Prosodic Phrasing from Normalised Acoustic Measures

W. N. Campbell
ATR Interpreting Telephony Research Labs

Abstract

Normalised measures of duration and power, averaged over each syllable, can provide clues to the prosodic patterns of emphasis and boundaries in spoken English. Results are presented from a test using normalised power values to help discriminate between syllables lengthened by stress or phrase-finally.

1 Introduction

There is considerable variation in segmental duration and power (rms amplitude) in English speech, much of which is systematic and related to linguistic events in the spoken signal, such as emphasis, focus, and phrase boundary marking. Previous work [1] has shown that after normalising for phone-specific differences, segmental durations can provide strong clues to assist in the detection of stressed and phrase-final syllables, with length differences between onset and coda segments within the syllable being used to differentiate between these two types of lengthening. This paper explores the extent to which correlations between duration and power can be used for differentiating syllables in these contexts in a database of readings of two-hundred phonetically balanced sentences by one male speaker of British English.

2 Materials

A digitised recording of the readings was cepstrally analysed to produce estimates of power (amplitude) for each 5 msec frame. The speech wave files had already been hand-segmented and labelled. Since the measures of duration are at the level of the phone, the power values were averaged locally between each phone boundary. To remove any phone-specific influences, all values were then normalised to unit variance about a zero mean for each phone class. To further remove any bias resulting from distributional differences specific to a particular phone type, these normalised values were then averaged between syllable boundaries. Thus a measure of the duration and power of each syllable was obtained.

To increase robustness and reduce any influence of outliers on the calculations, these normalised values were passed through a nonlinear sigmoidal transform to produce a distribution biased towards the center values, and within a range of 0 and 1. First differences (Δdur and Δpow) of these measures were also calculated for use in the analysis as an indication of the local contour.

3 Method

Each syllable was assigned a value for stress and boundary strength. Stress and boundary locations were determined by listening to recordings of the readings. Primary and secondary levels of stress were determined, giving three levels in all. Five levels of boundary strength were decided, based on the degree of prosodic continuity between each pair of syllables. A score of zero denoted a syllable boundary within an orthographic word, one: a cliticised word, two: a prosodic-word break, three: a minor-tone-group break, and four: a major tone-group break.

Analysis of variance for the factors stress (three levels) and boundary (five levels) both yield significant results for duration and power (stress: $F_{2,\ 3302}$ = 433.2 and 31.21; boundary: $F_{4,\ 3300}$ = 243.0 and 37.86 respectively). Student's t tests showed stressed syllables (n = 885) to be significantly different from unstressed syllabes for both parameters (t_{3303} = 27.3 (dur) and 7.87 (pow) respectively), but although all three stress levels were well discriminated by duration, power did not differ significantly between primary and secondary stress levels (t_{883} = 9.6 for duration, and 0.8, n.s. for power). The 'primary' and 'secondary' stressed levels were therefore combined to produce a binary stress category. A total of 885 out of the 3305 syllables were marked as stressed.

Final syllables (boundary value = 4, n = 254) were found to be significantly different from non-final syllables for both parameters (t_{3303} = 22.2 and 9.5 respectively). 254 syllables were marked as final, of which 112 were also stressed.

It is hypothesised that final-lengthening can be differentiated from stress-induced lengthening by differences in power; a syllable lengthened by gradual decay into a following pause may be expected to lack the power of one lengthened by emphatic articulation. Power can thus be expected to be low in final syllables, and high in stressed ones. The corresponding null hypothesis is that there is no difference between stress-lengthened and phrase-finally length-

正規化した音響的尺度を用いた韻律フレージング
ウイルヘルム N キャンベル

ened syllables with respect to power. It is hoped that any difference found will prove useful for disambiguation.

4 Results

Means and variances for the two data sets were significantly different ($t_{6608} = 4.57$), but both close to 0.5 (dur: 0.494, sd = 0.156, n = 3305 pow: 0.510, sd = 0.137, n = 3305). Table 1 shows the probability of a syllable having a value lower than 0.5 for each of the four conditions.

Table 1: Probability(*value* < 0.5)

	unstressed		stressed	
	f-	f+	f-	f+
dur :	0.71	0.21	0.31	0.01
pow :	0.45	0.79	0.28	0.68

(f-: non-final, f+: final)

Clear separation can be seen for both duration and power values when the data is factored into these four subsets. Table 2 shows that for both non-final and final syllables there was an average increase of 0.14 (duration) and 0.05 (power) between unstressed and stressed tokens. There was also a 0.19 difference in average duration for both stress levels between non-final and final states, but this was accompanied by an average *decrease* of 0.09 in power. Power does therefore increase on stressed syllables, but decreases, and more so, on final syllables, even when stressed.

Table 2: Average values for each condition

		unstressed			stressed			
		mean	sd	n	mean	sd	n	
f-	dur:	0.44	0.130	2278	0.58	0.139	773	+.14
	pow:	0.50	0.139		0.55	0.118		+.05
f+	dur:	0.63	0.150	142	0.77	0.109	112	+.14
	pow:	0.41	0.139		0.46	0.108		+.05
		(+.19, -.09)			(+.19, -.09)			

This difference suggests that power may be used as a clue in disambiguating durational lengthening. However, while this difference alone may be sufficient, the use of an absolute value may be misleading. A high value may be lower than a previous one, or a low value high in relative, or local terms, reflecting more global shifts in e.g., speaking style. As a check on this, the first differences of the normalised measures of duration and power (Δdur and Δpow) were also examined. It can be seen from Table 3 that the probabilities are similar to those for the absolute values.

Table 3: Probability(negative Δ)

	unstressed		stressed	
	f-	f+	f-	f+
dur:	0.61	0.44	0.23	0.05
pow:	0.53	0.76	0.37	0.61

In summary, while there is only a very small probability that a stressed syllable will be shorter (relatively) in duration than the previous syllable, there is a greater than chance probability that its power will be lower if it is also final. We can thus proceed to quantify the effect of incorporating power information as a filter.

5 A filter test

Using these measures as a filter, we can select candidate syllables as markers of prosodic events. Duration indicates stressed and final candidates; power may prove useful as a filter for distinguishing the final ones.

Table 4 shows that duration alone selects 84.8% of the stressed syllables, 88.9% of the final ones, with a false selection rate of 35.3%. Power (weak) selects 81.9% of the final syllables, and only 44.6% of the stressed, with false inclusions of 47.4%. We can see that the lw (long, weak power) filter may be optimal for selection of final syllables; it rejected 78.2% of the stressed syllables, and 86.8% of false candidates, but only included 65.3% of final syllables, including 75 stressed-and-final ones.

Table 4: Results of filtering by dur & pow:

lwrf	s+	s-	%	f+	f-	%	false	%
l...	656	117	85	226	28	89	1079	35
.w..	345	428	45	208	46	82	1447	47
.wr.	638	135	83	229	25	90	1824	60
.w.f	151	622	20	152	102	60	752	25
lw..	169	604	22	166	88	65	404	13
lwr.	539	234	70	204	50	80	848	28
lw.f	286	487	37	186	68	73	632	21

l: long, w: weak: r: posΔ, f: negΔ, s: stressed, f: final, +: selected, -:not included, %: percentage selected; false: number of other syllables included.

6 Discussion

Although power does differ significantly on final syllables, no combination of filters was entirely successful in distinguishing final syllables from stressed ones. The labelling of stressed and final syllables is not absolute, and these figures may be only indicative of the real performance of such filters. However, these results show that the inclusion of an additional measure (power) may be useful for the prosodic segmentation of speech signals. Further work will show how this can be used in conjunction with the clues from segmental duration differences within the syllable.

References

[1] Campbell, W. N., *Prosodic Segmentation of Recorded Speech,* in PERILUS, Working papers of Stockholm University Linguistics Department, 1992.