2—Q—13

Prosodic characteristics at pauses and restarts

W. N. Campbell
ATR Interpreting Telecommunications Research Labs

1 Abstract

This paper presents results from a preliminary study of spontaneous telephone speech of Japanese. It focusses on the prosodic environment of pauses and restart phenomena, and shows how variance in segmental lengthening and fundamental frequency contours across different speakers can be robustly expressed as normalised intonation profiles. It presents findings from an analysis of 50 one-minute excerpts from unprompted telephone recordings, examining profiles in the vicinity of silence in the waveform in order to determine which clues may be used to help identify and distinguish pauses to assist linguistic phrasing for a speech recognition system that includes prosodic information processing.

2 Introduction

For the robust processing of spontaneous speech, we need a method for incorporating prosodic as well as spectral information that is robust to speaker differences without the need for special training for a new speaker. An essential stage in the processing of such speech is the chunking of the input into phrase-sized units for linguistic processing. The pause provides a convenient clue to this, commonly occurring at linguistic phrase boundaries, but is not always reliable; pauses can occur mid-word, for example, or can be non-silent (filled), so a further independent source of information would be useful in identifying the valid phrase-dividing pauses.

The most readily available source of prosodic information is the fundamental frequency contour (f0), but this can vary considerably according to speaker-characteristics, and is not always reliably distinguishable throughout the utterance. The spectrogram filters out speaker-characteristics from the speech waveform; this paper proposes a method for similarly reducing the speaker-related variance in the intonation, while preserving lengthening and f0 contour shape information. We tested this transform using phrasal boundaries marked by pauses in an attempt to distinguish pauses located at valid phrasal boundaries from other kinds of pause.

ポーズの韻律的特徴
ニック キャンベル
ATR 音声翻訳通信研究所

2.1 Kinds of pauses

Interpausal runs have been cited as convenient units (or 'chunks') of speech for linguistic analysis [4]. They do frequently correspond to phrasal units [1], but the pauses that delimit them are not always valid interphrasal pauses. We can differentiate pauses *between* phrases from pauses *within* phrases, and further differentiate phrase-internal pauses into a) hesitation pauses (e.g., when the speaker is searching for an appropriate word), b) filled pauses (e.g., when holding the floor), c) restart pauses (e.g., during self-correction), and d) phonetic pauses (in plosives and glottalised portions of speech). In recording speech over telephone lines, there can also be unnatural pauses, such as when line problems occur or when a tape is faulty.

3 Reducing speaker variance

Materials for the analysis were selected to meet three criteria: a) the speech must be absolutely natural and spontaneous b) it must be representative of a wide range of speakers and c) the data must be publicly available, for replication of the results. The Oregon Graduate Institute's OGI telephone speech database includes recordings of spontaneous telephone monologues from 50 Japanese currently living in the USA. Manually transcribed narrow-phonetic labels provided the text source and segmental duration information. Fundamental frequencies (extracted from the speech waveform using Waves+ [2]) were also analysed in this study. There was considerable variety in speakers' sex, age, background, speaking styles, and topic.

To normalise between speakers, the following steps were taken: First, the segmental durations were z-score transformed and the F0 was smoothed using quadratic splines [3]. Then first derivatives of these values were taken to reduce both to a zero mean. Averages were taken over a 3-segment window (duration) and over a 250 msec window (pitch) to the left and right of each phone. To create the lengthening profile of each utterance, the averaged lengthening of the three phones following each segment (d_R) was subtracted from the averaged lengthening of the three phones preceeding it (d_L), and similarlly for the pitch profile, the averaged f0 values were expressed as a ratio ($\frac{p_L}{p_R}$). To bring them into the range around 0, they were reduced to $1 - \frac{p_L}{p_R}$.

3.1 Intonation profiles

The pitch profile $(1 - \frac{p_L}{p_R})$ has the following characteristics:
- zero-crossing: level part of the f0 contour
- positive values: rising f0
- negative values: falling f0
- depth/height from zero: steepness of fall/rise

The lengthening profile $(d_L - d_R)$ can be similarly interpreted:
- close to zero: no change in the lengthening
- abrupt zero-crossing: rapid change of state
- positive values: previous segments are longer
- negative values: later segments are longer

In both profiles, an abrupt rise through zero signals the most prosodically sensitive point, marking the end of a fall in the fundamental frequency, and showing phrase-final lengthening in the durations, hence indicating a prosodic phrase boundary.

The intonation profiles reduce the speaker-specific features of the prosody while preserving the structure of the lengthening and fundamental frequency contours well, indicating both the direction and points of change in rise and fall, and the steepness of the change at each point. Visual examination of the profiles shows clear demarcation of bunsetsu units.

4 Distinguishing pause types

Preliminary analysis of a subset of the data showed no reliable distinguishing characteristics in the prosody of the segments surrounding the different types of pause, detected spectrally; their distribution (in this data at least) seems to be independent of the prosody. This means that, for robust phrasing, other ways of distinguishing the pause types must be found.

Two attempts were made to use the profiles for pause scoring, weighting pauses according to their position relative to the contour. The first measured the depth and time below zero in both profiles before each pause, and the second measured the distance of each pause from a combined upward-crossing (i.e., the point where both lengthening and pitch marked a prosodic boundary). Although visual examination of the results appeared promising, a major difficulty in determining the 'right answer' to the presence and position of a phrasal boundary prevented a quantitative evaluation. In the case of self-correction pauses (e.g. ええ私がまれま ⇐ う ⇐生まれたのは日本の南のきゅう ⇐九州にあります), a binary choice can be easily made, but in a phrase such as (三年半 ⊗ ぐらい前に) or (テニスをする ⊗ ことが) how should a pause or boundary insertion be judged? Without a categorical definition of the appropriateness of a phrasal boundary, it is difficult to train and test for a quantitative evaluation.

Instead, on the assumption that valid inter-phrasal pauses should not occur in the rising portion of an f0 contour, nor without pre-pausal lengthening, the point where both upward zero-crossings coincide was tested against pause occurrences in the corpus. Pauses were more frequent than prosodically determined breakpoints, with 1085 pause-delimited units, compared with only 816 prosody-delimited units. 314 of these boundaries coincided exactly (from amongst 25898 possible positions), and a further 195 were within a single mora apart, giving a 62% agreement (509/816). Of the rest, 186 prosodically determined boundaries were found to be located at syntactically appropriate points in the speech, where the speaker could, but did not, insert a pause, thus bringing the overall agreement between the prosodic phrases and boundaries in the speech to 85% (694/816).

However, with the addition of linguistic processing, these results could be improved further, since a number of the remaining boundaries appear to have been detected too early because of an utterance-final f0 rise on verbs and particles, after the detected pitch valley, as in (桜島は爆発し↘↗ますと。。。), and (大都会の生活になれている訳↘↗ですが。。). It was also of interest to note that by far the majority of filled pauses (ええと、まあ、あのー) came immediately *after* a prosodically determined boundary, thus making them easier to detect under later linguistic processing.

5 Discussion

We tested a method of reducing the inter-speaker variance in intonation characteristics, and examined the use of this transform in the detection of valid pauses to assist in the phrasing of spontaneous speech utterances.

The phrase boundaries determined from prosodic analysis correspond well with linguistically useful breakpoints in the utterance, and they successfully ignore non-break silences, but they also fail to correspond with some valid-break pauses. They can be used positively to confirm the validity of a pause, where both coincide, but not to reliably distinguish restart or hesitation pauses from valid phrase breaks.

We can conclude that the addition of prosodic information to the spectral analysis is of some practical use, but that without the inclusion of linguistic analysis, the different types of pause cannot be easily distinguished. Further studies, incorporating linguistic filtering, are now being carried out.

References

[1] Entropic Waves+/ESPS signal processing package 1993

[2] Hirst, D. J., & Espesser, R., "Automatic modelling of fundamental frequency", in Press.

[3] Schriberg, L. & Lickley, R., "Intonation of filled pauses in spontaneous speech", Proc ICSLP-92.

[4] Seligman, M., Hosaka, J., & Singer, H., "Pauses and hesitations in spontaneous Japanese dialogues", submitted to COLING-94,