HALF-SYLLABIC UNITS FOR SPEECH PROCESSING - AN AUTOMATIC SEGMENTATION

Mamoru NAKATSUI

Radio Research Laboratory, Ministry of Posts and Telecommunications, 4-2-1, Nukuikita-machi Koganei-shi, Tokyo, Japan 184

INTRODUCTION

The half-syllabic units proposed here are units each of which has segment boundaries at steady portions and preserves a transition between two phonetic units. Segment boundaries are basically determined by the minima (valleys) of gross spectral variation measure. The spectral variation measure is defined as the root-mean-square value of the slopes of the weighted regression lines calculated from LPC cepstrum parameters over several frames. The maxima (peaks) of the measure will serve as the reference points for further processing.

In speech synthesis by rule, it is primarily important to select synthetic units that have reasonably small size of inventory to represent spoken utterances and, at the same time, are easily concatenated. In speech analysis-synthesis system at very low-bit-rates such as phonetic vocoding, the units must, further, be automatically segmented and be suitable for interpreting into or matching with the reference units. These requirements on segmentation and matching or labelling are expected to be satisfied for speech recognition system in many cases and for providing useful tools for automatic generation of the inventory of concatenative units.

Syllables and Half-Syllables

One of the selections for the unit to be used in concatenation-based speech processing is the syllable. There have been several discussions and experiments on syllable as recognition unit [1-4]. The syllable has been also used as a unit in synthesis by rule of Japanese [5]. One of the disadvantages to using syllables as units is that the size of inventory representing spoken utterance is large. This problem can be solved by introducing smaller units such as the half-syllabic units proposed here, since much of the co-articulation among phonetic units is associated with transition regions and since boundaries at the steady portions outside transitions are easily definable.

There exist similar units known as dyads [6], diphones [7], or demisyllables [8] which have the common concept of incorporating the transition between phonemes. The context-dependent diphones have been utilized in constructing a phonetic vocoding system [9]. The demisyllables originally proposed for use in a high-quality concatenative speech synthesis [8] have been successfully applied to constructing concatenative templates in the word recognition for large vocabularies [10].

Dynamic Spectral Feature

The gross spectral variation measure derived from a series of LPC cepstrum coefficients has been proposed as a dynamic measure investigating individuality of utterances [11]. This dynamic measure has been used in the study on Japanese CV-syllable perception and it has been shown that dynamic spectral feature plays a primary role in phoneme perception [12]. Usefulness of the dynamic measure in comparison with its static counterpart has also been shown in word recognition experiment [13]. The dynamic measure has also been applied to the segmentation in a very low-rate speech coding where boundaries of the pattern are defined by the maxima of the measure [14].

The half-syllable-like unit has not yet been applied to processing Japanese utterance as far as we know. Our expectation for the units proposed is in the relatively small size of inventory in representing Japanese utterances, since Japanese has relatively simpler syllable organization than that of English. Our ultimate objective is to provide nearly universal units suitable for processing spoken Japanese. As the first step to that goal, our current interest is in confirming whether the proposed units meet the basic requirements, that they would be

1) automatically and reliably segmented,
2) closely related to certain linguistic units, and
3) suitable to acoustic phonetic observations

in the course of constructing the analysis-synthesis system like segment vocoder. This paper reports a preliminary experiment on segmentation of speech signal into the units proposed and some observations of the result with respect to the above requirements.

SEGMENTATION ALGORITHM

Speech sample is bandlimited to 4 kHz and digitized to 12 bits at sampling frequency of 10 kHz. Linear prediction (LP) analysis is carried out on a frame-by-frame basis (100 frames/s). Additional acoustic parameters currently used are a log power P, a zero-crossing count Z, a count for sign change of waverform X, and the first order PARCOR coefficients k1. The spectral variation measure D(j) for j-th frame is calculated by

$$D(j) = \left[\frac{1}{12}\sum_{i=0}^{11}\{u(i)\cdot a(i,j)\}^2\right]^{1/2} \qquad (1)$$

where weight u(i) is currently one for all i and a(i,j) is the i-th coefficient of the weighted regression line of LPC cepstram parameter over several frames. A triangular weighting function is currently applied over seven frames.

With these acoustic parameters, signal processings on input speech are basically carried out in the following steps (descriptions in parentheses are associated with indications in Fig. 1):

1) appointing candidates for segment boundaries at local

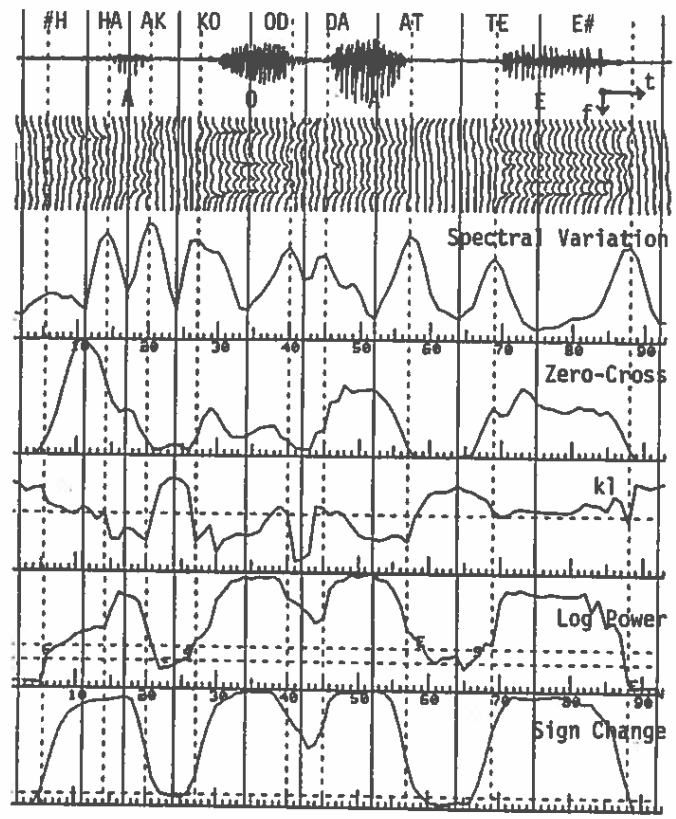

Fig. 1. An example of segmentation and acoustic parameters.

Table 1. Segmentation errors for 455 segments

Position	Initial	Middle	Final	Total
Deletion	7	7	0	14
Insertion	0	11	1	12
Total	7	18	1	26

maxima of spectral variation measure (vertical lines),

2) adjusting the segment boundaries by start and end points of speech interval (S and E),

3) classifying the boundaries into sub-groups of phonetic units and assigning candidates of vowel identity,

4) assigning the reference points at maxima of the ' variation measure for time arraignment in spectral matching with the reference patterns (dotted vertical line),

5) adopting weights for pattern matching inversely proportional to the normalized values of the spectral variation measure.

Among those steps, 3) to 5) are beyond the scope of this report. However some preliminary trials will be shown later. As for 2), a hysteresis characteristic is given to the decisions of speech interval (from S to E) providing two levels of thresholds for the log power P and the decisions for the non-speech interval associated with intervocalic unvoiced-stops are stabilized by referring the count of sign change X. The minimum (valley) just before S and that just after E were assigned as boundaries of the utterance.

RESULT OF PRELIMINARY EXPERIMENT

Sixty names of Japanese cities spoken by a male adult were used as the test material for segmentation process. It was estimated that the test material consisted of 455 half-syllabic units by our visual inspections.

Segmentation

Fig. 1 shows an example of segmentation where the segment boundaries are denoted by vertical lines and reference points for matching are denoted by the dotted vertical lines. Result of an automatic segmentation of the test material is summarized in Table 1. Correct rate of segmentation is more than 94 %. Most of the deletions of segment boundaries at word-middle are associated with intervocalic [r] and [g] sounds. These problems are going to be solved by the test material having wider spectral bandwidth. It is revealed that problems concerning deletions at word-initial and insertions at word-final are also due to inadequacy of the test material such as low signal-to-noise ratio and over-cuts at the beginning and the end of utterances. So, new test material suitable for our experiment is under preparation, because the current sample has been prepared for other experimental purpose.

Most of insertions of segment boundaries, extra boundaries than expected, are associated with nasal and unvoiced stop consonants. It is observed that extra segments correspond to nasalized vowels and aspirations after stop bursts. The detailed observation for much speech material from the point of view of acoustic phonetics should be made in order to give such solution and interpretation systematically.

Some Observation on Segments

Signal processings described below have not been fully automatized yet and, further, most of the observations have been based on a small set of test material. Alphabets at the top of Fig. 1 are our tentative labelling for the segments (units). Segment boundaries are first classified as either vowels or one of a consonantal group such as voiced-stops and unvoiced-fricatives using a set of acoustic parameters. Spectral distances between spectral frame of the boundary and single frame reference patterns including isolated five vowels and nasal murmurs were used as additional information in the classification.

Alphabets on the segment boundaries just below waveforms in Fig. 1 denote the first candidates of vowel identity showing minimum spectral distance. Ninety percent of vowel boundaries are identified as the first candidates and the remaining ten percent as the second for a sub-set of the test material having 40 vowels. Linear spectral matchings of the CV-type segments with the CV-syllable reference patterns were tried after pre-selections using those data on consonantal group and the first and second vowel candidates described above. In the matching, time arraignment between the segment and the reference pattern was adjusted in such a way that the reference points of both patterns coincide. It is observed that correct CV-syllable appears within the top three candidates for most cases in this arrangement.

CONCLUDING REMARKS

Although our experimental evidence is at quite a primitive stage, the half-syllabic units proposed seem to have potential to meet three basic requirements described above. Among many problems left to be solved, our current interests are in (1) preparation of speech material suitable for our objectives, including city-names at different speeds of utterence and conversational utterances, (2) improvement and tuning of the segmentation algorithm applicable for these speech data, and (3) the detailed observation of the units in acoustic phonetic aspect and systematic organization of classification algorithm.

REFERENCES

[1] O. Fujimura: "Syllable as a Unit of Speech Recognition," IEEE Trans. ASSP-23, 82-87, 1975.

[2] P. Mermelstein: "A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech," IEEE Trans. ASSP-23, 79-82, 1975.

[3] M. J. Hunt, M. Lennig and P. Mermelstein: " Experiments in Syllable-Based Recognition of Continuous Speech," Proc. ICASSP'80, Denver, 880-883 ,1980.

[4] H. Fujisaki, K. Hirose, T. Inoue and Y. Sato: "Automatic Recognition of Spoken Words from a Large Vocabulary Using Syllable Templates," Proc. ICASSP'82, Sandiego, #26.12 (Vol. 3), 1982.

[5] Y. Tohkura and Y. Sagisaka: "Synthesis by Rule using CV-syllable and Its Speech Quality," Trans. Commit. Speech Res. Acoust. Soc. Japan, #S80-47, Oct. 1980(Japanese Text).

[6] G. E. Peterson, W. S.-Y. Wang and E. Sivertsen: "Segmentation Techniques in Speech Synthesis," J. Acoust. Soc. Amer. 30, 739-942, 1953.

[7] N. R. Dixon and H. D. Maxey: "Terminal Analog Synthesis of Speech Using the Diphone Method of Segment Assembly," IEEE Trans. AU-16, 40-50, 1968.

[8] O. Fujimura, M. J. Macchi, and J. B. Lovins: "Demisyllables and Affixes for Speech Synthesis," Proc. 9th ICA, Madrid, #I-107, p. 513, 1977.

[9] R. Schwarts, J. Klovstad, J. Makhoul and J. Sorensen: A Preliminary Design of a Phonetic Vocoder Based on a Diphone Model," Proc. ICASSP'80, Denver, 32-35, 1980.

[10] A. E. Rosenberg, L. R. Rabiner, J. G. Wilpon and D. Kahn: "Demisyllable-Based Isolated Word Recognition System," IEEE Trans. ASSP-31, 713-726, 1983.

[11] S. Sagayama and F. Itakura: "On Individuality in a Dynamic Measure of Speech," Spring Meeting of Acoust. Soc. Japan, #3-2-7, 589-590, 1979 (Japanese Text).

[12] S. Furui: "On the Role of Spectral Transition in Phoneme Perception and Its Modeling," 12th ICA, Tronto, 1986 (in print).

[13] S. Furui: "Speaker-Independent Word Recognition Using Dynamic Features," IEEE Trans. ASSP-34, No. 1, 1986 (in print).

[14] Y. Shiraki and M. Honda: "Very-Low-Rate Speech Coding Using Time Space Spectrum Patterns," Trans. Commit. Speech Res. Acoust. Soc. Japan, #S84-06, Apr. 1984(Japanese Text).