Durational Cues to Prominence and Grouping

W. N. Campbell
ATR$^\pi$ Interpreting Telecommunications Research Labs
Hikari-dai 2-2, Seika-cho, Kyoto 619-02, Japan.
e-mail: nick@itl.atr.co.jp

ABSTRACT

This paper discusses the analysis of prosody in a speech-to-speech-through-machine environment and describes the types of processing that can be used to facilitate transfer of extra-textual information in automatic speech translation. In particular, it details the use and contribution of duration-based sub-routines in such a system.

INTRODUCTION

In order to improve the efficiency of automatic translation, we need to make use of prosodic information. A speech interface to machine translation typically only processes a textual representation of the utterance, and in the three main stages of the task (speech recognition, language processing, and speech synthesis) prosody is only considered in the third. However, speech encodes more than just segmental information, and to make use of this extra channel to enrich the communication, we need to code the meaning contained in the prosody for processing alongside the orthographic representation. In order to study the flow of prosodic information, we need a database of utterances coded for speech-act and analysed for corresponding acoustic correlates. This paper describes the first steps in building such a database.

Prosodic Information Processing

Decoding not just the sentence, but the issuing of an utterance in a speech situation, we need detailed annotation of prosodic events and their correlates at several levels of representation. Leaving aside the interesting area of language-specific prosodic realisations and the translation of prosodic gestures between different cultural environments, we can simplify the task initially by assuming monolingual domains, analysing Japanese and English aspects of prosody separately, and coding them into a language-neutral message structure[1]. We are then left with the question of what constitutes the basic data for analysis. This can be answered by consideration of a) the needs, and b) the resources available.

Needs:

There is not space here for an exhaustive list of the uses of prosodic information in translation, but necessary in the long term, and perhaps most difficult to achieve, is the recovery of the speaker's intention; the pragmatic content or illocutionary force of each utterance. More immediately realisable are attitudinal characteristics (Hirschberg & Ward, 1992; Murray & Arnott, 1993), clues to the discourse structure and role (Hirschberg & Pierrehumbert, 1986; Swertz *et. al.*, 1992), and to the focus and prosodic boundary relations (Veilleux & Ostendorf, 1993; Wightman & Ostendorf, 1993). A further requirement for translating into and out of Japanese is an understanding of the degree of politeness (Ogino, 1986); this too is largely carried in the prosody. Non-linguistic uses include repair detection and correction (Schriberg & Lickley, 1992; Nakatani & Hirschberg, 1993), and modelling of the speaker-specific speech characteristics.

[1] Japanese, for example, makes use of a post-particle system; while English, on the other hand, employs greater use of stress in the signalling of information relations. At the deeper level of linguistic and paralinguistic feature representation, we assume that the two languages are equivalent.

Resources:
Since we require the prosodic interpretation to be done by machine, the main part of the first stage of the project involves the development of automatic procedures for the detection of prosodic features, and the mapping of these features to elements of the message.. Two small 'spontaneous-speech' corpora of 12 prompted dialogues are available to us; one produced by 107 speakers of English (Wood, 1992), and one by 19 speakers of Japanese. Both have been analysed for linguistic structure in the same format as used in the language-processing modules (Fais & Kikui, 1991; Nagata *et al.*, 1992). They include the following categories: Syntactic: declarative, interrogative, imperative, person, tense. polarity, etc.; semantic: probability, possibility, ability, potential, volition, permission, desire; and pragmatic: request, suggest, invite, reject, offer, response, acknowledge, inform, question, permit.
An example structure is shown below:

```
[[M [[PRAG [[HEARER !X10[[LABEL *HEARER*]]]
            [SPEAKER !X9[[LABEL *SPEAKER*]]]
            [TOPIC [[FOCUS !X7[[RESTR that [[RELN [(J-1]]]]]
                    [SCOPE [[RELN is [(J-IDENTICAL]
                            [OBJE !X7]
                            [ASPT !X8 STAT]
                            [IDEN [[RESTR [[RELN NAMED]
                                           [IDEN conference-office-1]]]]]]]
                    [TOPIC-MOD HA]]]]]
     [SEM [[RELN QUESTIONIF]
           [AGEN !X9]
           [RECP !X10]
           [OBJE !X11[[RELN BE-VI-5]
                      [OBJE !X6[[RESTR [[RELN THIS-PRON-1]]]]]
                      [ASPECT [[PERF -]
                               [PROG -]]]
                      [ASPT !X8]
                      [IDEN !X2[[RESTR [[RELN NAMED]
                                        [IDEN CONFERENCE_OFFICE-IDIOM-1]]]]]
                      [TENSE PRESENT]]]]]
     [SYN [[CAT S-TOP]
           [INV -]]]]] ... ( 57 lines of syntactic tagging omitted ) ... ]]]]
```

The representation for 'conference office' in "Is this the conference office?".

Analysis-by-synthesis, or differencing from a generated default, provides a measure of the prosodic correlates of these features. To detect marked areas of the utterance, we first normalise pitch for height, range, and attack; duration for global and local variance; and energy for phone-type and context[2]. The correlations with fundamental frequency variation have been well studied in this respect. I shall next explain the contribution of segmental duration information.

CONTRIBUTIONS FROM DURATIONAL STRUCTURING

Previous work (Campbell, 1992, 1993) has shown that speakers signal the intended interpretation of an utterance not only through pitch-related intonational variation, but also through the durational structuring of their speech. Segmental durations are available to us from the initial speech-recognition stage, via post-processing. By normalising them to reduce phone-specific effects, the underlying prosodic structure becomes very clear. This can be used to assist in the automatic detection of pitch prominences.

[2] The energy component of the speech waveform can vary for reasons unrelated to the message (e.g., changing microphone-to-mouth distance), so may be less useful in an automatic analysis.

Underlying prosodic structure

Segments in prominent or focussed words and phrases in the utterance are typically lengthened, as are those preceding a prosodic phrase boundary. In English, reduced segments are significantly shortened. Differential lengthening of segments in onset and coda parts of the syllable enables us to distinguish between the two different lengthening contexts and therefore to distinguish prominence effects from pre-boundary effects. This information, which is easily obtained from the normalised durations, given the syllable structure, enables us to detect the intonational groupings in the utterance. Figure 1 diagrams the main steps of this process: raw durations are first z-score normalised to remove phone-type-specific durational effects, then the syllable-internal differences (slope of lengthening) are compared to distinguish prosodic boundary locations from stress-related lengthening. This is explained in more detail in the oral version of this paper and in Campbell '93. Boundary locations detected by use of duration differentials have revealed interesting inter-speaker differences concerning rhythmic vs. syntactic phrase structuring that go unnoticed on simple listening.

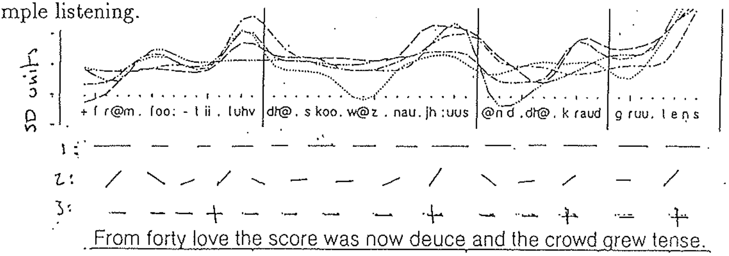

Figure 1: From normalised waveform through phone labels to boundary indication.

Pitch-prominence detection

We are testing several methods to extract meaningful phonological information from the raw fundamental frequency contour (Fujisaki & Sudo, 1971; Bagshaw, 1992; Taylor, 1993a). For Japanese, the Fujisaki-Sudo model specifies the location, strength and duration of each accent and phrase command in a small number of parameters that can be obtained by stochastic estimation procedures. These parameters include a speaker-specific/utterance-type-specific measure of the angle of onset/offset (attack) for each accent. This is normally fixed to a pre-defined value, but by estimating it along with the other parameters, we gain a useful indicator of speaking style and speaker characteristics. This, however, depends crucially on the accurate location of the accent and phrase commands; the numbers and locations of which must be specified in advance. Knowing the position and number of accents and phrases from the duration contour greatly increases the accuracy of this F0 decoding[3]. The residual from this estimation, along with the parameters specifying strength and type of accent form the raw data for our future analyses.

Unit selection for speech synthesis

In addition to facilitating pitch labelling, the prominences and boundaries detected in this way also improve the quality of our synthetic speech. The current method uses acoustic measures of goodness-of-fit when selecting units for concatenation from a labelled natural-speech source database (not enough is yet known about the fine details of segmental coarticulation to accurately predict the acoustic transitions by rule under parametric techniques). However, because no consideration is made of

[3] Unlike English, there is no durational correlate of accents in Japanese, but their detection with this method is relatively robust when given reliable phrase boundary location information.

the prosodic environment from which these units are selected, there is occasionally considerable distortion resulting from mismatch between the unit and the desired prosodic parameters. By labelling the source data according to prosodic features, and including these in the selection criteria, we can eliminate this cause of distortion. Looking to the future, with a corpus labelled for prosodic as well as segmental features, the task of speech synthesis may be even simpler. Since a small number of factors explain much of the variance in speech prosody and articulation, instead of predicting individual parameters directly, selecting appropriate units from a sufficiently large and well-labelled source database should yield natural durations and transitions that are optimal by default. Adequate labelling of the data should eliminate the need for prediction of the parameters.

CONCLUSION

This paper has introduced some recent work in the area of prosodic interpretation and highlighted the role of normalised segmental duration information in automatic labelling techniques for the acquisition and mapping of data. It showed how both pitch labelling and unit selection for synthesis can be helped by prominence and boundary detection from segmental lengthening contours. It is not yet clear what features of the utterance need to be mapped onto what higher-level aspects of the message, but we have laid the groundwork for an essential analysis.

ACKNOWLEDGEMENTS

I offer my sincere thanks to ATR for supporting this work, and to colleagues for helpful discussion.

REFERENCES

- P. C. Bagshaw (1992), An investigation of acoustic events related to sentential stress and pitch accents in English, pp 808-813, *Proc SST-92,* Brisbane, Australia.
- W. N. Campbell (1992), Multi-level Timing in Speech, *Unpublished PhD Thesis,* Sussex University, Department of Experimental Psychology..
- W. N. Campbell (1993), Automatic detection of prosodic boundaries in speech, *Speech Communication* (In Press).
- L. Fais & G. Kikui (1991), Determining surface forms for indirect speech acts in English, *ATR Technical Report,* TR-I-0235.
- H. Fujisaki & H. Sudo (1971) A model for the generation of fundamental frequency contours of Japanese word accent, Japan Acoustic Society, 27.9, 445-453.
- J. Hirschberg & J. Pierrehumbert (1986) The intonational structuring of discourse, *Proc ACL-86,* 136-144, Denver, USA.
- J. Hirschberg & G. Ward (1992) The influence of pitch-range, duration, amplitude, and spectral features on the interpretation of the rise-fall-rise intonation pattern in English, *Journal of Phonetics 20,* 241-252.
- I. R. Murray & J. L. Arnott (1993), Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *JASA 93,* 1097-1108.
- M, Nagata, M. Suzuki, & S Tachikawa 19920, First steps towards annotating illocutionary force types to a bilingual dialogue corpus, *ATR Technical Report,* TR-I-0298.
- T. Ogino & M. Hong (1992) 日本語音声の丁寧さに関する研究 (A study on politeness in Japanese speech) 215-258 in 日本語イントネーションの実態と分析 (The state-of-the-art and analysis of Japanese intonation), Mombusho, Tokyo, Japan.
- E. E. Schriberg & R. Lickley (1992), Intonation of filled pauses in spontaneous speech. *Proc ICSLP-92,* 991-994, Banff, Canada.
- M. Swertz, R. Geluykens, & J. Terken (1992) Prosodic correlates of discourse units in spontaneous speech, *Proc ICSLP-92,* 421-424, Banff, Canada.
- P. A. Taylor (1993), Automatic Recognition of Intonation from F0 Contours using the Rise - Fall - Connection Model, *Proc Eurospeech-93,* Berlin, Germany.
- N. M. Veilleux & M. Ostendorf (1993), Probabilistic parse scoring with prosodic information. *Proc ICASSP-93,* II-51, Minneapolis, USA..
- C. W. Wightman & M. Ostendorf (1992) Automatic recognition of prosodic features, *Proc ICASSP-92,* 321-324, San Francisco, USA..
- C. A. Wood (1992), The ATR-ITL/CMU Conference-registration task (Spontaneous speech), *ATR Technical Report,* TR-I-0328.