Exemplar-Based Production of Prosody: Evidence from Segment and Syllable Durations

Antje Schweitzer & Bernd Möbius

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

Abstract

We present results from experiments on the temporal properties of prosodic events, providing evidence that accumulations of exemplars implicitly define perceptual target regions in prosody production. We argue that z-scores of segment and syllable durations are the relevant perceptual dimension of these regions. To support this hypothesis, we present experimental results confirming that realizations of segments and syllables in different prosodic contexts show significantly different z-score distributions. Further experiments show that the relationship between syllable z-scores and the z-scores of the corresponding segments is significantly stronger for infrequent than for frequent syllables. We claim that this is due to the fact that infrequent syllables have to be assembled from smaller units because they are not represented by enough exemplars to establish a syllable-level target region.

1. Introduction

We have previously proposed an extension and generalization of Guenther and Perkell's [1, 2] speech production model. 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 [3, 4].

It has been claimed that internal phonemic models emerge from storing in memory representations of large numbers of perceived acoustic realizations [5, 6]. 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 production, these exemplars could serve as perceptual target regions in the sense of Guenther and Perkell if we assume that the accumulation of exemplars implicitly defines a corresponding region in perceptual space [7]. Thus, the speaker has access to stored representations of prosodic events, including their tonal and temporal structure, that serve as a reference in speech production. In this paper, we illustrate this view by means of experimental results obtained for temporal aspects of prosody in section 3.

Frequent syllables claimed to be stored in a mental syllabary [8] have been shown to exhibit more coarticulation than rare syllables, which are assumed to be assembled on-line from smaller units [9]. We will argue below (section 4) that this would necessarily follow given an exemplar-theoretic interpretation of Guenther and Perkell's speech production model: infrequent units are represented by considerably fewer exemplars; for the most infrequent units, there may be no exemplars stored at all. This implies that there are no target regions available for infrequent units, and that the speaker has to resort to smaller and therefore more frequent units. We present data from an experiment on durations of frequent and infrequent syllables that confirm differences in the production of very frequent and very infrequent syllables.

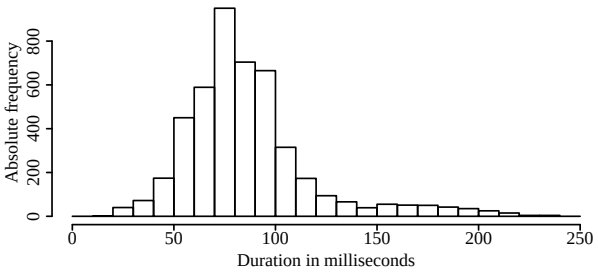

Figure 1: *Histogram of durations for [s].*

Before reporting details of the experiments, we briefly describe the speech corpus that we used as a database for our computations.

2. The speech corpus

The experiments reported in this paper are based on a large speech corpus originally recorded for unit selection speech synthesis. The corpus was read by a professional speaker. Each utterance was annotated on the segment, syllable and word level by forced alignment and manually checked afterwards. Prosodic phrases and pitch accents were manually annotated using GToBI(S) [10]. The data amounts to almost 160 minutes of speech and contains approximately 94,000 segments and 34,000 syllables. All statistics on this corpus throughout this paper were conducted using the R package [11].

3. Temporal target regions

According to [1], the only invariant targets in speech production are regions in perceptual space. When applying this model to the production of temporal aspects of prosody, three questions have to be answered: (i) what are the temporal dimensions of perceptual space, (ii) which are the relevant prosodic events, and (iii) which are the target regions corresponding to these events.

Question (ii) is beyond the scope of this paper. In analogy to the segmental domain, the relevant events should be prosodic categories that are perceptually different. For the present purpose, we will assume that the relevant events are phrase boundaries and pitch accents. We will distinguish between intermediate boundaries (ip) and intonation phrase boundaries (IP), but not between different types of pitch accents.

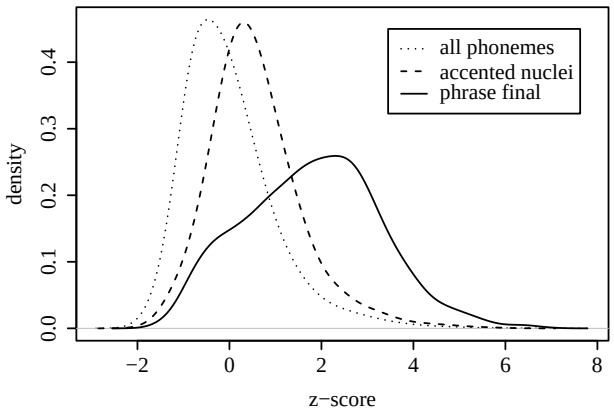

Figure 2: *Z-score density functions for all phonemes (dotted line), nuclei of accented syllables (dashed), and phrase-final phonemes (solid). Z-scores of accented nuclei and phrase-final phonemes are significantly higher than the average.*

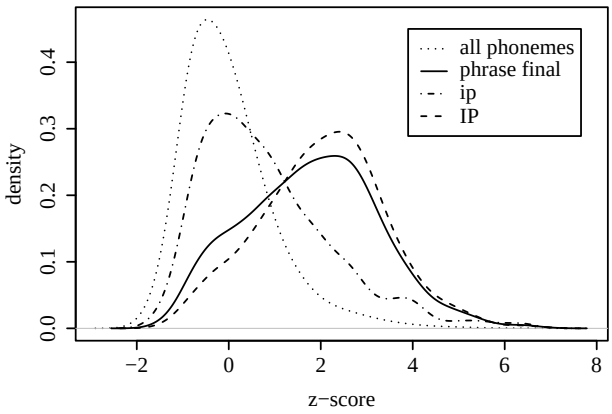

Figure 3: *Z-score density function for phrase-final phonemes (solid line), repeated from figure 2. The bump on the left slope is due to different distributions for phrase-final phonemes in ip (dot-dashed) and IP (dashed). For comparison, the z-score distribution for all phonemes is again indicated by the dotted line.*

Since we are focusing on temporal properties that are relevant for prosodic events, we are primarily interested in local changes in temporal properties in the vicinity of these prosodic events. This rules out more global temporal measures such as number of syllables per time unit or pause durations. To detect local changes, we need to examine the temporal properties of syllables and their constituents, i.e., syllable and segment durations.

If exemplars of phonemes are stored in memory, the distribution of instances of a particular phoneme according to their durations might look like the histogram of segment durations of [s] depicted in figure 1. Most exemplars of [s] have been realized with durations around 80 ms, but some are shorter than 50 ms, and several instances are up to 240 ms long. The position of a particular exemplar within the distribution can be seen as the distance of the respective exemplar from the distribution's mean. This distance can be interpreted as a measure for the extent of lengthening or shortening of the segment compared to other realizations of the same phoneme. Phoneme-specific constraints are visible in the distributions of exemplars for different phonemes: the distributions look similar but have different means and standard deviations. To assess the amount of lengthening or shortening pertaining to a particular exemplar not only with respect to other realizations of the same phoneme but with respect to all realizations of all phonemes, we adopt the concept of *z-scores* of segment durations from [12, 13] to eliminate phoneme-specific aspects of durations.

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 been used for prediction of segment durations in several text-to-speech systems. Reversing the original concept, we propose to take z-scores as a measure of a particular segment's position within the accumulation of all other exemplars. To answer question (i), the z-scores may then be regarded as the temporal dimension of the target regions in speech production, and the z-score of one specific exemplar would indicate its position on that dimension relative to all other realizations.

To address question (iii), viz. the identification of the target regions corresponding to the prosodic events, we have examined z-score distributions of segments in our corpus. We calculated z-scores for syllables and segments with mean durations and standard deviations taken from the corpus.

Effects of prosodic factors on segmental durations are clearly visible in the z-score distributions of phonemes in different prosodic contexts. Figure 2 shows that the distributions for all phonemes across all contexts (dotted line), for pitch-accented nuclei (dashed line), and for phrase-final segments (solid line) are different. The means of 0.00, 0.57 and 1.80, respectively, are pairwise significantly different (3 t-tests, $p \ll 0.0001$).

The distribution of z-scores for phrase-final segments in figure 2 shows a bump on the left slope. 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 3. Three more t-tests show that the means for intermediate phrases and intonation phrases (0.84 and 2.02 respectively) are significantly different from the overall means and from each other ($p \ll 0.0001$).

The influence of prosodic context is not limited to single phonemes. For instance, phrase-final lengthening can be observed for the z-score distributions of all segments in phrase-final syllables, as well as for the distributions for coda segments only. 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.

Turning to the z-scores of syllables, the problem arises that some syllables are extremely rare. For instance, 1,612 syllable types occur only once in our corpus, 5,433 syllables occur 5 times or less, 8,070 occur up to ten times. The smaller the number of instances of a particular syllable type, the less reliable is the z-score calculated on that basis, particularly because it is likely that there are still some segmentation errors in a corpus of this size even after manual checking. We have therefore only examined z-scores for the 326 syllables for which there are more than 20 realizations in our corpus. These syllable types add up to 22,638 syllable tokens. When looking at their z-score

distributions, we find the same effects of prosodic context as on the segment level. For instance, the overall mean z-score is 0.00, while the means for pitch-accented syllables and for phrase-final syllables are 0.65 and 1.41 respectively. The difference in means is again highly significant.

To summarize the experimental results presented in this section, and to answer question (iii), we have observed consistent effects of prosodic events on the z-scores both on the segment and on the syllable level. Different prosodic contexts produce significantly different z-score distributions. We conclude that the z-score distributions for units related to prosodic events can be regarded as target regions in the production and perception of temporal aspects of prosodic events.

4. Production of temporal properties of frequent and infrequent syllables

It is often assumed that the basic unit in articulation is the syllable. In [14] it is claimed that gestural scores for the articulation of syllables are stored in a mental syllabary. It is left open, however, whether gestural scores for *all* syllables are stored, even for languages like English or Dutch with approximately 12,000 syllables ([14], p. 111), or whether scores for infrequent syllables are computed on-line.

Under the assumption that accumulations of exemplars serve as targets in prosody production, there should easily be enough exemplars for frequent syllables to implicitly define target regions for these syllables independently of segmental target regions. In our corpus of 160 minutes of speech, the 326 most frequent syllable types, which occur, as mentioned above, more than 20 times each, account for 22,638 syllable tokens and thus cover approximately 67% of the corpus. These figures give an impression of the order magnitude of exemplars possibly stored in memory. We conclude that at least for very frequent syllables, there must be enough exemplars to be useful as a reference in speech production without resorting to the segment level.

As for determining which are the very infrequent syllables, we cannot rely on the frequencies observed in our corpus. Instead, the frequency classification of the syllables was based on syllable probabilities induced from multivariate clustering [15], which allows estimation of the theoretical probability even for unseen syllables. In [15], probabilities were obtained for a total of 41,711 German syllable types, ranging from $4.61*10^{-11}$ (for the syllable [R@sk]) to approximately 0.0259 (for the syllable [de:6]). Our corpus contains 3,793 syllable types, which means that there could be almost 38,000 syllable types missing in our corpus. Here the question arises how realistic the number of 41,711 is as an estimation of the number of different syllable types. It is worth noting that actually existing syllables can be found even among the least probable ones.

For comparison, Celex [16] contains only approximately 11,000 syllable types. But this is by far not the upper limit of different syllable types in German. For instance, our corpus contains syllables occurring in existing words that are not listed in Celex. Also, more words are used in German than can be expected to be listed in such a dictionary. German proper names for instance contain many types that are not listed in Celex. For example, according to a cliché, the most popular German surname is Schmidt, pronounced [SmIt], but [SmIt] as a syllable does not occur in Celex. This leads us to conclude that realistically, there are many more than 11,000 syllable types in German, with the upper limit being approximately 41,000. This means that the number of syllables missing from our corpus is somewhere between many more than 7,000 and 38,000.

Summing up these considerations, it can be said that there are many existing syllable types that are not represented by even one token in a speech corpus of 160 minutes of speech. It is therefore likely that for very infrequent syllables, there are not enough exemplars stored in memory to serve as a reference in speech production, and that the respective segments must be used as targets instead.

To assess the validity of this hypothesis, we exploited the specifics of our corpus. Since the corpus was designed for unit selection synthesis, one objective was to have a good coverage even of phonemes in infrequent contexts, and to have at least the same coverage as a diphone corpus. Therefore, after optimizing coverage for phoneme/context vectors, sentences containing diphone types that were not found in the corpus were manually added. As a consequence, the corpus differs from a randomly collected database in that it exhibits an unusual syllable frequency distribution with disproportionately many instances of some otherwise infrequent syllables. This allowed us to compute z-scores for realizations of these syllables even though they should not be represented by a sufficient number of exemplars in the speaker's memory.

Our assumption was that if, because of a lack of an appropriate syllable-level target, a very infrequent syllable is produced by concatenating segments, then the z-score of the resulting realization of the syllable should depend on the z-scores of the involved segments. There should be less dependency for very frequent syllables, because then the speaker does not access exemplars of the involved segments but directly uses exemplars of the syllable as a reference. To put it more simply, one could say that if a speaker intends to articulate a syllable lengthened by a z-score of 2, but does not have enough exemplars of the syllable to use as a reference, he will articulate the syllable using exemplars of the involved segments with a z-score of 2. Consequently, we expect more variation for frequent syllables than for infrequent syllables when looking at the relationship between syllable z-scores and the z-scores of the corresponding segments. This is reminiscent of results reported by [9], who found that syllables expected to be stored in the syllabary exhibit more coarticulation than rare syllables, which are assumed to be assembled on-line from smaller units.

To test our hypothesis, we calculated two linear regression models, one for very frequent and one for very infrequent syllables. Both models predict the syllable z-sore from the mean z-score of the involved segments. The criterion for very infrequent syllables was a probability of less than 0.00005 according to [15], and of more than 0.01 for very frequent syllables. To obtain reliable z-scores, we took only those syllables into account for which we had more than 20 realizations in our corpus. There were 114 very frequent and 16 very infrequent syllable types which met our requirements, adding up to 12,278 and 471 tokens, respectively. Figure 4 shows the mean z-scores of involved segments plotted against syllable z-scores for frequent (left panel) and infrequent (right panel) syllables.

For the two linear regression models, we obtained residual standard errors of 0.400 for frequent and 0.365 for infrequent syllables. This indicates that indeed the model for frequent syllables is less accurate in predicting syllable z-scores from mean segment z-scores, confirming that there is a stronger linear dependency between the two values for infrequent syllables. To determine whether this difference is significant, we applied the Bartlett test for homogeneity of variances to the residuals. The test confirmed that the variances are significantly different ($p \ll 0.0001$).

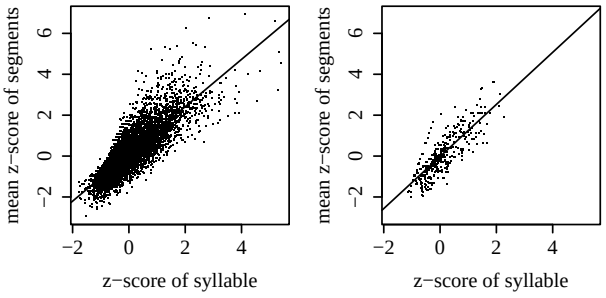

Figure 4: *Mean z-scores of segments within a syllable plotted against z-score of the syllable for frequent (left panel) and infrequent (right panel) syllables.*

5. Conclusions

We have presented an exemplar-theoretic interpretation of Guenther and Perkell's speech production model for the prosodic domain. We have suggested that z-scores of segment and syllable durations are the temporal dimension in the perception and production of prosody, and that z-score distributions are used as target regions in the production of temporal properties of prosodic events. This view is motivated by the fact that, if phonetic details of the exemplars are stored in memory, speakers have access to the durations of the stored exemplars and are likely to use them as a reference in production. Moreover, we have presented experimental results confirming that realizations of segments and syllables in different prosodic contexts show significantly different z-score distributions. We conclude that z-scores are an appropriate perceptual measure to make the target regions for different prosodic events sufficiently distinct from each other.

The starting point of our second experiment was the assumption that accumulations of syllable exemplars serve as a reference in production. We have argued that this cannot be the case for very infrequent syllables, which consequently have to be assembled from smaller units. This is supported by our experiments, which indicate that the relationship between syllable z-scores and the z-scores of the corresponding segments is significantly stronger for infrequent than for frequent syllables.

Taken together, our experiments on the temporal properties of prosodic events have provided further evidence that accumulations of exemplars implicitly define perceptual target regions in speech production.

6. Acknowledgments

This work was funded by the German Research Council (DFG, Grant DO 536/4-1) and 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.

7. 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 *Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark)*, 2001, vol. 1, pp. 665–668.

[4] B. Möbius and G. Dogil, "Phonemic and postural effects on the production of prosody," in *Proceedings of the Speech Prosody 2002 Conference*, B. Bel and I. Marlien, Eds., Aix-en-Provence, 2002, pp. 523–526.

[5] 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.

[6] 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.

[7] A. Schweitzer and B. Möbius, "On the structure of internal prosodic models," in *Proceedings of the 15th International Congress of Phonetic Sciences (Barcelona)*, 2003, pp. 1301–1304.

[8] W. J. M. Levelt and L. Wheeldon, "Do speakers have access to a mental syllabary?," *Cognition*, vol. 50, pp. 239–269, 1994.

[9] S. P. Whiteside and R. A. Varley, "Dual-route phonetic encoding: Some acoustic evidence," in *Proceedings of the 5th International Conference on Spoken Language Processing (Sydney)*, 1998, vol. 7, pp. 3155–3158.

[10] J. Mayer, "Transcription of German intonation—the Stuttgart system," Technical Report, Institute of Natural Language Processing, University of Stuttgart, 1995.

[11] R-Project, "The R project for statistical computing," Available online at [http://www.R-project.org/], 2001.

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

[13] 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.

[14] W. J. M. Levelt, "Producing spoken language: a blueprint of the speaker," in *The Neurocognition of Language*, C. M. Brown and P. Hagoort, Eds., pp. 83–122. Oxford University Press, Oxford, UK, 1999.

[15] K. Müller, B. Möbius, and D. Prescher, "Inducing probabilistic syllable classes using multivariate clustering," in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, 2000, pp. 225–232.

[16] H. Baayen, R. Piepenbrock, and L. Gulikers, "The CELEX lexical database—Release 2," CD-ROM, 1995, Centre for Lexical Information, Max Planck Institute for Psycholinguistics, Nijmegen.