AN EXEMPLAR-THEORETIC ACCOUNT OF SYLLABLE FREQUENCY EFFECTS

Michael Walsh Hinrich Schütze Bernd Möbius Antje Schweitzer

Institute for Natural Language Processing, University of Stuttgart, Germany
firstname.lastname@ims.uni-stuttgart.de

ABSTRACT

This paper presents an exemplar-theoretic computational model of syllable frequency effects which yields simulation results in keeping with experimental results found in the literature. The argument posited here is that syllable duration variability is a function of segment duration variability for infrequent syllables. However syllable duration variability for frequent syllables cannot be predicted from segment duration variability. The simulation results support the hypothesis that frequent syllables are accessed as units whereas infrequent syllables are more likely to be produced on-line from exemplars of their constituent segments.

Keywords: Exemplar theory; syllable frequency effects; computational modelling.

1. INTRODUCTION

Recent research in speech perception has provided considerable evidence indicating that the perception process is partly facilitated by accessing previously stored exemplars rich in phonetic detail. That is speakers accumulate exemplars over time and compare input stimuli against them. Exemplars are categorised into clouds of memory traces with similar traces lying close to each other while dissimilar traces are more distant. Exemplar-theorists posit that language comprehension and production are achieved via operations on these stored memory traces and there is now considerable evidence that vast amounts of linguistic tokens to which hearers are exposed are stored in memory. The exemplar approach is particularly attractive as it has been successful in treating phenomena across linguistic domains which more abstractionist approaches have found problematic [1],[2], [3],[4], [5],[8].

Schweitzer and Möbius [9] note that from a production perspective these exemplars could function as targets or plans of articulation, and that if this is the case then speakers should have a significant number of exemplars for high frequency syllables, which would then act as a production target region, and a small or negligible number of exemplars for low frequency syllables. Consequently they argue that low frequency syllables would have to be computed online from exemplars of their constituent segments or segment clusters [6]. They predicted, and observed, greater variation in duration for frequent syllables than for infrequent syllables when looking at the relationship between syllable duration z-scores (measure of standard deviations from the mean) and the duration z-scores of the constituent segments. The research presented here employs a computational model which goes some considerable way to explaining the underlying mechanism which causes this effect.

This paper presents an exemplar-theoretic account of the syllable frequency effects found in Schweitzer and Möbius [9] and employs a unique model of constituency interactions across the segment and syllable domains. The next section presents both an overview of the model and a detailed description of how it accounts for the effects discussed above. This is followed by sections 3. and 4., which present insights gained from the model and possibilities for future research.

2. SYLLABLE FREQUENCY EFFECTS MODEL

2.1. Model architecture

The aim of the research presented here is to elucidate, using a computational model, the underlying mechanism for the effect found in Schweitzer and Möbius [9]. In other words to explain how variability in syllable duration across syllable frequency categories (frequent and infrequent) can be explained using a computational process. The model is intended to represent a competition between syllables accessed as units and those that are produced as a result of accessing the exemplar clouds of their constituent segments. The intuition here is that until a particular threshold is reached, through frequent exposure to co-occurring segments, the duration of a particular syllable will reflect the sum of the durations of its constituent segments. Once the threshold is passed the constituent segments will tend to behave more as a single unit, a syllable. The computational model discussed below is applied to the data in the corpus employed by Schweitzer and Möbius. The corpus is a single-speaker speech database for unit selection speech synthesis, recorded by a professional male speaker of German, and contains approximately 160 minutes of speech (2601 utterances with 17489 words, 33800 syllables and 94300 segments). The corpus was manually annotated on the segmental, syllabic, word and prosodic levels. In their experiments Schweitzer and Möbius extracted the 326 most frequent syllable types, each with more than 20 tokens. In total this accounted for 22,638 syllable tokens, covering approximately 67% of the corpus. These syllable types were matched against categorisation criteria for frequency and infrequency based on analysis by Müller et al. [7], the criteria being that very infrequent syllables have a probability of less than 0.00005 and very frequent syllables have a probability in excess of 0.001. This further refinement yielded 114 very frequent and 16 very infrequent syllable types. According to Müller et al. syllable frequency information was based on syllable probabilities induced from multivariate clustering, which allows an estimation of the theoretical probability even for unseen syllables.

On each iteration of the model programme a syllable is produced. The model is initially seeded with exemplar duration values using the mean duration values of the segments in the corpus. Within the context of exemplar theory this can be viewed as segments possessing exemplar clouds along the duration dimension, i.e. they represent a memory of the previous occurrences of these particular segments and how long they lasted (syllables in the model also possess exemplar clouds). On each iteration a syllable, either frequent or infrequent, is selected for production according to two competing production pathways known as a *composite* production pathway and a *unit* production pathway. According to the composite pathway a duration for each of the syllable's constituent segments is randomly selected from their respective exemplar clouds and random noise is added, to reflect motor/articulatory perturbations. This syllable is known as a *composite* syllable. Its duration is the sum of the durations of its constituent segments. In parallel, an identical *unit* syllable is also selected for production. A duration for this syllable is randomly chosen and modified by noise in a similar fashion to the segment selection for the composite syllable. In the initial stage of the model however the unit cloud for a given syllable is empty, in which case the unit pathway does not produce output. The unit syllable has an associated activation based on the density of its cloud. If the unit syllable reaches a certain minimum activation it is chosen for production, otherwise the composite syllable is chosen. The winner is then produced, perceived, and stored. The architecture of the model is presented in figure 1

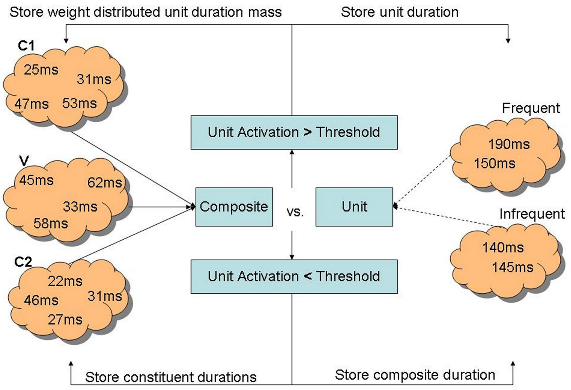

Figure 1: Architecture of the model.

2.2. The model in action

This model has an underlying hierarchical organisation of syllable structure based on Selkirk [10] which divides the syllable into onset and rhyme, and divides the rhyme into peak (or nucleus), and coda. Given the balanced nature of the corpus the syllables modelled here are of a variety of forms, including CV, CVC and more complex structures. For the purposes of illustration the model is discussed from the perspective of CVC production.

According to the composite pathway each iteration of the model selects a CVC syllable for production. A duration is then randomly selected for each constituent segment from their respective duration clouds. These clouds are initially seeded with the mean value from the corpus for each segment. It is important to note that these mean values are calculated from the corpus as a whole, not merely from the duration of segment tokens found in the frequency bins discussed above. Random noise (an initial assumption) is then added to each duration value:

(1) $C1_{dur} = C1_{ex} + e_{rd}$

(2) $V_{dur} = V_{ex} + e_{rd}$

(3) $C2_{dur} = C2_{ex} + e_{rd}$

The net effect is that each segment is either lengthened or shortened (by up to 5%) depending on the effect of the noise. The introduction of noise is intended to reflect the high degree of variability in speech production. The duration of the *composite* syllable is the sum of the segment durations.

(4) $CompSyll_{dur} = C1_{dur} + V_{dur} + C2_{dur}$

Recall that Schweitzer and Möbius [9] found evidence to suggest that infrequent syllables were constructed on-line from their constituent segments whereas frequent syllables were accessed as units, perhaps from a mental syllabary. According to the unit pathway, with each iteration of the model a *unit* syllable, nominally identical to the composite syllable, is also selected for production. The duration of the unit syllable is selected randomly and noise is added in a similar fashion to that described for the segments. As with the segments it is important to note that the level of additional noise is commensurate with the size of the unit. Thus at this point in the execution of the model there are two competing syllable hypotheses, one *composite* and one *unit*. The determining factor in deciding between the two is the level of activation of the unit syllable. In the simulations presented here the model initially has no exemplars of either the frequent or infrequent syllable stored as a unit. Thus the unit syllable activation is zero and no unit syllable is produced. However as the model is exposed to syllable productions the sizes of the syllable clouds (frequent and infrequent), and their corresponding activations, increase.

In the event that the unit syllable possesses an activation less than a unit-threshold θ, the durations of the composite syllable's constituent segments are stored in their respective duration clouds, and the composite syllable's total duration is stored in the duration cloud of its unit syllable equivalent. This represents the production, and perception, of a syllable using its constituent segments rather than accessing the syllable as a unit stored in exemplar memory.

On the other hand, if the activation of the unit syllable is greater than the threshold, then its duration is stored in its duration cloud and its duration mass is divided into the duration clouds of its constituent segments in a manner corresponding to their typical influence on syllable duration. This is achieved through the following normalisation process:

(5) $$Seg_{i\_weight} = \frac{Seg_{i\_mean}}{\sum_{j=1}^{n} Seg_{j\_mean}}$$

(6) $Seg_{i\_duration} = UnitDuration * Seg_{i\_weight}$

The reason for the distribution of the unit syllable's duration mass to the segments which would normally compose it (despite the fact that it is operating as a unit) is that intuitively, from an exemplar-theoretic perspective, the production/perception of a syllable should result in greater activation of that syllable's exemplar cloud, and in greater activations of the clouds of the segments that can be perceived within the syllable.

The size of all the exemplar duration clouds increases over multiple iterations of the model.

2.3. Experiment

The algorithm described above was executed to yield *n* productions per syllable based on the prior probabilities of frequency presented in the previous section. This was performed with a view to establishing a critical mass of durations which would facilitate inspection of the model. In order to carry out the inspection a further 500 iterations were performed, per syllable, using the enlarged clouds provided by the pre-inspection execution. In other words the model was forced to produce, on the basis of duration results yielded by the pre-inspection execution, 500 unit syllable tokens per syllable type (across both frequent and infrequent types), and 500 composite syllable equivalents. The threshold θ was manually set to 100 (see section 3. for further discussion). These inspection durations did not enlarge the original pre-inspection duration clouds, that is the resulting duration clouds for the frequent and infrequent syllables were stored separately. The duration z-score of each syllable token was calculated and plotted against the mean z-scores of the involved segments in the composite equivalent, where the z-score of a token (syllable or segment) is given by:

(7) $$Token_{z-score} = \frac{Token_{dur} - Type_{mean}}{Type_{standard-deviation}}$$

Aggregate results are presented for both frequent and infrequent syllables in figure 2.

As a function of the fact that the infrequent syllable is not produced in sufficient numbers to exceed the threshold the z-scores of each infrequent syllable will, to a significant degree, reflect those of the constituent segments as the durations in the syllable cloud will essentially be the durations of composite syllables. Given that each composite syllable is composed of segments to which noise has been added, it is likely that the net effect on the syllable duration will be small as the addition of noise to each segment will to some extent cancel each other out. In other words one segment might grow longer whereas another grows shorter. Overall the standard deviation (on which z-scores depend) of the composite syllable will not be that large, and, since the duration cloud of an infrequent syllable relies heavily on composite syllable durations the standard de-

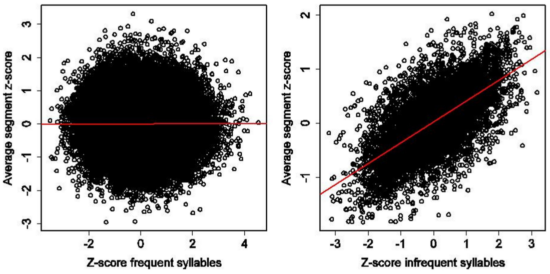

Figure 2: Mean z-scores of segments within a composite syllable plotted against z-score of the the unit syllable for frequent (left panel) and infrequent (right panel) syllables.

viations of infrequent syllable durations will be similarly restricted. Hence the z-score of an infrequent syllable will be quite well predicted by the z-scores of the syllable's segments, as illustrated in figure 2 (right panel). The syllables of the frequent category however exhibit greater variability. This is due to the fact that after a number of productions they possess activations greater than θ and hence much of the durations in their clouds correspond to previous *unit* durations with added noise. The noteworthy point here is that the noise is added to the syllable as a whole, not to its constituent parts. Thus there is no cancellation effect, and the syllable unit is more likely to vary more significantly from the mean. This is in keeping with the findings of Schweitzer and Möbius [9], and the dual-pathway theory posited by Whiteside and Varley [11].

3. DISCUSSION / CONCLUSION

The computational model presented here represents a novel exemplar-theoretic account of the syllable frequency effects discussed above and achieves this through a simple competitive model and threshold. One challenge for exemplar theory is to explain how exemplars of constituents interact with exemplars of compositions of several constituents into larger units. The model proposed here is the first to address this issue. Hence, while the architecture of the model is very preliminary it both models to a significant extent empirically observed syllable duration variability and represents a promising first step towards an exemplar-theoretic account of constituency.

4. FUTURE WORK

One difficulty with the model is the manual selection of the threshold value θ. The lower the value of θ the greater the variability at the *unit* level. However, although the threshold value employed here effectively prevents the syllables of the infrequent category from ever reaching *unit* status (i.e. infrequent syllables are not produced in sufficient number to exceed the threshold) and presents a clean separation with respect to the variability of infrequent and frequent syllables, the key point remains irrespective of threshold value: if syllable frequency exceeds the threshold syllable duration variability will increase. Nevertheless it would be desireable for the model to converge automatically on a threshold value. Other opportunities for future research include experiments on segment context effects, and examinination of exemplar cloud densities and the manner in which exemplar clouds are formed [8].

5. REFERENCES

[1] Abbot-Smith, K., Tomasello, M. 2006. Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review* 23, 275–290.

[2] Bod, R. 2006. Exemplar-Based Syntax: How to Get Productivity from Examples. *The Linguistic Review* 23, 291–320.

[3] Bybee, J. 2006. From usage to grammar: the minds response to repetition. *Language*, 84, 529–551.

[4] Goldinger, S. D. 1998. Echoes of Echoes?. An Episodic Theory of Lexical Access. *Psychological Review* 105, 251–279.

[5] Johnson, K. 1997. Speech perception without speaker normalization: An exemplar model. In: Johnson, K., Mullennix, J. W. (eds), *Talker variability in speech processing*. Amsterdam: Benjamins, 145–165.

[6] Levelt, W. J. M. 1999. Producing spoken language: a blueprint of the speaker. In: Brown, C. M., Hagoort, P. (eds) *The Neurocognition of Language*. Oxford, UK: Oxford University Press, 83–122.

[7] Müller, K., Möbius, B., Prescher, D. 2000. Inducing Probabilistic Syllable Classes Using Multivariate Clustering. *Proc. 38th ACL Meeting* Hong Kong, 225–232.

[8] Pierrehumbert, J. B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In: Bybee, J., Hopper, P. (eds), *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins, 137–157.

[9] Schweitzer, A., Möbius, B. 2004. Exemplar-Based Production of Prosody: Evidence from Segment and Syllable Durations. *Proc. of Speech Prosody 2004* Nara, Japan, 459–462.

[10] Selkirk, E. O. 1982. The Syllable. In: van der Hulst, H., Smith, N. (eds) *The structure of phonological representations (Part II)*. The Netherlands: Foris Publications, 337–383

[11] Whiteside, S. P., Varley, R. A., 1998. Dual-route phonetic encoding: Some acoustic evidence. *Proc. of the 5th International Conference on Spoken Language Processing* Sydney, Australia, 3155–3158.