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 [9].

Scope

Phoneme length adjustment is the aspect we wish to adress in this work. In particular, we want to find an algorithm for adjusting phoneme durations according to the variations in 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 rather within a specific speech rate. When this varies, model performance deteriorates, like van Santen states for his model [15]. 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 [15] 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 prediction by considering the speech tempo. Table 1.2 vizualises our challenge: we want to put all available/visible information from the table in relation so that we can replace the question marks with values that are close to the real ones. Consequently, the modified speech unit would sound more natural at a faster or a slower speech rate as when being modified with a linear function.

	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			

Tabelle 1.2: 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 contains our challenge: How long is the duration of each phoneme of the given word, at a speech rate of 0.11 sec, equivalent of a word length of 0.42 s ?

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]. 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 [6].

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.

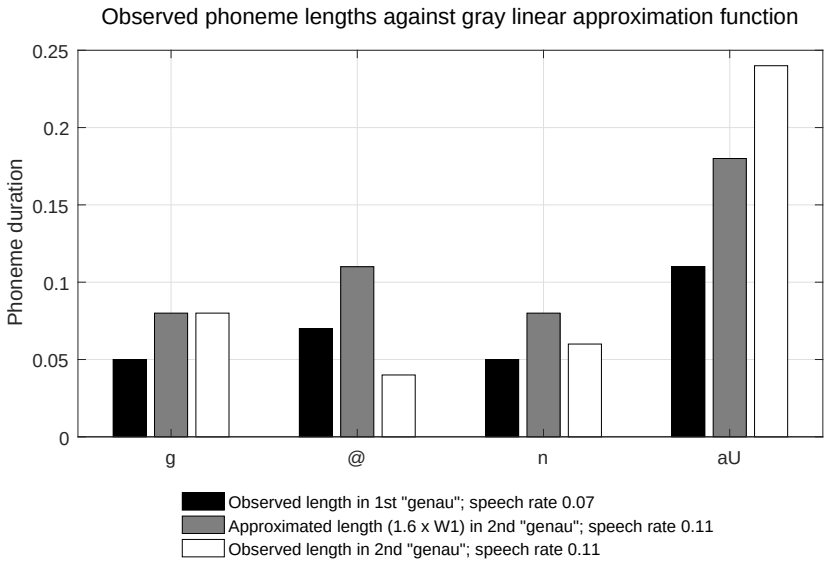

Fig. 1.1: This is a linear solution to the issue presented in table 1.2 for the German word “genau“

2 Basics

2.1 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. It consists of a so-called training phase, in which the training data is being processed by the selected machine learning method and hypotheses or models are formed. These can then be used to make predictions about new data. The training data consists of a set of instances, each of them being described by a feature vector. The training phase is followed by an evaluation phase, in which predictions about unseen data are being made and compared to the actual values.

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.

We used machine learning to solve a task involving both classification and numeric prediction. We used for this two model tree methods: M5P and REPTree. The reasons we chose them is because trees are easy to interpret, they can perform both classification and numeric prediction tasks, we could easily generate them using Weka, and are the closest to classification and regression (CART) trees, also used in the literature [**Brinckmann2003**] for phoneme duration prediction, so we have data to compare with.

2.1.1 Decision trees

Decision trees divide the attribute space into well defined clusters, identified by a class name or a numeric value. In the first case we are dealing with a classification tree, in the second one with a regression tree. If one needs to combine the two methods, he/she may use a classification and regression tree, called CART, or a model tree. The advantage of trees is that new instances are easily assigned to a class, or a numeric value, or a linear model which leads to such a numeric value, and they are easy to unterstand.

The **M5P model tree** [**Witten2011**] is a decision tree based on the M5 algorithm introduced by Quinlan (1992). It builds a usual classification tree out of the attributes, and retrieves a linear regression model in each leaf out of the attributes contained in that specific subtree. Branching is realised based on the reduction of variation within the resulting subset.

A **REPTree** is a decision and regression tree similar to M5P, but optimized for speed. This means it sorts values for numeric attributes only once.

Both trees cope well with missing data. However, both need a very large training set in order to perform well.

2.2 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 (bigger) training dataset, used for training the model, meaning discover patterns in data, and a (smaller) 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.

2.3 Performance Evaluation

Models produced by machine learning may have different performance levels, which may be evaluated in several ways. As they assess different features of the models, and treat outliers differently, we didn't want to rely only on one of them. Furthermore, we wanted to be able to compare the results with results presented in the literature. After studying the common evaluation methods used for other phoneme prediction models, we chose 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 **Pearson correlation coefficient**, a popular measure for continuous data, 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.

2.4 Tools

We used **Weka** 3.8 for building and training the model trees for the phoneme duration prediction. It is an open source freeware using a Java VM, and contains a collection of tools for all machine learning tasks we needed, from data pre-processing over classification and regression till model evaluation and data visualization.

In order to extract the needed information from our database, and put it in a form ready to use with Weka, we used self-made code written in Python 3.6. This is an easy to learn programming language with many features dedicated to statistical analysis and even machine learning. Most of the figures in this paper were produced in python as well.

3 Corpora

Language related databases are called corpora, and for our purpose we needed one containing segmented and annotated German speech as a set of text files, so-called "textual data". Our minimum requirement for such a corpus was that it contains 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. As phoneme duration may be influenced by several factors, and we did not intent to do any further processing of the audio files, we also needed that this textual database contains enough information to identify following attributes:

- phrase boundaries
- word boundaries
- syllable boundaries
- phoneme identification
- vowel stress
- pauses

There are three corpora which we had at our disposal: (1) The Kiel Corpus of Read (PhonDat) and of Spontaneous Speech, (2) Verbmobil, and (3) spoken Wikipedia. The first one contains two equal sized corpora of $>$ 4 hours each: KCoRS, which is a collection of read speech, and KCoSS, which is a collection of spontaneous speech, both being manually segmented, which is a good indicator for good segmentation quality. However, as Pfitzinger [10] also showed in his reassessment of the PhonDat-II corpus, for nowadays research the Kiel Corpus is rather small. Spoken Wikipedia is a new project comprising of a much larger collection of read speech. Unfortunately, the phoneme-level segmentation was not ready on time to evaluate in our project.

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" [14]. Our dataset included 604355 automatically labeled phoneme instances.

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. 3.1 for the case of `/a:/`.

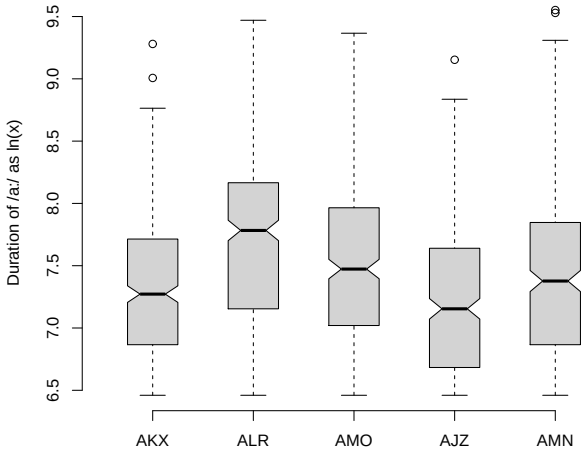

Fig. 3.1: Overview of `/a:/` length on the 5 most frequent speakers. Phoneme length is expressed as the natural logarithm of the sample frequency. While the minimum duration is technically limited to an equal value corresponding to 0.04 sec, all other coordinates vary, e.g. maximum value has a variation 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. However, only 49 of these occur in the examined corpus. The missing ones are /2/ as in *Ökonom*, and the nasals /E~/ and /O~/ as in the loanwords *Teint, Saison*. Using the classical phonetic conventions, we can group these phonemes using a tree structure. The first branching splits them into vowels, consonants and a glottal stop, although the latter one is sometimes treated as a consonant [5]. Vowels represent 24% of the total phoneme inventory analyzed. 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.

4.1 Vowels

Vowels are an important class, because they show a high degree of elasticity in terms of duration, and correlate well with the speech rate. The minimum value recorded for vowels lies at 0.03 sec, while their maximum is 2.3 sec. The variation inside the value group may be expressed in terms of standard deviation with a σ value of 0.07 sec. However, most values lie within a much smaller range, and 75% of the vowels are shorter than 0.1 sec. If we remove 1% of the data at the higher end, σ reduces to 0.5 sec and the maximum duration at 0.3 sec. After carefully analyzing some of the outliers, we decided to keep them. We observed that some vowels may last as long a one can hold breath, like in the case of /a:/. Due to this high elasticity, vowels are the best candidates to start working on any phoneme duration model, as they are extremely sensitive to duration influencing factors.

One can say that vowels generally tend to get proportionally longer when the words themselves are spoken slower. We investigated this assumption by comparing the vowel duration inside words containing a specific number of phonemes, as well as the variation of the word time proportion occupied by them in such words. The results confirmed this hypothesis, as you may see in fig. 4.1 and 4.2. This also means that it is a good idea to look for the correlation between vowels and speech rate. 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.

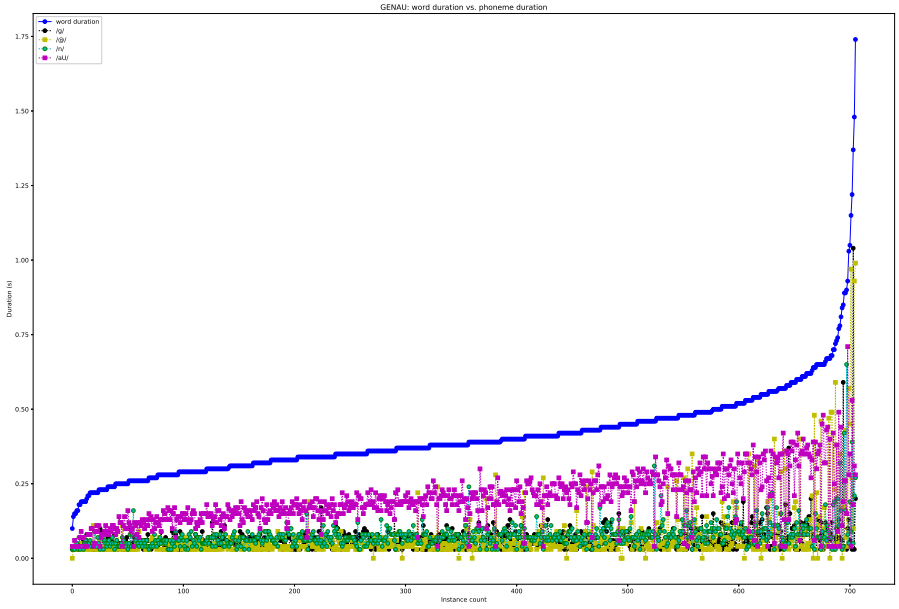

Fig. 4.1: We ordered ascendingly the 706 occurences of the word "genau" based on word duration and ploted recorded durations of the component phonemes against it. The picture suggests clearly that the diphthong `/aU/` occupies a rather fixed steak of the total word duration, while the others show no adjustment to the increasing word total duration.

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 [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/`.
- Long vowels. As long vowels usually carry the primary stress, as some german phonetics manuals such as the one of Kohler [5] suggest, we considered the long

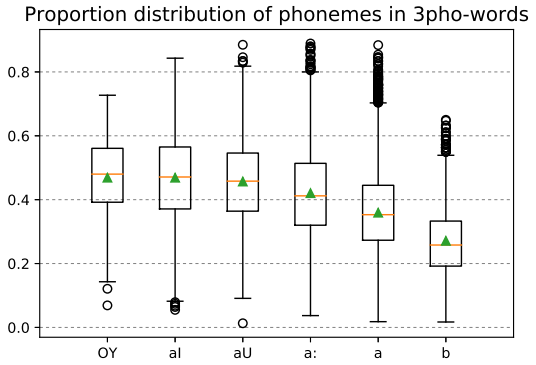

Fig. 4.2: The time slice occupied by vowels in words of specific length vary relatively little, indicating a possible linear correlation between phoneme duration and the variation of word duration. The y-axis measures time in sec.

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, group phonemes that are to be considered as vowels from the point of view of their production, but which show a relatively low elasticity with respect to duration, if we compare them to the other three vowel groups.

Table **??** below presents the main duration information regarding the above mentioned classes, sustaining our comments on them.

	Min value	Max value	Mean	St. Dev.	Q25	Q75	Count
Diphthongs	.04	.71	.12	.06	.08	.15	8495
Long vowels	.04	2.3	.11	.08	.06	.13	30425
Schwa	.03	1.02	.08	.07	.04	.09	21637
Short vowels	.03	.81	.07	.04	.04	.08	42570

The vowel tree can be continued further down, introducing several further quality related features from the domain of the articulation manner, which leads to single phonemes at the leaves. However, our purpose was to group vowels, in order to get an overview about the phoneme durations for the beginning, so we don't go further down the tree, but just show for the record fig. 4.3 on locating vowels according to their place of creation.

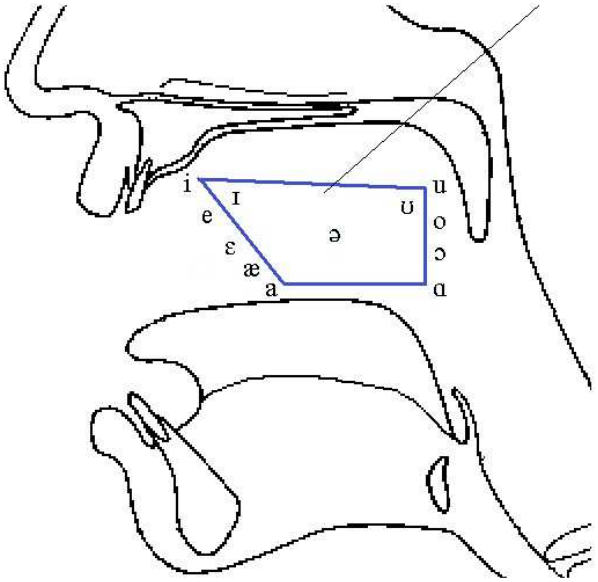

Fig. 4.3: Vowels are produced by the air flow in an open vowel tract, and they differ based on the modification we add to the shape of this cavity. The only way for humans to produce such modifications are by moving the thongue and/or the lips. These changes are usually represented in the so-called vowel square, which may be further simplified to a vowel triangle. Here you can see how to place a vowel square containing the main vowels in the mouth cavity.

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 [16]. 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 [16].

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

In general, we can say that consonants present different degrees of elasticity, which are, however, generally lower than the ones of vowels.

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.

Same as vowels, the consonant branch of the tree may be split further down into smaller categories. We chose to include them in our phoneme overview, as some of these categories proved to be relevant for the modelling of phoneme durations as well. Consequently, level 2 split shows for for the consonant node following categories:

- Plosives, which are generally short, due to their articulation characteristics. Their production involves two separate phases: closure, when the vocal tract is completely closed, and release, when the accumulated air flow is let out as in a burst. Considering this, it is straightforward, that the first phase may be longer, limited only by the human physiology, while the second one is technically extremely short. Some speech corpora, like PhonDat, label the two phases differently. Verbmobil

doesn't, so we had to take them as a whole. We keep these considerations in mind, when referring to the special case of `/t/` that we found in our corpus.

- Fricatives, which show a duration variation almost comparable to that of vowels. This fact can also be explained by their production process: the vocal tract is not completely closed, but the air flow is forced out through a narrowing at some point in the vocal tract, so that a friction sound is produced.

- Nasals also show an elasticity comparable to that of vowels. Again this is explainable through their production: the oral passage is blocked, while lowering the velum, so that the air can escape through the nose. Nasals may also constitute the nucleus of a syllable, when the vowel gets reduced. /m/, /n/, /N/.

- Liquids, also called approximants, are produced by partial closure of the mouth produced by the thongue. German knows two liquids `/l/, /r/`. One relevant feature is that they have the greatest freedom to occur in consonant clusters, that is a group of consonants belonging to the same syllable part. This also means that they must be particularly short, as we identified vowels as the greater steak holders in word duration. Another fact we considered, is the `/r/` is relatively seldom in German, being often reduced to a schwa, and when it is articulated, then it is generally extremely short, which durations comparable to the ones of the glottal stop.

Plosives and fricatives may be split further down into voiced, if the vocal cords are participating in the production of the sound, and voiceless, when the vocal cords do not participate. We didn't analyse these groups in particular, as we focused on vowels.

Table **??** below presents the main duration information regarding the above mentioned classes, sustaining our comments on them.

	Min value	Max value	Mean	St. Dev.	Q25	Q75	Count
Plosives	.03	1.50	.06	.05	.03	.07	45477
Fricatives	.03	1.89	.06	.05	.03	.08	51667
Nasals	.03	1.46	.08	.08	.04	.09	39936
Liquids /l/	.03	1.51	.06	.06	.03	.07	8019
Liquids /r/	.03	.46	.05	.03	.03	.06	5453

Tabelle 4.1

5 Speech rate

Speech rate is a measurement of the "tempo" of speech, that is how quickly someone speaks. This may be seen as a discrete/qualitative measure, with discrete values as *slow*, *normal*, and *fast* or as a continuous one, with continuous numerical values. We opted for the continuous version, because it is quantifyable and so can be a useful input into a regression, see Section 2.1.1.

Expressed mathematically, speech rate can be defined similarly to speed in spatial displacement, which would be distance covered by time. In the case of speech, distance is not counted in the spatial domain but in relevant linguistic units. Other than in the case of distance, the length of the speech units is not standardized (like kilometers, miles, etc.). Moreover it's size is also expressed in units of time: phoneme `/a/` has the length of x seconds, miliseconds, etc.

This is fine, if one has a fixed time interval and is interested in the number of units that may fit in. However, this way of looking at it does not consider the fact, that the length of the unit may vary. We are not interested in variing the number of units, which should actually stay constant. We are rather interested in variing the time interval while keeping the number of units constant. This is what happens when someone is speaking faster or slower than a specific norm. Therefore we inverted the typical speed-like calculating formula, as you can see in Eq. 5.1, so we can investigate the length variation in the speech units, with focus on units = phonemes.

$$SR = \frac{word \quad duration}{phoneme \quad count} \tag{5.1}$$

The next thing to think about is the size and identity of the fraction elements. Unit size may be phoneme, syllable or word. Variing the scope from local to globa, the segment on which we calculate speech rate may be a syllable, a word, a phrase, a speech turn or the entire corpus. In the literature [11] we also found mentions about a relative speech rate, which combines the speech rate calculated to two different unit sizes.

From the beginning, we did not consider the variant of calculating a speech rate for the entire corpus. Given its structure of dialogues recorded by many speakers, it is straightfoward that speech rate may vary tremendousely not only between speakers, but also between different speech turns, and consequently be a bad measure for characterizing our corpus. The size of a speech turn in a dialog proved to be a bad measure as well. The reason for this is again the nature of spontaneous speech as "live" creation of the human mind. Hesitations, thinking, and emotions have a much greater influence as in read speech, and result in a variing tempo at which we produce speech. In figure 5.1 we are illustrating an example of this phenomenon.

Fig. 5.1: This is the content of the speech turn number 22 of speaker AAJ in dialoge g001. The closer the points are to oneanother, the faster he speaks. Orange lines mark prosodic boundaries.

Seing this, there is no wonder that we got no good correlation values as long as we calculated speech rate on entire turns. This got better when we reduced the size of the segment to phrase, and then further down to words. An additional measure we used to evaluate how well a speech rate calculating formula may suit our purpose was the correlation value between speech rate and phoneme duration. As long vowels proved to be the most sensible to the variation of the speaking tempo, we present below the values we obtained for this group, together with the speech rate calculated at different segment sizes.

Segment	Formula	Correlation coefficient (Pearson)
Turn	$turn\_dur/phon\_count$	0.727178570436
	$phon\_count/turn\_dur$	-0.270392865503
Phrase	$phrase\_dur/phon\_count$	0.561266362391
	$phon\_count/phrase\_dur$	-0.377058296098
Word	$word\_dur/phon\_count$	0.794901825792
	$phon\_count/phrase\_dur$	-0.616539202577

Tabelle 5.1: Duration of long vowels vs. speech rate calculated at turn, phrase, and word level

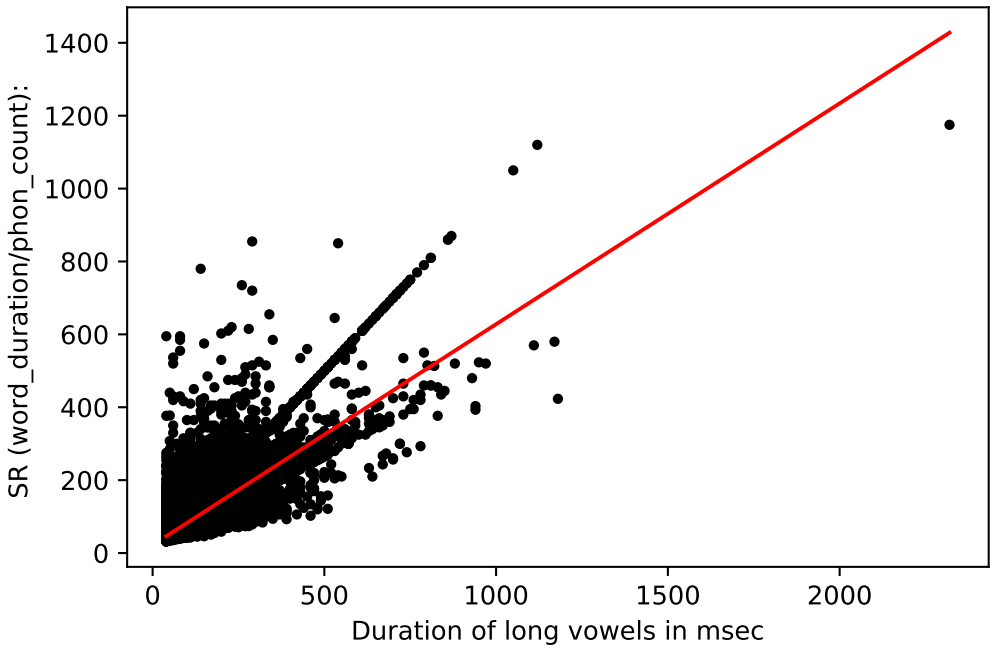

Fig. 5.2: Phoneme durations measured in msec plotted against speech rates.

6 Phoneme duration prediction models

There are several duration prediction methods in use:

- Klatt [3] 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. The drawback of this model is its intransparency.

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

Our starting premise was that phoneme durations are being influenced by the speaking tempo. In addition to this, we assumed that this influence is not equal on all phonemes, and on all situations, and therefore also not linear. So we formulated our aim as to investigate the influence of speech rate on the duration of different phonemes and find a way of using this information to improve the performance of phoneme prediction models.

In order to achieve this aim, we first examined the phonemes with respect to their duration. We found out that the classical phonetics book grouping of phonemes is not respected in terms of duration. There are consonants like `/N/`, and `/m/` which may be longer than some long vowels, and long vowels which are shorter than some short vowels or even several consonants, like `/y:/`. If one would apply a clustering method on phonemes to group them on seven classes based on their duration interval coordinates using, say k-means method, the resulting groups would certainly differ from those listed in Chapter 4. Fig?!

Therefore we decided to capture phoneme identity in our model as well, although it inflates the number of attributes and attribute values to compute. The relation between phoneme duration and factors other than speech rate was treated only marginally, as these have already been extensively examined in the literature on various corpora. For the German language, intuition aside, we have information on such factors provided by: Kohler [4], Riedi [12], Moebius [8], and Brinckmann & Trouvain [1], who applied some of the existent segment duration models to the German language.At this stage, we considered such factors only to the extent they might have helped identify groups of phonemes that may be best suited to evaluate the influence of speech rate. Such a factor is primary stress, which creates a group of very elastic vowels, which are therefore also sensitive to changes in speech tempo.

Then we looked for a definition of speech rate that can be used to model phoneme durations. This supposed the presence of a good correlation between speech rate and phoneme duration, at least for some phoneme groups. We explained our choice whith respect to this in Chapter 5. The fact, that speech rate doesn't show any correlation with some phonemes, like `/r/`, while it has a pretty good correlation with others, like `/a:/` suggests the need for classification in whichever model we choose.

6.2 Experiment

In order to test whether adding speech rate information really improves the predictions of a phoneme duration model, we tested the same model with and without the speech rate feature. We chose a tree model, because it is simple to implement, and to interpret.

One can easily examine this kind of model to see which features are being used, and which weights they bear.

For the sake of comparison, and to see whether we lie on the right track, we created at the beginning two basic models:

1. The first one predicted the phoneme duration as the mean of its durations across the whole corpus.

2. The second one build a two level tree, which split the root on the 49 realized phonemes, and then each phoneme on two specific speech rate values. The leaves contained three predicted values for each phoneme. These values were calculated using the mean duration of the specific phoneme across the corpus modified or not by a proportion of its standard deviation, based on the speech rate value split. Figure ... shows the structure of this tree.

As the results of the initial two models were encouraging, we proceeded with a "real" model. For this, we reconstructed to a great extent a previous experiment [1] that used a CART tree to predict phoneme duration, and reached good approximation values. By using this model we also had the advantage of having a reference for comparison. There is only one feature out of the 19 used in that experiment, which we could not recreate: degree of word accentuation. We lacked prosodic information for doing this. However, we tried to capture this kind of information by adding two extra word type classes: emphatic function words, like interjections and answer particles, and emphatic content words, like proper names and spelled items. For such words it is generally known that they usually carry phrasal stress.

Next, we pruned the used feature space by using two feature selection algorithms: forward and backward feature selection. Their principle is easy to understand and they are available in Weka. Consequently, following features were selected:

- PHONEME INSTRINSIC FEATURES
  - phoneme id: all the 52 phoneme definitions used in Verbmobil
  - phoneme type: vowel, or consonant
  - articulation manner: vowel, plosive, fricative, nasal, lateral, approximant, other
- SYLLABLE RELATED FEATURES
  - position of phoneme in syllable: initial, final
  - syllable part containing current phoneme: onset, nucleus, code, single-phoneme-syllable
  - lexical stress of the syllable (using the stress type of the nucleus vowel)
- PHONEMIC CONTEXT
  - previous phoneme type: vowel, consonant, or none

- following phoneme type: vowel, or consonant, or none
- articulation manner of following phoneme: vowel, plosive, fricative, nasal, lateral, approximant, other, none
- if following phoneme is voiced: yes, no, or none
- syllable part to which following phoneme belongs: onset, nucleus, code, single-phoneme-syllable

Note that all features consider only actually realised units in speech. We also had to reduce the amount of data used for the model to 58% of the total corpus, due to capacity limitations imposed by Weka on our system. After pruning, the *word* related features got completely eliminated, together with *syllable-place-in-word* and *articulation-of-realised-previous phoneme*.

To this collection of features we added *speech rate* in the second phase and compare the results.

6.3 Results

We trained and evaluated both a M5P and a REPTree model. The results were very similar, so present here only the results obtained using the model tree M5P, as they were slightly better. As can be seen in Table 6.1 adding the speech rate feature caused a significant improvement on all our models.

	Basic - SR	Basic + SR	M5P - SR	M5P + SR	CART
Correlation coef	.42		.50	.79	.86 / .83
MAE	31.02		34.21	25.19	22.46 / 21.40
RMSE	47.09		62.91	44.70	NA

Tabelle 6.1: Overview of the performance results of 3 models with (+) and without (-) speech rate (SR). CART is the model created by Brinckmann & Trouvain [1]. The two values in the CART cells represent the performance obtained for 2 different speakers.

However, the performance of our model tree is smaller than the performance of the CART built by Brinckmann & Trouvain [1]. This is explainable if we consider the nature of the used datasets. The CART tree used PhonDat, a rather homogenous dataset in terms of speech rate and number of speakers. Furthermore, the model was trained and tested separately on data produced by one speaker. The reported results show even for PhonDat a difference in performance between speakers. Another aspect that can have a great impact on the model result is the quality of the input data. While PhonDat was manually segmented and annotated, Verbmobil was processed automatically using the MAUS segmentation system. When evaluating outliers we found many examples of wrong segmentation, as we also discussed in Section 4.2.

On the other hand, our model tree shows robustness with respect to factors like inter-speaker variability and speech rate. The results obtained for a test set containing data

from a speaker not included in the training data were similar to those obtained on known speakers on terms of correlation coefficient, and showed less error rates, as can be seen in table 6.2. Moreover, if the model is to be evaluated on data produced by only one speaker, it produces good results even when using less than 10% of the amount of data used to train on the multiple speaker corpus. The most frequent speaker in our database still only has 185 recorded turns, consisting of 9315 identifyiable uttered phonemes, which represent 0.02% of the total phoneme instances in our corpus, so we could not test our model on a larger dataset of a single speaker. In comparison to this, the dataset used for CART contained 624 sentences, and a total of 4932 orthographically different words, which results in more than a double sized corpus.

	Train: whole Test: unseen speaker	Train&Test (+SR): speaker AKX	Train&Test (-SR): speaker AKX
Correlation coef	.76	.75	.63
MAE	22.91	22.47	26.95
RMSE	40.36	33.87	40.30

Tabelle 6.2: Overview of the performance of our M5P model tree on different training and test sets.

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 [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, 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] 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.

[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] H. R. Pfitzinger. "10 Years of PhonDat-II: A Reassessment". In: *Proc. ICSLP 2002.* Vol. 1. Denver, U.S.A., 2002, pp. 369–372.

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

[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] 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.

[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.

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