3—2—2

Crossing Feet: Syntactic versus Prosodic foot structure in English

© Nick Campbell
ATR Spoken Language Translation Research Laboratories
Hikari-dai 2-2, Kyoto 619-0288, Japan. (nick@slt.atr.co.jp)

Abstract

This paper[1] describes a system of prosodic labelling currently being tested for the annotation of a large corpus of spoken English. It shows that at many syntactically-determined phrase-boundaries, the supposedly "following" function word frequently groups prosodically with the previous phrase. This phenomenon provides support for the compound view of prosodic phrasing proposed by Ladd, as opposed to the Strict Layer Hypothesis proposed by Selkirk.

1 Introduction

Rules for segment and syllable duration remain one of the least satisfactory aspects of most speech synthesis-by-rule systems. Empirical studies [1,2] have established many of the factors that affect duration, including both segmental differences (manner and place of articulation, vowel height, etc.) and prosodic factors such as degree of stress and position in phrase. However, current models still fall well short of accurately reproducing the timing of natural speech.

There is reason to believe that part of the difficulty in modelling duration stems from theoretical shortcomings in the identification of the prosodic factors involved. A number of current issues in phonological theory concern the nature of prosodic structure and the relationship among different prosodic features. Obviously, if the definition of e.g. 'phrase' is open to debate, then this will affect the way 'phrase boundaries' are marked in any given corpus or text sample, which in turn will affect any empirical findings about the influence of phrase boundaries on syllable and segment duration.

In previous work, Ladd & Campbell [3] reported a study to evaluate Ladd's theoretical claim [4] that there is no principled limit to the depth of prosodic structure. We did this by comparing the durational effects of phrase boundaries *within* and *between* what Ladd calls 'compound' phrases, i.e. phrases that are themselves composed of two or more phrases, and showed two kinds of results: first, that there are significant differences on this comparison, and second, that inclusion of the distinction makes possible a significant improvement in the amount of variance accounted for. This work showed the relevance of issues in phonological theory for practical applications in phonetics and speech technology.

[1] 「英語における韻律構造と文法構造の違いについて。」
ニック キャンベル（ATR 音声言語通信研究所）

2 Corpora for Speech Synthesis

At ATR-SLT we are now working with speech corpora for concatenative synthesis [5] that are large enough to be "self-contained", i.e., no longer need generalised rules for the prediction of durations, since the databases themselves now contain enough samples to allow direct selection of naturally appropriate waveform segments through judicious control of feature contexts. i.e., instead of training a duration prediction model on the basis of repeated observations in the hope of generalising the findings to unobserved contexts, we can now assume that all relevant contexts (in terms of prosodic and phonemic environment) will be represented in the corpus so we can select units directly, without recourse to an intermediate numerical prediction stage. For this work, we are currently using a 16-hour corpus of read British English speech from one male speaker.

However, although the intermediate task of predicting an appropriate segmental or syllabic duration has now become redundant, the task of labelling the speech with an appropriate set of features that will allow us to retrieve appropriate segments according to their defining contexts has become even more important than before.

Following work on direct feature-based selection of waveform segments using a ToBI-coding [6], which replaces numerical pitch prediction by a higher-level non-numeric representation of the contour, we are now experimenting with a non-pitch-based higher-level representation of the speech in terms of phrasing and prominence relations. By replacing the traditional ToBI "tonal tier" with a "prominence-marking tier" (see Figure 1), we aim to annotate the relevant prosodic phrasing features that relate to the chunking and hierarchies in the speech.

3 P-BI prosodic transcription

The annotation differs from the traditional 'British school' analysis, in which utterances are composed of major tone groups, and major tone groups are composed of minor tone groups. This categorisation of prosodic phrases conforms to the 'Strict Layer Hypothesis' (Selkirk [7]), according to which the prosodic structure of any utterance consists of a hierarchical arrangement of a fixed number of prosodic domain types.

The Strict Layer Hypothesis is what is challenged in Ladd's work: specifically, Ladd has argued for the existence of 'superdomains' or 'compound prosodic

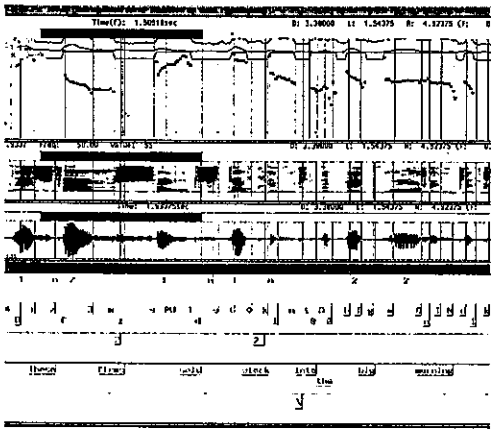

Figure 1: Sample waveform and transcription showing the inclusion of syntactically distinct segments in the previous prosodic phrase (as marked by the "V" in the bottom tier). From top to bottom, panels show the fundamental-frequency contour, spectrogram, waveform, prominence tier, segment labels, break indices, words, and the "miscellaneous" tier.

domains', in which two or more adjacent domains of a given type are gathered together in a larger prosodic constituent *which is itself of that type.* Evidence for this proposal includes studies of acoustic cues to 'boundary strength' (e.g Cooper and Paccia-Cooper [8], Campbell [2], Ladd [4]), in which considerable depth of structure is reflected in segmental duration and F0 properties in the vicinity of boundaries.

The Prominence and Break-Index marking (P-BI) system that we are currently evaluating, discriminates between three levels of prominence and four levels of prosodic boundary (none, minor, major, utterance), allowing for the possibility of compound prosodic domains by marking subordinate tone-group boundaries (with a minus-sign). As with traditional ToBI, a break-index level 0 can be used to mark phonetic transformations (e.g., in the case of elision), and break-index level 1 can be used to indicate a morphological boundary undistinguished by a prosodic break.

Thus, three levels of prosodic break are marked, using index numbers 2, 3, and 4 as in ToBI, but the use of "-" and "p" differ in our transcription system: The diacritic "p" is used to indicate a "continuation" boundary tone (which may be accompanied by a pause) that indicates list-type intonation or parenthesis. Question intonation is marked in the miscellaneous tier along with other speech-act information. The minus "-" is reserved to indicate compound domains, where a distinction is made with respect to the rhythm of the utterance that is not necessarily reflected in the intonation.

Figure 1 shows an example of this labelling, and the text below shows a reconstructed transcription of the sample database utterance "These firms sold stock into the big morning decline but, seeing the velocity of the market's drop, held back on their offsetting purchases of futures until the S&P futures hit the trading limit." In the transcription, plus-signs mark secondary prominence, and asterisks mark full prominence on the following word:

```
<breath> * These + firms 2 * sold * stock 2- <V> into
the + big + morning + decline 3 but + seeing the  *
velocity 2 of the + market's + drop 3 <breath> held
* back 2 on their * offsetting + purchases 2- of +
futures 2 <breath> until the + S&P * futures 2- hit
the  + trading limit 4 <breath>
```

4 Crossing dependencies

In many cases we found that syntactic boundaries (e.g., between a verb phrase and a prepositional phrase, as in the example above) were not reflected in the prosodic phrasing, and that the "initial" preposition was clearly included prosodically as the last word of the previous phrase:
e.g., *[{these firms sold stock} {into] [the big morning rush}]* (where {} shows syntax, and *[]* prosody).

For this and other similar examples, we propose using the "V" label as a marker of such "boundary crossing", and note that of the 45 cases found in a small sample of the corpus, more than 80% occur where there is a "2-". It can be confirmed from figure 1, that the falling F0 contour clearly includes the preposition.

We take this as evidence supporting the theory that there are four hierarchically arranged types of tone group boundary rather than, as in the traditional transcriptions, two. These data conform with Ladd's proposal that at least the following depth of prosodic structure is possible:

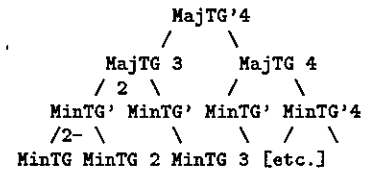

Figure 2. Layers of prosodic phrasing (after [4,3]), showing the proposed P-BI numbering codes.

References

[1] Klatt, D. H., (1976) *Linguistic uses of segment duration in English* JASA #59 pp 1208 - 1221,
[2] Campbell, W. N., (2000) *Timing in speech: a multi-level process*, pp 281-334 in **Prosody: Theory & Experiment**, ed Merle Horne, Kluwer Academic Publishers.
[3] Ladd D. R., & Campbell, W. N., (1990) *Theories of prosodic structure: Evidence from syllable duration.*, Proc ICPhS-90.
[4] Ladd, D. R., (1986) *Intonational phrasing: The case for recursive prosodic structure* Phonology Yearbook 3, 311 - 340.
[5] Campbell. W. N.,(1992) "Synthesis units for natural English speech". Technical Report SP 91-129, IEICE.
[6] Fujii, K., & Campbell. W. N., (2000) *ToBI-based unit selection.* Proc ASJ, Spring 00.
[7] Selkirk, E. O., (1984) **Phonology and Syntax: The relation between sound and structure.** Cambridge, Mass.: MIT Press.
[8] Cooper, W. & Paccia-Cooper, J., (1980) **Syntax and speech** Harvard Univ. Press, Cambridge MA.