Great Expectations – Introspective vs. Perceptual Prominence Ratings and their Acoustic Correlates

Petra Wagner

Institut für Kommunikationsforschung und Phonetik
Universität Bonn, Germany
pwa@ikp.uni-bonn.de

Abstract

In order to gain knowledge about the interaction between top-down expectations of listeners concerning prosodic prominence and its acoustic correlates, two exploratory empirical studies were carried out. First, native and non-native subjects rated prominences of speech read at normal and very fast —prosodically very different — speech. Later, these ratings were compared with introspective prominence ratings of different listeners. First results indicate a major influence of the introspection on prominence ratings, especially if acoustic cues are difficult to interpret, as it is the case in very fast speech. Compared to native subjects, non-natives rely less on their introspection and more on the acoustics.

1. Acoustic and Phonological Prominence Prediction and Perception

Algorithms of prominence prediction that are based on phonological and morpho-syntactic information perform quite well. Their evaluation on the basis of perceptual prominence ratings shows good results (e.g. [1]).

Also, the acoustic correlates of perceptual prominence are quite well-known – F0, duration and spectral tilt[1] have been claimed to be the most important cues for a number of languages, even though the relative proportion of each parameter may be language specific.

It is also well-known that the listeners' psychoacoustic sensitivity for variations in duration, F0 or intensity differs dramatically between speech and non-speech stimuli [2]. One of the reasons for this is that prominence and its verified acoustic correlates do not stand in a 1:1 relationship. Instead, there is a complex interaction between acoustic prosody and listeners' top down expectancies concerning lexical or sentence stress. The latter is based on their linguistic competence, the former based on their perception.

[3] detected high correlations between perceptual prominence ratings and well-known acoustic correlates. But they claim that:

"The results show that subjects can use vocal effort, the distinctness of F0-movements, and vowel duration as cues for rating syllable prominence. However, we can not tell which cues they actually used. A strategy based mainly on top-down processing could have produced a similar result".

It is still an unresolved question, to what extent perceived prominence is independent of top-down hypotheses based on the speaker's native language competence. In other words, we do not know whether a lexical stress is perceived because it has been produced or because the speaker expects it to be produced. This lack of knowledge may be an explanation, why it has been difficult finding clear evidence for well-established phonological concepts such as stress shift [4]. Also, it has been shown that the native language makes speakers more or less sensitive to the acoustic cues of prominence [5]. E.g., native speakers of a language with fixed lexical stress may have problems detecting it in a language with free lexical stress [6]. The strong influence of top-down expectations concerning prominence perception receives further support by research from speech synthesis. Here, it was often found that prominence prediction based on linguistic knowledge performs better than an acoustic prediction [7, 8]. The strong influence of top-down knowledge would also explain the high inter speaker and inter-listener correlation of prominence patterns that was sometimes found (e.g. [9]).

It would be of course nonsense to assume an independence of prominence perception and acoustic correlates. If this were the case listeners would be unable to detect deviances of the "expected" prominence pattern, as it happens frequently in contrastive utterances, corrections or faulty and - consequently - often hard to understand L2-productions. Also, without a clear correspondence between acoustic prominence and phonological words, listeners would not have a chance to train their top-down expectancies concerning prominence during the process of language acquisition.

2. Research Questions and Hypotheses

Prominence perception cannot be explained on the basis of either acoustic prosody or top-down expectancies alone. Thus, there must be a trade-off between both factors influencing prominence perception. It is likely, that listeners possess a certain tolerance concerning speaker specific or speaking style specific variations. For example, a native *listener may extrapolate* certain expected pitch excursions even if the pitch contour he listens to remains rather flat. However, *strong deviations* from an expected acoustic correlate *must be perceptible* since they may have an important communicative function or signal. An example for

[1] Spectral tilt or spectral slope is commonly regarded as the relative intensity of high versus low frequencies. No standardized measurement of spectral tilt has been established yet.

the former would be an utterance spoken at a high articulation rate, an example for the latter would be a correction.

As a first approach to this problem, two preliminary working hypotheses were formulated:

2.1. Hypothesis 1

Given two utterances of identical lexical content but *differing strongly* in the realisation of the prominence lending acoustic parameters, they ought to be perceived differently in terms of prominence.

2.2. Hypothesis 2

A non-native listener is more sensitive for a shift of the prominence lending acoustic correlates because s/he is less influenced by top-down expectancies.

3. Perception Study

It was tested whether native and non-native prominence perceptions differ and how they relate to acoustic correlates of prominence patterns.

3.1. Experimental Setup

The stimuli were two German sentences "Am nächsten Tag fuhr ich nach Husum" and "Es ist eine Fahrt ans Ende der Welt," each read in two versions by a female native speaker:

1. normal reading speed
2. the fastest reading speech possible according to the specifications stated in [10]

It was verified that the prosodic realization of the sentences read at normal speed matches the typical expectancies of a phonetician trained in prosodic annotation. That way, top-down expectations and acoustic realizations are matched as close as possible. The fast versions, however, have an almost completely flat F0-contour without any pitch accents. The durational variation minimal and seems to be mostly the product of articulatory constraints.

At first, the subjects listened to the fast versions in order to avoid any influence from the "standard prosody" in the sentence read at normal speed. They were asked to judge the perceptual prominence of each syllable on a free grid and a scale ranging from 0 to 30. In order to give the subjects some orientation on how to use the grid, they are asked to assign the prominence value of '0' for completely non-prominent syllables and '30' for extremely prominent syllables, e.g. corrections. No further instructions were given. After the fast versions, the normal speed versions had to be rated accordingly.

Each sentence was presented three times to the listeners through loudspeakers. During this time, the subjects had to assign prominence values to each syllable. The subjects were first and second year undergraduate students with little or no phonetic training. 42 subjects participated in the study, 24 were native speakers of German, 14 were non natives of very different linguistic backgrounds, most of them speaking a Slavic language or Chinese as their L1. All had a high level of linguistic competence in German.[2] Compared to German, Slavic languages tend to use a larger pitch range[3] and naturally, speakers of a tone language would have a different sensitivity for pitch movements. It is therefore expected that the non native listeners differ from the Germans concerning their prominence assignments, especially in the fast speech condition.

3.2. Results and Discussion

Correlation coefficients were measured between speaker groups (non-native vs. native) speaking styles (normal speed vs. fast) and some established acoustic prosodic correlates of syllable prominence (F0-distance to mean F0, normalized syllable duration, spectral tilt as described in [11].

The most striking result were the substantial listener correlations between all listener groups and speaking styles. Table 1 shows the correlations between listener groups for sentence 2. Correlations for sentence 1 are slightly lower but comparable.

Table 1: Mean Correlations (Spearman-Rho) between the prominence assignments of different listener groups and different speaking styles in sentence 2.

		non-nat.	natives	
		slow	fast	slow
non-nat.	fast	0.66	0.76	0.78
	slow		0.70	0.92
nat.	fast			0.73

It is also interesting that both listener groups agree almost perfectly in their prominence judgments for the slow speech, where acoustic correlates of prominence are in harmony with the top-down expectancies. Obviously, native and non-native speakers use similar cues for judging prominence. Interestingly, the lowest correlations can be found between the prominence ratings within the group of non-natives: their ratings differ much more between slow and fast speech. This might indicate that non-natives indeed base their ratings less on top-down knowledge than natives.

When comparing the prominence ratings and the acoustic correlates, slow speech has little surprises to offer: as expected, significant correlations can be found between duration and F0-variation for both listener groups (cc between 0.7 and 0.85), though not for spectral tilt.

Looking at the prominence ratings and the acoustic correlates of fast speech, only marginal correlations are found. The only exception is the parameter of duration. Non-native listeners based their ratings at least partly on durational variation (cc=0.75, $p<0.05$), whereas the native listeners seemed to rely more on their intuition.

3.3. Conclusions

The results once more confirm F0-variation and duration as good indicators for prosodic prominence. They also indicate

[2] The native languages were Bulgarian, Chinese, Indonesian, Latvian, Lithuanian, Polish, Russian.

[3] Grit Mehlhorn, personal communication

that given their absence, native speakers rely to a large extent on their top down knowledge when rating prominent syllables. Non-native speakers with a high L2 competence relied more on acoustic cues than the native speakers. But there still is a substantial correlation between native and non-native ratings and the ratings for fast speech. This indicates that the non-native speakers extrapolated some of their ratings from their top-down knowledge of German prosody as well. Summing up, hypothesis 1 is rejected in its strong version: prominence ratings kept relatively stable even given a strong deviation from the standard prosody. Hypothesis 2 is accepted — though not without hesitation — since at least a tendency for non-native listeners' comparatively high ability to rely on acoustic cues to prominence was found.

4. Introspection Study

In our perception study we showed a relative stability of speakers' prominence ratings even in the absence of acoustic cues. Instead of rating the acoustic prosody in the signal (in fast speech this would have been the assignment of a relatively equal low prominence throughout the utterance) they obviously decided to fall back to some top-down prominence pattern they had expected. In order to determine the degree of this influence for both natives and non-natives, additional prominence ratings were collected from a different group of listeners. This time, the subjects had to rate the syllable prominence based on an orthographic representation of the two sentences already used in the perception study. This way, it ought to be ensured that the ratings are as close as possible to some top-down expectancy concerning prosody. Again, the subjects were undergraduate students, 27 were native speakers of German, 15 were non-native with a high competence in German as an L2.

4.1. Results and Discussion

On average, the intuitive ratings correlated highly between the two listener groups (Spearman-Rho=0.9). Also, high mean correlations were found between all listener groups and prominence ratings of the previous perception study. Table 2 lists the results for sentence 2. Figure 1 shows a comparison of introspective and perceptual prominence ratings for non-natives, Figure 2 for natives. Correlations for sentence 1 were slightly lower but comparable.

Table 2: Mean Correlations (Spearman-Rho) between the introspective and prominence ratings based on speech perception in sentence 2.

	non-native		native	
	fast	slow	fast	slow
intro non-nat.	0.89	0.84	0.95	0.92
intro native	0.84	0.70	0.83	0.80

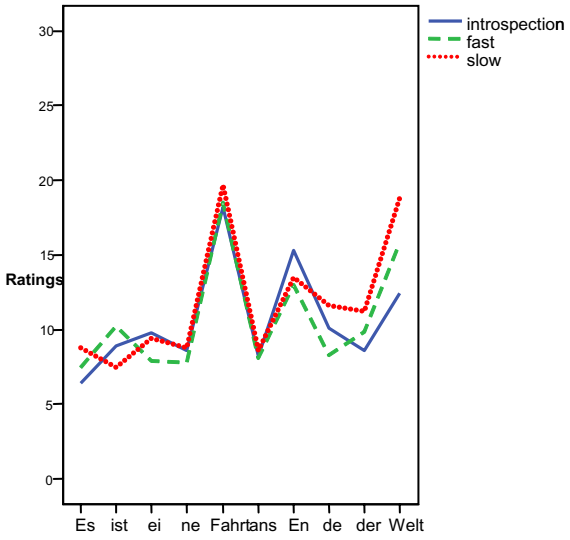

Figure 1: Mean prominence ratings based on introspection, fast and slow speech for non-natives.

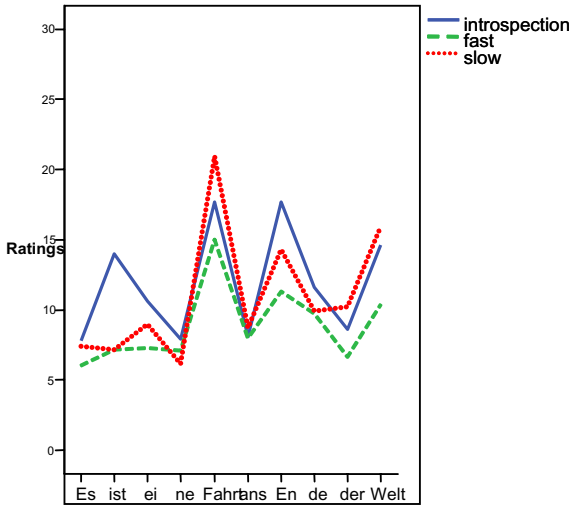

Figure 2: Mean prominence ratings based on introspection, fast and slow speech for natives.

Again, the relationship between acoustic prominence correlates and introspection is straightforward: For slow speech, the introspective ratings mirror the acoustic prosodic correlates of prominence duration and F0-variability for both listener groups (see Figure 3). For fast speech, as expected, no relationship between acoustic data and the acoustic parameters were found (see Figure 4).

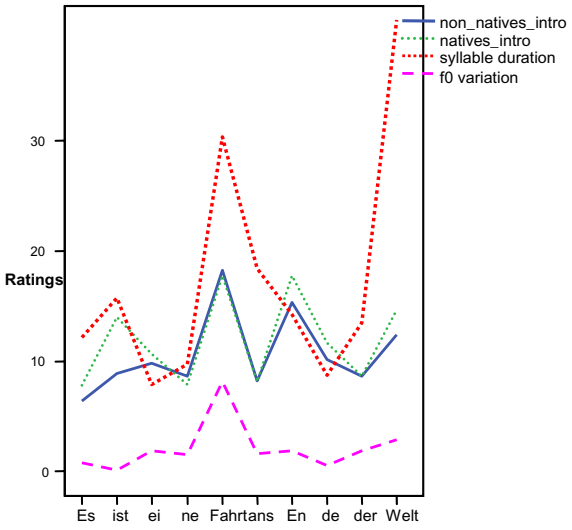

Figure 3: Comparison of Introspective Prominence Ratings and Acoustic Parameters for Slow Speech

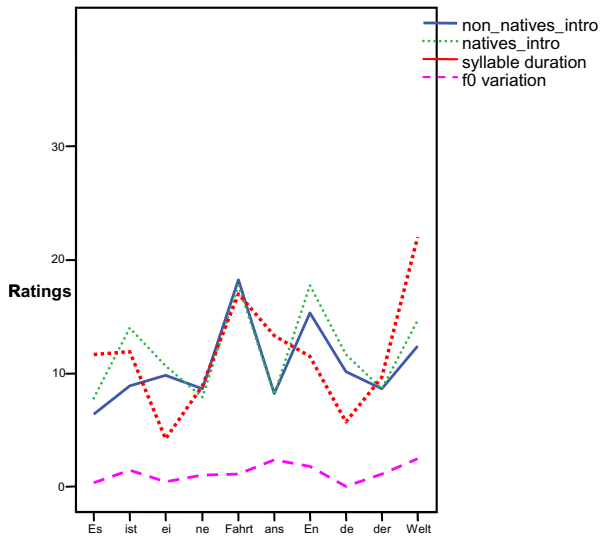

Figure 4: Comparison of Introspective Prominence Ratings and Acoustic Parameters for Fast Speech

5. Conclusions and Outlook

The assumption that top down intuitions guide listeners in their ratings of syllable prominence has received massive support. Likewise, non-native listeners with a high competence in the L2 base their prominence judgments on similar acoustic and linguistic cues as native speakers. Listeners use introspection as a fallback strategy if no reliable acoustic cues to prominence are present, as it is the case in very fast speech. Compared to native speakers, non-natives tend to rely more on acoustic cues. Since the acoustic parameters of prominence happen to correlate nicely with the native intuitions for speech read at normal speed, this does not pose a major problem to any theory of standard prosody. But theories of prediction or perception may run into problems when looking at overemphasized or monotonous speech. Here, the acoustic reality probably does not match listeners' impressions.

Future studies will look at more acoustic data for a regression analysis in order to determine the relative proportions of acoustic and introspective cues used by listeners for their prominence ratings.

6. References

[1] Wagner, P., *Wahrnehmung und Vorhersage deutscher Betonungsmuster*, Doctoral Thesis, Universität Bonn: 2002.

[2] Lehiste, I., *Suprasegmentals*, Cambridge, Mass., MIT Press: 1970.

[3] Eriksson, A., Thunberg, G. and Traunmüller, H. Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing, *Proceedings of Eurospeech 2001*, S. 399-402: 2001

[4] Wagner, P., Fischenbeck, E. Stress perception and production in German stress clash environments. *Proceedings of the Speech Prosody 2002 Conference*, Aix-en-Provence.

[5] Eriksson, A., Grabe, E., and Traunmueller, H. Perception of syllable prominence by listeners with and without competence in the tested language. *Proceedings Speech Prosody 2002*, Aix-en-Provence, 275-278.

[6] Peperkamp, S., Dupoux, E. and N. Sebastián-Gallés, Perception of stress by French, Spanish, and bilingual subjects. *Proceedings of EUROSPEECH'99*, 6, 2683-2686.

[7] Portele, T., Perceived Prominence and acoustic parameters in American English, *Proceedings of ICSLP'98*, Sydney, 3, 667–670.

[8] Streefkerk, B., Prominence. *Acoustic and lexical/syntactic correlates.* Doctoral Thesis. LOT, Utrecht: 2002.

[9] Fant, G. and Kruckenberg, A., Preliminaries to the study of Swedish prose reading and reading style. *STR-QPSR*, 2/1989 KTH, Stockholm, 1–83.

[10] Dellwo, V., Aschenberner, B., Wagner, P., Dancovicova, J., Steiner, I.. BonnTempo-corpus and BonnTempo-tools: a database for the study of speech rhythm and rate, *INTERSPEECH-2004*, 777-780.

[11] Heldner, M. (2001), Spectral Emphasis as an Additional Source of Information in Accent Detection. *Prosody in Speech Recognition and Understanding*, paper #10, Red Bank, NJ, USA.