

Victor Kuperman¹, Mirjam Ernestus^{1,2}, R. Harald Baayen³

¹Radboud University Nijmegen, ²Max-Planck-Institute for Psycholinguistics, ³University of Alberta
victor.kuperman@mpi.nl, mirjam.ernestus@mpi.nl, harald.baayen@ualberta.ca

Frequency Distributions of Uniphones, Diphones and Triphones in Spontaneous Speech

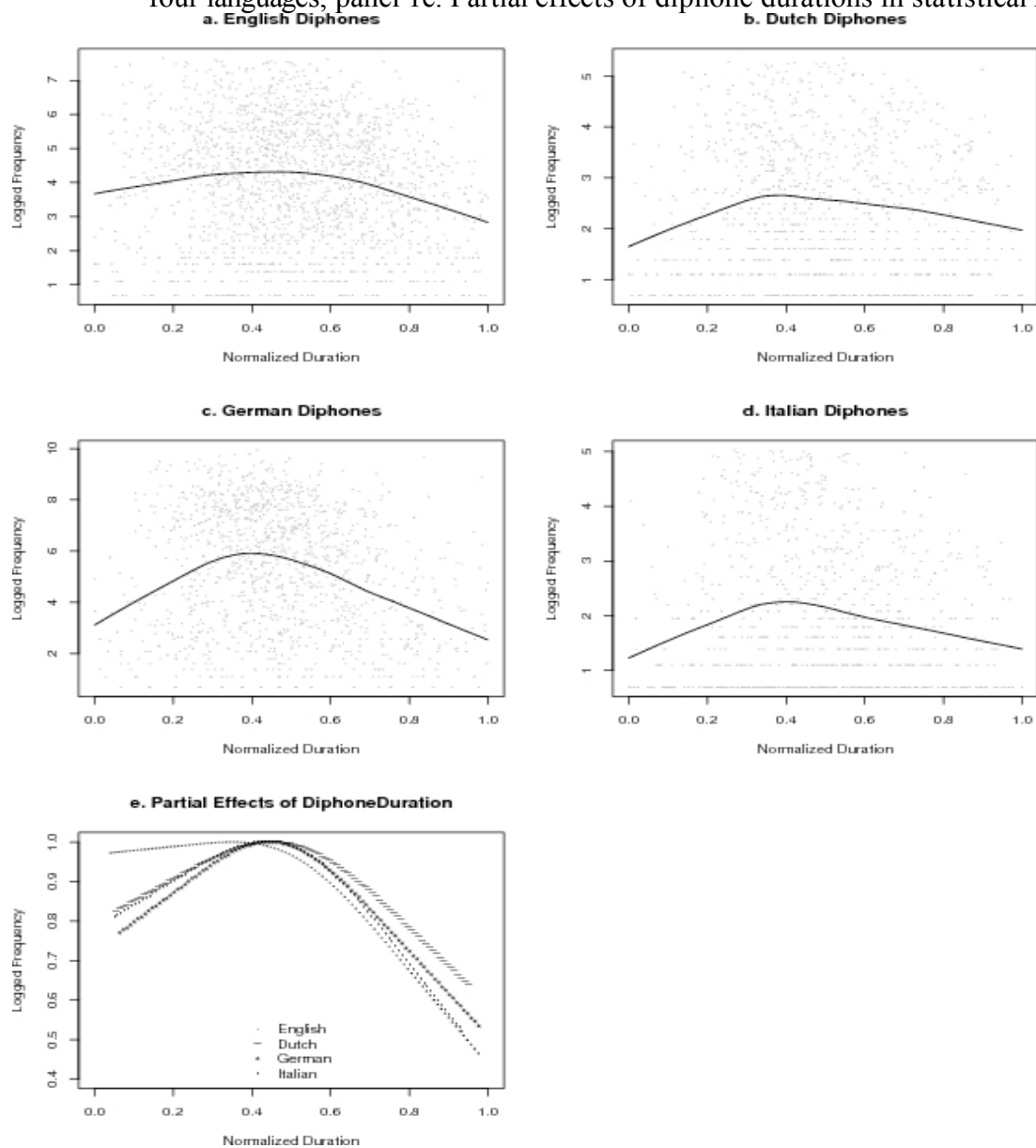
Starting with Zipf (1929; 1935), the overall frequency of occurrence of a speech unit has been argued to enter into a negative (linear or nonlinear) relation with the “degree of complexity” of that unit, which would include its acoustic duration. The more frequent, or otherwise predictable, a speech unit (an n-phone, a syllable, or a word) is, the easier its acoustic realization is claimed to be (cf. Jurafsky et al., 2001). This approach only takes into account the speaker-oriented principle of least effort, but fails to recognize the listener-oriented principle of maximal perceptual contrast as an additional factor that codetermines the relation between frequency of occurrence and production effort. We make the simplifying assumption that acoustic duration of a speech unit reflects (on average, among many other factors) the relative ease of articulating that unit. We hypothesize, along with Zipf (1935), that phonemic sequences with difficult pronunciation will be of a low frequency of use, due to the increased costs for the speaker. In addition, we argue that sequences with extremely easy articulation (e.g., very short ones) may be problematic for the listener and thus be of low frequency in the language as well. The demands of the speaker and the listener may be optimally satisfied by those sequences that are relatively easy to produce and also relatively easy to perceive, that is, by n-phones in the middle of durational range.

In the present paper we tested these hypotheses and explored the relation between frequency of occurrence and acoustic duration of uniphones, diphones and triphones in several languages with different phonemic inventories and different phonologies, namely, English, Dutch, German and Italian. We opted for exploring the relation in spontaneous speech, as several studies show that variation of acoustic duration is larger in this speech variety than, say, in careful speech (e.g., Johnson, 2004). We based our analyses on large (sub)corpora of spontaneous speech in those languages: The Buckeye Speech Corpus for American English, the IFA Spoken Language Corpus of Dutch, modules Verbmobil-I and -II of the Bavarian Speech Archive for German and the Spoken Italian Varieties Archive for Italian. The speech files of these corpora come with transcriptions at the phone level. Moreover, these transcriptions provide temporal boundaries for each phone in the signal (i.e., phone-level aligned segmentation). Except for the IFA corpus, which was labeled manually, all collections were labeled automatically with subsequent manual verification of the alignment.

We defined diphones (or triphones) as sequences of two (or three) phones without an intervening pause, end of turn, noise, laughter, a non-speech sound, a phone marked as incomprehensible by the transcribers, or a segment extraneous to the phonetic inventory of that language. Notably, in identifying the diphone or triphone sequences we ignored word or utterance boundaries. This approach treats the speech signal as a continuous stream, in which word segmentation is not a given, but rather a task for the listener.

Across the four languages, we found consistent patterns in frequency distributions of diphones and triphones, such that the shortest and the longest n-phones had the lowest frequency of occurrence. In other words, the functional relation between (log-transformed) frequency of occurrence of diphones and triphones as a dependent variable and their (log-transformed) acoustic duration as a predictor, has an inverse-U, concave shape, rather than the monotonically decreasing shape predicted by Zipf’s approach, see Fig. 1.

Figure 1, panels a-d: Log frequency of diphones as a function of (normalized) diphone duration across four languages; panel 1e: Partial effects of diphone durations in statistical models.



This set of findings is in line with our hypothesis. For each dataset (e.g., diphones and triphones in each language) we compared the performance of the Zipfian models (that predict a monotonic negative relation) and our models (which predict an inverse-U shape relation). To this end, we used multiple regression models while modeling non-linearities with the restricted cubic splines method. In all cases, our models explained more variance than models based on Zipf's predictions: The average R^2 value of our models was 2.6%, while the average R^2 value of the Zipfian models was 0.2%. The binomial sign test shows that the probability of our models outperforming their counterparts by chance in eight model pairs (four pairs for diphones and four for triphones) is less than 0.008.

N-phone duration can be influenced by a number of factors, including word frequency and speech rate. Can the patterns we observed be explained by those factors? We fitted mixed-effects multiple regression models to each dataset with n-phone duration as a dependent variable, with as fixed effects word frequency, the sum of mean durations of uniphones in the n-phone, mutual information of

uniphones, the position of the n-phone in the word and the phrase, and with speaker as a random effect. We then considered the residuals of those models as a measure on n-phone duration, from which other factors of influence were regressed out. Finally, we considered n-phone frequency as a function of the residual n-phone duration to test the performance of our models, and the residual n-phone duration as a function of n-phone frequency to test Zipfian models. The effects of predictors on corresponding dependent variables were statistically significant in all models. Crucially, the advantage that our models showed in fitting the mean durations of diphones and triphones across languages is still preserved when the influence of multiple other predictors is statistically partialled out.

We also tested for whether the inverse-U shape patterns might be an artifact of the so-called sampling error and in fact represent a normal distribution of data points around the mean n-phone duration. For each dataset, we simulated 5000 samples from the normal distribution with the size, mean and the standard deviation equal to those observed in the distribution of residual n-phone durations in the given dataset. The Kolmogorov-Smirnov test invariably showed that the simulated and the observed distributions are significantly different across datasets. We also used the one-sample t-test to estimate the probability that data points in the observed distribution follow the normal distribution (with the mean and standard deviation equal to those of the observed distribution). For over a half (over two-thirds) of data points in each dataset this test showed that their probability of being part of the normal distribution is above the 5% (1%) level of significance. We conclude that the observed distribution patterns cannot be fully accounted for by the statistical fact that values closer to the population mean tend to have higher frequency of occurrence than extreme values.

In order to obtain a better understanding of the observed cross-linguistic patterns, we implemented the hypothesis about the interacting demands of efficient speech production and effective speech comprehension mathematically in a theoretical function based on Job and Altmann (1985). The function is based on assumptions that (a) the relative amount of change in frequency is proportional to the change in the difference in efforts for the interlocutors and (b) language as a self-organization system tends to reach an equilibrium between conflicting processing demands, such as demands of easy production and easy comprehension of speech. The function provides good fits to the distributions of frequency of diphones and triphones over their acoustic durations supporting our hypothesis.

Our data document the existence of consistent frequency distribution patterns in several languages, as revealed via large corpora of spontaneous speech. These patterns demonstrate the emergence of global cross-linguistic regularities from the individual instances of communication that operate on a microscopic scale and provide evidence for processes of self-organization in language.

References

- Job, U. and Altmann, G. 1985. Ein Modell fuer anstrengungsbedingte Lautveraenderungen. *Folia Linguistica Historica* , VI:401-407.
- Johnson, K. 2004. Massive reduction in conversational American English. In *Spontaneous speech: data and analysis. Proceedings of the 1st session of the 10th international symposium* , pages 29-54, Tokyo, Japan. The National International Institute for Japanese Language.
- Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In *Frequency and the emergence of linguistic structure*, Bybee, J. and Hopper, P. (eds.), pages 229-254. John Benjamins, Amsterdam.
- Zipf, G. K. 1929. *Relative frequency as a determinant of phonetic change*. Harvard Studies in Classical Philology , 15:1-95.
- Zipf, G. K. 1935. *The Psycho-Biology of Language*. Houghton Mifflin, Boston.