Spectral emphasis as a perceptual cue to prominence

Heldner, M.

journal: TMH-QPSR
volume: 42
number: 1
year: 2001
pages: 051-057

http://www.speech.kth.se/qpsr
Spectral emphasis as a perceptual cue to prominence

Mattias Heldner
Department of Philosophy and Linguistics, Umeå University
and Centre for Speech Technology (CTT), KTH

Abstract
This paper is a first attempt at investigating whether spectral emphasis has any relevance for the perception of accented words. In particular, this study focuses on (i) whether an increase in spectral emphasis will cause accented words to be perceived as more prominent, and (ii) whether modeling of spectral emphasis in connection with accents is likely to improve the quality and naturalness of speech synthesis. A method for experimentally manipulating spectral emphasis in natural and synthesized speech without adverse effects on speech quality is proposed. This method is subsequently used to create stimuli for two perceptual experiments. The first experiment involved increasing spectral emphasis in accented words in natural speech and asking listeners to compare the prominence of the manipulated and original words. Similarly, the second experiment included comparisons of the naturalness of manipulated and original words, this time generated by an mbrola synthesis. The results of these two experiments were on the whole quite negative. Increased spectral emphasis, at least as it was implemented here, did not cause words to be perceived as more prominent. Neither did it improve the naturalness of speech synthesis. A few possible explanations of these results are discussed. However, it remains an open question whether other and perhaps more realistic implementations of increased spectral emphasis may produce more salient effects.

Introduction

It has long been recognized that fundamental frequency (f0) and duration are the most important cues for the perception of accents and of prominence distinctions in general. Recent work seems to indicate, however, that these two features may, after all, be insufficient to model the characteristics of different levels of prominence. Among additional features to take into account, the slope of spectrum has emerged as an interesting candidate.

Part of the evidence to support such a hypothesis comes from production studies. Various measures of spectral slope (e.g. spectral balance, spectral tilt or spectral emphasis) have thus been shown to be reliable acoustic correlates in several languages for distinguishing between stressed and unstressed syllables (Sluiter & van Heuven, 1996), or (focally) accented and non-accented words (e.g. Sluiter & van Heuven, 1996; Campbell & Beckman, 1997; Fant, et al., 2000; Heldner, forthcoming). What this amounts to in practice is that stressed syllables and accented words may be expected to have relatively more energy in the higher frequency bands than comparable unstressed syllables and non-accented words do – in addition to the differences in duration and f0.

The crucial evidence for the hypothesis that spectral slope features ought to be included in the modeling of prominence, however, should be sought in perception studies. It must be shown that the observed acoustic differences in spectral slope also have a perceptual relevance. However, perceptual experiments with manipulations of spectral slope are rare, perhaps due to the fact that such features have been more difficult to modify compared to f0, duration or overall intensity without noticeable degradation of speech quality. One important exception, though, is the elaborate experiments performed by Sluiter et al. (1997). They showed that spectral balance – implemented in terms of increasing the levels of the frequency components above 500 Hz – provided a relatively strong perceptual cue for lexical stress, that is, for distinguishing stressed syllables from unstressed. They moreover showed that spectral balance was almost as important in this respect as duration. It deserves to be stressed here that they studied a prominence distinction at the lower end of the scale. A Dutch reiterant
nonsense word nana, concatenated from duplicates of an unaccented and unstressed syllable na, was used as the starting point for the acoustic manipulations. Furthermore, the subjects were instructed to determine the position of the stressed syllable in the target word, that is, whether the word was Nana or nana.

Taken together, the observations from production and perception studies suggest that differences in spectral slope may contribute to the perception of prominence. The energy in the higher frequency bands increases with prominence in natural speech and the observed acoustic differences seem to have a perceptual relevance, at least for distinguishing stressed and unstressed syllables. However, it remains to be shown that the acoustic differences have perceptual relevance also for distinctions at the upper end of the prominence scale, for example for distinguishing (focally) accented and unaccented words. Furthermore, the stronger the evidence for perceptual relevance, the more meaningful it will be to test the importance of spectral emphasis for speech synthesis. Given that spectral emphasis contribute to the perception of prominence an explicit modeling of spectral slope in connection with prominence could possibly improve the quality and the naturalness of speech synthesis (cf. Campbell & Beckman, 1997; Sluijter et al., 1997).

The current paper and the experiments reported in it have a double goal. A first aim has been to investigate whether spectral emphasis – a specific implementation of changes in spectral slope – contributes to the perception of prominence and especially to the perception of focal accents. In particular, it has been tested whether an increase in spectral emphasis causes words to be perceived as more prominent given an increase of the same magnitude as that found when comparing focally accented and non-accented words in natural speech. A second aim has been to explore whether modeling of spectral emphasis in connection with focal accents is liable to improve speech synthesis as far as quality and naturalness are concerned.

## Perception experiment I

### Method

#### Material

The speech material used in the first experiment was a read-aloud passage from a short story about a Robinson Crusoe not wanting to be rescued from his desert island. One male Swedish speaker read this passage and rendered it as six shorter phrases (henceforth Phrases 1 – 6) and accented the words indicated by capitals in Table 1.

The recording was made in a sound-treated room with a high quality condenser microphone. The signal was digitized and stored on hard disk (16 bit, 44.1 kHz).

For each phrase an alternative version was subsequently created involving increased spectral emphasis in each one of the accented words. The fact that spectral emphasis was increased in words that were already accented

---

1 The title of the short story originally written by Jean Ferry is Robinson and it first appeared in 1953 in Le mécanicien et autres contes (Swedish translation by Claes Hylinger).
deserves to be stressed here. The original phrases together with the manipulated version of each phrase yielded a total of twelve stimuli.

Increased spectral emphasis was implemented in terms of amplifying the frequency components above the fundamental frequency by 4 dB, while attenuating the fundamental approximately 2 dB. As a result overall intensity increased by roughly 3 dB. The spectral slices seen in Figure 1 can be used to illustrate the effect of the variations. The manipulations were brought about by means of digital filtering with a time varying high shelf filter whose corner frequency followed the course of the fundamental frequency at 1.5 times \( f_0 \). A high shelf is a filter that either amplifies or attenuates everything above its corner frequency by equal amounts. The digital filtering was carried out by means of software that was primarily intended for music production (Renaissance Equalizer from Waves Ltd. used as a plug-in module in ProTools Free from Digidesign Inc.). It should be noted that this method of increasing spectral emphasis does not degrade speech quality and that implementations with other filter characteristics are easily accomplished using the same tools.

**Subjects and procedure**
The strength of spectral emphasis as a perceptual cue to prominence was assessed by means of a paired comparisons procedure. The listeners were exposed to pairs of stimuli with the original version of each phrase constituting one member of the pair and the other member a version of the same phrase where spectral emphasis had been increased in the accented words. To neutralize a possible bias of order of presentation within each pair (concerning the order of the original version relative to the manipulated one), as well as any contextual effects between pairs, these two presentation orders were balanced, randomized and different for all subjects. Stimuli were presented over headphones and the listeners could repeat each pair as many times as they wished. Simultaneously with the presentation of each pair the corresponding text – where accented words had been capitalized (as in Table 1) – was shown on a computer screen. The listeners were instructed to concentrate on the accented words and to determine whether the accentuation was stronger in the first phrase of the pair or in the second (binary forced choice).

Ten native speakers of Swedish (five males, five females) participated in the experiment. All were employees, students or visitors at the Department of Philosophy and Linguistics, Umeå University, or at the Department of Speech, Music, and Hearing at KTH. Each listener had to judge each pair four times. In all, 40 judgements were obtained of each pair. Each session lasted approximately 15 minutes.

---

**Figure 1.** Spectral slices to illustrate the effect of the filtering in an open-mid vowel. The slice drawn with a solid line shows the original version and that drawn with a dashed line is the version provided with increased spectral emphasis.
Results

The responses from the first listening test were coded according to the number of times a given version of a phrase was judged to have stronger accentuation than the other version of the same phrase. These figures are shown in Table 2. If an increase in spectral emphasis does make accented words appear more prominent, the listeners can be expected to pick out the higher spectral emphasis version as having stronger accentuation more often than the original version. If they fail to do so, this means that increased spectral emphasis does not in fact increase word prominence and possibly also that the listeners have not been able to discriminate between the two versions of each phrase.

As is evident from Table 2, increased spectral emphasis did not prove to be an unambiguous cue to prominence in this experiment. Although the accented words of the phrases with increased emphasis were, on the whole, perceived as more strongly accented than those of the original versions (cf. Totals column in Table 2), there were also two phrases (Phr 1 and 5) for which the original version was perceived as stronger in the indicated sense. Furthermore, the Chi-square tests across all speakers for the individual phrases were significant for the third and sixth phrases only:

\[ \chi^2(1) = 0.01, \ p = 0.75; \ \chi^2(1) = 1.6, \ p = 0.21; \ \chi^2(1) = 6.4, \ p = 0.01; \ \chi^2(1) = 0.9, \ p = 0.34; \ \chi^2(1) = 0.4, \ p = 0.53; \ \chi^2(1) = 22.5, \ p < 0.01. \]

Discussion

The results presented above did not reveal any dramatic effect on perceived prominence of an increase in spectral emphasis. As a matter of fact, spectral emphasis, as implemented here, did not cause already accented words to be perceived as more prominent in any straightforward manner, since perceived strength was significantly affected only in two out of six comparisons. This result may seem somewhat unexpected in the light of those obtained in previous listening experiments involving manipulations of spectral slope (Sluijter et al., 1997).

It is true that several factors concerning the speech material were not controlled for in this experiment. Nevertheless, there seems to be no simple ad hoc explanation why an increase in spectral emphasis would not yield a significant effect on perceived prominence in the remaining four comparisons. First of all, the observed results are unlikely to be accounted for by properties inherent in the accented words. The reason is obviously that the phrases showing a significant difference between the original version and the increased emphasis version on the one hand and the ones where such a difference is lacking on the other are quite comparable both as for the distribution of word length in accented words and of vowel height in the stressed vowels (cf. Table 1). Nor does it seem very probable that the differences could be accounted for by phrase characteristics, such as length (in terms of number of words) or complexity (in terms of embeddedness). In fact, both groups of phrases – whether the differences were significant or not – contained phrases of about the same length and complexity. Furthermore, all the phrases were longer than the one used by Sluijter et al. (1997).

The explanation for the weak effect of spectral emphasis on perceived prominence, however, could of course be the way spectral emphasis was increased. Firstly, the method chosen in the present study might just not have been realistic enough to affect perceived prominence. The uniform amplification of components above \( f_0 \) was perhaps to crude an approximation of increased spectral emphasis. Secondly, the spectral change caused by this particular implementation may not have been large enough. Indeed, the implementation was slightly different from the one used by Sluijter et al. (1997). They increased the levels of the frequency components above 500 Hz by up to 9 dB, whereas in the present experiment, the lower limit of the amplified frequency range was variable (i.e. determined by 1.5 times \( f_0 \) in each instant) and the components above this limit were amplified by 4 dB. Thus, the lower limit of the amplified frequency range as well as the amount of amplification was generally lower. As

| Table 2. Number of judgements of stronger accentuation for each phrase across all speakers. |
|---------------------------------|-----|-----|-----|-----|-----|-----|
|                                | Phr 1 | Phr 2 | Phr 3 | Phr 4 | Phr 5 | Phr 6 |
| Increased emphasis             | 19    | 24    | 28    | 23    | 18    | 35    |
| Original                       | 21    | 16    | 12    | 17    | 22    | 5     |
| Totals                         | 40    | 40    | 40    | 40    | 40    | 240   |

54
a consequence was the spectral change smaller and therefore possibly more difficult to perceive in the present experiment. It should be emphasized, however, that the increases in spectral emphasis operated on the test material were of the same magnitude as, or even slightly larger than, those typically found when comparing focally accentted and non-accentted words in production studies (e.g. Heldner, forthcoming).

Furthermore, the stimuli used in Sluiter et al. (1997) may have varied both in formant frequency and in spectral balance, the reason being that an amplification of components above 500 Hz may also affect the formant frequencies in the vicinity of 500 Hz. This effect is non-negligible. In a preliminary test of our own, F1 in the beginning of an unstressed /a/ uttered by a male speaker increased from about 590 Hz to 650 Hz, that is about 0.4 barks, as a result of the amplification of components above 500 Hz by 9 dB. Thus, the vowel quality changed towards a more open vowel. Such an effect is likely to be perceivable in itself under normal listening conditions (e.g. Kewley-Port & Zheng, 1999). The positive results, therefore, may have been due to additional formant frequency differences among the stimuli rather than the variations in spectral slope. Spectral emphasis, as implemented in the present experiment, may also affect F1. However, the actual changes in F1 were minor and generally within a range of ±10 Hz.

Finally, there may possibly be limitations to the increase in prominence to be expected when raising spectral emphasis in already accented words where, presumably, spectral emphasis has already been elevated as compared to some non-accented baseline. If this is so, it might simply be unfair to make comparisons with the previously mentioned listening experiments, where positive results were reported. In fact, Sluiter et al. (1997) started from a much lower baseline in their experiments by increasing the spectral balance in non-accented and unstressed syllables.

We must not forget, however, that there were two comparisons where the increase in spectral emphasis had a significant effect on perceived prominence. If nothing else, this shows that an increase of the magnitude used in the present experiment may be above a perceptual threshold.

In the following experiment, mbrola synthesis utterances were manipulated in the same way as in the first experiment. In these utterances there was no previous modeling of spectral emphasis in connection with the accented words in the original utterances.

### Perception experiment II

The aim of the second experiment was to examine whether modeling of spectral emphasis in connection with accents will improve the quality and the naturalness of synthetic speech. To this end, utterances produced by a mbrola speech synthesis were manipulated in the same way as in the first experiment and listeners were asked to compare the naturalness of original and manipulated utterances.

### Method

#### Material

The phrases from the first experiment were also used in the second experiment (cf. Table 1). However, this time they were produced by a male mbrola synthesis voice. The current version of the synthesizer contains no modeling of spectral emphasis in connection with prominence, so accented words in the second experiment were marked by durational and tonal means only. The synthetic utterances were prepared using the software WaveSurfer (Sjölander & Beskow, 2000) with a text-to-speech plug-in.

A version with increased spectral emphasis in the accented words was subsequently created of each phrase using the same digital filtering technique as in the first experiment. Thus, the frequency components above the fundamental frequency were amplified by approximately 4 dB while the fundamental was attenuated by 2 dB, and the overall intensity level in the vowels was raised about 3 dB. All in all, the six original phrases in combination with the manipulated version of each phrase yielded a total of twelve stimuli.

#### Subjects and procedure

In a listening test it was then explored whether any gain in naturalness had been obtained from the inclusion of a realistic increase in spectral emphasis in the accented words. The details of the procedure were practically identical to those of the first experiment. The only difference concerned the actual stimuli – as mbrola utterances were used here – and the task given to the subjects. As in the first experiment, the listeners were instructed to compare the accented words in two versions of a phrase. This time, however, they were asked to judge which version sounded more natural (binary forced choice).

Another ten native speakers of Swedish (five males, five females) participated in the second experiment. None of them had participated in
the first experiment. All of them were employees, students or visitors at the Department of Speech, Music, and Hearing at KTH. Each listener had to judge each pair four times. Thus, 40 judgements were obtained for each pair. In this experiment, too, each session lasted approximately 15 minutes.

Results

The responses from the second experiment were coded according to the number of times a given version of a phrase was judged to have a more natural accentuation than an alternative version of the same phrase. The figures obtained can be studied in Table 3. In case an increase in spectral emphasis causes accented words in the mbrola synthesis to be perceived as more natural, the listeners could be expected to judge the version with elevated spectral emphasis to be more natural more often than the original version.

As is shown by Table 3, manipulation of spectral emphasis did not affect the perceived naturalness of the accented words to any great extent. Rather, the original versions seem to have been perceived as slightly more natural (cf. Totals column in Table 3). Moreover, none of the Chi-square tests for the individual phrases were significant: [Phr 1: \( \chi^2(1)=0.0, p=1 \); Phr 2: \( \chi^2(1)=0.1, p=0.75 \); Phr 3: \( \chi^2(1)=0.4, p=0.53 \); Phr 4: \( \chi^2(1)=0.9, p=0.34 \); Phr 5: \( \chi^2(1)=0.1, p=0.75 \); Phr 6: \( \chi^2(1)=1.6, p=0.21 \)].

Discussion

All in all, the second experiment yielded clearly negative results. As we have just seen, there was no significant improvement in the perceived naturalness of accented words in mbrola utterances provided with an addition of a realistic increase in spectral emphasis. In other words, although an increase in spectral emphasis may in principle be a reliable acoustic correlate for distinguishing various levels of prominence (e.g. Sluijter & van Heuven, 1996; Campbell & Beckman, 1997; Fant et al., 2000; Heldner, forthcoming), the inclusion of this particular feature did not manage to make synthesized utterances appear any more natural to the listeners in this experiment.

Now, this result is perhaps not very surprising. After all, the perceptual influence of spectral emphasis found in the first experiment was already extremely feeble, in spite of the fact that the manipulated natural speech it contained had a high sound quality as compared to that of the mbrola speech in the second experiment. Since subtle acoustic differences are likely to be more difficult to perceive in speech of degraded quality, one would expect those used in the second experiment to have an even weaker effect on perception. Supposing, furthermore, the actual implementation of spectral emphasis in the first experiment to be somehow problematic, results of the second experiment would have been equally affected, since the stimuli of both experiments were created using the same technique.

On the other hand, the second experiment had the merit of making the manipulated versions more similar to natural speech, as realistic variations of spectral emphasis in connection with accents were introduced. In contrast, the manipulations carried out in the first experiment diverged from naturalness, as the increases in spectral emphasis were performed on already accented words. Therefore, the second experiment ought to have been the better test of whether realistic increases in spectral emphasis also increase the naturalness of accented words. Apparently, however, spectral emphasis as implemented here did not improve the naturalness of accented words in mbrola utterances. A possible explanation for this could be that the listeners were not able to perceive the difference.

Conclusions

This study has presented a flexible method for increasing spectral emphasis without degrading the quality of the output. It works equally well for natural and for synthesized speech. The technique also allows for local increases in spectral emphasis, as for example in accented words or in stressed syllables. Potentially this

Table 3. Number of judgements of more natural accentuation for each phrase across all speakers.

<table>
<thead>
<tr>
<th></th>
<th>Phr 1</th>
<th>Phr 2</th>
<th>Phr 3</th>
<th>Phr 4</th>
<th>Phr 5</th>
<th>Phr 6</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased emphasis</td>
<td>20</td>
<td>19</td>
<td>22</td>
<td>17</td>
<td>19</td>
<td>16</td>
<td>113</td>
</tr>
<tr>
<td>Original</td>
<td>20</td>
<td>21</td>
<td>18</td>
<td>23</td>
<td>21</td>
<td>24</td>
<td>127</td>
</tr>
<tr>
<td>Totals</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>240</td>
</tr>
</tbody>
</table>

56
method ought to be useful for exploring spectral emphasis in connection with the perception of prosodic phenomena.

Furthermore, this study has reported on two listening experiments in which this technique was used. Interestingly, both experiments yielded negative results, which was somewhat unexpected given the results of previous production and perception experiments. Manipulations of spectral emphasis of the magnitude found when comparing the production of accented and non-accented words – i.e. what could be considered a realistic level – did not make accented words more prominent in any straightforward fashion. Nor did they improve the naturalness of accented words in synthetic utterances.

So, a tentative conclusion must be that spectral emphasis, at least as implemented here, seems to be fairly weak as a cue to prominence at the upper end of the prominence scale, and moreover, of little value for improving the quality of speech synthesis. Future research will have to show if other ways of implementing spectral emphasis may produce more salient effects on perceived prominence and on perceived naturalness of accented words in speech synthesis.

Acknowledgements

This research was carried out when the author was a guest at the Centre for Speech Technology, a competence center at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH, and participating Swedish companies and organizations. The author would like to thank Eva Strangert and Inger Karlsson for invaluable comments on earlier drafts.

References


Heldner M. Spectral emphasis as a perceptual cue to prominence