Objective, Subjective and Linguistic Roads to Perceptual Prominence How are they compared and why?

Petra Wagner[1], *Fabio Tamburini*[2], *Andreas Windmann*[1]

[1]Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Germany
[2]Dipartimento di Studi Linguistici e Orientali, Università di Bologna, Italy
petra.wagner@uni-bielefeld.de, fabio.tamburini@unibo.it, andreas.windmann@uni-bielefeld.de

Abstract

Prosodic prominence denotes the perceptual salience of linguistic units. There exists no agreement on (1) adequate methods for its subjective measurement, (2) its objective acoustic correlates and (3) its relationship to linguistic structure. A traditional approach for evaluating any of these descriptive layers is an inter-level comparison, e.g. between a perceptual and an acoustic model of prominence. However, (1) there exists no standard procedure for such a comparison, and (2) such a comparison is misleading if both layers are expected to be symmetrical, given the neglected influence of linguistic top-down expectancies. We propose an evaluation procedure for prominence models relying on tripartite correlations of perception, its acoustic correlates and linguistic expectations. We suggest a novel correlation metric and test its usefulness on a prosodic corpus of German.

Index Terms: prominence, evaluation, prosody

1. Introduction: Model evaluation by comparing descriptive layers of prominence

Prosodic *prominence* is commonly regarded as the *perceptual salience of a linguistic unit relative to its environment*. However, we are far from having a consensus on how it is measured subjectively and how it relates to objectively measurable acoustic events or linguistic structures such as lexical and sentence stress or prosodic focus. There is wide agreement that prominence perception is influenced by both top-down expectations (mostly shaped by linguistic structures but also “paralinguistics”) and bottom-up processing and interpretation of the acoustic signal [1, 2, 3, 4, 5].

Given the fact these three layers of prosodic prominence — subjective perceptual, objective acoustic, expectancy based — are at least partly independent of each other, we obviously cannot expect them to stand in a 1:1 relationship — still, many models make the implicit assumption that they should be mirror images.

Despite this, we do expect a resemblance of prominence patterns between the various descriptive layers. It still is inherently problematic to interpret mismatches between two layers as “mistakes” in one of the descriptions: A mismatch between subjective prominence perception and an objective acoustic prominence model (henceforth *A-model*) does not necessarily mean that we have been looking at the wrong acoustic measures of prominence. It could be the case that subjective perception was based primarily on linguistic expectations or that the annotation design, i.e. the subjective perceptual model (henceforth *P-model*) was inadequate. If a reanalysis shows that a lot of the non-correspondences between acoustics and perception are explicable by top-down effects, the acoustic model may still be adequate. Still, if such a reanalysis fails, but we do find a good resemblance of the perceived patterns in our expectancy based linguistic model (henceforth *L-model*), we have indirect evidence that our acoustic model is inadequate.

Once all three models have reached a suitable — but evidently not perfect — quality, a similar mismatch analysis may yield interesting insights based on the ”missing link”: Mismatches between P-model and L-model can be a good diagnostics for acoustically driven influences on prominence perception, possibly caused by certain “non scripted” speaking styles deviating from a “citation form” represented in the expectancy based models. Mismatches between P-model and A-model may indicate the precise influence of expectancy based perceptual prominence judgements, while mismatches between L-model and A-model may uncover strategies of various listener groups, e.g. native vs. non-native listeners (see Figure 1).

In the next section, we will propose a method for comparing prominence profiles of various model layers. The main aim of this model of comparison is a diagnostic method for detecting weaknesses in all descriptive layers thus having a better idea where the models need further improvement.

2. A method to compare continuous prominence descriptions

We treat prominence as a continuous variable, not inherently limited to a predefined number of levels [3, 5, 7] . In

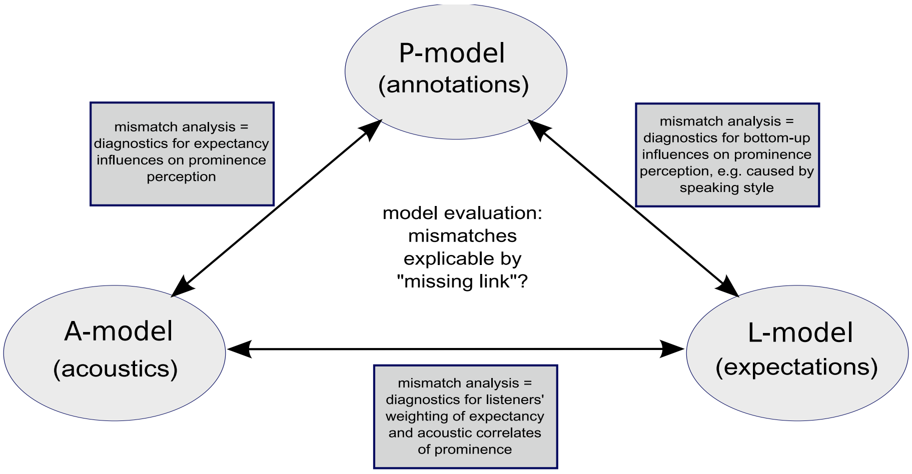
P-model
(annotations)
mismatch analysis = diagnostics for expectancy influences on prominence perception
mismatch analysis = diagnostics for bottom-up influences on prominence perception, e.g. caused by speaking style
model evaluation: mismatches explicable by "missing link"?
A-model
(acoustics)
L-model
(expectations)
mismatch analysis = diagnostics for listeners' weighting of expectancy and acoustic correlates of prominence

Figure 1: *Triangular prominence model evaluation based on mismatch: If one of the descriptive models does not correspond with the other two while these two do, we have evidence for a lack in precision for the regarded model. Mismatches between two models are explicable by the model layer located vis-à-vis, the "missing link", e.g. mismatches between A-model and P-model inform us about influences of L-model etc.*

order to compare two prominence profiles, seen as continuous functions over discrete values $P : \{0, N-1\} \rightarrow [0,1]$, we have to define a specific metric function able to capture the kind of comparison linguists have in mind when judging prominence profiles. We are interested in developing a measure able to (a) verify that the local maxima in the profiles are located on the same syllables and (b) the different heights of these local maxima in the two profiles draw similar pictures. Requirement (a) regards the concept that, a local maximum being a point where we perceive a prominence, prominent syllables in the two profiles should match. Requirement (b) ensures that the relative importance of these maxima is respected, evaluating their relative height and prominence strength. These two constraints are very different and require different approaches for measuring the degree of congruence between two profiles: The first asks for a local measure, while the second implies a global measurement and comparison of the two profiles.

We will try to develop this idea by using the examples in Figure 2, where the reference profile A is compared with other profiles. Qualitatively, the linguist would expect that B1, when compared to A, will obtain a medium score, because the two maxima are in the same position but have different heights. B2 should obtain a very low score, because there is no correspondence between the maxima at all, while B3 should get a high score because there are only slight differences in the two profiles.

There are various methods for comparing two functions that span over continuous values, the most common ones certainly being correlation coefficients. In particular, the Pearson Correlation Coefficient (PCC) and Spearman Rank Correlation Coefficient ($SRCC$) have been used in various studies for comparing prominence profiles [8, 9, 6]. Unfortunately, as shown in Figure 2, these measures are influenced by the distribution of values in the whole utterance and fail to capture the local correspondence of maxima in the two profiles; they represent good measures of the *global* matching of the profiles and the degree of matching between the heights of the local maxima. Comparing profile A with B1, we expect, despite the difference in height of the maxima, to have a medium value of correlation, but both SRCC and PCC return low correlation values. The other two examples, namely B2 and B3, behave as expected, providing low values for B2 and high scores for B3.

As mentioned, correlation coefficients capture global properties of the profiles in the correct way (for B2 and B3), but are inadequate for measuring local similarities, as in example B1. We therefore consider (1) as a measure of the global similarity between two profiles A and B.

$$G(A,B) = PCC_{\{0,N-1\}}(A,B) \qquad (1)$$

We can define a local similarity measure by averaging the contribution of the different portions of utterances, compared through a sliding window of length 2K+1 and a SRCC as in (2).

$$L(A,B) = \frac{\sum_{j=0}^{N-1} SRCC_{[j-K,j+K]}(A,B) \cdot W_N[j-K,j+K]}{\sum_{j=0}^{N-1} W_N[j-K,j+K]} \qquad (2)$$

where W_N weights the contribution of each window, accounting for the fact that the initial and final windows cross the borders of the utterance and thus contains fewer values.

$$W_N[p,q] = \min(q, N-1) - \max(p,0) + 1 \qquad (3)$$

As is shown in Figure 2, L(A,B) provides a good measure of the local matching between the profiles, but it is

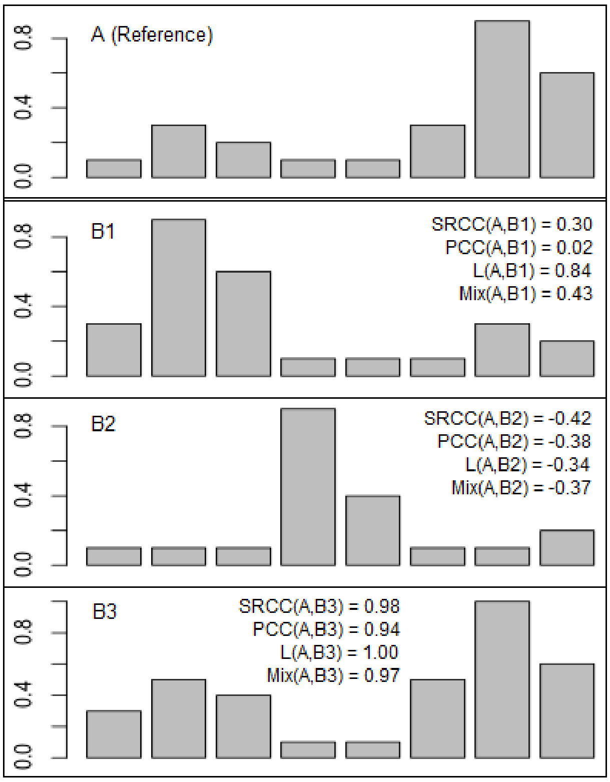

Figure 2: *Prominence profiles.*

insensitive to the absolute values, in particular the maxima heights.

Given the two measures of local and global matching between profiles we can combine them into a unique matching measure by averaging them as in (4).

$$Mix(A,B) = (L(A,B) + G(A,B))/2 \tag{4}$$

By simply averaging the different contributions of global and local similarity, we can interpret the final score provided by Mix again as a correlation coefficient, thus simplifying result interpretation.

When comparing the behaviour of the Mix similarity measure on the examples in Figure 2, we see that it resembles the linguists' intuition of prominence profile comparison, providing a mid similarity score to the first example, B1, and a low and a high score respectively to examples B2 and B3.

3. Prominence model evaluation

The reference corpus [8] we are dealing with has been annotated for prominence by experts on all three different layers. The corpus comprises 285 utterances of read speech produced by three speakers. The material has been carefully designed for containing a lot of typical prosodic variations such as various types of questions, imperatives, focus constructions, expressions of agreement and disagreement, comparisons, lists and stories and is phonetically balanced. The corpus has been used for studying the acoustic realisation of German prominence patterns for the purpose of prosody modelling in speech synthesis [10] and has been used for building rules predicting prominence patterns based on linguistic representations [9]. These rules were used to improve the prosodic adequacy of a unit selection speech synthesis system [11]. The objective prominence annotations were built on the model developed in [12] and adapted to German in [6]. The objective annotations were also used to automatically annotate a speech corpus used as a database for unit selection speech synthesis.

3.1. Subjective Prominence Model - P-model

The prominence annotations were carried out by three phonetically trained annotators using a 31-point scale [1]. The annotations were performed on the level of the syllable. The median of each syllable was used as the subjectively perceived prominence value of each syllable. This method of annotation resulted in a quasi-continuous prominence profile for each utterance. Inter-annotator agreement ranged between $\rho = 0.74$ and $\rho = 0.86$ (Spearman-Rho).

3.2. Objective Acoustic Model - A-model

The objective model has been previously described in and yields a continuous prominence profile for any examined utterance, attaching a prominence value to each syllable in the corpus using only acoustic information. We can define the prominence function

$$\begin{aligned} Prom^i = \; & W_{FA} \cdot \left[SpEmph^i_{SPLH-SPL} \cdot dur^i \right] + \\ & W_{PA} \cdot \left[en^i_{ov} \cdot \left(A^i_{ev}(at_M, at_m) \cdot D^i_{ev}(at_M, at_m) \right) \right] \end{aligned} \tag{5}$$

where $SpEmph_{SPLH-SPL}$ is the spectral emphasis, dur is the nucleus duration, en_{ov} is the overall energy in the nucleus and A_{ev} and D_{ev} are the parameters derived from the TILT model [13] as a function of the maxima alignment type – at_M – and the minima alignment type – at_m. All parameters are referred to the generic syllable nucleus i. Refer to [12, 6] for any detail about this model.

3.3. Expectancy Based Linguistic Model - L-model

The algorithm for the expectancy based model has been described in [9]. It predicts metrical grids based on POS information, utterance boundaries and rhythm rules in order to fill metrical gaps. It returns a grid with metrical strengths for each syllable within an utterance, the grids ranging between a "metrical strength" 1 and 7. Thus, unlike in the continuous subjective and acoustic descriptions, only 6 levels of prominence are differentiated. Given that this model only takes into account the structural linguistic influence on perceptual prominence, it is expected to perform poorer on spontaneous speech. However, on a prosodically and segmentally varied corpus of read speech, produced by three speakers of German, its performance on average reaches a very good agreement ($\rho \approx 0.8$) with the perceptual prominence patterns.

3.4. Results

We calculated correlations between acoustic, perceptual and expectancy driven prominence descriptions for all utterances in the database. A comparison between the local, global and mixed comparisons proposed in 2 shows that the mixed approach provides correlations between those of the local and the global metric. We concluded that while it is tempting to use the local measure (given the higher correlations), the mixed metric does indeed capture different aspects of each prominence profile. All subsequent analyses were based on the mixed measure for comparison.

Table 1: *The proposed correlation measures applied to evaluate the different annotation models.*

	A-model vs P-model	A-model vs L-model	P-model vs L-model
L(A,B)	0.652298	0.596236	0.810729
G(A,B)	0.577096	0.459507	0.770268
Mix(A,B)	0.614697	0.527872	0.790499

These results indicate that P- and L-model are generally in closer agreement than any of them is with the A-model. For a more in-depth evaluation, we identified those utterances where all models are in good agreement ($mix > 0.6$ for all comparisons). As argued above, a perfect match cannot be expected given the interactions between the descriptive layers. This analysis yielded that all three models were in agreement for 38% of all utterances contained in the database. The remaining utterances lacked correspondence between at least two descriptive layers. In order to diagnose specific problems of all three models, we calculated those cases where *one* model does not correlate with two others. This analysis clearly showed a comparatively large number of cases (11%) where the A-model disagrees with both L-model and P-model while these two correspond. The L-model clearly disagrees with both other models in only 3% of all cases, the P-model in only 2% of all cases. We have ample evidence that the acoustic model needs further improvement. However, it ought to be kept in mind that the style of speech (carefully read speech) is likely to be in accordance with our linguistic expectations. We therefore predict a closer agreement between A- and P-model and more cases where the L-model diverges from the two others when regarding less controlled, spontaneous speech showing more stylistic diversity. Furthermore, investigating the behaviour of less trained annotators may uncover problems with the robustness of the P-model.

4. Conclusions

We suggested a tripartite method for evaluating subjective, objective and linguistic models of prosodic prominence based on a mismatch analysis. We developed a method for comparing prominence profiles taking into account both global and local similarity and evaluated prominence models using an annotated corpus of read German. The evaluation procedure detected a lack of correspondence between the used acoustic prominence model and the perception and expectancy models. This result may be interpreted either as deficits of the acoustic model, or provide information to the integration of top-down and bottom-up information. It is likely that this result is at least partly determined by the examined speaking style.

5. Acknowledgements

Petra Wagner is supported by the BMBF funded “Professorinnenprogramm”, FKZ 01FP09105A. Andreas Windmann is supported by the Center of Excellence for Cognitive Interaction Technology CITEC, Bielefeld, Germany.

6. References

[1] Fant, G., Kruckenberg, A. “Preliminaries to the study of Swedish prose reading and reading style”, STL-QPSR, 30(2):1–80, 1989.

[2] Streefkerk, B.,“Prominence. Acoustic and lexical/syntactic correlates ”, LOT Series, 2002.

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

[4] Wagner, P., ”Great Expectations – Introspective vs. Perceptual Prominence Ratings and their Acoustic Correlates”, Proceedings of Interspeech, Lisbon, Portugal: 2381–2384, 2005.

[5] Arnold D., Wagner, P., Möbius, B., ”The effect of priming on the correlations between prominence ratings and acoustic features”, Prosodic Prominence: Perceptual and Automatic Identification (Speech Prosody 2010 Workshop), Chicago, USA: 2381–2384, 2010.

[6] Tamburini F., Wagner P., ”On Automatic Prominence Detection for German”. Proceedings of InterSpeech 2007, Antwerp, 1809-1812, 2007.

[7] Todd, M., Neil, P., Brown, G., “Visualization of Rhythm and Meter”. Artificial Intelligence Review, 10: 253–73, 1996.

[8] Heuft, B., Portele. T., Wagner, P., Widera, C., Wolters. M. “Perceptual Prominence”. In: Sendlmeier, W. (ed.). Speech and Signals. Aspects of Speech Synthesis and Automatic Speech Recognition. Hektor, Frankfurt a.M., 97–116, 2000.

[9] Wagner, P. Vorhersage und Wahrnehmung deutscher Betonungsmuster. Doctoral Dissertation, Universität Bonn, URN:: urn:nbn:de:hbz:5-00548, 2002.

[10] Portele, T., Heuft, B. “Prominence-driven speech synthesis”. Speech Communication, 1998.

[11] Windmann, A., Jauk, I., Tamburini, F., Wagner, P. “Prominence-Based Prosody Prediction for Unit Selection Speech Synthesis”. Proceedings of Interspeech, Florence, Italy: 325-328, 2011.

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

[13] Taylor, P. “Analysis and synthesis of intonation using the tilt model”. JASA 107: 1697–714, 2000.