Acoustic Correlates of Perceived Syllable Prominence in German

Hansjörg Mixdorff [1], *Christian Cossio-Mercado*[2], *Angelika Hönemann*[3], *Jorge Gurlekian*[2], *Diego Evin*[2], *Humberto Torres*[2]

[1] Department of Computer Science and Media, Beuth University Berlin, Germany
[2] LIS, University of Buenos Aires, Argentina
[3] University of Bielefeld, CITEC, Germany

mixdorff@beuth-hochschule.de, ccossio@dc.uba.ar, ahoenemann@techfak.uni-bielefeld.de, anagraf99@hotmail.com, diegoevin@gmail.com, hmtorres@hotmail.com

10.21437/Interspeech.2015-11

Abstract

This paper explores the relationship between perceived syllable prominence and the acoustic properties of a speech utterance. It is aimed at establishing a link between the linguistic meaning of an utterance in terms of sentence modality and focus and its underlying prosodic features. Applications of such knowledge can be found in computer-based pronunciation training as well as general automatic speech recognition and understanding. Our acoustic analysis confirms earlier results in that focus and sentence mode modify the fundamental frequency contour, syllabic durations and intensity. However, we could not find consistent differences between utterances produced with non-contrastive and contrastive focus, respectively. Only one third of utterances with broad focus were identified as such. Ratings of syllable prominence are strongly correlated with the amplitude *Aa* of underlying accent commands, syllable duration, maximum intensity and mean harmonics-to-noise ratio.

Index Terms: prominence, perception, automatic speech recognition, Fujisaki model, *F0*

1. Introduction

The information structure of an utterance is reflected in the relative saliency of its lexical constituents. At the acoustic level we observe that accented syllables serve as anchoring points of this structure. They are emphasized or toned down by acoustic means. The perceptual correlate of this process is the so-called prominence. Various segmental and supra-segmental factors have been shown to affect prominence, cf. [1][1][3], such as *F0* excursions, segment durations, intensity as well as vowel type and syllable coda structure. In an earlier study [9], the first author and his co-worker investigated the relationship between perceived syllable prominence and the *F0* contour in terms of the parameters of the Fujisaki model [1]. The model was used to parameterize a subcorpus of the Bonn Prosodic Database [1]. Analysis showed that prominences labeled on a scale from 0-31 strongly correlated with the interval of *F0* movements, but only when it was anchored to accented syllables. This indicates that the prominence judgment is partly guided by linguistic considerations. Evidence in support of this assumption has been presented for many languages, including German which is the language of the present study.

Despite efforts to integrate prosodic knowledge in the process of automatic speech recognition [7][8], state-of-the-art technology still makes little or no use of prosodic information. Our work is intended to explore to what extent linguistic information such as focus and sentence modality can be retrieved from the prosodic features of an utterance. Since a direct link between the acoustics and the meaning of an utterance seems difficult to establish, we decided to first derive syllable and word prominences from the acoustic signal and then relate these prominence ratings to the focal and sentence mode conditions. Part of this effort is a perceptual evaluation regarding the ability of humans to retrieve the intended focal condition from isolated utterances. Only those acoustic differences which are perceptually salient could also be exploited by a speech understanding system. This is a joint German-Argentine work in which we aim to apply the same approach to German and Argentine Spanish.

In the current paper we only present results for the German data. We first perform acoustic analysis of single-phrase utterances produced with varying sentence mode and focus and examine how prosodic features such as *F0*, duration, intensity and voice quality are affected by the underlying linguistic information. Then we link these results to the outcomes of a perceptual experiment in which subjects were asked to determine the sentence mode and focus of the same utterances as well as rate the prominence of each syllable.

2. Stimuli Recording and Perception Experiment Design

We employed six short sentences created by Andreeva et al. [9][1] for their studies on focus distinctions in German which are part of a multilingual corpus:

Der Mann fuhr den Wagen vor.	*The man drove the car up.*
Das Mädchen soll ein Bild malen.	*The girl must draw a picture.*
Der Peter kann den Film gucken.	*Peter can watch the movie.*
Das Kind sollte im Bett sein.	*The kid should be in bed.*
Das Bild soll nicht hässlich sein.	*The picture mustn't be ugly.*
Mein Vater kann Türkisch lesen.	*My father can read Turkish.*

The potential locations of narrow focus are the two critical words underlined in the text. Andreeva provided speech data from six German native subjects, all produced in declarative mode. At Beuth University we recorded ten additional subjects, also adding sentences in interrogative mode.

Copyright © 2015 ISCA

The declarative conditions include:

1) Broad focus
2) Narrow focus on first critical word (non-contrastive)
3) Narrow focus on second critical word (non-contrastive)
4) Narrow focus on first critical word (contrastive)
5) Narrow focus on second critical word (contrastive)

Subjects were presented with the sentences on a laptop screen preceded by a question either asking for the whole sentence, e.g. "What did you say?" to elicit broad focus, or part of it, e.g. "Who drove the car up?" Contrastive focus was elicited using echo-questions presupposing a fact contradicting the one conveyed by the sentence: "The woman drove the car up?" Questions were elicited by having subjects listen to recordings of narrow focus statements whose content the subjects were supposed to question using echo-questions of the same wording. Each item was uttered three times by each subject. Recordings were performed in a lightly sound-treated room using a close-talk microphone. Hence we had recordings by a total of 16 subjects. These were cut from the session audio and checked auditorily for appropriate focus type and sentence mode, that is, we controlled whether the audio contained a plausible reaction, given that the intended condition was known to us. For the ensuing perception experiment we chose the last version of each kind, yielding a total of 528 utterances.
In the perception test we intended to examine the following research questions:

1) Are subjects able to identify the intended sentence modality?
2) Can subjects identify the intended focal condition?
3) How do subjects rate the prominence of the syllables associated with the utterance?

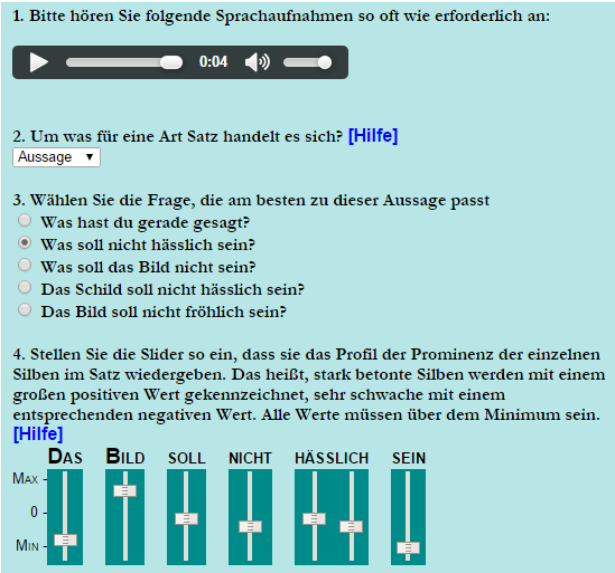

Figure 1: *GUI of perception experiment. Subjects are prompted to play back each stimulus, and then decide whether they perceived a statement or question. In the case of statements, subjects need to choose the most appropriate question. Finally, they rate each syllable prominence using the sliders provided.*

Figure 1 shows the GUI of the perception test. It was performed online and hosted on a server at LIS Buenos Aires. The experiment was preceded by a verbal explanation of the task. Then six examples were presented which had already been solved. Due to the large number of stimuli the task was rather taxing. Therefore we recommended that participants should only rate a maximum number of 100 stimuli in one session. A total number of 14 (seven male, seven female) native listeners of German took part. Subjects, most of them students of Beuth University Berlin, took up to 3.5 h hours to complete the whole task, not counting pauses. They were paid for their time. Using an array of sliders for rating prominence was inspired by works of Anders Eriksson et al. [10] who employed a similar paradigm in their prominence rating experiments.

3. Acoustic Analysis of Stimuli

As mentioned above, for the additional ten speakers, target utterances were cut out of the contexts and checked for the intended meaning. Recordings were down-sampled to 16 kHz and force-aligned with the WEVOSYS LINGWAVES aligner on the syllable level [11]. Automatic syllable segmentations were checked and corrected manually in the *PRAAT* TextGrid Editor [12]. We subsequently calculated syllable durations based on the segmentations.
F0 values were extracted at intervals of 10 ms and contours checked and if necessary corrected in the *PRAAT* PitchEditor. All *F0* contours were then subjected to Fujisaki model parameter extraction [13] with *alpha* of 3 and *beta* of 20. The base frequency *Fb* was determined automatically. Although we usually treat *Fb* as a speaker-dependent constant, applications of speech recognition need to operate on individual utterances without prior knowledge of the speaker.

Results were checked and if necessary corrected in the *FujiParaEditor* [14]. In this way, we obtained a smooth, interpolated model *F0* with the accent command amplitudes *Aa* as a measure of the underlying *F0* gesture magnitude. The alignment of accent commands with syllables was performed based on linguistic information about content words and their lexically stressed syllables from a TTS front-end. In addition, the final syllable was scanned for high boundary tones also associated with accent commands.

Intensity contours were extracted in *PRAAT* with default settings, and mean intensities in dB, as well as maxima employing parabolic interpolation were determined for each phone. Syllabic mean harmonics-to-noise levels were also calculated within *PRAAT* applying default settings.

4. Results of Acoustic Analysis

Figure 2 displays examples of Fujisaki model-based analysis of the stimulus utterance "Der Mann fuhr den Wagen vor" produced by male speaker SP11 for the focus conditions (1) broad, (2) narrow early, (3) narrow late, (4) narrow early-question and (5) narrow late-question. All narrow foci are non-contrastive. Each of the five panels displays, from the top to the bottom: the speech waveform, the F0 contour (extracted and modeled), and the underlying phrase and accent commands. The syllable segmentation is indicated by the dotted vertical lines. Syllable texts are provided in German SAMPA transcription. As can be seen, F0 contours differ clearly for all five conditions, and the narrowly focused words are associated with accent command of high amplitude. Questions are marked by high levels of F0 after the focused item and a trailing high boundary tone at the end of the utterance.

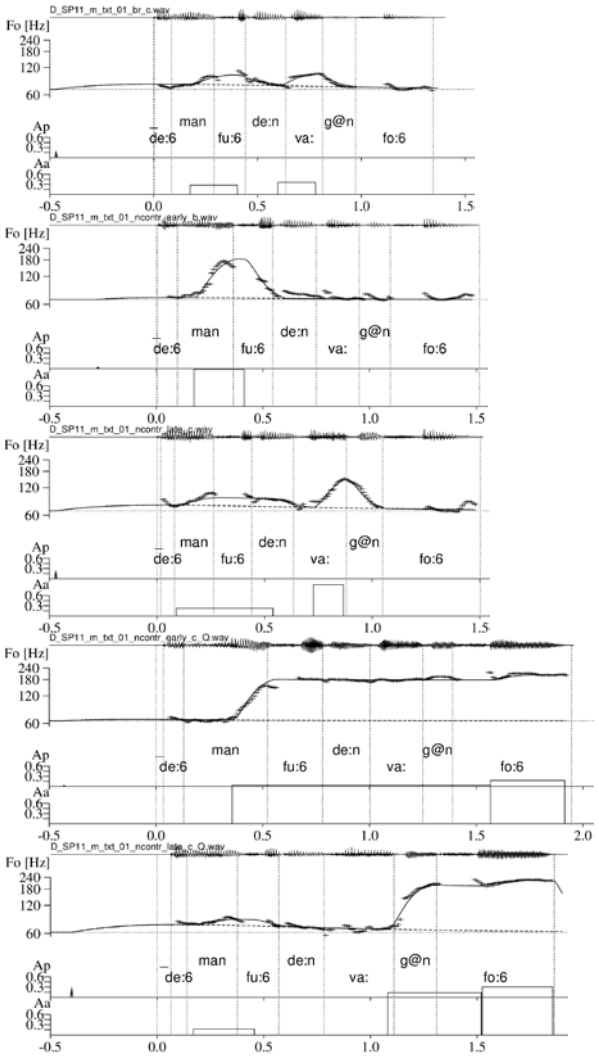

Figure 2: *Results of analysis for male speaker SP11. From the top to the bottom: statements of broad focus, early focus on 'Mann', late focus on 'Wagen', questions of early focus on 'Mann', late focus on 'Wagen'.*

Table 1: *Means and s.d. of accent command amplitude Aa for the lexically stressed syllable of the first and second critical word depending on the focus condition, statements.*

word	focus		mean	s.d.	N
first	broad		.35	.13	83
	early	non-contr.	.51	.19	83
		contrastive	.59	22	90
	late	non-contr.	.20	.14	81
		contrastive	.21	.15	83
second	broad		.36	.22	83
	early	non-contr.	.08	.08	83
		contrastive	.06	.07	90
	late	non-contr.	.50	.22	81
		contrastive	.55	.17	. 83

Table 1 shows values of accent command amplitude *Aa* for the five different focus conditions, averaged over all lexically stressed syllables in the two critical words that received narrow focus in our data, that is, "Mann" (first) and "Wagen" (second) in the aforementioned sentence. As can be seen, the effect of focus location is clearly visible, however, with little difference between contrastively and non-contrastively focused items except for a slight increase in Aa on the narrowly focused item.

Whereas the distinction regarding focus width is significant (Kruskal-Wallis test of independent samples, $p < .001$) for both of the critical words, it is not for the contrast distinction (Mann-Withney-U test for the first word with $p < 0.09$ and $p < 0.94$ for the second word, respectivly). Likewise, syllable durations and max intensity in the aforementioned syllables are significantly affected by the focus width (Kruskal-Wallis test of independent samples, $p < .001$). In contrast, the mean harmonics-to-noise ratio of the syllables does not seem to be affected by the focus location (Kruskal-Wallis test of independent samples, $p < .096$).

Table 2: *Means and s.d. of accent command amplitude Aa for the lexically stressed syllable of the first and second critical word, as well as the pre-boundary syllable depending on the focus condition, questions.*

word	focus	mean	s.d.	N
first	early	.68	.27	54
	late	.24	.15	54
second	early	.67	.32	54
	late	.66	.34	54
boundary	early	.91	.27	54
	late	.89	.36	54

As could be seen in Figure 2, *F0* patterns of questions can be quite distinct as *F0* rises to a high plateau in the narrowly accented syllable, only to be further boosted by a question-final rise on the ultimate syllable of the utterance. This pattern also shows up in the *Aa* means presented in Table 2. It becomes clear that the second critical word is always produced at a high *F0* regardless of focus. Therefore *Aa* cannot serve as a cue (Mann-Whitney-U Test, $p < .521$), but the onset time of *T1* of the underlying accent command relative to the onset of the syllable, as well as duration and, interestingly, mean harmonics-to-noise ratio can (Mann-Whitney-U Test, $p < .001$ for all of these parameters). The effect on HNR can be explained by the sometimes rasping voice quality marking the questioned item in (incredulous) echo-questions. This is confirmed by a drop in mean HNR by almost 4dB when the critical word is in focus.

5. Focus Identification and Prominence Ratings

As expected, questions were correctly identified in 94% of judgments. We examined the reliability of our subjects at determining the focus condition of statement utterances. The results are given in Table 3.

Table 3: *Proportions of correct identification of the focus condition, statements.*

focus intended	broad	non-contr. early	non-contr. late	contr. early	contr. late
broad	.34	.19	.37	.01	.02
contr early	.03	.63	.02	.24	.00
contr late	.12	.06	.62	.01	.11
non-contr early	.03	.66	.04	.20	.00
n.contr late	.14	.06	.60	.01	.09

As explained, subjects were asked to select the most suitable question for the statement utterances they listened to and we hoped to identify the perceived focus from their choices. As can be seen, proportions for contrastive and non-contrastive foci were rated almost identical with more than 60% being classified as non-contrastive. In almost all of these cases the

focus location was identified correctly. This result seems plausible given the results of acoustical analysis presented above. However, only one third of broadly focused utterances were classified as belonging to this category, even more were taken to be non-contrastively focused on the second critical word. Still 19% were classified as non-contrastively focused on the first critical word.

Now we turn to the perceptual prominence ratings of our evaluators. The slider values were mapped onto an integer scale from -5 to +5. In the scope of this paper we do not normalize the subjects' responses, but simply pool the responses for each syllable in every stimulus utterance and relate them to the underlying acoustic features of the stimulus.

Table 4: *Correlations between perceptual prominence and selected prosodic features of syllables, as well as correlations between these features. Pearson's r, two-sided significance value and number of instances.*

feature		*Aa*	max. intensity	duration	mean HNR
perc. prom.	r	.738**	.458**	.317**	.123**
	sign.	0.000	.000	.000	.000
	N	3151	3151	3151	3141
Aa	r		.418**	.369**	.088**
	sign.		.000	.000	.000
	N		3151	3151	3141
max intensity	R			.021	.044*
	sign.			.237	.014
	N			3151	3141
duration	r				.168**
	sign.				.000
	N				3141

*Significance levels: ** p<.01, * p < .05.*

As can be seen, the prominence values are most strongly correlated with the interval of *F0* transitions (tone switches) which are associated with the syllable and expressed by the accent command amplitude *Aa*. Other correlations are moderate to weak, but still significant. The fact that in this analysis mean syllable HNR shows a weak correlation corresponds to earlier finding that syllable sonority [15] among other factors influences perceived prominence. There is also a small negative correlation between perceived prominence and the index of the syllable concerned (Pearson's r=-.137, p < .01), suggesting that syllables later in the utterance are assessed as weaker.

A regression model based on the aforementioned factors for predicting prominence is able to explain 76.1% of the variance (see Table 5). When we add the base frequency *Fb* to this model, the explained variance rises to 76.4%. This suggests that higher pitched voices also yield higher prominence ratings. Introducing the index of the syllable in the phrase as a factor, however, does not yield any improvement, as the index itself is negatively correlated with *Aa* and max. intensity (Pearson's r of -.109 and -.501, respectively, p < .001). When we apply the same model to all utterances and include the echo-questions explained variance drops to 72%. The influence of mean HNR becomes non-significant. This might be to do with the special connotation of incredulity of our echo-questions which decreases HNR in the focused item as we have seen in the acoustic analysis. There is also a practical issue here regarding the slider-based assessment of prominence. As we have seen, the influence of *F0* on prominence is considerable. In utterances of statements the prominence profile set up with the sliders resembles a stylized *F0* contour. This paradigm does not work as well with the echo-questions and their high *F0* plateaus, as post-focal items are less prominent and at the same time exhibit extremely high pitch.

6. Discussion and Conclusions

We presented results from a production and perception study aiming at determining the effect of focus and sentence mode on several prosodic parameters, as well as the connection between perceived focus and syllable prominence with these parameters. The ultimate aim is to determine this kind of linguistic information in automatic speech recognition and enrich the word hypothesis.

Table 5: *Regression model for predicting perceptual syllable prominence based on the factors listed in the left column.*

factor	coefficients		t	sign.
	B	standard error		
(constant)	-7.221	.397	-18.187	.000
Aa	4.950	.110	45.062	.000
duration	1.420	.242	5.873	.000
max.intensity	.070	.005	14.601	.000
mean HNR	.014	.004	3.963	.000

Our acoustic results are in line with earlier studies with respect to effects on *F0*, duration and intensity, that is, *F0* range expansion, increased duration and intensity in focused items and the reverse for the de-focused ones.

They confirm that contrastive and non-contrastive foci are not separated by acoustic features. Broadly focused utterances are clearly distinguished from those with narrow or late focus on one of the two critical words. However, this result was not matches by our perceptual outcomes. Utterances with intended broad focus were only identified in about one third of cases. This can possibly be explained by our approach of having subjects determine the focus indirectly by choosing the most appropriate question. The broad focus question was generic whereas the others related to one of the critical words, either by asking for it, or contrasting it with conflicting information. Subjects were possibly drawn to the questions that appeared "more specific". This idea is supported by the observation that errors pulling the decision towards "broad" in the non-broad cases were relatively few and almost exclusively occur in "late" cases (compare Table 3). This latter result points to the fact that by rule [16] broad (or default) focus always implies a prominence on the last accentable item, in our case the second critical word, making the choice "late" a plausible one, whereas narrow focus on the first word is much more marked acoustically, as we saw in Figure 2. We will have to examine to what extent the actual acoustic properties of particular broad focus utterances influenced the listeners' decisions. We also found that prominence ratings can be fairly well predicted based on the prosodic properties of the underlying syllables. Although our perception results somewhat question the ability of human listeners to reliably detect focus, if ASR were able to enhance the word string with prominence ratings, this would already be a step forward. In future work we will compare our results with those from Argentine Spanish and aim to implement prominence detection in an ASR system.

7. Acknowledgements

This work was funded through German DLR research grant 01DN13007 ARG. Thanks go to Bistra Andreeva for supplying the data of six speakers and valuable discussion.

8. References

[1] Fry, D.B., “Experiments in the perception of stress”, Language and Speech 1, 126-152, 1958.

[2] Gay, T., “Physiological and acoustic correlates of perceived stress”, Language and Speech 21, 347-353, 1978.

[3] Koreman, J., Van Dommelen, W., Sikveland, R., Andreeva, B., Barry, W.J., ”Cross-language differences in the production of phrasal prominence in Norwegian and German”, in M. Vainio, R. Aulanko, O. Aaltonen (eds.): Nordic Prosody X. Frankfurt: Peter Lang, 139-150, 2009.

[4] Mixdorff, H., Widera, C., “Perceived prominence in terms of a linguistically motivated quantitative intonation model”, Proc. Eurospeech 2001, Aalborg, Denmark, 403-406, 2001.

[5] Fujisaki, H., Hirose, K., “Analysis of voice fundamental frequency contours for declarative sentences of Japanese”, Journal of the Acoustical Society of Japan 5, 233-241, 1984.

[6] Heuft, B., “Eine prominenzbasierte Methode zur Prosodieanalyse und –synthese”, in W. Hess and W. Lenders (eds.): Computer Studies in Language and Speech, Vol. 2, Peter Lang, Frankfurt am Main, 1999.

[7] Nöth, E., Batliner, A. et al.., “The Use of Prosody in the Linguistic Components of a Speech Understanding System”, in IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 8, NO. 5, SEPTEMBER 2000, p. 519-532.

[8] Hasegawa-Johnson, M., Borys, S. and Chen, K., “Experiments in Landmark-Based Speech Recognition”, Sound to Sense: Workshop in Honor of Kenneth N. Stevens, June, 2004 (NSF 0132900)

[9] Andreeva, B., Barry, W. and Koreman, J., “A Cross-language Corpus for Studying the Phonetics and Phonology of Prominence”, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA), 326-330.

[10] Eriksson, A., Gunilla Thunberg, G. Traunmüller, H., “Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing”, Proc. Eurospeech 2001, Aalborg, Denmark.

[11] http://www.wevosys.com/products/lingwaves/lingwaves.html

[12] Boersma, P., “Praat, a system for doing phonetics by computer”, Glot International 5, 341-345, 2001.

[13] Mixdorff H., “A novel approach to the fully automatic extraction of Fujisaki model parameters”, Proc. ICASSP 2000, vol. 3, 1281-1284, Istanbul Turkey, 2000.

[14] Mixdorff, H., “FujiParaEditor”, http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html, 2009

[15] Mixdorff, H., Niebuhr, O., “The influence of F0 contour continuity on prominence perception”, Proc. Interspeech 2013, Lyon, France, 230-234, 2013

[16] Stock E., Zacharias, C., “Deutsche Satzintonation“, VEB Verlag Enzyklopädie, Leipzig, 1982.