A systematic framework for studying the contribution of F0 and duration to the perception of accented words in Hebrew

Noam Amir[1], Bat-Chen Almogi[1] and Hansjörg Mixdorff[2]

[1]Department of Communication Disorders
Tel-Aviv University, Israel
[2]TFH Berlin University of Applied Sciences
noama@post.tau.ac.il

Abstract

In this study we defined a methodology for examining the combined effect of F0 and duration in accenting a syllable, leading to the perception of narrow focus in a short sentence. The F0 and duration variables were manipulated independently in one word in each of three four-word sentences. F0 was manipulated through use of the Fujisaki model of intonation, whereas duration was manipulated using Praat software. Subjects were asked indirectly to determine whether any word was accented. Initial results showed that perception was influenced by changes in F0 more than by changes in duration, with some degree of interplay between the two variables. Identification curves show a categorical perception of accent as a function of F0.

1. Introduction

The prosodic structure of natural discourse is extremely complex. The different strategies that speakers employ in natural conditions, in order to create the impression of focus are varied and multifaceted, and studying them on spontaneous speech creates many methodological difficulties because linguistic, para-linguistic and non-linguistic information is invariably present In the present study we therefore set out to conduct a systematic study of the prosodic cues that lead to the perception of a single word as accented, in the framework of short isolated sentences.

It has been found previously that word focus is achieved by a rise in pitch, duration and intensity of the stressed syllable [2, 4, 5, 6, 8]. This present study follows a previous work by the authors on the *production* of narrow focus in short Hebrew phrases [7] which showed for instance typical lengthening ranging from 40 to 100 ms. Cooper et al. [1] found that English speakers typically raised their pitch by 20-30 Hz in accented syllables, as did Eady and Cooper also in a later study [3]. Production studies can demonstrate typical values for the above parameters, whereas perception studies must deal with a larger range, in order to find thresholds of perception and upper limits above which the speech sounds unnatural.

Attempting a systematic study of this type raises many issues. Deciding on the range of pitch changes, the range of duration changes, and how to create these changes, are issues which can be approached in different ways – from having a speaker utter the same phrases repeatedly, to synthesizing them purely from scratch.

Many software packages available today (e.g. Praat) enable modifying the intonation and duration of recordings freely. Using such software in an effort to change accent is problematic, in the sense that too many degrees of freedom are available. Without any underlying model and an ensuing parametrization, it becomes difficult to make any systematic variations.

Based on the authors' extensive experience with the Fujisaki model of intonation, it was decided to base the present research on this model, using software made freely available by one of the authors. This is further in keeping with the production study mentioned above. The well-known Fujisaki model was therefore employed to decompose a given F0 contour into phrase and accent contributions, the former modeling the slowly falling declination line, and the latter for creating the F0 humps associated with accented syllables.

2. Method

The overall aims of this study were as follows:

- To make some initial observations on the pitch and duration strategies employed by a native Hebrew speaker when asked to accent a word.
- To create a series of modified utterances, in which the strategies found above were implemented as a series of stimuli with manipulated pitch and/or durations, both independently and in all their combinations.
- To carry out listening tests that could give an indication of threshold values for the perception of narrow focus, and reveal whether this perception is truly categorical.

A great deal of care was taken to make the manipulations as naturally sounding as possible, and to create an experimental design which could give as clear an answer as possible to the issues we wished to examine.

2.1. Stimuli

3 four-word sentences were used in this study, recorded in Hebrew by a native Hebrew-speaking male. The sentences were (capitalization indicates lexical stress):

1. "bikarnu xaVER axarei halimudim" (we visited a friend after school)
2. "baxeder hasmaLI yoshev hamenahel" (in the left room sits the manager)
3. "haxatul hakaTAN barax maher" (the little cat escaped quickly)

Our first objective was to obtain some observations on the production strategies employed by the speaker to create the effect of accent. To this end, each sentence was initially uttered by the first author several times in a neutral manner. The same sentences were then uttered with the second word in each sentence moderately accented.

The Fujisaki parameters (base F0, phrase command and a

single accent command) were then fitted manually, in order to observe how the accent was created by the speaker in terms of accent command placements and vowel prolongation. Two typical examples are presented in Figures 1 and 2. Typically, Neutral utterances are characterized by wide and low accent commands, as shown in Figure 1, whereas short sentences with narrow focus are characterized by the presence of a short and high accent command on the accented syllable, as shown in Figure 2.

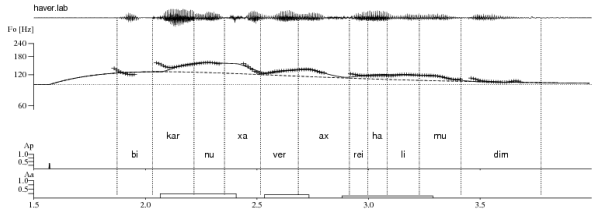

Figure 1: *Fujisaki parameters for phrase 1 with broad focus. From top to bottom: signal, pitch contour (original – '+', modeled – solid line), transcription, phrase command and accent commands.*

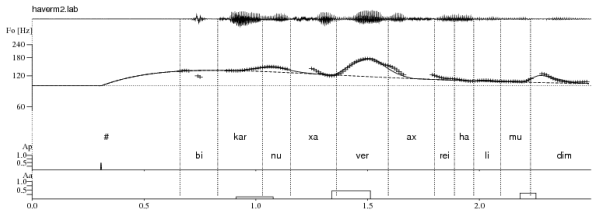

Figure 2: *Fujisaki parameters for phrase 1 with narrow focus. Syllable 'ver' is accented. From top to bottom: signal, pitch contour (original – '+', modeled – solid line), transcription, phrase command and accent commands.*

We found that in all the utterances with accented words, the accent command started at the beginning of the opening consonant of the accented syllable, and ended either in the middle of the vowel in the accented syllable (phrases 1 and 2) or at the end of this vowel (phrase 3).

The neutral sentences, along with their associated phrase command, were used as a baseline for further manipulations. Theoretically, the narrow-focus utterances could be used, with the accent command removed, but in these utterances the accented syllable already contained a measure of lengthening.

These neutral utterances were then manipulated as follows: an accent command, having a width matching the accent command taken from the narrow-focus utterance, was inserted, to raise the pitch of the accented syllable. The amplitude of the accent command was determined so that the pitch was raised in a series of steps that were determined in a pilot study. These steps were 0, 15, 30, 45 and 60 Hz above their values with no accent command. Next, the accented syllable was lengthened, using the manipulation editor of Praat, also in a series of steps, by 0, 25, 50, 75, 100 and 125 ms. For each lengthening, the above pitch steps were repeated. This gave us, for each sentence, 30 variants, with all the possible combinations of pitch rise and syllable lengthening. The procedure also ensured that the above variations were carried out as systematically as possible, without any need for additional decisions as to the precise form of the pitch contour, since this was ensured by the Fujisaki editor. Further interactions between word placement in the phrase and perception of focus were avoided by manipulating only the second word in each phrase.

All the above manipulations were carried out on the same word in each phrase. To avoid a situation where listeners would become aware of this, 6 additional stimuli were created for each phrase. In these stimuli the three *other* words, in turn, were in narrow focus, each one in two variants. However, the results for these stimuli were not taken into account in the statistical analyses.

2.2. Subjects

21 subjects participated in this study, aged 18 to 30 (mean: 25.3), 14 females and 7 males. All participants were native Hebrew-speakers. Prior to the experiment, the subjects underwent screening for hearing difficulties, separately for both ears. All had thresholds below 20 dB for 500, 1000 and 2000 Hz, except 3 subjects which had PTA levels of 25 dB at one frequency in one ear.

2.3. Procedure

A Matlab Graphic User Interface (GUI) was programmed to run the experiment. For each stimulus, the subjects had to answer two questions: "What is the additional information in the sentence?" and "Did the utterance sound natural?"

The answer to the second question was simply yes or no, whereas the answer to the first question was given indirectly, through a set of multiple choices, adapted to each of the phrases. For instance, in response to the phrase "the little cat escaped quickly", the subjects could choose one of four responses:

1. no additional information was present
2. the cat, not a dog, escaped quickly
3. the little cat, not a big cat, escaped quickly
4. the little cat escaped quickly, not slowly

These responses represent all the possible forms of changing the focus word in the phrase. When the presentation was perceived as neutral, the expected response would be (1), when the focus word was "cat", the expected response would be (2), and so on.

It is accepted procedure to present each stimuli several times in this type of paradigm. In this study we presented each of the 36 stimuli per phrase 5 times, giving 180 presentations per phrase. To avoid listener fatigue we therefore limited each listener to two phrases out of the three. Thus each phrase was heard by 14 out of the 21 subjects. All stimuli for a single phrase were presented before moving on to the next phrase, and presentations were in random order.

3. Results

Overall rates for recognition of focus were tallied for each phrase, for each combination of pitch and duration manipulation. Results for the three phrases appear in tables 1-3.

A complete graphical and statistical analysis is beyond the scope of this paper, however some representative graphs will be brought here to exemplify some of the results.

Table 1: *Accent recognition rates for phrase 1, mean and STD.*

Pitch [Hz] / Dur. [ms]	0	15	30	45	60
0	2.9 (7.3)	12.9 (21.6)	40 (30.4)	70 (30.1)	71.4 (31.1)
25	7.1 (16.8)	22.9 (31.2)	52.9 (37.3)	65.7 (29.8)	74.3 (38.8)
50	7.1 (14.9)	17.1 (29.2)	55.7 (35.2)	60 (34.2)	75.7 (32.5)
75	4.3 (16.0)	17.1 (28.1)	57.1 (37.5)	74.3 (33.7)	75.7 (28.5)
100	5.7 (9.4)	18.6 (31.8)	60 (33.3)	77.1 (34.1)	82.9 (28.1)
125	14.3 (22.8)	22.9 (33.1)	52.9 (36.5)	78.6 (33.7)	87.1 (30.0)

Table 2: *Accent recognition rates for phrase 2, mean and STD.*

Pitch [Hz] / Dur. [ms]	0	15	30	45	60
0	10 (18.8)	25.7 (31.8)	57.1 (36.7)	84.3 (27.4)	91.4 (10.3)
25	7.1 (16.8)	32.9 (27.9)	82.9 (23.3)	91.4 (17.0)	94.3 (12.2)
50	17.1 (24.6)	35.7 (23.8)	85.7 (14.5)	91.4 (12.9)	97.1 (7.26)
75	15.7 (16.0)	52.9 (32.0)	82.9 (15.4)	91.4 (17.0)	95.7 (8.5)
100	25.7 (24.1)	51.4 (30.1)	90 (10.4)	94.3 (12.2)	97.1 (7.3)
125	41.4 (25.4)	75.7 (19.5)	94.3 (9.4)	94.3 (12.2)	92.9 (9.9)

Table 3: *Accent recognition rates for phrase 3, mean and STD.*

Pitch [Hz] / Dur. [ms]	0	15	30	45	60
0	14.3 (14.5)	22.9 (20.5)	41.4 (26.6)	74.3 (31.8)	77.1 (22.0)
25	17.1 (23.3)	25.7 (19.9)	47.1 (33.9)	70 (33.0)	87.1 (27.9)
50	25.7 (30.8)	41.4 (29.8)	60 (27.2)	82.9 (23.3)	82.9 (24.6)
75	44.3 (26.2)	52.9 (36.5)	68.6 (28.0)	82.9 (23.3)	88.6 (17.0)
100	44.3 (32.5)	67.1 (30.0)	68.6 (28.0)	84.3 (26.2)	90 (15.2)
125	52.6 (25.5)	60 (27.1)	81.4 (16.6)	64.3 (28.5)	90 (15.2)

Figures 3-6 show recognition scores for phrase 1,as a function of one variable only, with the other held fixed. Observing Figures 3 and 4, where pitch shift is held constant (0 in Figure 3 and 60 in Figure 4), it is apparent that for this phrase, duration of the accented syllable has a much smaller effect on perception of focus than pitch.

Figures 5 and 6, where lengthening is held constant (0 in Figure 5 and 125ms in Figure 6), we observe typical categorical perception as a function of maximal pitch in the accented syllable. Though this is more marked when it is present in conjunction with syllable lengthening, the crossover point is between 30 and 40 Hz in both cases. Similar observations were made for the other phrases, though perception in phrase 2 was more affected by duration. It is interesting to observe that all phrases were not affected identically.

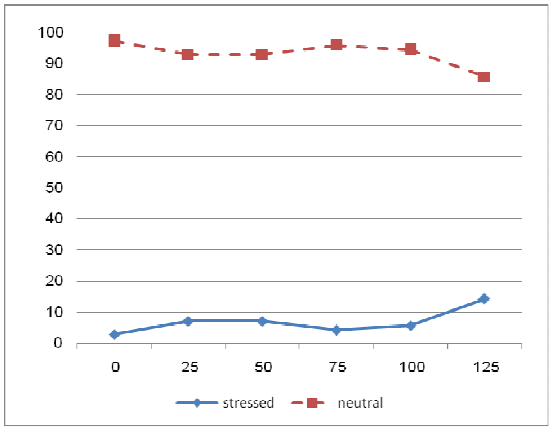

Figure 3: *recognition rates as a function of lengthening with zero F0 shift. Even when lengthening is extreme, very few instances are classified as narrow focus.*

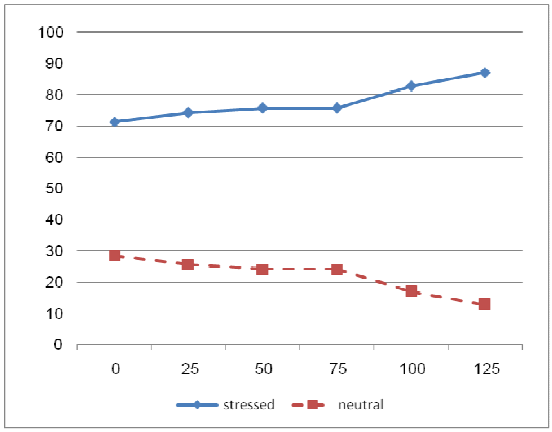

Figure 4: *recognition rates as a function of lengthening with 60Hz F0 shift. Even when no lengthening is present, over 70% of the utterances were recognized as narrow focus.*

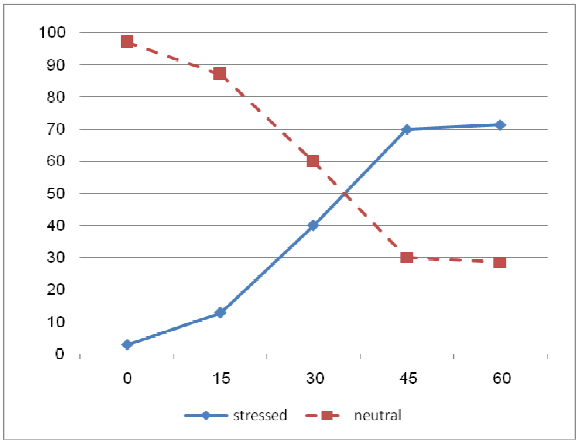

Figure 5: *recognition rates as a function of F0 shift with zero lengthening. The graphs indicate categorical perception.*

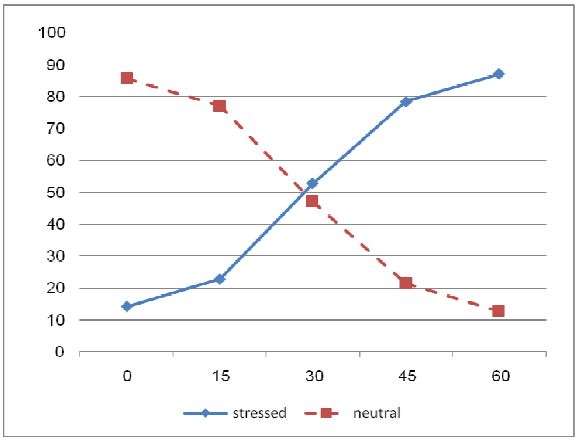

Figure 6: *recognition rates as a function of F0 shift with 125ms lengthening. The graphs indicate categorical perception.*

Further statistical analyses are necessary to determine whether the perception of accent is truly categorical, and to what extent this is a function of F0 shift and lengthening in combination. It is also interesting to analyze the relationship between the perception of naturalness and accent. These are presently being carried out.

4. Discussion and Conclusions

We have shown that combining the use of Praat for syllable lengthening, and the use of the Fujisaki model of intonation for raising the pitch on a selected syllable, it is possible to attain a systematic manipulation of these two parameters in order to examine their effect on the perception of syllable accentuation, leading to narrow focus on a selected word.

Listening tests based on these manipulations give results that are complicated to analyze, both in terms of perception of accent and perception of naturalness. Some initial conclusions can be drawn from the results presented here. First, though a pilot study clearly indicated that lengthening above 125ms sounded unnatural, lengthening by shorter amounts is rarely enough to create accentuation, as demonstrated in Figure 3. On the other hand, when F0 was raised by 60Hz, lengthening contributed only mildly to accentuation. It is clear that for the phrase analyzed here, raising F0 has a much larger influence than lengthening, with a threshold value in the vicinity of 30Hz.

Further statistical analysis is called for to verify whether the perception of accent as a function of pitch raise is actually categorical, in order to compare the results across the different phrases, and in order to analyze them also in terms of naturalness. The suggested methodology, however, is an interesting complement to production studies, and provides interesting data that can be used in speech synthesis and as a baseline for examining the perception of pathological populations.

5. References

[1] Cooper, W. E., Eady, S. J. and Mueller P. R. (1985) Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America,* 77 (6), 2142-2156.

[2] de Jong, K. and Zawaydeh, B. (2002) Comparing stress, lexical focus and segmental focus: Patterns of variation in Arabic vowel duration. *Journal of Phonetics* 30, 53-75.

[3] Eady, S. J. and Cooper, W. E. (1986) Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America,* 80 (2), 402-414.

[4] Fry, D. B. (1955) Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27 (4), 765-770.

[5] Fry, D. B. (1958) Experiments in the perception of stress. *Language and Speech,* 1, 126-152.

[6] Kochanski, G., Grabe, E., Coleman, J. and Rosner, B. (2005) Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America* 118 (2), 1038-1054.

[7] Mixdorff, H., Amir, N. (2002) The prosody of modern Hebrew – a quantitative study, *Proceedings of Speech Prosody 2002*, Aix en Provence.,

[8] Tamburini, F. and Caini, C. (2005) An automatic system for detecting prosodic prominence in American English continuous speech. *International Journal of Speech Technology*, 8, 33-44.