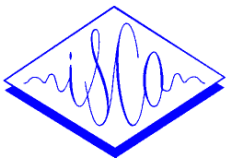

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

CAN 'LEVEL WORDS' FROM ONE SPEAKING STYLE BECOME 'PEAKS' WHEN SPLICED INTO ANOTHER SPEAKING STYLE?

Florien J. Koopmans-van Beinum

Institute of Phonetic Sciences, University of Amsterdam
Herengracht 338, 1016 CG Amsterdam, The Netherlands

ABSTRACT

Is the listener using fixed reference targets, when perceiving either read or spontaneous speech, or is he/she attuning his/her perception mechanism to the specific speaking style, applying some kind of style normalization. To answer this question we spliced function words from three different speaking styles into two types of lexically identical short sentences, either from free conversation, or from the same text read aloud. To avoid F0-influences the sentences were made unvoiced by LPC-resynthesis. It turned out that in a number of cases function words segmented from more carefully pronounced speaking styles were heard as accented when spliced into more reduced speaking styles, merely because of their longer duration and/or more spectral contrast than the original words.

10.21437/ICSLP.1992-326

1. INTRODUCTION

In normal conversation every speaker will automatically aim at an optimal balance between careful pronunciation and sloppiness: the communicatively important words in conversation do not tolerate any misunderstanding, whereas less important or already known parts of the utterances do not need the same amount of carefulness and can be understood with half a word.

Our acoustic studies with respect to notions like 'new' versus 'given' [4], 'focus' versus 'non-focus' [5], and 'content words' versus 'function words' [1] all point into the same direction: words with a high load of semantic information are pronounced with a longer duration, more spectral contrasts, with more intonational movements, and more energy than words with a low semantic load. These results fit in very well with our so-called peak-level model [5]. This peak-level model was introduced in order to account for the fact that the acoustic *differences* (or *contrast*) between focus and non-focus words in read speech productions are comparable to those in spontaneous speech conditions, although the acoustic realisations in spontanous speech situations as a whole are considerably more reduced than in read speech forms.

However, the question to be put next is: does the listener actually make use of these acoustic differences between peak and level, when perceiving either read or spontaneous speech? Does he/she always use some fixed reference target or is he/she attuning his/her perception mechanism to the specific speaking style, applying some kind of style normalization, in order to cope with the differences in peak and level values? And if so, what are the implications in the field of speech technology? How can automatic speech recognition and text-to-speech synthesis systems take account of style?

Originally we planned to test our peak-level model perceptually by means of diphone synthesis. We intended to change peak and level values with the use of both full and reduced diphones, within the Dutch diphone speech synthesis system [2]. However, as was demonstrated before [5], the available set of reduced diphones displays still more mutual vowel contrasts than full vowels in focus words in natural continuous *read* speech of the same speaker, let alone the differences with the amount of reduction in natural *spontaneous* speech conditions. We therefore decided to start our perceptual tests with the related question: can level words from a more carefully pronounced speaking style evoke peak or focus responses, when spliced into the same context in another more reduced speaking style?

2. A WAY OF PERCEPTUALLY TESTING THE PEAK-LEVEL MODEL

In Fig. 1 a schematic overview is represented of possible variants of the peak-level model. However, on the basis of our acoustic measurements thus far, the possibilities I, III, V, and VI have to be rejected. The remaining possibilities II and IV differ from each other only in absolute values of peak-to-level contrast, but it is clear that these values highly depend on measuring parameters and type of scaling. In order to test some aspects of the model we tackled the question whether 'level' words from the one (more carefully pronounced) speaking style become 'peak' words when spliced into another (more sloppily pronounced) speaking style, if the differences between the two speaking styles are large. To test this statement we first had to make the notions 'peak' and 'level' operational. In our previous research it turned out that the notions 'accentedness' and 'unaccentedness' were highly related to the terms 'in focus' and 'not in focus'. So with respect to our present perception tests we asked our listeners to indicate the accented words in short neutral sentences. By asking our subjects for accentuation, these experiments can be considered as a continuation of the perception experiments as done before [7]. In that study the degree of spectral reduction of synthesized vowel stimuli was varied by moving the vowels in equal steps along vectors in the formant field from the cardinal vowels /u/, /i/, and /a/ to the centre being a schwa-like, neutral vowel. Vowels were presented in sequences of three-syllabic nonsense words of the form /fVfVfV/, in which one of these three vowels V was less centralized than the other two. It turned out then that, with all other parameters held constant, just this greater amount of decentralization or spectral contrast alone, was a significant cue for the perception of stress. Nevertheless listeners afterwards reported to have been guided in their decisions, apart from by vowel quality, also either by intensity, by duration, or by intonation!

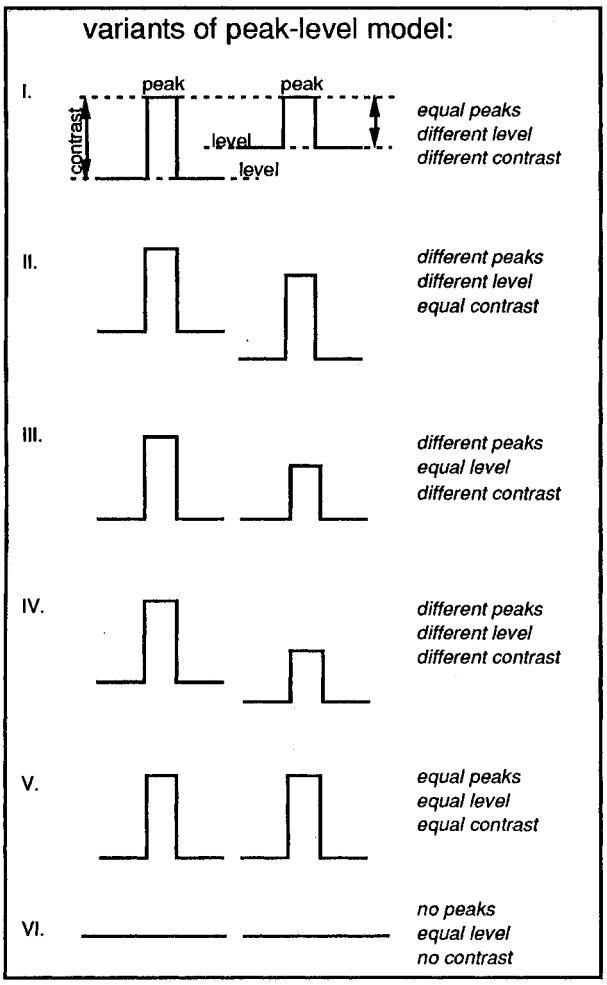

Fig. 1. Six possible variants of the peak-level model for production and perception of acoustic contrasts, which may concern any acoustic parameter, e.g. duration, energy, fundamental frequency, or spectral contrast. Peaks represent focus words and level stands for the average characteristics of all remaining non-focus words. Arrows indicate degree of contrast between peak and level.

In the present study we used manipulated natural speech stimuli, LVS-resynthesized. Again a part of the sequences presented to the listeners is more carefully pronounced (since excised from a more carefully pronounced speaking style) than the neighbouring context and therefore may give rise to stress or accent perception as well. It is apparent that in natural speech a number of parameters are responsible for the perception of accent: F0-movements are an important cue, but next to this also temporal, spectral, and intensity cues play their role in combined action [3]. Since we want to concentrate right now on the role of the spectral and temporal cues, a method had to be found to get rid of the accent-lending F0-movements. A small pilot listening experiment was set up using stimuli sentences provided with a fixed F0-declination (from 160 to 110 Hz in all sentences), presented over loudspeakers in a class room. It turned out that more than 70 % of the originally accented words were still correctly identified as being accented, although F0-movements could not have been a cue here. So fixing the F0-declination might be a practicable solution in our actual listening tests, but since the listeners reported to be annoyed and distracted by the unnatural declination, we decided henceforth in subsequent experiments to make use of unvoiced LPC-resynthesized sentences, sounding whispered and hoarse, but much more natural. Also with this procedure F0-movements cannot play any role in the identification of accentedness.

3. STIMULUS MATERIAL

In contrast with the synthetic speech material of the perception experiments mentioned above, we now started from natural speech material as stimuli for our present listening tests. In a number of recent studies within the scope of the Dutch national speech synthesis program, speech material of one professional male talker has been recorded and analysed. This provides us with a rather extended, quite accessible database of standard Dutch. This material includes 'spontaneous', i.e. free conversational speech recorded in laboratory situation, and the same text material read out afterwards in an identical recording session, so that two parallel versions of the same lexical material were created. In this way the acoustic data for the material of the two speaking styles was completely comparable.

In another study [1] concerning acoustic aspects of vowel reduction and the influence of sentence accent, word stress, and word class, the same talker was involved as one of the speaking subjects. To construct the stimulus sentences for the present listening tests, we selected five short sentences from our own spontaneous speech material and from the concurrent read material, together containing nine of the function words that were used in [1] as well, providing us with a third condition, i.e. the function word as part of an accented content word in short sentences read aloud, so a carefully pronounced speaking style.

The words were selected in such a way that the vowels in these CVC function words represent nine of the twelve Dutch vowels, only the scarcely occurring vowels /y/, /ø/, and /œ/ were lacking. Each of the nine function words, segmented from three different conditions, could be spliced into two types of carrier sentences (spontaneous and read), resulting in 54 short sentences (9 words x 3 conditions x 2 types of carriers). If indeed the degree of contrast between peak and level is used by the listener in identifying the words in focus (or the accented words), level words from one condition may be identified as peaks when spliced in another condition.

The actual stimulus tape was composed in the following way.

1) All 27 CVC-syllables (function words) were segmented and copied from the original, digitized speech material (9 words in 3 speech conditions).
2) The 10 carrier sentences (5 sentences in two conditions, viz. spontaneous and read aloud) were segmented to be combined with the three types of function words.
3) All sentences were LPC-resynthesized and made unvoiced in order to get rid of all F0-movements, so that F0 could not provide any accentuation cue.
4) All sentences were copied in random order on a audio tape. Each stimulus sentence was recorded three times with one second in between. Interstimulus time between the different sentences was three seconds. The whole set of stimuli was recorded a second time in a different random order. So the resulting tape consisted of 108 stimulus sentences in total.

Table 1. Overview of number of times each of the nine function words has been marked 'accented', summed over the ten listeners (max. score = 20). Type of stimulus consisted of function words from originally spontaneous (sp.), read aloud (rd.) speech material or from content words (co.) in short read sentences, spliced into five spontaneous sentences (sp. sent.) or five lexically similar read sentences (rd. sent.)

type of stimulus	heb	niet	voor	moet	was	meer	daar	wil	kon
sp. word in sp. sent.	8	8	2	0	0	1	10	2	2
rd. word in sp. sent.	7	11	3	0	1	0	20	12	18
co. word in sp. sent.	12	17	16	10	18	19	20	20	15
sp. word in rd. sent.	4	4	3	0	0	0	7	0	0
rd. word in rd. sent.	3	9	2	1	0	0	20	0	13
co. word in rd. sent.	13	16	18	8	7	19	20	13	9

4. LISTENING EXPERIMENT

The stimulus tape was presented to ten listeners, some of them experienced and some unexperienced in listening tests, in two different orders to prevent a too strong order effect on the results. Half of the listeners started with the first randomized version, the other half started with the other version. Before the actual listening test all listeners were presented with one version (in threefold) of each of the five carrier sentences, to get accustomed to the lexical content of the sentences, to the whispered sound, and to the tempo of presentation.

Listeners were asked to indicate for each stimulus sentence the word which had been accented (or put in focus) by the speaker. Next they had to mark whether this was done by using a more or less neutral sentence accentuation or by using extra accentuation on a certain word, giving the sentence a meaning deviating from the normal, informative statement (e.g. extra: *"I cannot do **that**"* and *"I cannot do that"* versus neutral: *"I cannot do that"*). After the test the listeners were asked to write down their comments about task and scoring strategy. The listening test took about twenty minutes per subject.

5. RESULTS

In Table 1 the overall results are given for the three types of function words spliced into carrier sentences from spontaneous speech and read aloud. It is clear that in almost all cases the CVC-syllables segmented from a content word evoke accentuation effect, when spliced as a function word into a spontaneous or read sentence. The CVC-syllables (function words), segmented from read speech and spliced into spontaneous sentences, evoke this effect in a number of cases as well. So far these results are in agreement with our peak-level model.

The data, however, that have been used as stimuli in the listening experiment, require a much more detailed inspection of the results, since each sentence in its original form had its own accentuation pattern. This means that these patterns could have influenced the responses and therefore have to be taken into consideration when interpreting the results. So for each of the five carrier sentences in both speaking styles we compared the given responses in relation to the responses when the original word was presented in its original context. In Fig. 2 an example is given of scores concerning the spontaneous spoken sentence "of ik daar wilde komen" (in Eng.: "whether I would come there"), with the read and the content version of the word "daar" spliced into this sentence (left-hand graph) and next to this the word "wil" into the same sentence (right-hand graph). The white bars in the figure indicate the distribution of accentuation responses in the original sentences.

As can be seen in the left-hand graph the word "daar" always attracts many accentuation responses, but the right-hand graph shows how this is overruled when the word "wil" is spliced into the same sentence, both as part of a content word and as read function word. In this example a level word has clearly become a peak.

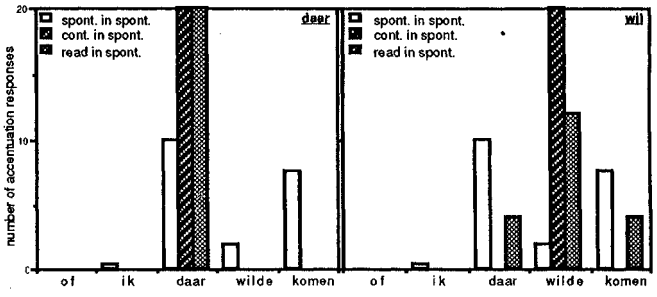

Fig. 2. Example of the perception results when content words and read function words are spliced into a spontaneously spoken carrier sentence. For more details, see text.

6. ACOUSTIC DATA OF THE STIMULI

Inspection of the acoustic data of the stimuli may provide some explanations for the results. To start our perception experiments with, we modified as little as possible the words to be spliced into the carrier sentences; therefore durational as well as spectral parameters had been left unaltered. With respect to the durational and spectral information the following data are the most relevant ones.

The durations of the five original spontaneous and read sentences, to be used as carrier sentences displayed a correlation of 0.93 (respectively 710, 729, 962, 708, 1085 msec for the spontaneous sentences, and 772, 910, 1170, 955, 1232 msec for the concurrent read sentences). So in all cases the durations of the read sentences are longer, as could be expected from our previous measurements [4].

In Fig. 3 word and vowel durations are given of the nine function words in the three conditions.

As can be seen from the diagrams, the durational behaviour of these words and vowels is rather varied: for some words vowel and/or word durations are almost equal in all three conditions, whereas in some other words large discrepancies exist. Correlation between vowel duration and word duration for the three conditions grouped together is 0.70, which in our opinion is lower than might be expected.

In Fig. 4 vowel formant frequencies (F1 and F2) in Hz, measured at the most stable part of the vowel, are represented in the formant field for each of the nine vowels in each of the three conditions (for details on measuring methods, see [1]). It is clear that for all nine words quite different spectral vowel realisations occur in the three conditions, in spite of the identical consonantal context for all three conditions, and even identical sentence context as far as function words in the spontaneous and the read speech are concerned. Again large differences exist in the behaviour of the nine words.

From the acoustic data presented in Fig. 3 and Fig. 4 it will become clear that not in all cases an explanation of the behaviour of the listeners can be found in either durational or spectral parameters. Although duration in most cases might account for the high accentuation scores of the CVC's from content words, this does not work so much for the read function words. For instance the scores on "wil" in the example of Fig. 2 are more likely to be explained by the spectral differences.

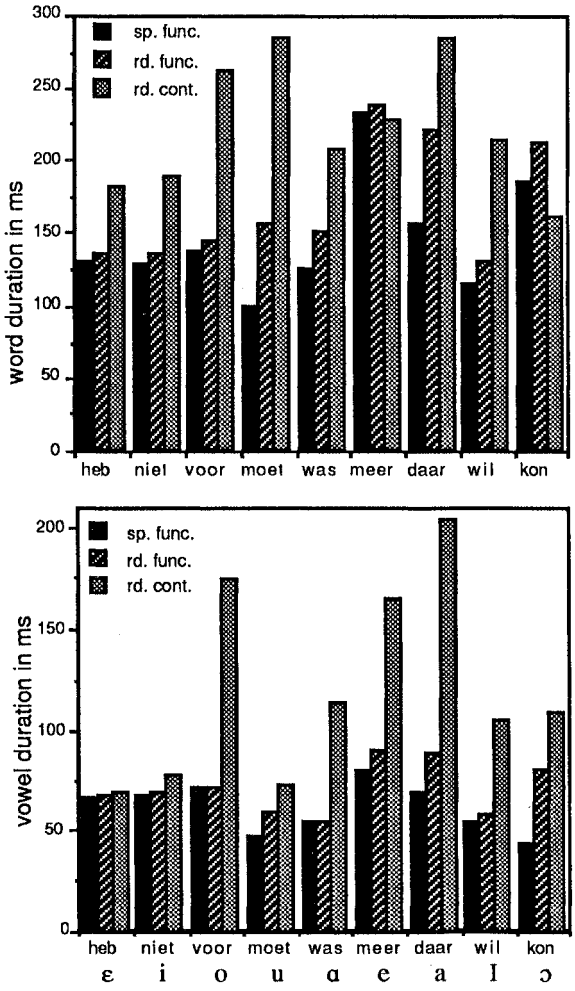

Fig. 3. Word durations (upper graph) and vowel durations (lower graph) in msec for nine function words in three speech conditions.

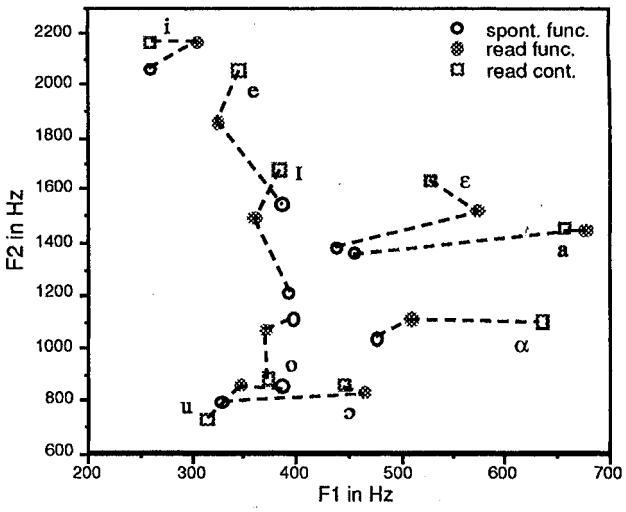

Fig. 4. Vowel formant frequencies (F1 and F2 in Hz) for the nine vowels in function words in three speech conditions.

Systematic manipulation of the stimuli by means of e.g. PSOLA has to provide more insight into the interacting cues. The only cue excluded in this experiment was F0, although most listeners reported to have been guided in their strategy mainly by intonation!

7. CONCLUSIONS

In this study we tried to tackle the question whether listeners actually make use of the acoustic differences between peak and level, when perceiving either read or spontaneous speech. It turned out that 'level' words from a more carefully pronounced speaking style became 'peak' words when spliced into a more sloppily pronounced speaking style. Although the results still have to be analysed into more detail, the outcome so far provides good arguments to conclude that the listener indeed attunes his/her perception mechanism to the speaking style in question.

REFERENCES

[1] Bergem, D.R. van, "The influence of sentence accent, word stress, and word class on the quality of vowels". *Proc. Eurospeech 91*, Genova, Vol. 5, pp. 1455-1458, 1991.

[2] Drullman, R. and Collier, R., "On the combined use of accented and unaccented diphones in speech synthesis". *J. Acoust. Soc. Am.*, Vol. 90, pp. 1766-1775, 1991.

[3] Hart, J. 't, Collier, R. & Cohen, A., *A perceptual study of intonation; an experimental-phonetic approach to speech melody*. Cambridge University Press, 1990.

[4] Koopmans-van Beinum, F.J.,*Vowel contrast reduction. An acoustic and perceptual study of Dutch vowels in various speech conditions*, Ph. D. Diss. (University of Amsterdam), 1980.

[5] Koopmans-van Beinum, F.J., "The role of focus words in natural and in synthetic continuous speech: acoustic aspects". *Speech Communication*. J. Llisterri & D. Poch (Eds.), Special Issue on the Phonetics and Phonology of Speaking Styles, to appear.

[6] Koopmans-van Beinum, F.J. & Bergem, D.R. van, "The role of 'Given' and 'New' in the production and perception of vowel contrasts in read text and in spontaneous speech". In: Tubach, J.P. & Mariani, J.J. (Eds.), *Proc. Eurospeech '89*, Paris, CEP Consultants Ltd, Edinburgh, Vol. 2, pp. 113-116, 1989.

[7] Rietveld, A.C.M. and Koopmans-van Beinum, F.J., "Vowel reduction and stress". *Speech Communication* 6, pp. 217-229, 1987.