Improving the modelling of tempo and phoneme duration

Alexandra Krah

06. April 2017

Inhaltsverzeichnis

1 Introduction

The attempts to improve the quality of speech synthesis when tempo changes include linguistic processing, and phoneme length manipulation before the actual synthesis takes place. As mentioned in a patent documentation of Fujitsu Limited, phoneme length manipulation plays an important role in this context and one cannot rely on a linear adjustment function [6]. Phoneme length adjustment is the aspect we wish to adress in this work. In particular, we want to find an algorithm to modify the phoneme length which takes into consideration the local speech rate. The result should be an approximation of the phoneme length change which is close to the actual change that occurs when the speaking rate alters. Consequently, the modified speech unit would sound more natural at a faster or a slower speech rate than being modified with a linear function.

Hoequist and Kohler observed already back in 1986 that the change in speaking tempo does not produce a linear change of the acoustic segments of an utterance (i. e. phonemes) [1].

Syllable and foot structure of the utternace belong to the important factors influencing the segment length modification [1].

More important as the segmentation of the utterance in words is its segmentation in prosodical segments. In spontaneous speech, some word parts may fall, others may be uttered together with other words so that the resulting phoneme duration tends to relate more to the structure of the resulting prosodic segment as to the original word [4]. This is also an alternative for calculating speech rate.

1.1 Goals

1.2 Outline

2 Fundamentals

2.1 Data Mining

2.2 Machine Learning

2.3 Performance Evaluation

2.4 Preparing the Data

3 Corpora

3.1 Verbmobil

The Verbmobil corpus is a database of spontaneous speech containing a collection of appointment making dialogs, fully transliterated and annotated.

4 Phoneme durations and the dynamics of speech

We omitted pauses at the beginning and at the end of the turns.

Long a shows a strong negative correlation with the local speech rate for secondary stressed and unstressed instances.

4.1 Glottal stop

4.2 Vowels

4.3 Consonants

5 Defining speech rate

Local speech rate (check window size) vs. global speech rate vs. relative speech rate.

Pfitzinger [7] showed that syllable count per second is a better approximation of the real speech rate than calculating the number of phones per second. He also proposed a formula combining the two rates, which should further improve this result for the calculation of the local speech rate, which was confirmed in later studies [8].

We decided to calculate global speech rate as a ratio of number of syllables per second. As syllables were not annotated in our database, we decided to use the number of (realized?) vowels as an approximation for the number of syllables [2] [3].

Koreman showed [5] that the perceived speech rate is also influenced by the listener's knowledge of the expected articulations for a particular utterance, therefore it makes a difference if all expected phones are also articulated or not. Consequently, we did compare speech rate considering only articulated vowels/syllables, as well as considering both realized and intended syllables, and tested both methods in our model.

Calculated also local speech rate.

We obtained the best approximation when using speech rate calculated as ...

6 Phoneme duration prediction models

There are several standard duration prediction methods in use:

- Klatt is a rather simple method developped originally for American English. It multiplies a so-called inherent duration of given segment with a context dependent factor value and add this to a segment specific minimal duration. The challenges of using this method reside in finding an appropriate inherent duration, adapting the Klatt factor values to the German language, and chosing the minimal duration to use.
- CART [**Riley1992**] is a binary classification and regression tree with questions about the influencing factors at nodes and predicted values at the leaves. It allows consideration of both categorical and continuous factors, is quite simple and easy to interpret, in contrast to e.g. neural networks, and has a good time complexity of N*logN. The challenge of selecting an appropriate feature set is solved statistically. The sole drawback is that it needs a huge amount of training data in order to perform well. ge is that it needs a large amount of training data.
- neural networks

Brinckmann and Trouvain [**Brinckmann_2003**] compared the Klatt and CART methods for predicting segment duration in synthetic speech using the PhonDat database. Their results show a significantly better performance of CART over Klatt. However, they also report a strong influence of the quality of the input data on the model performance.

7 Conclusion and Outlook

Analyze duration changes of vowels in interrupted words. Consider actual realized syllables in speech instead of syllables based on the canonical transcription. In spontaneous speech, the word and syllable boundaries change, and this change influences other phonetic aspects such as stress, syllable duration, phoneme duration.

Bibliography

[1] Charles E Hoequist and Klaus J Kohler. “Summary of Speech rate Perception Research at Kiel”. In: *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK)*. Vol. 22. Universität Kiel, Institut für Phonetik, 1986, pp. 5–28.

[2] Yishan Jiao et al. “Convex Weighting Criteria for Speaking Rate Estimation”. In: vol. 23. 9. Institute of Electrical and Electronics Engineers (IEEE), 2015, pp. 1421–1430. DOI: 10.1109/TASLP.2015.2434213. URL: https://doi.org/10.1109%2Ftaslp.2015.2434213.

[3] Klaus J. Kohler. *Einführung in die Phonetik des Deutschen*. 2nd ed. Berlin: Erich Schmidt Verlag, 1995. ISBN: 3-503-03097-2.

[4] Klaus J Kohler. “Parameters of Speech Rate Perception in German WWord and Sentences: Duration, Pitch Movement and Pitch Level”. In: *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK)*. Vol. 22. Universität Kiel, Institut für Phonetik, 1986, pp. 137–175.

[5] Jacques Koreman. “Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech”. In: *The Journal of the Acoustical Society of America* 119.1 (2006), pp. 582–596. DOI: 10.1121/1.2133436. URL: https://doi.org/10.1121%2F1.2133436.

[6] R. Nishiike et al. “Text-to-speech apparatus”. Pat. US Patent App. 12/213,792. 2008. URL: https://www.google.com/patents/US20080319755.

[7] Hartmut R. Pfitzinger. *Local Speech Rate As A Combination Of Syllable And Phone Rate*. 1998. DOI: doi=10.1.1.34.7749.

[8] Hartmut R. Pfitzinger. “Local Speech Rate Perception in German Speech”. In: *Proc. ICPhS 1999*. Vol. 2. 1999, pp. 893–896.