Phone Elasticity in Disfluent Contexts

Jana Vosse[1], Simon Betz[123], Petra Wagner[13]

[1] *Phonetics and Phonology Workgroup, Bielefeld University, Bielefeld, Germany*

[2] *Dialogue Systems Group, Bielefeld University, Bielefeld, Germany*

[3] *CITEC, Bielefeld, Germany*

Speech Disfluencies

Disfluencies in speech are instances of hesitation or correction that affect both the speaker and the listener. Typical surface forms of disfluencies are filled pauses (such as *uh, uhm*), silences at places in the utterance where syntax would not predict them, or repetitions of parts of the utterance (like *(I mean + I mean) we should go*). Disfluencies carry a folk notion of erroneousness or badness. Voice coaches often advise to eliminate disfluencies such as filled pauses from speech. Research in the past decades began to put disfluencies into a more positive light, as it was discovered that they provide the listener with important meta-information about the current speech from their interlocutor. Listeners can infer difficulties of reference or problems of forming the right speech plan when they hear the speaker hesitate (See [1] for an overview). We largely follow the view of [2], who formulated the new perspective on disfluencies as them being solutions, not problems. Recent studies started investigating the question if disfluencies are suitable for human-machine communication as well, cf. [3]. One of the possible places of applicability is incremental spoken dialogue systems. Compared to non-incremental systems, these systems can prepare their response to human speech input in real-time while the user is speaking and can thus respond with only milliseconds delay, cf. figure 1. This, however, can cause problems such as running out of things to say or uttering erroneous output which needs correction. It is an open question whether disfluencies can remedy these issues by equipping incremental spoken dialogue systems with human features of dialogue management in order to solve these problems in addition to sounding more natural. With the aim in mind to improve the quality of incremental spoken dialogue systems, we currently investigate the issue of disfluent lengthening. In many cases of hesitation, speakers slow down their speech rate momentarily. Sometimes a brief and hardly noticeable period of slower speech is all there is in the surface of the signal. The interesting aspect of lengthening for dialogue systems is that it can buy valuable dialogue time without being detrimental for sound quality [4]. In a previous study, we investigated in a corpus study of spontaneous human-human dialogue where in the syllable lengthening preferably manifests itself [5]. This study aims at the last piece of information required to synthesize lengthening properly, that is, how to distribute the duration increase resulting from the slowdown of speech over the individual sound segments of the syllable.

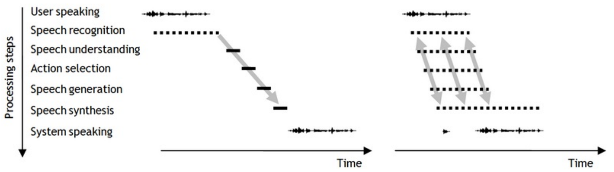

Abbildung 1: Non-incremental vs. incremental dialogue systems; figure from Skantze & Hjalmarsson 2013 [3].

Elasticity Hypothesis

The idea of using a segment's elasticity to predict its duration is first introduced by [6]. Here, elasticity refers to a segment's flexibility in duration: a highly elastic segment, such as a long vowel or a liquid, can be lengthened or compressed to a great extent, whereas less elastic elements, such as short vowels or stops, have a smaller range of possible durations. The following formula adapted from [6] represents the duration of a bi-segmental syllable as composed of the mean duration plus k times the standard deviation of each of its segments.

$$syldur_{2\ segs} = (\mu_{seg1} + k * \sigma_{seg1}) + (\mu_{seg2} + k * \sigma_{seg2}) \quad (1)$$

Here, every bracket term calculates the duration of a syllable's segment and represents its measure of elasticity. k is a factor within this measure of elasticity and is segment-independent. That means, k is constant within a single syllable, but may have different values for different syllables. From a more formal point of view, k describes the average Z-Score of the phonemes within a syllable. That is, k delivers information about the average normalized distance of the syllable's segments from their mean duration.

$$K_{syll} = \frac{(dur_{syll} - \Sigma_{i=1}^{N} \mu_i)}{\Sigma_{i=1}^{N} \sigma_i} \quad (2)$$

It has to be stated that the elasticity hypothesis in its 'strong' form, as depicted in (1), is not tailored to represent special contexts. Campbell [6] notes that for such environments, specialized 'weak' forms of the hypothesis have to be constructed:

"Weaker forms of the elasticity hypothesis would state that statistics have to be gathered separately for syllables in different positions in the sentence (e.g. finally versus non-finally), for segments in different parts of the syllable (e.g. for those in the onset and ryme), and in different phonetic contexts (e.g. for vowels before voiced and unvoiced stops)."

Following our clarification of the context-sensitivity of conversational phenomena such as disfluent lengthening, we consider a context-sensitive approach as promising and develop a weak elasticity hypothesis specified for disfluent lengthening. With this, we aim to provide a more accurate segment duration prediction.

Methods

In order to evaluate the prediction accuracy of the strong, baseline form and the weak, specialized form of the hypothesis, we first compile a subset of GECO, a corpus of spontaneus German speech [7]. This subset contains all syllables of the corpus that have been labeled as containing disfluently lengthened segments. Second, we use the baseline elasticity hypothesis (1) to predict the segmental durations of the subset. The total syllable duration and the means and standard deviations for each segment based on the full corpus serve as input. The equation can then be solved for k and each segment's duration can be calculated. Third, we repeat this process, but this time with the specialized elasticity hypothesis, which differs only in one aspect: the means and standard deviations are now based on the subset rather than the full corpus. Finally, we repeat the last step, but this time we compare predictions to instantiations in a different corpus. To evaluate this approach, we calculate root mean square errors between the measured durations and the two types of predicted durations. The question is, whether this specialized form of the hypothesis outperforms the baseline. It is based on the same kind of data it is supposed to predict; however the formula uses abstracted mean values to predict concrete values, which is not to be confused with a machine-learning technique trained on the same data it is to predict. If the specialized form outperforms the baseline, future work on this matter can use small specialized corpora rather than costly-to-compile big data.

Evaluation

[8] introduced the concept of K-Deviation, which is a measure of a syllable's deviation from the elasticity hypothesis. It is defined as the root mean square Z-Score deviation from the syllable's K-Score. This deviation is calculated over all phones within the syllable. Thus, a syllable's perfect fit to the elasticity hypothesis would result in zero K-Deviation, which is always the case if the present syllable is monophonic. (Equation (3) is taken from [8]).

$$Kdev_{syll} = \sqrt{\frac{\Sigma_{i=1}^{N}(K_{syll} - Z_i)^2}{N}} \tag{3}$$

[8]'s approach bases on their observation that syllables do not exactly fit the elasticity hypothesis in reality. The implication that all phones within a syllable have the exact same Z-Score, and therefore the syllable K-Score is identical to all phone Z-Scores within the syllable, could not be confirmed.

The concept of K-Deviation has the following implications for our investigation: First, we follow [8]'s approach and take the root mean square deviation - also called root mean square error - as a measure of deviation for our investigation. As we are primarily interested in deviation regarding concrete segment durations, our formula differs from the K-Deviation in terms of the applied variables:

$$Ddev_{seg} = \sqrt{\frac{\Sigma_{i=1}^{N}(D_{seg} - D_{seg\ pred})^2}{N}} \tag{4}$$

Using the actual (D_{seg}) and the predicted ($D_{seg\ pred}$) syllable duration, we calculate the root mean square duration deviation from the true syllable duration.

Second, the gold standard for our hypothesis testing should not be a perfect fit of a syllable into the elasticity hypothesis, as this does not correspond to natural language. Hence, a small root mean square error is to be aimed, but it has to be considered that a result of zero cannot be attained within natural speech.

Results

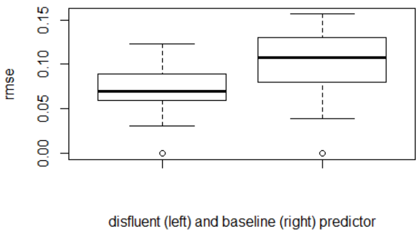

Abbildung 2: Deviation from observed duration is smaller when predictions are based on disfluent data.

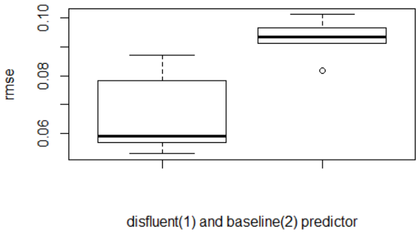

Abbildung 3: Preliminary tests on a different corpus with the same underlying means yield similar results.

For each of the 750 syllables in our dataset that contain a phone with disfluent lengthening, we compare each of its phones' durations to the durations predicted using the disfluency-based form of the hypothesis and the baseline form. As can be seen in figure 2, the disfluency-based form of the elasticity hypothesis exhibits a lower root mean square error, i.e. it differs less from the observed duration distributions. Not only the disfluent phones are predicted more accurately, also each non-lengthened phone is predicted more accurately using the disfluent form. The performance of the disfluent form is significantly better than the baseline form: paired t(30) = -6.7, $p < 0.001$. To check whether this prediction method is transferable to other data, we conducted a preliminary check on a different set. Using the same means and standard deviations from GECO as before, we now extract instances of disfluent lengthening from the DUEL-Dreamappartment corpus [9] and investigate which form of the elasticity hypothesis predicts these instances more accurately. As can be seen in figure 3, the tendency is the same: The specialiced form exhibits a lower rmse score. Since the lengthening analysis of the DUEL corpus is still in the beginning phase, we do not have enough data points (40 instances of disfluent lengthening) to perform reasonable statistic tests for this part.

Discussion

It is not too surprising that predictions based on a specialized dataset outperform those based on an average dataset. However, the important finding in this study is that a disfluent context is another special context to be taken into account and that it yields significantly more accurate predictions than the baseline prediction, even when sentence position, syllable position and phonetic context are not taken into account individually.

For modeling disfluencies, and probably other conversational speech elements, it appears thus adequate to compile small-scale corpora that are rich in conversational speech phenomena, which can be elicited specifically; such as hesitation phenomena that can be elicited by delaying the flow of information a speaker has to present to a listener. This is interesting in the sense that no cost-intensive gathering of big speech corpora is necessary to improve conversational features of spoken dialogue systems. As in this study, existing large-scale corpora can be used, however, to create sub-sets of the conversational speech element in question.

In earlier studies on disfluent lengthening, the authors showed that there are phonetic preferences where the lengthening centers [10]. In summary, hesitation lengthening prefers a long vocalic syllable nucleus, but can also evade to the syllable coda if that is elastic and the nucleus is not. For the final inclusion of hesitation lengthening into dialogue systems, a synthesis of the findings is aimed for - prediction of the duration distribution using the adapted elasticity hypothesis presented here with a modifier that gives the most adequate phone in the syllable extra weight.

Acknowledgements

This research was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

Literatur

[1] M. Corley and R. Hartsuiker, "Hesitation in speech can... um... help a listener understand," in *Proceedings of the twenty-fifth meeting of the Cognitive Science Society*, 2003, pp. 276–281.

[2] H. Clark, "Speaking in time," *Speech Communication 36*, 2002.

[3] G. Skantze and A. Hjalmarsson, "Towards incremental speech generation in conversational systems," *Computer Speech and Language 27*, 2013.

[4] S. Betz, P. Wagner, and D. Schlangen, "Micro-structure of disfluencies: Basics for conversational speech synthesis," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015, Dresden)*, 2015, pp. 2222–2226.

[5] S. Betz and P. Wagner, "Disfluent Lengthening in Spontaneous Speech," in *Elektronische Sprachsignalverarbeitung (ESSV) 2016*, O. Jokisch, Ed. TUD Press, 2016.

[6] W. N. Campbell and S. D. Isard, "Segment durations in a syllable frame," *Journal of Phonetics*, vol. 19, no. 1, pp. 37–47, 1991.

[7] A. Schweitzer and N. Lewandowski, "Convergence of articulation rate in spontaneous speech," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013, Lyon)*, pp. 525–529.

[8] L. Molloy and S. Isard, "Suprasegmental duration modelling with elastic constraints in automatic speech recognition," 1998.

[9] J. Hough, Y. Tian, L. de Ruiter, S. Betz, D. Schlangen, and J. Ginzburg, "DUEL: A Multi-lingual Multimodal Dialogue Corpus for Disfluency, Exclamations and Laughter," in *10th edition of the Language Resources and Evaluation Conference*, 2016.

[10] S. Betz, P. Wagner, and J. Vosse, "Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data," in *Phonetik und Phonologie 12*, 2016.