A FOCUS DETECTOR USING OVERALL INTENSITY AND HIGH FREQUENCY EMPHASIS

Mattias Heldner[1], Eva Strangert, and Thierry Deschamps
Department of Philosophy and Linguistics, Umeå University, SE-901 87 Umeå, Sweden
[1] Also at Centre for Speech Technology (CTT), KTH, SE-100 44 Stockholm, Sweden

ABSTRACT

The ultimate goal of the present study is to develop a focus detector for Swedish integrating several sources of information. To that end we have examined the extent to which different measures of overall intensity and high frequency emphasis can be used to detect focus in read and spontaneous Swedish. It is shown that focused words can be detected with a rate approximating 70% under favorable conditions. Furthermore, as the effectiveness of the measures varies between speakers and the position of the focused word in the phrase, we believe that it would be advantageous to include both measures of intensity and of emphasis in a focus detector.

1. INTRODUCTION

Detection of semantic focus is important for humans as well as for automatic speech recognition systems in order to achieve a full interpretation of spoken utterances. One possible way to detect focus is through its prosodic manifestation in the acoustic signal, as focused words are likely to be pronounced accented in many languages and accents have acoustic correlates.

It is generally agreed that the primary correlates of accents are fundamental frequency (F0) movements. Consequently, F0 information is crucial to many focus or prominence detectors [e.g. 1,2,3]. However, accents also have other acoustic correlates that might be used by humans as well as by machines. For example, work by Sluijter and van Heuven [4,5,6,7,8] has shown that accented words in Dutch and American English in addition to the F0 movements tend to be produced with longer segmental durations, higher overall intensity and a high frequency emphasis. Accordingly, it has been confirmed by e.g. Campbell [9,10,11] and Wightman, Ostendorf and Ross [12,13,14] that segmental duration, intensity and spectral information are reliable cues for the automatic detection of prominence in English (see also [15] for German).

The primary goal of this study is to evaluate the extent to which overall intensity and high frequency emphasis can be used to detect focus in read and spontaneous Swedish speech. A secondary goal is to evaluate the performance of different measures of overall intensity and high frequency emphasis as focus detectors. A future goal is to integrate information about known correlates of accents and focus such as overall intensity, high frequency emphasis, duration and F0 movements in a focus detector for Swedish.

2. METHOD

2.1 The focus detector

The focus detector was based on the assumption that a focused word is the most prominent word in the phrase, i.e. that there is only one focused word per phrase. It was also assumed that these prominence relations would show up in the measures of overall intensity and high frequency emphasis. Thus, the detector should pick up the word containing the highest value in the phrase for a given measure and classify it as focused.

A number of different measures of overall intensity and high frequency emphasis were evaluated as predictors of focus. First, four measures of overall intensity were derived from the speech signal. The measures, differing in the time over which intensity was integrated, included the maximum value in the phrase as well as the RMS mean across the duration of each segment, each syllable and each word.

Second, the measures of overall intensity were compared with measures of high frequency emphasis. High frequency emphasis (sometimes called spectral tilt or spectral balance) can be measured in several ways [e.g. 11,16,17,18]. We used a method proposed by Traunmüller [19] calculating the difference between the overall intensity and the intensity in a signal low-pass filtered at 1.5 times the F0 mean for each utterance. For a subset of the test material, three additional measures were calculated where the low-pass filter frequencies were fixed at 0.5kHz, 1kHz and 2kHz instead of being determined by the F0 mean in the utterance. As for the intensity measures, the maximum value in the phrase as well as the RMS mean across the duration of each segment, each syllable and each word were calculated for each of these four measures. Thus, in all, 16 different measures of high frequency emphasis were evaluated (4 different low-pass filter frequencies x 4 different integration times).

The detector was implemented in a UNIX and *ESPS/Waves+™* environment. As mentioned above, it should pick up one word candidate per phrase. Thus, for detection based on the maximum value in the phrase the specific word containing this value was selected. For measures based on mean values, the word containing the highest mean for the total of segments, syllables and words, respectively, in each phrase was selected. The performance of the detector was evaluated on the basis of comparisons with prosodic transcriptions of the test material (see 2.2 below).

2.2 Speech material for testing the detector

As a first step, the detector was tested on a database containing 1,105 three-word sentences [20]. Each sentence consisted of one

phrase. These sentences were produced as answers to questions designed to elicit narrow focus on each of the three words in turn. The sentences were read by four female and four male Swedish speakers. The recordings were made in a sound-treated room with a microphone mounted on a headset, so that a constant distance from the mouth to the microphone was maintained.

To facilitate a comparison between the measures differing in integration time, information about segment-, syllable-, word- and phrase-boundaries was required. Therefore, since no automatic methods (such as automatic speech recognizers or boundary detectors) for procuring this information were available, the speech files fed into the detector were accompanied by an orthographic transcription where these boundaries were indicated. A prosodic transcription where the focused words were indicated was used as a target to evaluate the performance of the detector.

In a second step, the detector was tested on less controlled material. Two recordings were used. The first was a read-aloud news text, 233 words and 33 phrases long. The other was a retelling of the content of the same news text, 252 words and 58 phrases long. The same male Swedish speaker rendered both recordings. The recordings were made in a sound-treated room, but in this case, in contrast to the read sentences material, there was no control over the distance between the speaker and the microphone.

The recordings of the read-aloud news text and the spontaneous retelling had been prosodically transcribed by a group of experts in a study of labeling consistency [21]. These transcriptions contained perceived phrase boundaries and different prominence levels, including focused words. The transcriptions by the expert with the longest experience of prosodic transcriptions were used as targets to evaluate the performance of the detector.

There were no indications of segment or syllable boundaries (only word boundaries) in the label files fed to the detector, and thus only a subset of the intensity and emphasis measures (the maximum values in the phrase and the RMS means across the duration of each word) was evaluated on the basis of the read-aloud news text and the retelling of its content.

3. RESULTS

3.1 Three-word sentences

The output of the detector produced one candidate from each acoustic measure for each sentence = phrase. These candidates were compared with the prosodic transcriptions and the percentage of correct detections was calculated. Table 1 shows the number and percentage of correct detections obtained using the different overall intensity measures. Table 2 shows the number and percentage of correct detections obtained using the four different high frequency emphasis measures.

	Overall intensity			
	Max	Mean/seg	Mean/syll	Mean/word
Counts correct	801	775	737	742
Percent correct	72%	70%	67%	67%

Table 1. Counts and percentage of correct detections in the three-word sentences for the maximum values in the phrase (Max) and the RMS means across the duration of each segment (Mean/seg), syllable (Mean/syll) and word (Mean/word) for overall intensity. N=1105.

Tables 1 and 2 show that the best measures of overall intensity and high frequency emphasis by themselves could be used to detect about two-thirds of the focused words in the three-word sentences. Moreover, Tables 1 and 2 show that the measures differed in strength. In general, the scores were higher for overall intensity than for high frequency emphasis and comparing the different measures for high frequency emphasis, the one where the low-pass filter was determined by the F0 mean produced the best results. Also, the means per syllable and word (and to some extent also the means per segment) gave consistently lower scores than the maximum values for intensity as well as for emphasis measures.

	High frequency emphasis			
F0 mean	Max	Mean/seg	Mean/syll	Mean/word
Counts correct	744	736	701	675
Percent correct	67%	67%	63%	61%
0.5kHz	Max	Mean/seg	Mean/syll	Mean/word
Counts correct	613	632	583	575
Percent correct	55%	57%	53%	52%
1kHz	Max	Mean/seg	Mean/syll	Mean/word
Counts correct	643	663	615	603
Percent correct	58%	60%	56%	55%
2kHz	Max	Mean/seg	Mean/syll	Mean/word
Counts correct	447	435	398	413
Percent correct	40%	39%	36%	38%

Table 2. Counts and percentage of correct detections in the three-word sentences for the maximum values in the phrase (Max) and the RMS means across the duration of each segment (Mean/seg), syllable (Mean/syll) and word (Mean/word) for the high frequency emphasis measures where the low-pass filter was determined by the F0 mean in each utterance (F0 mean), and for the fixed low-pass filters at 0.5kHz, 1kHz and 2kHz. N=1105.

As mentioned, the scores were higher for overall intensity than for high frequency emphasis and moreover, the maximum values of intensity in the phrase gave the highest scores. However, a closer examination revealed considerable speaker differences. The maximum value of intensity was the best measure for five of the speakers, while the maximum value of high frequency emphasis was the best for the other three. Furthermore, the percentage of correct detections for the different speakers ranged between 47% and 91% for the maximum value of intensity and between 49% and 89% for the maximum value of high frequency emphasis. As a group, the percentage of correct detections was higher for the male speakers than for the female speakers, but since this difference

was mainly due to two of the speakers (one male and one female), it seemed to be a matter of individual speaker rather gender. Another observation was that the overall best measure, the maximum value of intensity, were achieved in initial and medial position (86% correct detections in initial and 87% in medial position) rather than in final position (44%) in the sentence. The corresponding figures for maximum high frequency emphasis were 62%, 79% and 61% in initial, medial and final position, respectively.

3.2 Read-aloud news text and spontaneous retelling

In the read-aloud news text and the spontaneous retelling of the news text, the assumption of only one focused word per phrase was sometimes proved false. There were phrases transcribed as having no focused words as well as phrases with more than one focused word. Thus, a relaxed criterion of correct detection had to be used. If one of the words transcribed as focused in a phrase was detected, this was taken as a correct detection. Phrases without any focused words were excluded from the analysis. Moreover, phrases consisting of only one focused word were also excluded, since they could not be falsely detected.

Since the label files fed into the detector were restricted to word and phrase boundaries for this part of the material, the output of the focus detector consisted of word candidates selected on the basis of four measures, which candidates were selected by the maximum values in the phrase and the means across the duration of each word. Moreover, detection based on high frequency emphasis was restricted to the F0-mean related measure. Table 3 shows that the performance of the detector was very poor for the read-aloud news text. With an average phrase length of 7 words, the reported values were only slightly above what you would expect to find by chance. The results from the spontaneous retelling were somewhat more promising. Here, the average phrase length was 5 words and the percentage of correct detections for maximum intensity and emphasis were well above chance levels. Moreover, the percentage of correct detections for maximum intensity for the retelling was close to the result for the three-word sentences. Once again, intensity produced higher scores than emphasis, and the maximum values higher scores than the values integrated over a longer stretch of speech.

	Overall intensity		**High frequency emphasis**	
Read-aloud	Max	Mean/word	Max	Mean/word
Counts correct	6	5	4	0
Percent correct	25%	21%	17%	0%
Spontaneous	Max	Mean/word	Max	Mean/word
Counts correct	25	22	18	12
Percent correct	66%	58%	47%	32%

Table 3. Counts and percentage of correct focus detections in the read-aloud news text and in the spontaneous retelling of the news text for maximum values (Max) for overall intensity and high frequency emphasis and RMS means across the duration of each word (Mean/word). N=24 read news text; N=38 spontaneous retelling.

4. DISCUSSION

This study shows that overall intensity and high frequency emphasis can be used to detect focused words in Swedish to an extent comparable to that reported for English, Dutch and German [c.f. 6,9,11,15]. About two thirds of the focused words in the three-word sentences were correctly detected using the maximum values of either overall intensity or high frequency emphasis in the phrase.

Furthermore, it was shown that the predictive power of the different measures of overall intensity and high frequency emphasis varied considerably. The maximum values in the phrase gave more correct detections than the integration of data over longer stretches of speech, i.e. the RMS means across the duration of each segment, syllable and word. Moreover, low-pass filters adjusted to the fundamental frequency of the speaker gave more correct detections than fixed low-pass filter frequencies (0.5kHz, 1kHz and 2 kHz) in the high frequency emphasis measures.

However, the performance of the detector depended on the speaker and on the position of the focused word in the phrase. While the maximum value of overall intensity in the phrase was the best predictor of focus in general, the maximum value of high frequency emphasis was the best predictor for some of the speakers. (There were no obvious differences in the performance of the detector between male and female speakers.) Moreover, overall intensity was less robust than high frequency emphasis in phrase-final position.

From these findings, we conclude that it is advantageous to include both measures of overall intensity and of high frequency emphasis in a focus detector integrating several sources of information. Moreover, maximum values are better than integration over longer stretches of speech and the emphasis measure using a speaker adaptive low-pass filter is better than the measures with fixed low-pass filters [19]. On the basis of these results, it would be interesting to make comparisons with a measure using a dynamic low-pass filter following the fundamental frequency even closer.

As to the reason why focused words were harder to detect in the read-aloud news text than in the spontaneous retelling of the same content, we have no definite answer. Our results, however, seem to indicate that speaking style or rather, speaking style variation, is reflected in the use of overall intensity and high frequency emphasis. We know that it was not harder to detect focused words in the read news text for human listeners. The consistency measures obtained in the previously-mentioned labeling study for a group of expert transcribers using the same material were even higher for the read-aloud news text than for the spontaneous retelling [21]. Obviously, the differences cannot be attributed to speaker differences, as the same speaker rendered both recordings. Nor can they be attributed to the distribution of the focused words in the phrase; the spontaneous retelling contained more phrase final focused words (where detection was less robust than in initial and medial position in the phrase) than the read-aloud news text. The good news, however, is that the spontaneous speech worked better than the read-aloud speech.

5. CONCLUSIONS

Overall intensity and high frequency emphasis are probably not as significant correlates of focus as F0 movements and segmental lengthening phenomena. Nevertheless, this study shows that the measures evaluated can be used to detect about two-thirds of the focused words under favorable conditions and thus possibly enhance the performance of focus detectors integrating several acoustic measures.

ACKNOWLEDGMENTS

The work was supported by a grant from the Swedish Council for Research in the Humanities and Social Sciences.

REFERENCES

[1] House, D. and Bruce, G., 1990. Word and focal accents in Swedish from a recognition perspective. In Wiik, K. and Raimo, I. (Eds.), *Nordic Prosody V*. Turku University.

[2] Sautermeister, P. and Lyberg, B., 1996. Detection of sentence accents in a speech recognition system. *Journal of the Acoustical Society of America*, 99 (4, pt 2), 2493.

[3] Elsner, A., 1997. Focus detection with additional information of phrase boundaries and sentence mode. In *Eurospeech '97 Proceedings*. Rhodes, Greece: ESCA.

[4] Sluijter, A. and van Heuven, V., 1995. Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch. *Phonetica*, 52, 71-89.

[5] Sluijter, A. and van Heuven, V., 1996. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100 (4, Pt 1), 2471-85.

[6] Sluijter, A. and van Heuven, V., 1996. Acoustic correlates of linguistic stress and accent in Dutch and American English. In *Proceedings ICSLP 96*. Philadelphia: Applied Science and Engineering Laboratories, Alfred I. duPont Institute.

[7] Sluijter, A., van Heuven, V., and Pacilly, J., 1997. Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, 101 (1), 503-13.

[8] Sluijter, A., Shattuck-Hufnagel, S., Stevens, K. *et al*., 1995. Supralaryngeal resonance and glottal pulse shape as correlate of stress and accent in English. In *Proceedings ICPhS 95*. Stockholm: Department of Speech Communication and Music Acoustics, KTH and Department of Linguistics, Stockholm University.

[9] Campbell, N., 1992. Prosodic encoding of English speech. In *ICSLP 92 Proceedings*. Edmonton, Alberta: University of Alberta.

[10] Campbell, N., 1994. Combining the use of duration and F0 in an automatic analysis of dialogue prosody. In *Proceedings ICSLP 94*. Yokohama: The Acoustical Society of Japan.

[11] Campbell, N., 1995. Loudness, spectral tilt, and perceived prominence in dialogues. In *Proceedings ICPhS 95*. Stockholm: Department of Speech Communication and Music Acoustics, KTH and Department of Linguistics, Stockholm University.

[12] Wightman, C. and Ostendorf, M., 1992. Automatic recognition of intonational features. In *Proceedings ISCASSP-92*. San Francisco: IEEE.

[13] Wightman, C. and Ostendorf, M., 1994. Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, 2 (4), 469-481.

[14] Ostendorf, M. and Ross, K., 1997. A multilevel model for recognition of intonation labels. In Sagisaka, Y., Campbell, N., and Higuchi, N. (Eds.), *Computing Prosody*. New York: Springer-Verlag.

[15] Nöth, E., Batliner, A., Kuhn, T. *et al*., 1991. Intensity as a predictor of focal accent. In *Proceedings ICPhS 91*. Aix-en-Provence: Université de Provence.

[16] Campbell, N. and Beckman, M., 1997. Stress, prominence, and spectral tilt. In Botinis, A., Kouroupetroglou, G., and Carayiannis, G. (Eds.), *Intonation: Theory, models and applications*. Athens: ESCA.

[17] Stevens, K. and Hanson, H., 1995. Classification of glottal vibration from acoustic measurements. In Fujimura, O. and Hirano, M. (Eds.), *Vocal fold physiology: Vocal quality control*. San Diego: Singular Publishing Group.

[18] Sluijter, A., 1995. *Phonetic correlates of stress and accent*. The Hague: Holland Academic Graphics.

[19] Traunmüller, H., 1997. Perception of speaker sex, age, and vocal effort. In Bannert, R., Heldner, M., Sullivan, K., and Wretling, P. (Eds.), *PHONUM 4*. Umeå: Department of Phonetics, Umeå University.

[20] Heldner, M. and Strangert, E., 1998. On the amount and domain of focal lengthening in Swedish two-syllable words. In *Proceedings of FONETIK 98*. Stockholm: Department of Linguistics, Stockholm University.

[21] Strangert, E. and Heldner, M., 1995. Labelling of boundaries and prominences by phonetically experienced and non-experienced transcribers. In Bannert, R. and Sullivan, K. (Eds.), *PHONUM 3*. Umeå: Department of Phonetics, Umeå University.