JUNCTURE SEGMENTATION OF JAPANESE PROSODIC UNIT BASED ON THE SPECTROGRAPHIC FEATURES

Kitazawa Shigeyoshi   Itoh Toshihiko   Kitamura Tatsuya

Department of Computer Science
Faculty of Information, Shizuoka University

{Kitazawa, t-itoh, kitamura}@cs.inf.shizuoka.ac.jp

ABSTRACT

A very detailed segmentation of prosodic phrase was carried out in order to construct a Japanese prosodic database. Boundaries correspond to junctures between phrases including C|C and V|V clusters. The "prosodic phrase" we introduced as a unit of the segmentation was defined and regarded as a unit of language speech perception. For the exact segmentation, the wide-band spectrum, the narrow-band spectrum, fine speech wave and fundamental frequency shapes and transition of amplitude of the higher order formants were adopted to enumerate the candidate points for the segment boundary. Fine time adjustment by the steps of the respective fundamental period of the speech determined the exact boundary. To maintain the consistency of the segmentation, one person ascertained the entire segment carefully. The database, referred to here as "Japanese Multtext", contains read style speech and spontaneous style speech by three male speakers and three female speakers in Tokyo dialect.

1. INTRODUCTION

Although information processing-research has been performed about each segmental sound of the phonetic system of the language, more research on bigger units, i.e., prosody, is necessary for further innovations to occur in the speech synthesis and recognition technology. Although we are going to develop a data base used as the foundation for prosody research, it is important to divide exact segmental sounds. For example, although labeling by Tone/break indices (ToB) has been widely used as a technique in prosody research, it should be assumed there that the automatic segmentation pause by HMM used in conventional voice-recognition studies is inadequate [1]. The purpose of this research is to find an accurate method for defining the boundary of a phoneme division. Although manual labeling is more exact than automatic labeling, for prosody research, an exact segmentation standard needs to be introduced also into the kind of phoneme series conventionally made ambiguous. Moreover, this segmentation method should incorporate the prosodic features together.

1.1. Problems of Conventional Method

In the DARPA-TIMIT database, the phoneme is assumed to not overlap in time, and label attachment is carried out by DARPAbet by inspection from the spectrogram [2]. Takeda et al. of ATR engaged in the label attachment of Japanese, and collected the details in a report [3]. Takeda etc. added a detailed segmentation of the gliding portion, the event, and the allophone. Detailed inspection was performed about plosives, fricatives, and nasals. Semivowels and vowels, however, were not clearly segmented, and the maximum changing point of the spectrum was used by the automatic segmentation. Moreover, when the same vowel continued (VV juncture), the method for dividing individual vowel sounds was not described. In recent years, the technique of automatic syllable segmentation by HMM has been established. However, even with correction by hand later, the problem with a difficult segmentation has yet to be solved. Difficult cases are carried out assuming isochrony as an expedient solution. However, this "isochrony" tends to be verified in the "prosody corpus" rather than assumed in foresight.

1.2. Problems to be Solved Here

In Japanese, the same vowel cluster and a consonant cluster often appear as the boundary of a prosodically important unit, that is, as a juncture. Junctures constitute indispensable information for the prosody database in terms of providing a means to measure these units. Normally, VC junctures are not very difficult to separate. However, the segmentation of CC junctures following an unvoiced vowel and a moraic nasal or of VV junctures poses considerable difficulty. Segmentation will be especially difficult when the vowel succession is between similar or identical vowels. Further research is also called for to address the problems posed by semivowels and gliding sounds that are difficult to segment using the conventional spectrogram method. These phoneme chains can be clearly segmented by using prosody information together.

Judgment of the disjunction by listening became easy dividing per prosodic phrase. Apart from that, for the moment, CV boundary is not so important prosodically, but this is not taken into consideration in this study.

2. PROSODIC UNIT AND JUNCTURE

The attributes of prosody are height of pitch, loudness of sound, and length of sound. Additional attributes include rhythm (isochrony of stress, mora, or syllable) and melody (intonation).

2.1. Prosodic Phrase

The prosodic phrase is the minimum unit of prosody control from the point of speech generation. It is the unit that corresponds to the so-called "chevron" fundamental-frequency pattern in rule synthesis [4]. It is the minimum unit that may bear an accent phonologically. In continuous speech, since the
reduction of an accent occurs, many prosodic phrases that do not contain an accent kernel also appear. The prosodic phrase is considered when it corresponds to an independent-word + an attached-word, i.e. a clause, in Japanese syntax theory. When it becomes too long unit such as a name of a place or a name of a person, it is divided for purposes of analysis into suitable composition elements (so as not to exceed 5 or more morae).

2.2. Perceptual Unit

We define a prosodic phrase as a unit that is perceived in language speech. Since a "prosodic phrase" is the minimum speech interval in the speech intervals that we can hear without so much sense of incongruity compared with the time of hearing in a continuous speech when taken out and listened to the speech interval, here we claim that a prosodic phrase is "a perceived unit ". Even when we cut out and hear the speech section of some length situated between two silent sections, there is probably no sense of incongruity. Although it is further subdivisible, a prosodic phrase uses this speech interval as the portion that nearly corresponds to the above-mentioned clause, for the time being.

2.3. Prosodic-Phrase Segmentation

In order to define the prosodic phrase, an example of the segmentation of Section 1 (fhk_o3r_sp) of our database is shown. A space shows a partition of a segment and # shows a pause. In the experiment, the subject hears segmented and isolated prosodic phrases, which are listened to to confirm that the segments are perceptually separated units.

In DARPA-TIMIT, the stop consonant is clearly divided into

5.1. Stop Consonant

In DARPA-TIMIT, the stop consonant is clearly divided into

4. METHOD OF SEGMENTATION

The labeling used Wavesurfer [8] free software. The prosodic phrases to be labeled were segmented with reference to the speech waveform and the spectrogram of wide-band and narrow-band, and then checked by listening to the speech segment.

4.1. Spectrogram

Usually, with a wide-band spectrogram suitable for inspection of the formats used for spectrogram reading, a narrow-band spectrogram is used for inspection of fundamental frequency. As fundamental-frequency data extracted by the algorithm is not necessarily be trusted, a narrow-band spectrogram gives abundant information to spectrogram readings through simple processing.

For wide band spectrogram purposes, the parameters in Wavesurfer are set for 256 FFT window length, a 128 point Humming window, and analysis band width of 375Hz.

The parameters for narrow band spectrogram analysis in Wavesurfer are set for 4096 FFT window length, a 2086 point Humming window, and analysis band width of 23Hz.

4.2. Voice Waveform

Speech signals are 48kHz and 16 bits. The amplitude of a speech waveform gives information important for the segmentation of a phoneme unit. A delicate change of the periodic waveform was observed when expanding a time-axis. This also provides important information about a phoneme boundary.

4.3. Section Listening

The determination of the phoneme interval by listening is not very accurate. However, in the junction interval of a vowel and a semivowel, judgment of final boundary adjustment must be based on listening. Such phoneme boundaries are especially ambiguous because their division is gradual, and not distinct. In this experiment listeners attempted to identify phoneme crosspoints.

5. PRACTICAL TECHNIQUE FOR EACH PHONEME ENVIRONMENT

Formerly, the guideline of DARPA-TIMIT segmentation has been the standard [2]. We refined the standard further by prosodic supplementary information, and performed the segmentation of junctures such as {moric nasal or devocalized vowel}-consonant junctures, identical nasal-nasal junctures, identical vowel-semivowel junctures, vowel-semivowel junctures or semivowel-vowel junctures, etc.

5.1. Stop Consonant

In DARPA-TIMIT, the stop consonant is clearly divided into
three acoustic-events; a burst; and an aspiration (delay at the voice onset time). Unless vocal-cord vibration appears clearly on a spectrogram, it is assumed to precede a 50ms closure interval with the stop consonant in the initial position of utterance. In any case, a burst and an aspiration are sustained intervals of a plosive. Emphasis was put on a segmentation clarifying a closure interval.

5.2. Fricative Juncture

The boundary of a fricative is easily distinguishable with the energy of a high frequency band. Segmentation of a fricative juncture may have to be carried out when a vowel, which is sandwiched between voiceless fricative consonants, devoices. In this case, the difference in a spread of a spectrum can divide. A part of the utterance /katakusijinnyuHde/ ("because of housebreaking" where "house" and "breaking" are divided), for example, composes a consonant juncture since a /sji/ follows a devoiced /ku/.

5.3. Nasal Juncture

In DARPA-TIMIT, a nasal is detectable by a fall of the high frequency energy in a zone upper 500Hz, the discontinuity of formant frequency, and a fall of formant amplitude. It is divided consonant duplication (/n/-/n/ of example "in nine") to which a vague boundary is set near the middle point. For our segmentation in the following example, however, /m/-/m/ of /teNmade/ ("till the point" where "the point" and "till" are divided; the moraic nasal /N/ is assimilated to the following nasal /m/) had become a consonant juncture. The rise of the 2nd formant amplitude in the latter-half consonant /m/ was a cue to the boundary. Similarly, /syaNkusyosjimadejaga/ ("as far as the junction") is signaled with a less evident cue, however, considering the bending point of the 2nd, the 3rd, and the 4th formants, it was able to divide the juncture clearly. In this example, the boundary of /m/-/m/ is not a middle point but inclines to the second half. From this, it can be said that it is inaccurate to set as a midpoint by the conventional method when a boundary is unknown. In other examples, a long closure of a stop consonant may come after the short vowel that follows after a moraic nasal, consequently duration of a mora is not simply isochronous.

5.4. Liquid and Glide

In DARPA-TIMIT, the segmentation of a liquid and a glide is presupposed to be very hard. In /y/ F2 formant transitions, F1and F2 serve as a key in /l/ and /w/; and , /r/ in F3. In such cases, DARPA-TIMIT were considered to be the segmentation point at the halfway of a formant transition. However, it is not exact to adopt this middle point, so we are adjusting by hearing the boundary to every 1 fundamental period.

5.4.1. /y/ juncture

The following example /cyoHdoyoiyoruno/ ("just good night") includes /oy/ and /iy/ junctures. At the /oy/ juncture, it sets the boundary at the onset where the F2 went up slightly from the /o/ level, not at the midpoint, with the linear frequency axis of spectrogram. The same is said of /al/ junctures. Concerning the /iy / juncture, it is the falling corner of F2 and it is the minimum point of the fundamental frequency in this case. In this case, /cyoHdoiyoiyoruno/ correspond to the boundary of a prosodic phrase. Hearing either side of this speech, this boundary point was the best suited.

5.4.2. /r/, /w / juncture

It may be hard to determine the phoneme of /r/ and /w/ by the formant transition visible on a spectrogram. The case /cumawaraigetu/ ("wife next month") where the feature was less clear is hard to read from the spectrogram. However, the principle that it is the lowest point of a fundamental frequency and the sudden decrease in amplitude of voice and higher-order formant can be determined. Finally, making the boundary move to right and left for every one fundamental period or one vertical line of a wide-band spectrogram, the sound of the section on either side is heard and the point of best separation is determined. Also a /w/ juncture in /watasjigawazjimasjina/ ("I am to Wajima city") was able to be treated similarly.
5.5. V1-V2 and VV Juncture

Vowels are easy to be segmented roughly but it is very difficult to do this precisely. There is no clear standard; researchers must rely on experience.

In a Japanese juncture, a vowel cluster occurs frequently that includes identical vowel chaining. With a phoneme /o/ as a start, a clause (a prosodic phrase) continues very often in contexts such as /*wo|o*/, /*to|o*/, and /*no|o*/ etc.. Moreover, /*ga|a*/ also appears frequently (although this /ga/ is usually produced as a voiced nasal sound).

Since these were difficult to pin down, they were not dealt with by voice recognition techniques until now. However, for prosody research, it is very important to disjoin /*o|o*/ and /*a|a*/ strictly. Segmentation performed through inspection of transition of fundamental frequency, verification by listening to both sides the split prosodic phrases, and observation of detailed feature of the waveform and the spectrogram are all necessary. When the two vowels standing side by side on the juncture mix mutually and cannot disjoin completely, we divide for symmetrical leakage on both sides of the boundary.

The portion of /jikogaiji/ ("there was an accident") is shown as an example of /*ga|a*/ juncture. Here, transformation of the amplitude of speech waveform and disappearance of the nasal formant from a nasalized /ga/ to an unasalized /a/ were made into the key.

![Figure 6: /ga|a/ juncture /jikogaiji/](image)

An example of /*o|o*/ is /waribikiwo|okonaQte/ ("are there any discount"). A stairs-shaped descent in fundamental frequency is observable on the narrow-band spectrogram.

![Figure 7: /wo|o/ juncture /waribikiwo|okonaQte/](image)

6. DISCUSSION

6.1. Consistency and Accuracy

Although limited to 480 small paragraphs, one person carefully investigated all segmentations consistently, realizing the perfect segmentation of all junctures. In an example of a typical phoneme automatic labeling by HMM, 70% is reported within 10 ms error and 95% is within 50 ms error. This error is too large for setting as the basis of prosody research. In our manual labeling, the boundary was adjusted in a fundamental period (that is, a male voice 0.8 ms and a female voice 0.4 ms of steps). Labeling work of about 30 seconds of speech has taken about 1 hour.

6.2. Remaining Difficulties

If a certain feature were found out by acoustic analysis, it is evident that segmentation has been performed with this as its basis. It was possible for us to collect such data thanks to high quality acoustic recording. A technique here may be unable to be adapted for acoustic recordings of low quality. Although segmentation is going to be detailed one by one, it is unknown how far detailing can be taken by this technique. It is considered to be difficult to carry out segmentation per phoneme. Supposing it is not suitable as a boundary if segmentation is difficult, it is also interesting to find cases where it is possible. The unit here segmented is considered to be a portion (prosodic unit) which the speaker and a listener are both aware of as a segment. This prosodic unit is a phenomenon corresponding to some levels described by BI (Break Index) of J-ToBI.

7. CONCLUSION

In the segmentation by the conventional method for inspecting a conventional waveform and the conventional spectrogram, if the phoneme chain that was difficult used prosody information together, it could divide clearly. Judgment of the segmentation by listening became easy by carrying out segmentation per prosodic unit.

8. ACKNOWLEDGEMENT

This research is based on the domain research specific (B) (2) subject number 12132204.

9. REFERENCES


