Improving the modelling of tempo and phoneme duration

Alexandra Krah

06. April 2017

Inhaltsverzeichnis

Abbildungsverzeichnis

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 [9]. 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) [2].

Syllable and foot structure of the utternace belong to the important factors influencing the segment length modification [2].

More important as the segmentation of the utterance in words is its segmentation in prosodical segments, so-called feet. 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 [6].

Accross the paper we will use the German SAMPA notations when referring to phonemes.

1.1 Goals

1.2 Outline

2 Fundamentals

2.1 Data Mining

Data Mining means in foreground looking for patterns in data with the help of computers. Patterns are helpful for humans to understand data. They reduce the amount of data, and of available features to the relevant ones. Computers help process within resonable time amounts of data that would need years of manual processing.

2.2 Machine Learning

Machine Learning means building a program (model) which based on data mining results from a given dataset can make predictions about an unseen dataset (fill in missing information).

2.3 Preparing the Data

A very important and time consuming step in data mining is data preparation.

2.4 Performance Evaluation

Considering machine learning, it is straightforward that one needs to split the available dataset at least into two parts: a training dataset, used for discovering patterns (data mining) in data, and a test dataset, used for testing the performance of the model (result of machine learning process). Model performance may be evaluated in several ways. We decided for the following three:

1. **RMSE** or **root mean squared error** measures the differences between the numeric values predicted by a model and those actually observed: $RMSE = \sqrt{\frac{\sum (predicted - actual)^2}{n}}$. It is a very popular method for evaluating errors of a model, but it is very sensible to outliers.

2. **MAE** or **mean absolute error** is similar to RMSE, but penalises outliers less. It is computed using the average of the absolute errors: $MAE = \frac{\sum |predicted - actual|}{n}$

3. The **correlation coefficient** is a number between -1 and 1 that quantifies the correlation between two variables, each having a separate value set. We can interpret this as how much the variable movement through its value set corresponds to the

movement of the other variable in its own value set. Values close to 1 mean that greater values of the one variable imply greater values for the second variable, negative values mean the two variables move in opposite directions, and values close to 0 mean there is no correlation between the two variables. However, one must keep in mind that correlation does not imply causality, so the presence of correlation between two variables does not mean that one of them has any influence on the other one. If we compare red Ferrari cars to blue Renault cars, one may find a correlation between car color red and car speed. However, this does not imply that red cars are faster.

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. This was an important point in selecting this database for our purpose, as phoneme duration is influenced by many factors, and the Verbmobil textual database provides most of them. We used 286 recordings from the last verbmobil phase (Verbmobil II) created by 131 speakers. The total VMII corpus used amounts to 11783 conversational turns representing recorded speech time of 15,5 hours. The “symbolic data“ corresponding to each turn (e.g.phonetic transcription, segmentation, labeling, etc.) is captured in so-called “par files“, representing files in the “BAS Partitur Format“ [**Burger**].

Drawbacks of this database come mainly from the characteristics of spontaneous speech. Following natural rules, speech rate often varies significantly inside a speech turn. One starts his turn at a normal speech rate, then gets to a point where he/she hesitates and gets slower, like when weighting different transport means to get to the appointment they are to agree upon. This results in the challenge of finding an appropriate description for the speech rate to be used for modeling phoneme durations across the corpus. We handle this issue of speech rate separately in chapter 5. Another challenge resulting from the structure of this corpus as a collection of dialogs is speaker variation. The exact coordinates of the phoneme durations vary across speakers, as one can see in Fig. 3.1 for the case of /a:/.

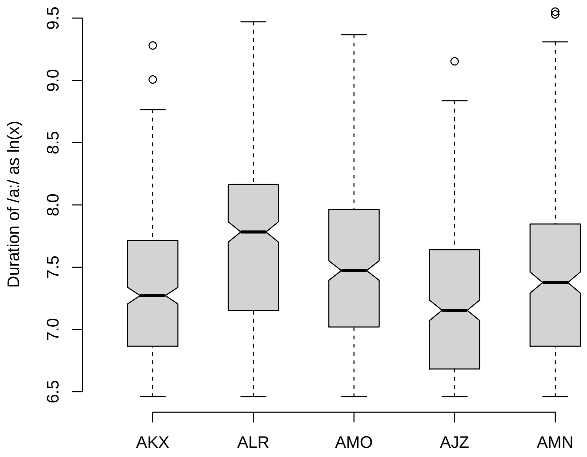

Fig. 3.1: Overview of /a:/ length on the 5 most frequent speakers. Phoneme length is expressed as the natural logarithmus of the sample frequency. While the minimum duration is technically limited to an equal value corresponding to 0.04 sec, all other coordinates vary, e.g. maximum value has a variation of 0.14 sec between speakers AKX and ALR. The greatest median variation is 0.07 sec between speakers ALR (0.15 sec) and AJZ (0.08 sec).

4 Phoneme durations and the dynamics of speech

The phoneme inventory used in Verbmobil includes 52 classes defined according to the German SAMPA. Using the classical phonetic conventions, we can group these phonemes according to their phonetic features using a tree structure into vowels, consonants and a glottal stop on a first level, although glottal stop is usually treated as a consonant. Further relevant groupings are:

- Vowels
  - Diphthongs: when two vowels are articulated together in a syllable. They are generally characterized by longer durations. German knows three diphthongs: /aI/, /aU/ and /OY/.
  - Schwa: most frequent vowel in German as well as in English. It is also called a reduction vowel [5], as most of the other monophthongs tend to be articulated as schwa in non-stressed vowels. Verbmobil differentiates two types of schwa for the German language: /@/ like in the last syllable of `lesen` and /6/ like in `Leser`, where the reduced vowel is followed by an /r/.
  - Monophthongs, which may be further split according to the short - long pairs or based on the articulation point like open - closed vowels or back - front vowels. We shall verify the relevance of these subcategories for our purpose in the section dedicated to vowels.
- Consonants
  - Plosives, which are generally short, is due to their articulation characteristics.
  - Fricatives, which show a legth variation almost comparable to that of vowels.
  - Nasals: /m/, /n/, /N/.
  - Laterals: /l/
  - Specials: /h/, /C/, /j/, /r/

Pauses may be caused and/or quantitatively modified both by linguistic (e.g. phonological) and non-linguistic factors, like technical, psychological, dialog dynamics, etc. Consequently pause length variation needs to be treated separately from phoneme duration variation, as a not negligeable large set of other factors are to be considered for this, which extends beyond the purpose of this work. Moreover, we omitted pauses at the beginning and at the end of the turns completely from our analysis, as we supposed the influence of non-linguistic factors which we cannot evaluate whithin the limits of this work to be greater in these positions.

Consistent with the study of Rosen [14] on the application of the lognormal distribution for modeling the variation of speech segment duration, we opted for a log-transformation of the observed phoneme durations using the natural logarithm, which proved to be a better description of the data in our database as well. (Include grafics?!)

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

There have been presented in the literature many influencing factors for the duration of phonemes, which vary not only from one author to another but also from one language to another. For the German language, intuition aside, we have information on such factors provided by: Kohler [4], Riedi [12], ...

We selected ... because...

4.1 Vowels

4.2 Glottal stop

4.3 Consonants

5 Defining speech rate

Under speech rate one always tends to understand a fraction of something divided by total time ocupied by that something. The German word “Sprechgeschwindigkeit“ involves the concept of “speed“ in the equation, which would be space divided by time. In the case of speech this would be number of units divided by total time ocupied by those units. However, this approach doesn’t consider the contribution of the single phoneme durations to the whole segment, but rather supposes an equal length for all phonemes in the numerator.

Speech rate may be seen as a discrete or as a continuous variable. We opted for the continuous version, as it is better suited for our purpose. Speech rate may be calculated as global, local or relative speech rate.

Pfitzinger [10] 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 [11].

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 [3] [5].

Koreman showed [7] 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. Our tests showed that ...

In our study we avoided expressing speech rate neither as a fraction, nor as a number. We took the segment duration (i.e. denominator) alone, varied it, and tried to figure out what happends in the numerator so that the proportion stays. Expressed mathematically, if the durations of the constituent phonemes are d_1, d_2, and d_3, SR is the resulting speech rate and *time* represents the total duration of the 3 elements, this corresponds to the follwoing transformation:

$$\frac{d_1 + d_2 + d_3}{time} = SR \iff (d_1 + d_2 + d_3) = SR \cdot time$$

6 Phoneme duration prediction models

There are several duration prediction methods in use:

- Klatt [**Klatt1979**] 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 adds this to a segment specific minimal duration. Then the duration is estimated by adjusting the change coeficient successively for different factors, 11 of them being suggested by Klatt. These account for 84% of observed total variance in segmental durations. 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 [13] 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. One drawback is that it needs a huge amount of training data in order to perform well. Also its performance deteriorates significantly with noisy or sparse training data [8].

- Sums-of-products (SoP) proposed by van Santen [15] generalizes the formula proposed by Klatt and uses a decision tree to group factors acting in the same direction, either lengthening or shortening the phonemes, so their effects add up (amplificatory interactions). In fact, the most important part and main challange of this model is to find an appropriate combination of factors to be used for specific phoneme groups. He didn't use speech rate as a factor, but observed that the performance of the model deteriorates when the speech rate changes. The advantage of this solution is that is copes well with noisy or missing data, and works well with much less training data than CART [8]. The application of this model for the German language by Möbius and van Santen using the PhontDat database resulted in an overall correlation coefficient of 0.896.

- Neural networks. Riedi [12] implemented a feed-forward neural network with two hidden layers for the prediction of phoneme durations and obtained a slightly better correlation coefficient than CART of 0.89 vs. 0.86.

Brinckmann and Trouvain [1] compared Klatt and CART methods for predicting segment duration using the PhonDat database. Their results show a significantly better performance of CART over Klatt, with values of 0.86 vs. 0.79 for the correlation coefficient, consistent with previous results. However, they also report a strong influence of the quality of the input data on the model performance.

All these models have an upper asymptote for their accuracy situated at less than 100%. It has been suggested, that the influence of so-called “macroscopic“ [**Cummins1999**] or para-linguistic [15] factors should be examined in order to move this asymptote further up. We focused on the speech rate.

6.1 Method

Whichever phoneme duration prediction model one takes, and whichever size of the segment one takes in the denominator, in terms of syllable, word, foot or phrase duration... Larger the segment size, larger the number of influencing factors for phoneme duration. We opted for a progressive approach, starting at the syllable level.

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] Caren Brinckmann and Jürgen Trouvain. "On the Role of Duration Prediction and Symbolic Representation for the Evaluation of Synthetic Speech". In: *International Journal of Speech Technology* 6.1 (2003), pp. 21–31. DOI: 10.1023/A:1021043804581. URL: https://doi.org/10.1023%2Fa%3A1021043804581.

[2] 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.

[3] 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.

[4] Klaus J. Kohler. "Dauerstrukturen in der Lesesprache. Erste Untersuchungen am PHONDAT-Korpus." In: *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)* (26 1992), pp. 175–196.

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

[6] 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.

[7] 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.

[8] Bernd Möbius and Jan van Santen. "Modeling segmental duration in German text-to-speech synthesis". In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*. Institute of Electrical and Electronics Engineers (IEEE), 1996. DOI: 10.1109/icslp.1996.607291. URL: https://doi.org/10.1109%2Ficslp.1996.607291.

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

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

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

[12] Marcel Riedi. “A neural-netword-based model of segmental duration for speech synthesis”. In: *EUROSPEECH*. 1995. URL: `http://www.tik.ee.ethz.ch/spr/publications/Riedi:95.pdf`.

[13] Michael D. Riley. “Tree-based modelling of segmental duration”. In: *Talking Machines: Theories, Models and Designs*. Ed. by G. Bailly. Ed. by C. Benoit. Ed. by T.R. Sawallis. Elsevier Science Publishers, 1992, pp. 265–273.

[14] Kristin M. Rosen. “Analysis of speech segment duration with the lognormal distribution: A basis for unification and comparison”. In: *Journal of Phonetics* 33 (2005), pp. 411–426. DOI: `10.1016/j.wocn.2005.02.001`. URL: `https://doi.org/10.1016%2Fj.wocn.2005.02.001`.

[15] JPH van Santen. “Assignment of segmental duration in text-to-speech synthesis”. In: *Computer Speech and Language* 8.8 (1994), pp. 95–128. DOI: `10.1006/csla.1994.1005`. URL: `https://doi.org/10.1006%2Fcsla.1994.1005`.