SIGNIFICANCE OF RHYTHMIC DURATION PATTERNS IN GERMAN READ CORPUS USING CLUSTER ANALYSIS

Zaheer Hussain

Institute for Acoustics and Speech Communication
Dresden University of Technology, 01062 Dresden, Germany
zaheer.hussain@mailbox.tu-dresden.de

Abstract: Synthesis of natural sounding speech is the greatest challenge in a Text-to-Speech Synthesis (TTS) system. In natural speech, duration, intensity and pitch are dynamically varied which is manifested as rhythm or prosody of speech. If these variations are not recreated, the synthesized speech will sound robotic. Synthesis of good quality speech depends on how well the duration and intonation patterns are imposed on speech segments. The best way to improve naturalness in speech is to mimic the way human brain imposes rhythm. We speak in a particular style by varying the duration of the speech segments in words and phrases as per certain specific duration patterns. Brain might be retrieving the corresponding patterns at the time of speaking for generating a discourse in a particular style such as news reading, bible reading, story telling etc. The main objective of this work is to investigate the existence of rhythmic duration patterns in natural speech using cluster analysis. Speech uttered in German language was taken for analysis. Cluster analysis was done on isolated words in continuous german read speech. Results of cluster analysis when investigated using silhouette plot showed the existence of rhythmic duration patterns in German read speech. The result is evidence to the existence of finite number of clusters of duration patterns in natural German read speech.

Index Terms: speech synthesis, rhythmic duration patterns, cluster analysis, k-means clustering

1 Introduction

Text-To-speech Synthesis (TTS) system converts natural language in text form to spoken form or, it is a system that reads any text aloud. The synthesized speech will appear to be natural, when it closely resembles the rhythm or prosody of speech of a native speaker of the language. In natural speech, when we speak or read continuously, the speech segments vary dynamically which is manifested as rhythm of speech. The variations in durations, intensities and stress are the prosodic features that results in the naturalness of speech. To produce natural sounding speech, TTS system should recreate the durational characteristics properly, failing of which makes the synthesized speech sound robotic and even unintelligible at times. Durational variation is incorporated in text to speech synthesis systems using duration models, which predict the duration of individual segments to be used for synthesizing speech, by considering various factors affecting duration of the speech units. The state-of-the-art duration models are generally developed using rule based or statistical approaches. Rule based methods require an intense analysis of segment durations for framing rules and are preferred for small database. Since the duration of individual segments depend on a large number of interacting factors, framing rules to capture these variations for a large database is a difficult task. Klatt developed the first rule based model in 1976 [8]. Statistical models make use of a large amount of recorded speech data to train models such as Sum of Products (SOP) models

[5], Neural network model [6], Classification and Regression Trees (CART) [7], Support Vector Machines (SVM) [4] etc. The main motivation behind the current work is taken from [1]. When we speak continuously, we utter words, phrases and sentences as a whole, and hence our brain imposes sequence of duration values on longer speech segments and not to individual syllables or phonemes. This sequence of duration values forms a particular pattern pertaining to that speech segment. Duration patterns of each language are different, and even in one language the pattern varies for different styles. Hence, it is very difficult for duration models to capture the duration patterns of a language for a particular style, beyond a certain limit. At the same time, human brain is able to generate any speech construct in a particular style, if we have sufficient exposure to it. The durational patterns can be captured, if we mimic the way human brain acquires different styles. Literature [1,8] reveals when a person learns a language, he/she store the rhythm of commonly used words, phrases or sentences in that particular style and recollect it while speech production. This points out that, brain might be storing a finite number of duration patterns to be used for generating the rhythm in speech. I have incorporate durational variation in words by fetching duration patterns from the clusters of duration patterns stored in the brain. In natural speech, duration patterns in speech can be identified using cluster analysis. Naturalness can be obtained in synthesised speech by storing duration patterns of words, phrases and sentences of the required style, which can be retrieved at the time of speech synthesis for varying time duration of speech units. Duration patterns are the vector formed by the duration values of speech units.

2 Motivation

2.1 Speech rhythm and parameters

Speech rhythm correlates with linguistic structure, but is not fully determined by it. Rhythm therefore does not carry much linguistic information, other than helping to signal the language of the speaker; without rhythmic organization, however, the linguistic message would be difficult to transfer. Rhythmical patterns provide us with information concerning important linguistic boundaries. It was believed that speech rhythm contributes highly to the intelligibility of speech and that it is thus one of the main acquisition targets for non-native speakers. Recent findings on consonants or vocal interval variability are motivated by the functions of speech rhythm in a language acquisition environment. The main argument is that infants use rhythmical cues (i.e. cues of durational consonantal or vowel interval variability) in a bilingual environment to help them distinguishing between two languages (when they are members of different rhythm classes). The studies argue that the number of children growing up in a bilingual environment is constantly underestimated (it is the rule rather than the exception) and that one of the major tasks of infants in these particular environments to separate between two languages during their language acquisition process. In literature [2] it is claimed that listeners may use acquired rhythmic structures of a language in any speech perception situation to perform preliminary low level segmental segmentation of the signal.
Rhythm is the systematic organization of prominent (higher fundamental frequency, higher duration, higher intensity) and less prominent speech units in time. Rhythm is "that property of a sequence of events in time which produces on the mind of the observer the impression of proportion between the duration of the several events or groups of events of which the sequence is composed. The analysis of speech rhythm mainly involved with four parameters correlated with the perception of prominence (F0, duration, Vowel quality, Energy). Literature shows that perceived rhythm is not just a product of durational structuring, but also it occupies a high position in the rhythm index hierarchy. Change in F0 is a strong factor. The

analysis of these parameters F0, duration, quality, energy refers to a sentence accent in a citation form utterance, which is phonologically marked for stress. The most essential thing for the production of rhythm in speech of whatever type and in any language is the global temporal bracketing of the speech signal into chunks that have recurring patterns. These global variables are patterns of syllabic timing of fundamental frequency and energy recurrent with some degree of regularity over time. In all sounds and utterances rhythm is understood as the quality of having certain beats and of being measurable by regular intervals. Rhythmic unit is roughly corresponding to the syllable combined with an optional stress pattern which plays a significant role as an intermediate level of perception between the acoustic signal and word level. Three major types of features may help to identify the language rhythm. Segmental features (acoustic properties of phonemes and their frequency of occurrence, manner and place of articulation, presence of nasal vowels etc..).Supra-segmental features (phonotactics, prosody, intonation, tones). High-level features (lexicon, morpho-syntax).

3 Investigation Methodology

Cluster analysis is performed on recorded speech to analyse the existence of duration patterns in natural speech. Analysis was performed on isolated words comprising of rhyming words on words in continuous German read speech. Results of clustering were investigated using silhouette plot. Duration patterns of the words were analysed to observe the similarity among these patterns, using cluster analysis. Cluster analysis of rhythmic duration patterns in continuous read German speech is not reported in literature. I have considered syllable level database for the analysis of rhyming words in continuous speech because, German language is syllabic in nature. The features influencing these clusters were identified as word length, the order of succession of syllables, and word position in the phrase.

4 Speech Corpus

Isolated words comprising of sets of rhyming words were selected from Verb Mobil (VM) database. Liebe-Triebe-Hiebe-Diebe-Siebe, klein-fein-rein-allein-sein-Bein-ein, Dose-Hose-Rose, Wiese-Riese-fiese-miese-diese, Kasten-Tasten-fasten, Becher-Verbrecher, Brot-Not - Kot-Boot, Tisch-Fisch, Licht-Sicht-nicht, Kerzen-Schmerzen, Messer-besser, Glas-Fraß - Spaß-Maß, Tasse-Kasse-nasse, Tomate-Granate, Feuer-teuer-Ungeheuer, Schrank-Bank - krank-Dank, Kram-lahm-zahm, Müll-Gebrüll, Bett-nett-adrett-fett-kokett. The rhyming word speech database is transcribed to phoneme level and the duration of phonemes is measured using the open source speech analysis software, PRAAT [7]. Rhyming Words in continuous speech database is transcribed to syllable level. Transcription is done by identifying the phoneme/syllable boundaries, listening to the audio and labelling them. Duration of each phoneme/syllable is then measured and tabulated. The duration values of words were formulated into duration vectors, with each element corresponding to duration of a single phoneme/syllable. These duration vectors form the word database.

5 Analysis of Clusters

Cluster analysis is an unsupervised learning method, which helps to find the natural clusters within a given data [10]. Among the different clustering algorithms, I have used a greedy iterative algorithm called K-means clustering algorithm [6]. Squared Euclidean distance is chosen as the similarity measure to accurately express the degree of similarity or dissimilarity between data objects. Since we have chosen squared Euclidean distance as the similarity

measure, words having almost the same sequence of duration values will be grouped under the same cluster. Once I get such clusters, then I can use a representative duration pattern (centroid or mean of the cluster) to represent all the elements in a particular cluster. Clustering results are observed using silhouette plot [5]. Silhouette refers to a method of interpretation and validation of clusters of data. This technique provides a graphical representation of how well each object lies within its cluster. The algorithm also returns a similarity measure called silhouette value, s(i), where $-1 \leq s(i) \leq 1$. s(i) close to one means that the datum is appropriately clustered. If s(i) is close to minus one, then it indicates a wrong clustering. s(i) near to zero means that the data is on the border of two natural clusters. A large average silhouette value of $s(i) \geq 0.5$ indicates that, the clustering algorithm has discovered a very strong clustering structure. Phrase position in the sentence, average phrase length and the order of succession of syllables are identified as the factors influencing the clusters.

6 Results and Discussion

The analysis results showed the existence of finite number of duration patterns in German speech. Results of clustering rhyming words signify the evidence for the existence of duration patterns for the rhyming words. Rhyming words showed common duration patterns since they have the same order of succession of certain phonemes. The plots and averages may be used to determine the natural number of clusters within a dataset. The results account for the existence of finite sets of rhythmic duration patterns in word level. From the result of cluster analysis it is investigated and observed that the factors influencing clustering was the word length, vowel position in the word, type of vowel (short or long) present in the word, word position in the sentence and the pronunciation style of the words.

6.1 Clustering of rhyming words (RW-1)

For the first phase of analysis, I have considered 5 sets of German rhyming words (RW-1).

* Set-1: liebe-triebe-hiebe-diebe-siebe
* Set-2: licht-siecht-nicht
* Set-3: weise-reise-fiese-miese-diese
* Set-4: tisch-frisch
* Set-5: klein-fein-rein-allein-sein-bein-kein

When cluster analysis was performed on this data, it was observed that similar sounding words fell into the same cluster (Figure 1.) The analysis was performed by varying the number of clusters and, average silhouette value was found for each case. The highest value for the overall average silhouette value, s(i) = 0.7897 was obtained when the number of clusters was set to five. s(i) > 0.7897 , which indicates that, we can represent the 15 rhyming words using only 5 representative patterns (cluster centroids). The 5 clusters(C-1, C-2, C-3, C-4, C-5) of rhyming words obtained are:

* C-1: w**ie**se-r**ie**se-f**ie**se-m**ie**se-d**ie**se
* C-2: f**ie**se
* C-3: lie**be**-hie**be**-die**be**-sie**be**- tri**ebe**
* C-4: klei**n**-fei**n**-rei**n**-allei**n**-sei**n**-bei**n**-kei**n**
* C-5: li**ch**t-si**ch**t-ni**ch**t

All the words in set 1 are grouped under cluster 3. They all have the same length (total no. of phonemes present = 5), and ends with the syllable '/be/'. Similarly set 2 words are grouped under cluster 5. The presence of a long vowel in the second position of the word and the total length of the words can be identified as the reasons for grouping set 4 words under cluster 3. The word /fiese/ is not clustered along with the other words in set 3, as the duration of the first phoneme '/fi/'is strikingly different from the duration of first phoneme of other words as seen from the plot of duration patterns (Figure 2.).

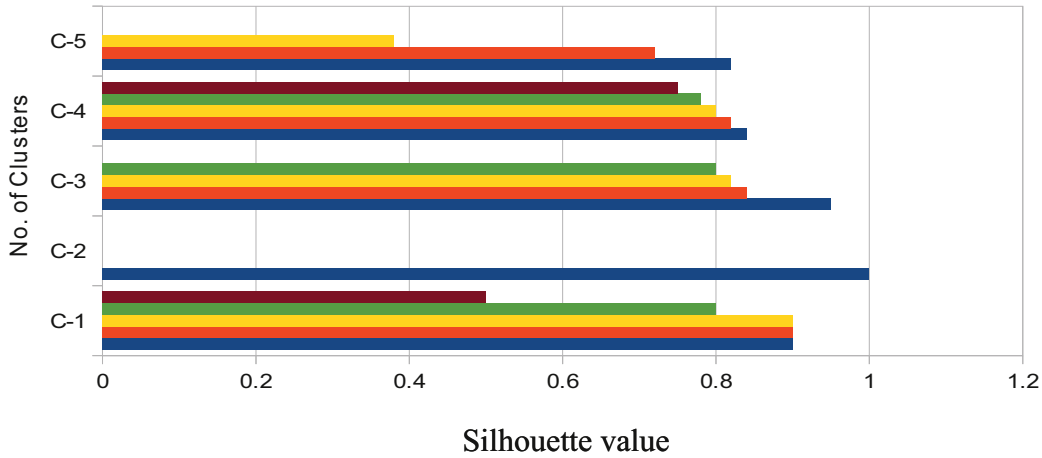

Figure 1. Silhouette plot showing results of clustering of German Rhyming words (RW-1)
Average Silhouette value obtained is s(i) =0.7897

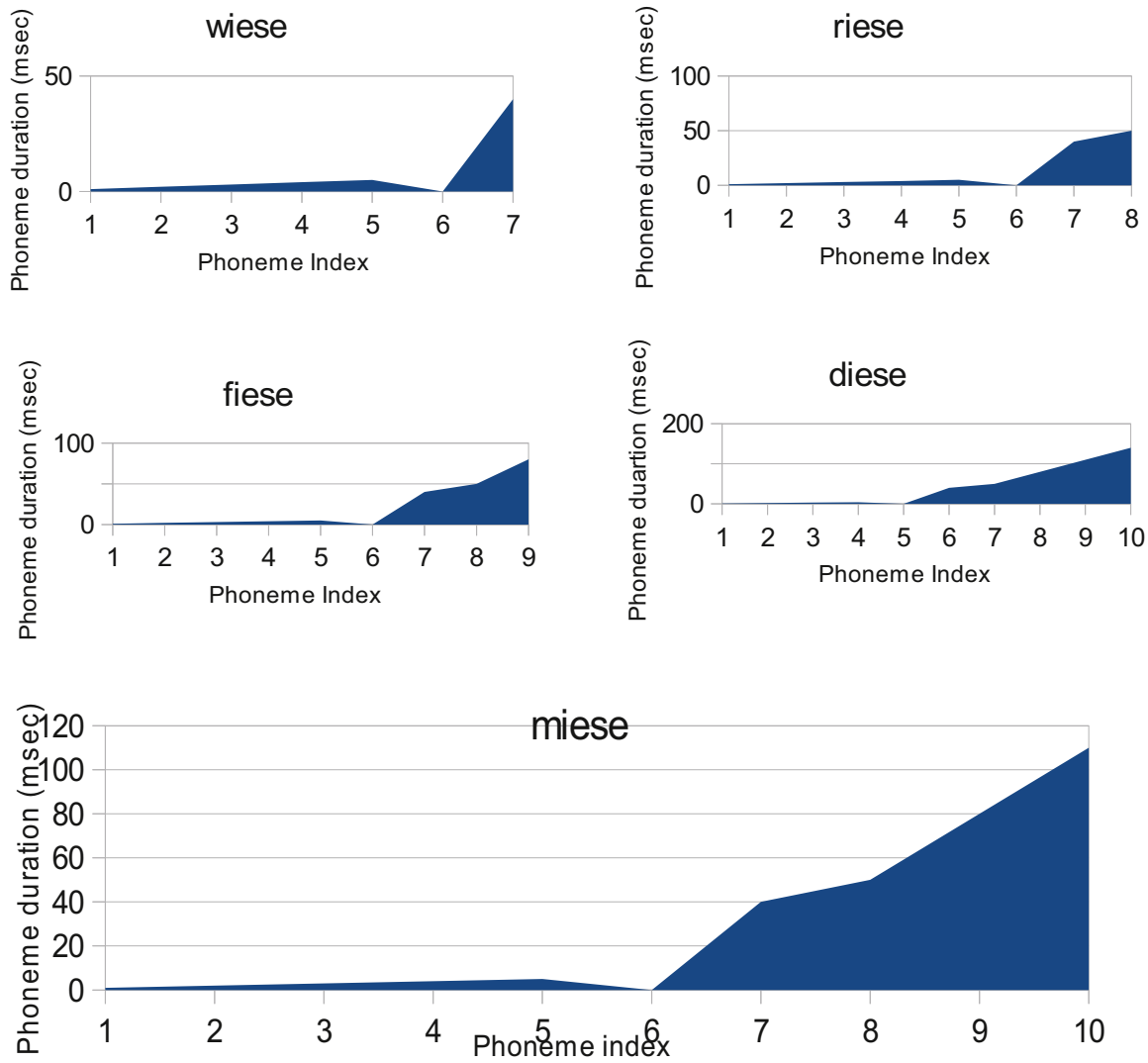

Figure 2. Duration patterns of rhyming words (wiese-riese-fiese-miese-diese). Phoneme durations are plotted against their position in the word

6.2 Clustering of rhyming words (RW-2)

For the second phase of analysis I have considered 8 sets of German rhyming words (RW-2).

* set-1: schrank-bank-krank-dank
* set-2: kasten-tasten-fasten
* set-3: Brot/Not/Kot/Boot
* set-4: kram/lahm/zahm
* set-5: feuer/teuer/ungeheuer
* set-6: tasse/kasse/nasse
* set-7: fraß/spaß/maß
* set-8: kerzen/schmerzen

In this case, best result for clustering was observed when the number of clusters was selected as 8 with an overall average silhouette value of s(i) = 0.9589 (Figure 3). The 8 clusters (C-1, C-2, C-3, C-4, C-5, C-6, C-7, C-8) of rhyming words obtained are:

* C-1: t**asse-**k**asse-**n**asse**
* C-2: k**asten-**t**asten-**f**asten**
* C-3: Br**ot-**N**ot-**Bo**ot-**K**ot**
* C-4: fe**uer-**te**uer**
* C-5: kram-lahm-zahm-schmerzen
* C-6: kerzen-ungeheuer
* C-7: fr**aß**/sp**aß**/m**aß**
* C-8: schr**ank-**b**ank-**kr**ank-**d**ank**

The words in set 2, Kasten /Tasten and /fasten are grouped under cluster 2, since they all have the vowel '/a/' in the starting position. The three words also have the same word length (no. of phonemes = 6) and the same pronunciation style due to the presence of the bisyllable '/asten/'. All the words in set 7 were grouped under cluster 7, since these words share the common characteristics such as same word length (number of phonemes) and the presence of the bisyllable '/aß/'. The presence of short vowel '/a/'in the word /Tasse/ can be identified as the reason for separating out that word from set3 and grouping it separately under cluster1, while the rest of the words were in cluster 3. This can be clearly observed from Figure 4., which depicts the duration patterns of words in set 3. The factor which brought all the words in cluster5 together is the total number of phonemes in the words. Even though these 4 words are grouped under the same cluster, there are some dissimilarities among these which is clear from the silhouette plot shown in Figure 3. All the 4 words are having almost similar duration patterns. Centroid of the cluster is the representative duration pattern for all elements in the cluster.

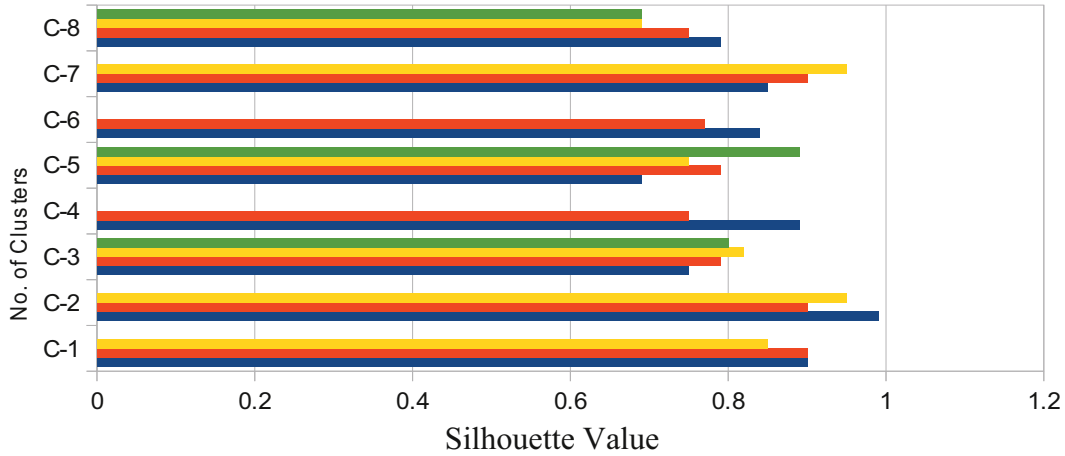

Figure 3. Silhouette plot showing results of clustering German Rhyming words (RW-2). The average Silhouette value obtained s(i)= 0.9589

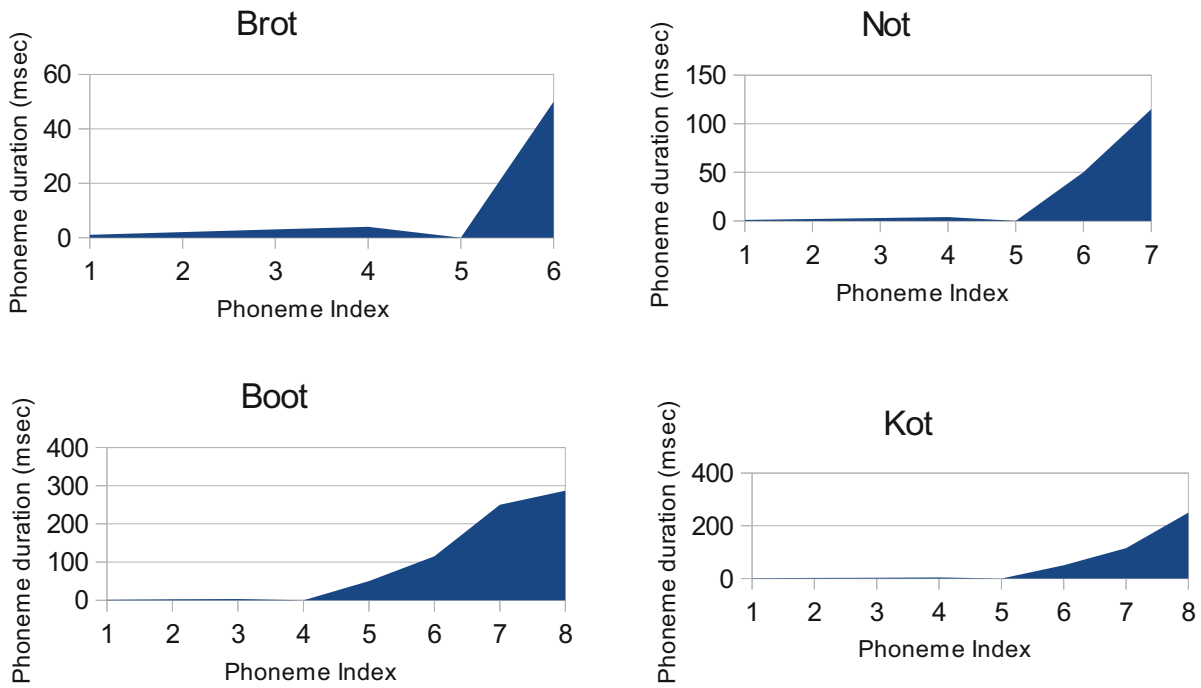

Figure 4. Duration patterns of rhyming words (Brot-Not-Boot-Kot). Phoneme durations are plotted against their position in the word

7 Conclusion

When cluster analysis was performed on the individual sets of words, it was observed that each set of similar words fell into separate clusters. Similar sounding words were grouped automatically. It was also observed that, similarity of words and phrases is determined by the order of succession of basic sounds. From this we can understand that, it is not the textual meaning that matters to form clusters but, rather the style of pronunciation, the order of succession of phonemes/syllables and the length of words. The result is evidence to the existence of finite number of clusters of duration patterns in natural German speech. Naturalness can be obtained in synthesised speech by storing duration patterns of words, of the required style, which can be retrieved at the time of speech synthesis for varying time duration of speech units. The analysis can be extended for identifying the frequently occurring duration patterns in different styles of speech, and modelling a duration model utilising the frequently occurring representative patterns, so that the TTS system yield good quality speech in terms of naturality. If the typical duration patterns in a particular style of speech can be identified, then naturality of synthesized speech can be improved by imposing durational variation according to these patterns. While we learn the speech constructions, we are not taught about their durations, but the brain memorises the frequently occurring duration patterns and the factors influencing it and reuses it when the same situation arise.

8 Future work

Clustering of words in continuous speech can be further carried out in next phase of the analysis. The results are clear indication on the existence of commonly occurring patterns in a particular speaking style at word level. Similar analysis can be extended to phrase or sentence

level. The features influencing these clusters were identified as, word length, the order of succession of syllables / phonemes, and word position. Clustering of phrases in continuous speech can be performed on continuous read news programme which can be transcribed to syllable level as the database. Phrase position in the sentence, average phrase length and the order of succession of syllables may be considered and identified as the factors influencing the clusters in continuous German speech.

9 Acknowledgements

I would like to express my sincere thanks to Prof. Dr.-Ing. habil. Rüdiger Hoffmann for his technical suggestions and to Dipl.-Slaw. Rainer Jäckel for his valuable suggestions and clear description of German language accent, short and long vowel pronunciation variations.

10 References

[1] Zaheer Hussain, Ruediger Hoffmann"Prosodic analysis of German read corpus for a CALL system for Rehabilitation Purposes" *ICT for Language learning 6th edition*, Florence, Italy., November 14-15, 2013.

[2] Rouas,J., Farinas,J., Pellegrino F., Obrecht R A., "Rhythmic unit extraction and modelling for automatic language identification" *Speech Communication*, Vol.47 *Elsevier,* 2005. pp. 436 – 456

[3] Jokisch, Oliver., Ding, Hongwei., Kruschke, Hans.,"Towards a multilingual prosody model for text to speech " *IEEE Proceedings* Vol.1: *ICASSP,* 2002 . pp. I:421-424

[4] K.Srinivasa Rao and B.Yegnanarayana,"Modeling syllable duration in Indian languages using support vector machines", *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP'04*, 2005, pp. 258-263.

[5] Peter J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J.of Computational and Applied Mathematics*, 20, Nov.1987, 53-65.

[6] http://en.wikipedia.org/wiki/K-means clustering

[7] http://www.ling.ohiostate.edu/welby/PRAAT/welby-ito-praat-tutorial.pdf

[8] D .H.Klatt, Linguistic uses of segmental duration in English :Acoustic and perceptual evidence, *J. of Acoustic Society of America*, 59, 1976, 1209-1221

[9] Jan P. H. Van Santen, Assignment on segmental duration in Text to Speech Synthesis, *J. of Computer Speech and Language* 8, 1994, pp. 95-128.

[10] http://home.dei.polimi.it/matteucc/Clustering/tutorial

[11] Ghitza,O. ,Greenberg, S. "On the Possible Role of Brain Rhythms in Speech Perception: Intelligibility of Time-Compressed Speech with Periodic and Aperiodic Insertions of Silence, *Phonetica 2009* ; 66 (No.12): pp. 113 -126