Improving the modelling of tempo and phoneme duration

Alexandra Krah

06. April 2017

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Introduction

Some people speak faster, others speak slower, and all people speak sometimes faster or slower as they usually do. This "speed" at which one speaks is what we call "speech rate". The fact that we understand people at all natural speech rates is certainly largely due to the quality features of sounds (phonemes) in the given language. However, if we have a perfect audio recording of speech, and play it faster or slower, we immediatly notice how naturalness deteriorates. Synthetic speech shows a similar behaviour when changing the speaking rate. In both cases phoneme quality doesn't need to change. It is its duration that changes, along with some other suprasegmental features like pitch, phrasal accents, etc., and that do the trick.

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 [10].

Scope

Phoneme length adjustment is the aspect we wish to adress in this work. In particular, we want to find an algorithm to adjust the phoneme duration according to the speech rate. The resulting phoneme duration should be close to the actual duration that occurs when the speaking rate alters. This would be the same as playing an audio file faster or slower.

There are several models for phoneme duration approximation, that work pretty well for some languages, German inclusive. However, they consistently don't take the speech rate into consideration, their performance level being acquired within a specific speech rate. When this varies, model performance deteriorates, like van Santen states for his model [17]. Moreover, 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 [17] factors should be examined in order to move this asymptote further up. So we focused on speech rate.

The purpose is therefore not to find another duration approximation model for TTS-systems, but to improve the algorithm of phoneme duration adjustment, given known reference phoneme and word durations, and a target word duration, as presented in Table 1.1. 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)

	genau				*genau*			
Phonemes	g	@	n	aU	g	@	n	aU
Durations (sec)	0.05	0.07	0.05	0.11	?	?	?	?
Speech rate *dur/#phon*	0.07				0.11			

The first column shows our inputs: duration of each phoneme in German word "genau", given a speech rate of 0.07 sec, or a word duration of 0.28 sec. The second column presents our challenge: how long is the duration of each phoneme of the given word, when it is spoken slower, taking e.g. 0.42 sec, speech rate of approx. 0.11 sec?

[2]. Indeed, if one simply "stretches" all phonemes in the same manner to accomodate a new speech rate, the result is rather dissapointing, as one can see in Fig. 1.1.

Notizen

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 [7].

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

Outline

We start our approach by defining our methods in chapter 2. In the following chapter we examine our database so that we can proceed with the actual task in chapter 4, where we analyze the German phonemes. We dedicate chapter 5 to the challenge of finding an adequate speech rate definition for our database and purpose. Chapter 6 finally deals with phoneme duration prediction models and the challenge of finding a solution to our above presented issue.

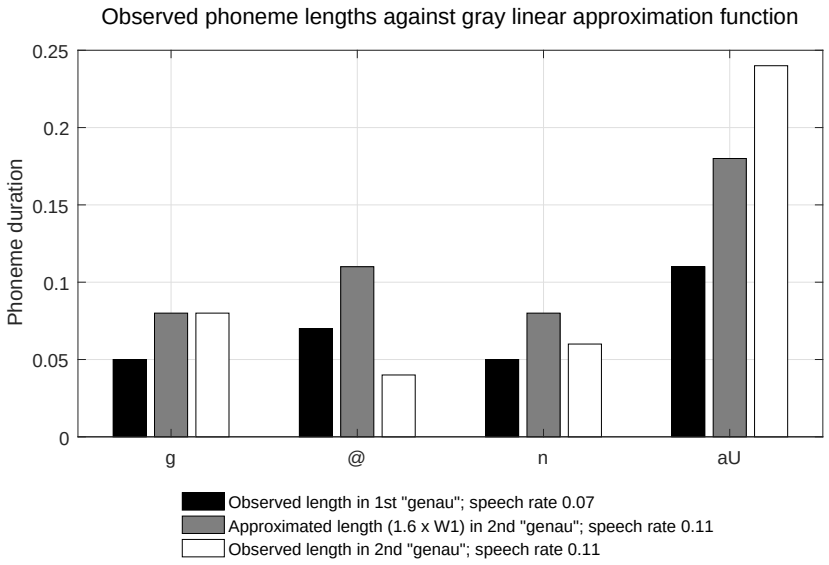


Fig. 1.1: This is a linear solution to the issue presented in table 1.1 for the German word “genau“

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 or model which can make predictions about an unseen dataset, like filling in missing information, based on data mining results from a given dataset.

2.3 Preparing the Data

A very important and time consuming step in data mining is data preparation, which is the next step after data collection. At this stage one needs to asses the quality of the data, expressed as noise ratio, and strategies for dealing with it. If the data used has been originally collected for another purpose, then it needs to be assessed in terms of features it contains, actually needed features, and organization resp. reorganization of data to serve new purpose.

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 in data through data mining, and a test dataset, used for testing the performance of the model resulting from the machine learning process.

We explain how we prepared our data in chapter 3, dedicated to our database.

2.4 Performance Evaluation

Models produced by machine learning may have different performance levels, which 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.

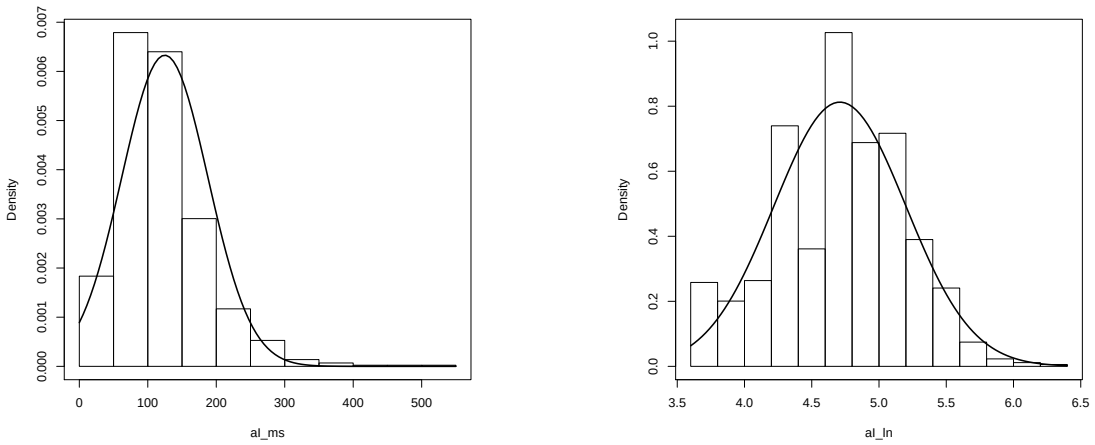
2. **MAE** or **mean absolute error** is similar to RMSE, but penalises outliers less. It is computed using the average of the absolute errors: $MAE = \frac{\sum|predicted-actual|}{n}$

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

Segmented and annotated German speech corpora represent sets of text files, so-called "textual data", containing information about the speech data. Our minimum requirement for such corpora was that they contain speech segmented at phoneme level, allowing us to calculate the duration of each phoneme occurence. The unit in which this duration is expressed is not relevant.

For further processing of this information from our side, we considered the study of Rosen [15] on the application of the lognormal distribution for modeling the variation of speech segment duration. In accordance to this, during the data mining tasks 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 (fig. 3.1).



(a) Duration of /aI/ measured in seconds.

(b) Duration of /aI/ measured in ln(sec).

Fig. 3.1: Phoneme durations measured in seconds have a not normal distribution shape. However, when applying a logarithmic transformation in terms of ln(msec), the distribution is better shaped.

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" [16].

Drawbacks of this database come mainly from the characteristics of spontaneous speech. Following natural rules, speech rate often varies significantly inside a speech turn. This means 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 phoneme durations vary across speakers, as one can see in Fig. 3.2 for the case of `/a:/`.

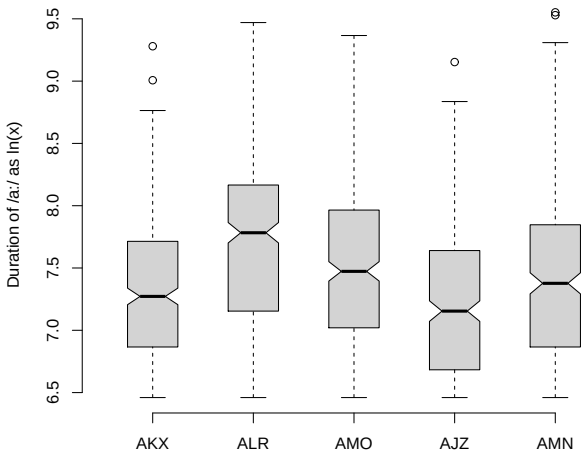


Fig. 3.2: Overview of `/a:/` length on the 5 most frequent speakers. Phoneme length is expressed as the natural logarithmus of the sample frequency. While the minimum duration is technically limited to an equal value corresponding to 0.04 sec, all other coordinates vary, e.g. maximum value has a variation 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 on a first level according to their phonetic features using a tree structure into vowels, consonants and a glottal stop, although the latter one is sometimes treated as a consonant [6]. Further relevant groupings are being examined under sections "Vowels" and "Consonants".

Pauses may be caused and/or be 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.

The literature records on many potential 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 [5], Riedi [13], Moebius [9], and Brinckmann & Trouvain [1], who applied some of the existent segment duration models to the German language.

We selected ... because...

4.1 Vowels

Vowels are an important class, because their variation in terms of duration correlates well with the speech rate. One can say that vowels generally tend to get proportionally longer when the words themselves are spoken slower. We obtained an overall correlation coefficient of 0.72 for vowels against speech rate. Inside the vowel group we could identify the primary stressed vowels as the actual correlation owners, with a group correlation value of 0.75, while the unstressed and secondary stressed ones showed rather modest correlation values of 0.56, resp. 0.47.

We identified following vowel categories as being relevant for our purpose:

- Diphthongs represent two vowels that are articulated together in a syllable. They are generally characterized by longer durations. German knows three diphthongs: `/aI/`, `/aU/` and `/OY/` as in *Haus* `/haUs/`, *heiß* `/haIs/` and *moin* `/mOYn/`. We considered them as a separate group, as they are generally longer than all other phonemes.

- Schwa: most frequent vowel in German as well as in English. It is also called a reduction vowel [6], 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/`.

- Long vowels. As long vowels usually carry the primary stress, as some german phonetics manuals such as the one of Kohler [6] suggest, we considered the long vowels as a standalone subclass of vowels. The correlation value for the long vowel group is even greather than that of the primary stressed one, amounting to 0.79.

- Short vowels in opposition to the above mentioned vowel categories.

4.2 Glottal stop

The glottal stop occurs 13 348 times and only word initially in our database, marked with `/Q/`. We decided to treat it separately because of its particular usage in German: it may occur only at the beginning of a word or a word stem and in front of a vowel. Some phonetic manuals don't event consider it a phoneme [18]. Furthermore, it's acoustical properties make it look like a (filled) pause in speech, as you may see in Fig. ..., which represents a challenge for the transcription. As our files were segmented automatically, there is place for doubt about the accuracy of the collected data for `/Q/`.

However, unlike pauses we did consider it for analysis, as it is has an important communicative function as delimiter for words and morphemes. It is in fact one of the important clues which help hearer understand speech [18].

The information collected from our database show a very large variation of it's duration from 0.04 sec. to 2.05 sec., with $\sigma = 0.06$, consistent with the official data on the complete VMII. However, if we remove the upper 0.8 % of the duration data, we obtain a dataset having only half of the mentioned variation with $\sigma = 0.03$ and a maximum duration of 0.3 seconds. We checked manually some of the largest outliers and they proved to be a result of wrong segmentation. Their real duration was within the first 2 thirds of the data.

4.3 Consonants

Based on the correlation between phoneme duration and speech rate we created a subgroup of consonants containing the nasals and `/x/, /C/, /h/` with a relatively good correlation value ≥ 0.7, and considered plosives to be a subgroup because of their consistently low correlation ≤ 0.42.

A big surprize in the Verbmobil dataset was the huge duration variation of the plosive /t/, which was even greater than most of the fricatives. Evaluation of some samples showed this to be a result of wrong segmentation in the context /ts/, giving /t/ the longer duration and /s/ the shorter one. Another explanation would be that the phase before the release phase may indeed be relatively long.

Below an overview of the classical consonant categories.

- 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/.
- Specials: /h/, /l/, /C/, /j/, /r/

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 [11] 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 [12].

(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] [6].)

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. In other words, what we do is to say that when the speech rate changes, the duration of changes as well. Now given one has an algorithm to predict the duration change of single words in a phrase when speech rate changes, we only care about phoneme duration change inside word of given durations.

6 Phoneme duration prediction models

There are several duration prediction methods in use:

- Klatt [4] 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 [14] 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 [9].

- Sums-of-products (SoP) proposed by van Santen [17] 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 [9]. 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 [13] 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.

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 ?.

Our approach is based on the observation that specific phonemes, respectively phoneme groups, occupy a specific percent of the total word duration (fig. 6.2), and that certain phonemes seem to modify their duration according to that of the segment they're in to a much higher degree than others, as you can see in fig. 6.1.

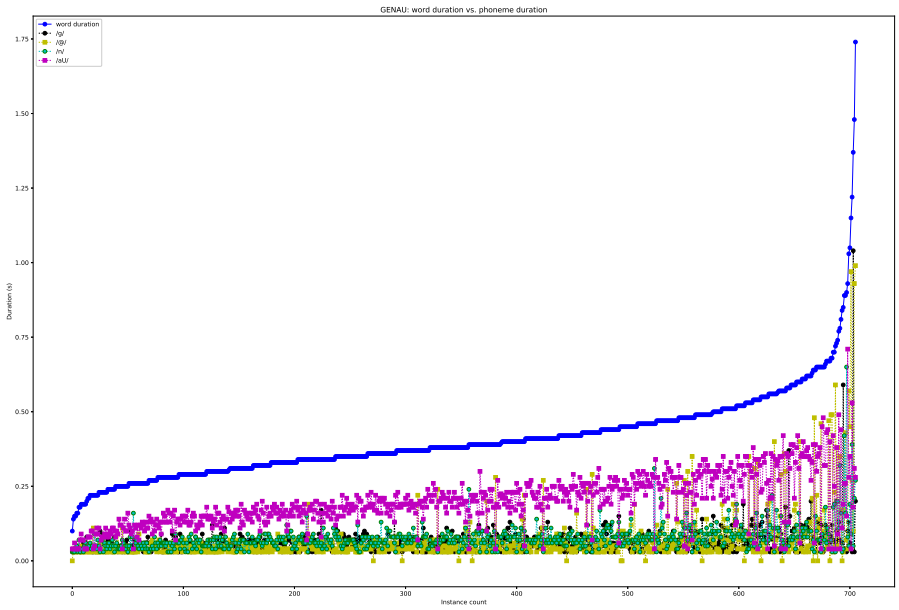


Fig. 6.1: We ordered ascendingly the 706 occurences of the word "genau" based on word duration and ploted recorded durations of the component phonemes against it. The picture suggests clearly that the diphthong `/aU/` occupies a rather fixed steak of the total word duration, while the others show no adjustment to the increasing word total duration.

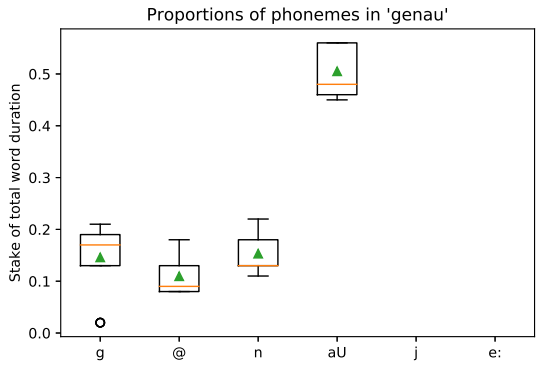


Fig. 6.2: Proportions taken by the component phonemes of "genau" suggest the possibility of grouping them in disjunct classes. Here we explored the phoneme distribution for a relatively short occurence of our reference word, of only 0.3 sec found 25 times. This one, as well as others, correlates well with the overall phoneme distribution on all occurences.

7 Conclusion and Outlook

Our model calculates an approximation for the phoneme duration when its corresponding speech segment changes in terms of duration. It does this no matter what was the cause of the duration change in the segment. However, when speech rate changes, there are also other speech characteristics, like prosody and phoneme quality which change, and which cannot be simulated only by modifying the phoneme duration. Therefore, making speech sound natural while modifying the speech rate needs to consider and adjust these other features too.

Koreman showed [8] 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, one could compare speech rate considering only articulated vowels/syllables, as well as considering both realized and intended syllables, and test both methods with a given model. One possible result of such an approach would be the prediction of phonemes reaching $duration = 0$ at specific speech rates, which means predict phoneme drop phenomena.

Notizen

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] D. H. Klatt. "Synthesis by rule of segmental durations in English sentences". In: *Frontiers of Speech Communication Research*. Ed. by B. Lindblom. Ed. by S. Ohman. New York: Academic Press, 1979, pp. 287–300.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

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

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

[13] 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.

[14] 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.

[15] 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.

[16] F Shiel S Burger K Weilhammer and HG Tillmann. "Verbmobil Data Collection and Annotation". In: *Verbmobil: Foundations of Speech-to-Speech Translation*. Ed. by W. Wahlster. Springer Nature, 2000, pp. 537–549. DOI: 10.1007/978-3-662-04230-4_39. URL: https://doi.org/10.1007%2F978-3-662-04230-4_39.

[17] 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.

[18] Elmar Ternes. *Einführung in die Phonologie*. Ed. by WBG. 3rd. Darmstadt, 2012. ISBN: 978-3-534-25578-8.