LOUDNESS, SPECTRAL TILT, AND PERCEIVED PROMINENCE IN DIALOGUES

W. N. Campbell

ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan.

ABSTRACT

This study explores the correlation between spectral tilt and perceived prominence in the continuous speech of simulated conference-registration dialogues. It builds on previous work showing that syllable prominence and focus marking can be detected automatically, using differences in normalised segmental duration and energy, by introducing spectral information that compensates when the prosodic clues are weak or absent.

INTRODUCTION

There has been continuing debate about the relation between loudness and stress (see Beckman [1] for a summary). Early theories presented stress as having fixed phonetic levels (*c.f.,* Bloomfield, Trager & Smith, Chomsky & Halle), as being related to *force of utterance*, as tonetic (Kingdon), or as dependent on *pitch accents* (Bolinger). Beckman highlights the role of pragmatics in determining the accentual organisation of an utterance. In examining the phonetic correlates of stress and non-stress accent, she shows that syntagmatic accentual contrasts divide an utterance into a succession of shorter phrases in larger groupings, defining *stress* as a phonologically delimitable type of accent in which the pitch shape of the accentual pattern cannot be determined from the lexicon.

Early experimental evidence (from Fry and others) shows energy to be the weakest clue to stress, and fundamental frequency the strongest. Beckman, on the other hand, found metrical stress (at the level of the *prosodic word*), to be best explained by relative loudness (*i.e.,* temporal summation of waveform energy through the syllable nucleus, expressed relative to its duration), and she emphasises the trading relation between energy and duration in the perception of prominence.

Duration and energy

Previous work has confirmed that segmental duration and energy are both reliable cues for the automatic detection of prominence in read speech [2], and the present paper extends that work to show that spectral information is also present in the marking of prominence, and that it exhibits a trading relation with duration in interactive speech.

For the analysis presented in [2], a set of sentences extracted from a corpus of conference-registration dialogues was marked (by capitalising certain words) to show shift of focus, resulting in different stress patterns on the same sequence of words (as in *"Please take the SUBWAY to Kyoto station."*). Each sentence (30 in all) was given three or four different patterns, and a set of 100 of these was produced in three different utterance contexts. To study the way focus is marked in speech, we first asked the speaker to read the sentences in sequential order, and to "emphasise" the capitalised words. Here, with each set of interpretations grouped, the shift of focus was clearly contrastive. For the second reading, we asked her to read the same sentences in randomised order.

Finally, we recorded an interactive dialogue, where the same focus marking was produced by eliciting emphatic corrections of feigned misinterpretations.

The corpus of 300 focus-shifting utterances was then stress labelled to indicate perceived prominence. In order to remain somewhat theory neutral, I had the corpus labelled for accent type and for prominence location in three ways, by different labellers: (1) by simply marking the syllables perceived as prominent (an either-or decision), (2) using an O'Connor & Arnold variant of tonetic stress marks, and (3) more recently, using the ToBI system of tones and break indices. I then took the common subset of these three labellings as defining 'stressed syllables' for the purpose of this study. (However there was a high correlation between all three, and the different labellers seem to be identifying the same feature.)

DETECTION OF PROMINENCE

Because of the use of fundamental frequency in signalling more complex relations than simple prominence, this was not included as a factor for analysis (though it certainly plays a significant part in marking prominence). Instead correlations were examined between stressed syllables and measures of energy and duration normalised by segment type. Viewing the two acoustic measures independently, rather than combined as an energy integral over time, allowed better understanding of their individual contributions, and of the trade-off between them. Absolute values were not examined, but rather, for each phone class (as defined by label type in the segmentation), durations were normalised by expressing deviation from the class mean in terms of standard deviations of the distribution of that class. Similar segment-type normalisation was applied to the waveform energy, measured as average rms amplitude across the duration of each segment. Because these unit-less normalised scores have a zero mean (and a typical range of ± 3) a combined measure of their joint effect was derived by simply adding them. Taken separately, durational lengthening information detected 54% of the prominent syllables, and energy information detected 55%. Combined by summation, this detection improves to the average of 76% across all three speaking styles [2].

Table 1: Stress and focus detection

	A	B	C
stress detection:	92%	78%	72%
focus detection:	79%	78%	74%

Key: A: read grouped, B: read in randomised order, C: interactive

Between the three utterance styles, there was no significant difference in the detection of marked focus (*i.e.,* in identifying the syllable carrying the intended prominence) from amongst the set of syllables detected as stressed, but the initial detection of stressed syllables did vary as a function of speaking style. Table 1 shows that stressed syllables in read speech of grouped sentences were more easily detected than the equivalent syllables in randomly presented sentences or in interactive speech. Although also perceived as prominent, the latter were less easily discriminated by measures of duration and energy. Error analysis confirmed that in the more conversational interactive speaking style the prominences were still easy to discriminate by ear, but the acoustically-derived measures of stress were weaker. This paper attempts to explain why this may be so.

SPECTRAL FEATURES

Of the prominences not detected, 26% were clearly prominent to the ear but showed no significant excursion from the mean in duration, energy, or fun-

damental frequency. This implies that there are also phonation-style differences which can serve as clues to prominence and which may also be of use to automatic detection. This would be particularly useful since although durational information is robust, raw waveform envelope magnitude is *not* a robust measure, as it can vary considerably with distance from the microphone, or more globally reflecting changes in environmental noise.

Lindblom, in sketching the H&H theory [3], suggests a notion of sufficient discriminability to explain the continuum of hyper- and hypospeech observed in interactive dialogues, by which speakers tune their production to communicative and situational demands. This might account for the differences in results relating to speaking style, since in the interactive dialogues the speaker knows the extent of common knowledge with the hearer, and in the grouped presentation of contrasting pairs of utterances, she is more aware of the need to stress the contrast. Lindblom refers to Sundberg's work, on the long term average spectra of singers, in explaining possible mechanisms for the range of clarity of phonation. More recently, Sluijter & van Heuven [4], also citing such work on overall "vocal effort" as Gauffin & Sundberg [5], showed that, in Dutch, stressed sounds are produced with greater local vocal effort and hence with differentially increased energy at frequencies well above the fundamental.

We can measure such spectral tilt in several ways. At the lower end of the spectrum, as the difference in energy between the first and second harmonics, or at the upper end of the spectrum as a general increase of overall energy. A pilot study examining energy across 26 ERB-scaled spectral bands [6] confirmed that at least for read lab speech of English, spectral tilt significantly correlates with linguistic prominence under both high and low tones, for three different vowel types, and confirmed Sluijter's findings of increased energy in the higher spectral regions. This paper shows that for dialogue speech too, spectral information can be very helpful in discriminating prominences.

Table 2: Analysis of variance detection

df=(1,10048)	mean sqr	F
spectral tilt	113.35	603.5,
harmonic ratio	11.63	60.8469
energy (fund)	1.41	7.3856

Extraction of spectral data

Because segment labelling was done for all the dialogues, acoustic measures derived from the waveform can be related directly to individual syllables. To estimate spectral tilt, the fundamental frequency was extracted and then used to index into an fft of the speech waveform for each utterance so that a) the energy at the fundamental, and b) the harmonic ratio could be calculated. As a further measure, the average energy in the top third of the ERB-scaled spectrum (between 2kHz and 8kHz) was measured relative to the overall energy of each spectral slice as a measure of energy-normalised tilt.

These three indicators, normalised by phone type as for duration and energy above, were computed for the sonorant peak of each syllable and compared with the labelled prominences.

RESULTS

Of the 10,049 syllables in the 300 sentences, 2,951 were marked as prominent. There were 16 classes of vowel, none with less than 110 tokens. All had a representative number of prominent variants. Analysis of variance from a linear discriminant analysis predicting prominence as a binary feature on the basis of the three spectral factors showed all to make a contribution (significant at $p < 0.001$, see Table 2.). There were great differences though in

the amount of the contribution of each, and energy in the upper areas of the spectrum was by far the clearest predictor of stress.

Interestingly, further factorisation of spectral tilt (as measured by the ratio of high-frequency energy to overall energy in the spectrum) according to speaking style, revealed that the greatest distinction between prominent and non-prominent syllables could be made for the spontaneous speech. See Table 3.

DISCUSSION

The above results confirm the correlation between spectral and prosodic information, and suggest that speakers also change their phonation according to the discourse context and type of information they impart. In style A (the grouped-presentation read-speech), the distinction between prominent and non-prominent syllables was clearly marked to accord with the capitalisation of the focussed word in the text. In the interactive case, when an interlocutor elicited the focus shift by misunderstanding selectively, the speaker was more personally involved in clarifying the meaning. This, too, resulted in a clearer articulation. However, for the intermediate case, where the focal shift was less markedly obvious, the distinction was less clear.

In all speaking styles, relative energy in the higher spectral regions proved the best correlate of prominence, and loudness (as measured by energy at the fundamental) the weakest. It is interesting that although the prosodically-based measures of duration and waveform envelope magnitude (amplitude) were weakend by the greater variation found in the more spontaneous rendition of the dialogues, the spectral measure was apparently strengthened. We can suppose that this trade-off is not coincidental, and in future work, include the spectral measures as well as the prosodic ones in the detection of prominence.

Table 3: ± prominent spectral tilt

	student's t	df
read grouped	35.63	7676
read randomised	19.01	6110
interactive	42.76	6974

Showing the separation in mean spectral tilt between prominent and non-prominent syllable peaks.

ACKNOWLEDGEMENTS

I am particularly grateful to Mary Beckman and Osamu Fujimura for their helpful discussion and advice, and apologies to the many authors whose relevant work would have been cited given more space.

REFERENCES

[1] Beckman, M. (1986) **Stress & Non-Stress Accent**, Floris Publications.

[2] Campbell, W.N. (1992) "Prosodic encoding of English speech", *Proc ICSLP 92,* pp 663-666, Banff, Canada.

[3] Lindblom, B. E. F. (1990) "Explaining phonetic variation: A sketch of the H&H theory". **Speech Production and Speech Modelling** edited by H. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp 403-409.

[4] Sluijter, A., & van Heuven, V., (1993) "Perceptual cues of linguistic stress: intensity revisited", *Proc. ESCA workshop on Prosody,* pp 246-249. Lund University,

[5] Gauffin, J. & Sundberg, J. (1989) "Spectral correlates of glottal voice source waveform characteristics", **JSHR 32,** pp 556-565.

[6] Campbell, W. N., & Beckman, M. (1995) "Stress, Loudness, and Spectral Tilt", *Proc Acoustical Soc. Japan,* Spring meeting, 3-4-3.