Stress, Loudness, and Spectral Tilt *

Nick Campbell †& Mary Beckman ‡
†ATR Interpreting Telecommunications Research Laboratories
‡ATR Interpreting Telecommunications & Ohio-State University.

1 Introduction

Traditional linguistic descriptions define prosodic prominence in Germanic languages, such as English, in terms of differences in loudness. These descriptions concur with strong native speaker intuitions about the nature of stress or accent: other things being equal, a stressed or accented syllable should be louder than a corresponding unstressed or unaccented syllable. Work by Fry [3] and others shows that stressed or accented syllables do tend to have larger overall RMS amplitude (see review in Beckman, 1986 [1]), and in computer speech synthesis it is common practice to modify the RMS amplitude contour to mimic these differences. However, simple multiplicative modification of the waveform envelope does not seem to accurately reproduce the salient differences in "loudness" as measured, e.g., by ISO 532 [4], and corresponding work on the perception of stress contrasts does not show these differences to be the most salient perceptual cue to linguistic prominence.

Sluijter & van Heuven [5, 6] have recently suggested that in Dutch, the expected loudness differences are accomplished by concommitant changes in spectral tilt. Citing such work on overall "vocal effort" as Gauffin & Sundberg [2], they suggest that stressed sounds are produced with greater local "vocal effort" and hence with differentially increased energy at frequencies well above the fundamental. This paper examines this hypothesis in a corpus of English sentences. The results suggest that spectral tilt is affected by linguistic prominence, but that the relationship is not so simple, and depends also on vowel type, but is independent of tone type.

Although the averaged frequency bands show an increase in energy in higher regions of the spectrum, finer discrimination of the lower frequency bands reveals a peak of increased energy around the fundamental, balanced by a trough of lower energy in the more prominent cases, further emphasising the difference in spectral tilt.

The implications for speech synthesis are that we must either improve our signal processing techniques to model such changes in spectral energy, or must include the prominence level as a factor in selection of appropriate source units for concatenation.

2 Methods and Data

Three speakers produced five tokens of each of 36 sentences, which put target vowels [ae], [i], and [u] in three stress conditions (accented, unaccented primary-stress, and unstressed) in both high- and low-pitched regions of the intonation contour. For example, the vowel [i] was produced in the 12 sentences shown in Figure 1, with target sentence accents indicated by capital letters on the accented word, and the target intonations indicated using the ToBI transcription system. Sentences (1)-(6) put the target vowel in a [v–s] context and sentences (7)-(12) put it in a [b–d] context. In "eastward" and "beadwork" the target vowel is the primary stress of the word, whereas in "East Warsaw Street" and "bedazzled" the following syllable has the primary stress. Sentences (1) and (7) put a high nuclear accent on the stressed target vowel, whereas (4) and (10) put a low nuclear accent on it. Sentences (2) and (8) put the target vowel in the unaccented region after the nuclear stress, in the high-pitched intonational "tail" of the "yes-no question" contour, whereas sentences (5) and (11) put it in a low-pitched intonational "tail" after a previous nuclear high accent in a "declarative" or "contrastive statement" contour. In sentence (3), we expected the unstressed target vowel to be in a high-pitched region just preceding the nuclear H* accent on "Warsaw", and so on.

Not all of the tokens were produced as intended, however. For example, all three speakers often produced an emphatic L+H* on the following nuclear stressed syllable in sentence (8), so that the target vowel was usually low-pitched, rather than high-pitched. The second author checked the utterances to note the actual intonation pattern, and modified the codings accordingly.

(1) You should drive EASTWARD from here.
H* L- L%
(2) [Walking to work is hard at this time of year, because the sun shines right in my eyes.] Isn't it harder to DRIVE eastward then?
L* H- H%
(3) It's the other side of East WARSAW Street?
H* L- L%
(4) Should I drive EASTWARD to get there?
L* H- H%
(5) [I know it's on this side of Moscow Blvd, but] is it this side of East WARSAW Street?
L* H- H%
(6) No, it's the OTHER side of East Warsaw Street.
H* L- L%
(7) I hear their BEADWORK is good at least.
H* L- L%
(8) I was simply BEDAZZLED by it.
H* L- L%
(9) [Well, their beadwork may be bad, but...] have you seen JOEL'S beadwork?
L* H- H%
(10) Have you seen their BEADWORK yet?
L* H- H%
(11) No, they're TERRIBLE at beadwork.
H* L- L%
(12) Did they all seem BEDAZZLED by it?
L* H- H%

Fig. 1 Examples sentences for vowel /i/.

*スペクトル傾斜およびラウドネスに対するストレスの影響
ニック・キャンベル（ＡＴＲ音声翻訳通信研究所）、メアリー・ベックマン（ＡＴＲ音声翻訳通信研究所・オハイオ州立大学）

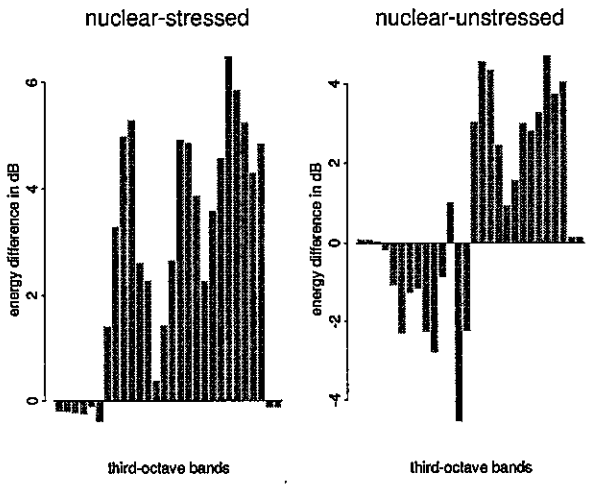

Fig. 2 Third-octave-band analysis

3 Results

Target vowel portions excised from each of these 540 utterances were analysed for spectral energy differencecs in the four bands used by Sluijter (0-500Hz, 500Hz-1kHz, 1kHz-2kHz, 2kHz-4kHz (also 4kHz-8kHz)) and across 26 third-octave bands calculated as for ISO532B [4].

After differencing averaged energy levels for stressed and unstressed segments in each class, it was confirmed that the influence of prominence on spectral tilt is better discriminated in higher regions of the spectrum (see Table 1.). However, analysis of variance showed significant interactions with both tone and vowel-type. The effect of speaker difference was only significant in the 2kHz-4kHz band.

Since higher tone may correlate with spectral shift, thus accounting for increased energy at the higher end of the spectrum, and since there is often a correlation between prominence and raised F0, we factored separately for tone, and found that whereas the energy differences are significant, they do not vary much in relation to frequency band (see Table 2). For one speaker (MB, female) the direction of the energy difference, though still remaining significant, changed for lo-tone prominences with low F1.

Table 1 Showing that prominence is of more influence in higher regions of the spectrum.

	band(Hz)	diff(dB SPL)	t(df=530)	
1	0-500	0.25	0.629	ns
2	500-1k	2.38	2.762	**
3	1k-2k	3.14	2.991	**
4	2k-4k	4.03	4.433	**
5	4k-8k	4.52	6.533	**

Energy difference (prominent - non-prominent)

However, further analysis of the spectrum divided into 26 third-octave bands (see Fig 2.) showed that the relative lack of effect for prominence in the lower bands can be accounted for by an area of increased energy around the fundamental, masked by surrounding areas of lower energy for the prominent segments in these bands. This conforms with the hypothesis of increased spectral tilt as a function of prominence, suggesting that the underlying spectrum may display even more tilt, though masked somewhat by the formant structure of the different vowels.

Table 2 Showing energy difference (prominent - non-prominent) for other factors. Note that tone shows equivalent differences regardless of band.

	hi-tone			lo-tone		
band	diff	t		diff	t	
1	2.13	3.622	**	1.41	4.023	**
2	2.20	4.371	**	1.00	2.951	**
3	3.13	6.496	**	0.53	1.565	ns
4	2.31	4.419	**	1.70	4.901	**
5	3.20	5.581	**	2.31	6.085	**

	/i/			/ae/			/u/		
band	diff	t		diff	t		diff	t	
1	7.34	1.841	ns	3.13	1.006	ns	0.33	0.494	ns
2	6.36	1.697	ns	6.27	1.912	ns	4.02	3.679	**
3	5.16	1.464	ns	10.46	3.186	**	2.73	2.648	**
4	10.18	2.585	**	7.25	2.280	*	2.42	2.235	*
5	12.17	3.317	**	5.20	1.770	ns	4.21	4.405	**

	spkr-1			spkr-2			spkr-3		
band	diff	t		diff	t		diff	t	
1	1.03	3.595	**	3.53	1.158	ns	6.27	1.516	ns
2	1.68	1.090	ns	4.41	1.337	ns	10.36	2.566	ns
3	2.15	1.287	ns	5.54	1.549	ns	10.59	2.749	ns
4	2.01	1.559	ns	4.61	1.384	ns	14.1	3.752	ns
5	5.03	5.674	**	6.20	2.040	ns	10.34	2.942	ns

4 Conclusion

It has been confirmed that prominent vowels exhibit different spectral energy characteristics from their non-prominent equivalents. It has also been shown that these differences are not due to differences in fundamental frequency alone (*c.f.* hi-tone, lo-tone contexts), and we can infer that they may arise from difference in phonation style for prominent syllables. From this, we must question whether simple multiplicative modification of the waveform envelope is adequate for the modeling of prominence in speech synthesis.

Bibliography

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

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

[3] Fry, D. B. The Physics of Speech, C.U.P., 1975.

[4] International Standard ISO 532-1975(E): "Acoustics - Method for calculating loudness level". 1975.

[5] Sluijter, A. M. C., & van Heuven, V. J., "Perceptual cues of linguistic stress: intensity revisited", pp 246-249 in Proc. ESCA Prosody W/S, Lund 1993.

[6] Sluijter, A. M. C., & van Heuven, V. J., "Spectral tilt as a clue for linguistic stress", presented at 127th ASA, Cambridge, MA. 1994.