtain these advantages in speech processing, we devoted to developing a formant based speech analysis and synthesis system [1]. In this system, formants are automatically determined and temporal, intensity, and spectral properties of speech are described by time sequences of formant frequency and its intensity parameters.

Several approaches for articulatory system and inversion have been presented [2], [3]. In the application of these systems to real speech processing, they employed no adaptation factor. It is however a necessary condition to adapt the model to input speech. Therefore, we introduce an adaptation factor that relates to the vocal tract length. We previously developed a formant-based speech synthesis system that employs nonlinear effect of vocal folds vibration and confirmed by an evaluation test that our system is superior to conventional LPC-based speech synthesis systems in voice quality [1]. The present method is intended to improve this formant-based system by introducing the nomograms that are generated by a computational vocal tract model. 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.

The focus of this paper is to present a flexible vocal tract model and get compact representation of speech by encoding only formant frequency allocation information, instead of a frequency spectrum, into an articulatory code string. The following sections describe the analytic investigation of vocal tract model, generation of the codebook, and analysis-re-synthesis experiment results.

2. ARTICULATORY CODING METHOD

2.1 Five Parameter Description of the Model

As a vocal tract model related with formant allocation information will become a stylized and simple model. 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 five variables is shown in *Figure 1*.

e 1. A vocal tract model described variables, A1, X3, A3, X5, and x.

The first three parameters, *A1, X3, A3* are the same as those in FANT-model. Parameter *A5* is the cross sectional area function at the constriction formed by the tongue tip and the hard palate. Parameter x is a multiplier of the vocal tract length that is calculated to minimize a matching error between formants generated from the model and formants extracted from input speech. Parameter *X5, L0, L3,* and *L5* are kept in constant. Parameter *L1* is a function of *A1*.

1.2 Mapping Algorithm

Parameter x causes a linear movement in the logarithmic frequency domain. The error function *E* is defined in equation (1).

Parameter x is calculated by equation (2) in advance.

where E_i is an error function at *i*-th frame of running analysis; $F_{i,j}$ is *j*-th formant extracted at *i*-th frame; and $G_{\boldsymbol{P},j}$ is *j*-th formant produced by vocal tract computation with a given parameter set $\boldsymbol{P}$={*A1, X3, A3, X5*}. Symbol $\boldsymbol{w}$ is a weighting function to emphasize lower formants. The best-matched parameter set $\boldsymbol{P}$ for an input formant vector $\{F_{i,1},F_{i,2},\ldots,F_{i,N}\}$ is obtained by minimizing the error function E_i among combinations of four numerical values in the given ranges.

1.3 Active ranges for the parameters

We have conducted analysis experiments for estimating active ranges of the parameters using speech samples in TIMIT and ASJ databases [5]. After a preliminary experiment, the ranges for the parameters, *A1*, *A3*, and *X3* were given as follows.

2.3.1 The range of x

For estimating x, parameter *L3* is 5cm and *L5* is out of consideration.

The procedure for estimation of x is as follows:

- Extraction of formants trajectories for voice parts of a speech sample.
- Computation of the error function (1) for all combination sets of quantized parameters in equation (3).
- Determination of x for each frame at the minimum of the error function.

If one takes the range of activity to $0.6 \le x \le 1.4$ for 370 sentences uttered by 37 speakers (TEST-set, DR1 and DR2) in TIMIT database, sample frames of 99.7% are included in the range.

2.3.2 Parameter L3

Next, we estimate the most fitted length *L3* for the databases using the procedure same as that in the previous section. Parameter *L3* ranged from 1 to 8cm. The best fitted value of *L3* for an input sentence is defined by the value at the minimal average error. The distribution of *L3* concentrates almost between 5.0 and 5.5 and the average is 5.2cm for speech samples of 126 sentences uttered by 126 speakers in ASJ and TIMIT databases.

1.4 Generation of the Nomograms

The nomogram is a codebook representing the functional relationship between a formant frequency

vector ***G*** given by vocal tract computation and an articulatory parameter vector ***P*** of {*X3, A1, A3, A5*}. Parameters *L3, L5, X5,* x in *Figure 1* are currently given as 5.2cm, 1.5cm, 11.5cm, and 1.0 respectively. The area function of the model described by ***P*** in *Figure 1* is smoothed and sampled at an equal spatial rate for constructing a cross sectional area series. The formant vector ***G*** is given as a polynomial root set for the denominator of the vocal tract transfer function. Table 1 shows a bit allocation scheme for making the codebook. The number of entries in the codebook is 76544 with excluding overlaps of *X3* and *X5*.

Figure 2. Block diagram for the articulatory coding of speech.

Table 1. A bit allocation scheme for the co

Parameter	Max.	Min.	bits
X3	12.5	2.5	6
A1	8.0	0.2	4
A3	5.0	0.2	4
A5	8.0	0.2	4

1.5 Experimental System

Figure 2 shows the block diagram for the experiments. 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. The second is the mapping stage between formant frequencies and articulatory parameters, in which a parameter vector ***P,*** {*X3, A1, A3, A5,* x}, minimizing equation (1) is determined frame by frame.

Figure 3. An example of articulatory parameter time patterns for a voice part "all year" in a sentence SA1 *in* TIMIT *database, uttered by a male speaker. Vertical lines are boundaries between consecutive phone segments.*

3. EXPERIMENTS

3.1 Segmental Features Extraction

Speech samples for the experiments are sentences uttered by 126 speakers included in TIMIT and ASJ databases. Average formant distortion e in the following equation for an input sentence ranged from 0.22 to 0.56dB/frame.

$G_{i,j}$ is *j*-th formant frequency calculated from the articulatory parameter vector ***P*** at *i*-th frame.

Figure *3* shows an example of articulatory parameter time patterns extracted for voice parts of a sentence uttered by a male speaker in TIMIT

database. Some segment boundaries are clearly observed in the articulatory parameter domain whereas the formant patterns are smooth in the corresponding sections.

1.2 Low Bit Rate Coding

We have applied directly the articulatory coding method to a low bit rate coder. *Table 2* shows a bit allocation form for the experiment. Formant allocation information described by the first 4 parameters in the table spends 18 bits a frame. Sound source parameters of *F0*, *BH*, and formant intensity vector $\boldsymbol{I}=\{I_1, I_2,...,I_N\}$ spend 54 bits in total. We have conducted articulatory encode-decode experiments using a system in *Figure 4* with TIMIT and ASJ speech databases. All formant distortions for the same samples in the previous section go into the range from 0.49 to 0.85dB/frame. The resulting bit rate is 7.2kb/s at the frame interval 10ms. We have confirmed that good quality speech synthesis is achieved in this condition (o010-e.wav and o010-j.wav are reconstructed speech samples for English and Japanese speakers respectively).

Table 2. A bit allocation scheme for the experiments of low bit rate coding of speech.

Parameter	Bits/frame
X3	6
A1	4
A3	4
x	4
F0, BH	16
I	38

2. Discussion

As indicated in *Figure 3* segmental movements are conspicuous by the time patterns of *X3* through the test data set. Within broad segments of *X3*, relatively smooth movements are observed on the patterns of *A1*, *A3*, and *A5*. These segmental features do not directly imply real human vocal tract movements. However, we think that this segmental property is methodologically profitable for extracting and organizing speech sound characteristics.

Figure 4. Speech analysis and re-synthesis system.

From the experiments of low bit rate coding using Japanese and English speaker data set, it will be expected that articulatory parameter representation of speech sounds has the advantage of providing a compact code set for applications to multi-lingual synthesis.

The vocal tract model and the coding system are still under development. It will be needed to introduce a method for constraining the model on the time-varying characteristics of the parameters for more minute presentation and improve the system to provide appropriate patterns for the articulatory parameter extraction of speech sounds.

3. ACKNOWLEDGMENT

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

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

[3] P. Badin, C. Abry, "Articulatory Synthesis from X-Ray and Inversion for an Adaptive Speech Robot", Proc. ICSLP96, pages 1125-1128, 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.