On the Structure of Internal Prosodic Models

Antje Schweitzer and **Bernd Möbius**

Institute of Natural Language Processing, University of Stuttgart, Germany
{antje.schweitzer,bernd.moebius}@ims.uni-stuttgart.de

ABSTRACT

We investigate the structure of internal prosodic models. It is hypothesized that accumulations of perceived exemplars implicitly define prosodic target regions that are used in speech production. Results from two experimental studies on tonal and temporal events, respectively, are reported. The first study concerns the categorical status of high and low boundary tones in German in a categorical perception paradigm. The aim of this experiment is to establish a methodology for testing the categorical status of prosodic events. In the second experiment we use z-scores as a measure of local speech rate. We demonstrate that this is an appropriate measure for the temporal dimension of perceptual target regions that are related to prosodic categories.

1 INTRODUCTION

We have previously proposed an extension and generalization of Guenther and Perkell's speech production model [1, 2], by integrating its segmental perspective with a new theory of the production of prosody [3, 4]. Guenther and Perkell's model posits that speech production is constrained by perceptual requirements. The only invariant targets of the speech production process are assumed to be regions in auditory perceptual space.

In analogy to the segmental domain, we interpret speech movements in the prosodic domain as tonal and temporal gestures that are planned to reach and traverse perceptual target regions. The targets are characterized as multidimensional regions in the perceptual space. Gestures that are successfully executed by the speaker produce acoustic realizations of perceptually relevant prosodic events, such as those predicted by intonational phonology. Examples of mapping relations between the target regions and tonal gestures have also been given [3].

In the present paper we further investigate the structure of internal prosodic models by following two interrelated avenues of our research program. Our first goal is to establish a methodology for testing the categorical status of prosodic events. For this purpose the adequacy for prosody of the classical categorical perception paradigm is assessed. Here we present a study concerning the categorical status of high and low boundary tones in German (section 3). Our second goal is to explore the dimensions of perceptual target regions that are related to presumed prosodic categories. We propose to use z-scores as a measure of local speech rate and demonstrate that the characteristic effects of the presumed categories on the temporal structure of speech can be modelled by this measure (section 4).

2 INTERNAL PROSODIC MODELS

According to the speech production model there is a unique phonetic target region in auditory-temporal space for each phoneme of a given language [5]. The process of language and speech acquisition involves establishing the mappings between abstract phonemes and their corresponding auditory targets. Once learned, such phonemic settings tend to be stable and resistant to change. We posit that these properties of speech pertain to the prosodic domain too. The role of stable internal models of prosody as a cause of intonational foreign accent, for instance, indicates that language learners acquire internal prosodic models along with segmental ones.

Recently it has been claimed that internal phonemic models emerge from storing in memory representations of large numbers of perceived acoustic realizations [6, 7]. There is evidence that what is used in speech perception is these exemplars themselves, including their phonetic detail, rather than more abstract representations built from the exemplars.

In speech perception, new tokens are perceived in identification tasks as belonging to the category that comprises the highest number of similar exemplars. Discrimination sensitivity depends on the local variation in the number of exemplars from competing categories [8]. This model can account for the warping of the perceptual space attributed to the perceptual magnet effect as well as for phoneme boundary effects: peaks in discrimination sensitivity are expected between overlapping categories, whereas lower discrimination sensitivity is expected in within-category regions.

In speech production, exemplars can serve as perceptual targets. This account is compatible with the concept of perceptual target regions outlined above if one assumes that the accumulation of exemplars implicitly defines a corresponding region in perceptual space. The z-score measure introduced below also requires that the listener has access to a large number of acoustic realizations of segments occurring in different prosodic contexts. Before presenting the details of the pertinent experiment in section 4, we first address the categorical status of boundary tones.

3 CATEGORICAL STATUS OF TONAL EVENTS

The categorical status of prosodic events is generally less agreed upon than that of segmental phonemes. It is not even obvious which prosodic events are relevant. In a series of identification and discrimination experiments by our group the categorical status of high (H%) and low (L%) boundary tones in German was investigated [9].

For this purpose the experimental design developed by [10] was adopted, which had been shown to succeed in verifying the categorical status of boundary tones in Dutch. The rationale behind these experiments was the assumption that the categorical contrast between high and low boundary tones is the least disputed prosodic contrast. If the categorical perception paradigm is not sensitive enough to serve as a diagnostic tool for determining the linguistic status of this major contrast, then there is little hope for its applicability to less clear tonal categories, such as types of pitch accents.

In the present perception experiment the phrase-final fundamental frequency (F_0) contour of a test sentence was systematically manipulated using PSOLA resynthesis [11]. The test sentence, produced by a professional male speaker in a dialog context, was carefully selected to satisfy the following criteria: (i) it ended on an ambiguous, i.e. neither low nor high, boundary tone; (ii) there was no pitch accent on the final syllable; (iii) the final syllable included a full vowel. From this test sentence, 11 stimulus sentences were generated with systematically varying final F_0 movements, such that they represented a quasi-continuum between typical low and high boundary tones as estimated from the recorded dialog. The final F_0 values of any two adjacent stimuli differed by 0.3 ERB.

These stimuli were presented in randomized order to a group of 24 listeners participating in two tasks, first an identification and subsequently a discrimination task. The result of the identification task showed clearly S-shaped curves for each of the subjects. Typically, the perceptual change from statement to question occurred within one stimulus step. However, the exact location of the switch varied between subjects; but even after averaging over all subjects, the category switch occurred within just a few stimulus steps. The correlation between the perceived category boundary and the discrimination peak was low but significant. These results suggest that high and low boundary tones are indeed intonational events with categorical status in German. More details of this study and its results are presented elsewhere at this conference [9].

4 TEMPORAL TARGET REGIONS

The relevant dimensions in the perceptual space and their respective acoustic correlates are not uncontroversial. While it has been demonstrated that the perception of fundamental frequency can be adequately modeled using the ERB scale [12], the relation between segment or syllable durations and perceived local speech rate is less well-established.

We propose to use z-scores of segment durations, i.e., segment durations normalized by phoneme-specific mean duration and standard deviation, as a measure for local speech rate. Analysis of speech data shows that well-known effects of prosodic context on local speech rate, e.g. phrase-final lengthening, are clearly visible in the z-score distributions for the corresponding units, both on the segmental and on the suprasegmental level.

4.1 USING Z-SCORES FOR NORMALIZING PHONEME DURATIONS

Z-scores have first been employed for duration modeling in text-to-speech synthesis (TTS) to predict individual segment durations for a given syllable duration according to an "elasticity principle" [13, 14]. The idea is that different phonemes can be lengthened or shortened to different extents and that this phoneme specific "elasticity" manifests itself in the distribution of observed segment durations and their standard deviation, respectively.

Formally, the z-score of a segment p_i is the factor that has to be applied to the corresponding phoneme's standard deviation $\sigma(p)$ such that it sums up to the observed segment duration together with the phoneme's mean $\mu(p)$. In other words, the z-score indicates by how much a particular segment deviates from the phoneme's mean duration. The formula is given in (1).

$$duration(p) = \mu(p) + z\text{-}score(p_i) * \sigma(p) \qquad (1)$$

Z-scores have since been used for prediction of segment durations in many different TTS systems. Reversing the original concept, we propose to take z-scores as a means to analyze the temporal structure of speech irrespectively of phoneme-specific durations.

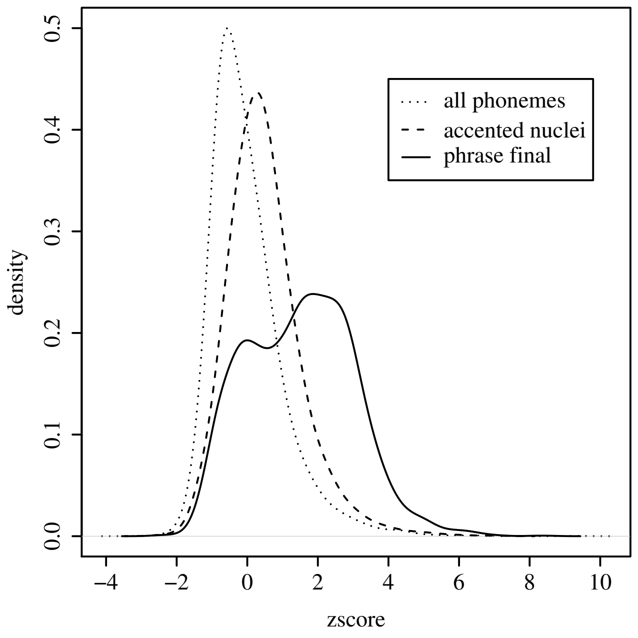

Figure 1: Z-score density functions for all phonemes (dotted line), nuclei of accented syllables (dashed), and phrase-final phonemes (solid). Z-scores of accented nuclei and phrase-final phonemes are significantly higher than the average.

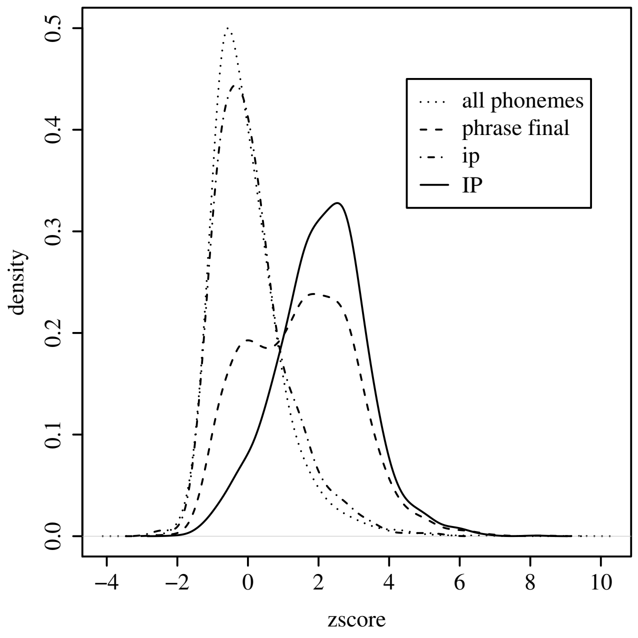

Figure 2: Z-score density function for phrase-final phonemes (dashed line), repeated from Figure 1. The bimodal distribution is due to different distributions for phrase-final phonemes in intermediate phrases (ip, dot-dashed) and intonation phrases (IP, solid). For comparison, the z-score distribution for all phonemes is indicated by the dotted line.

We investigated Z-score distributions in a speech corpus recorded for unit selection synthesis. The corpus was designed to cover at least all German diphone types and all phoneme types in different contexts. To this end, phoneme/context vectors for each sentence in a large collection of newspaper articles were predicted using the IMS German Festival TTS system [15]. From these sentences a subset with the same coverage in terms of diphone types and a good coverage of different phoneme/context vectors was extracted. Sentences containing diphone types that were not found in the corpus, but theoretically allowed by German phonotactics, were manually added, as was some application specific text material.

The corpus was read by the same professional speaker as in the previous experiment (section 3). Each utterance was annotated on the segment, syllable and word level by forced alignment with manually corrected transcriptions. Prosodic phrases (distinguishing between intermediate phrases and intonation phrases) and pitch accents (without distinguishing between pitch accent types) were automatically annotated using the TTS prosody prediction component [16]. The data amounts to almost 160 minutes of speech and contains approximately 94,000 segments and 37,000 syllables. Z-scores were calculated for syllables and segments with mean durations and standard deviations taken from the corpus.

4.2 RESULTS

Effects of prosodic factors on segmental durations are clearly visible in the z-score distributions of phonemes in different prosodic contexts. Figure 1 shows that the distribution for all phonemes across all contexts differs from the distribution for pitch-accented nuclei. Z-scores for the latter tend to be higher than the average. The effect is even stronger for phrase-final phonemes. The significance level is the same in both cases ($p \ll 0.0001$ for the Wilcoxon rank sum test with continuity correction and a confidence level of 0.999), but the difference of mean z-scores in the two prosodically marked contexts, compared to the mean z-scores across all contexts, is higher for phrase-final segments than for accented nuclei: the means are -0.014, 0.497 and 1.489 for all phonemes, accented nuclei and phrase-final phonemes, respectively.

The distribution of z-scores for phrase-final segments in figure 1 is bimodal. This is evidently due to the fact that there is almost no lengthening for phrase-final phonemes in intermediate phrases, whereas there is substantial lengthening in intonation phrases, as illustrated in figure 2.

The influence of prosodic context is not limited to single phonemes. For instance, phrase-final lengthening can be observed for all segments in phrase-final syllables, as well as for all coda segments in phrase-final syllables. The significance level is the same, but the means differ less for the coda segments and even less when all segments in phrase-final syllables are taken into account. The same holds for syllable-level z-score distributions.

5 CONCLUSION

In this paper we have investigated the structure of internal prosodic models. We have discussed the possibility that accumulations of perceived exemplars implicitly define target regions that are used in speech production. Under this assumption prosodic categories are characterized by regions with an increased density of exemplars in perceptual space.

Results from two experimental studies on the tonal and temporal dimensions of prosodic events, respectively, were reported. The first experiment aimed at establishing a methodology for testing the categorical status of prosodic events. The general adequacy for prosody of the classical categorical perception paradigm was demonstrated. The results of the experiment confirm the categorical status of one particular prosodic contrast in German, viz. that between high and low boundary tones.

When interpreting the results of the second experiment, a picture emerges where the z-score distributions are influenced directly and characteristically by the prosodic context. In an exemplar-based model it is plausible that both the speaker and the listener have access to typical duration and duration variability for these contexts to reliably distinguish stimuli along the z-score dimension in the temporal domain.

ACKNOWLEDGMENTS

This work was funded in part by the German Research Council (DFG, Grant DO 536/4-1) and in part by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the SmartKom project under Grant 01IL905K7. The responsibility for the content lies with the authors.

REFERENCES

[1] F. H. Guenther, M. Hampson, and D. Johnson, "A theoretical investigation of reference frames for the planning of speech movements," *Psychological Review*, vol. 105, pp. 611–633, 1998.

[2] J. S. Perkell, F. H. Guenther, H. Lane, M. L. Matthies, P. Perrier, J. Vick, R. Wilhelms-Tricarico, and M. Zandipour, "A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss," *Journal of Phonetics*, vol. 28, no. 3, pp. 233–272, 2000.

[3] G. Dogil and B. Möbius, "Towards a model of target oriented production of prosody," in *Proc. Eurospeech-2001 (Aalborg)*, 2001, vol. 1, pp. 665–668.

[4] B. Möbius and G. Dogil, "Phonemic and postural effects on the production of prosody," in *Proc. Speech Prosody 2002 (Aix-en-Provence)*, 2002, pp. 523–526.

[5] J. Perkell, F. Guenther, H. Lane, M. Matthies, J. Vick, and M. Zandipour, "Planning and auditory feedback in speech production," in *4th Internat. Speech Motor Conf. (Nijmegen)*, 2001, pp. 5–11.

[6] K. Johnson, "Speech perception without speaker normalization: An exemplar model," in *Talker Variability in Speech Processing*, K. Johnson and J. W. Mullennix, Eds., pp. 145–165. Academic Press, San Diego, 1997.

[7] J. Pierrehumbert, "Exemplar dynamics: Word frequency, lenition and contrast," in *Frequency and the Emergence of Linguistic Structure*, J. Bybee and P. Hopper, Eds., pp. 137–157. Benjamins, Amsterdam, 2001.

[8] F. Lacerda, "The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory," in *Proc. 13th Internat. Congr. Phonet. Sci. (Stockholm)*, 1995, vol. 2, pp. 140–147.

[9] K. Schneider and B. Lintfert, "Categorical perception of boundary tones in German," in *Proc. 15th Internat. Congr. Phonet. Sci. (Barcelona)*, 2003.

[10] B. Remijsen and V. van Heuven, "Gradient and categorical pitch dimensions in Dutch: diagnostic test," in *Proc. 14th Internat. Congr. Phonet. Sci. (San Francisco)*, 1999, vol. 3, pp. 1865–1868.

[11] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, 1990.

[12] D. J. Hermes and J. C. van Gestel, "The frequency scale of speech intonation," *Journal of the Acoustical Society of America*, vol. 90, pp. 97–102, 1991.

[13] W. N. Campbell and S. D. Isard, "Segment durations in a syllable frame," *Journal of Phonetics*, vol. 19, pp. 37–47, 1991.

[14] W. N. Campbell, "Syllable-based segmental duration," in *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoît, and T.R. Sawallis, Eds., pp. 211–224. Elsevier, Amsterdam, 1992.

[15] IMS Festival, "IMS German Festival home page," [www.ims.uni-stuttgart.de/phonetik/synthesis/].

[16] A. Schweitzer and M. Haase, "Zwei Ansätze zur syntaxgesteuerten Prosodiegenerierung," in *Proc. Konvens-2000 (Ilmenau, Germany)*, 2000, pp. 197–202.