International Conference "ICT for Language Learning" 6th edition

Prosodic Analysis of a German Read Corpus for a CALL System for Rehabilitation Purposes

Zaheer Hussain, Rüdiger Hoffmann
Chair of System Theory and Speech Technology, Dresden University of Technology (Germany)
zaheer.hussain@mailbox.tu-dresden.de. ruediger.hoffmann@tu-dresden.de

Abstract
Now-a-days Computer Assisted Language Learning (CALL) systems for rehabilitation purposes involving rhythmic analysis and prosodic analysis which are inter-related have gained much importance in speech synthesis and recognition systems. The main objective is to develop a language learning system and content data corresponding to the pronunciation training of German and Slavic languages. The call system AzAR (German acronym for 'automat for accent reduction') which was developed at our IAS laboratory provides adequate feedback regarding the pronunciation of the learner as well as the prosodic quality. With regard to learners and teachers, the new feature is the integration of large corpus and multilingual data bases. The motivation for this kind of analysis is that prosodic features carry a substantial part of the language identity that may be sufficient for humans to perceptually identify some languages. The main objective of the present work is to demonstrate the significance of limited German phoneme rhythmic based prosodic variations in frequency f0, duration and intensity and to develop rhythmic unit extraction model for speech synthesis and recognition. This work starts with rhythmic based prosodic analysis, an accent and de-accent experiment using resynthesis and a perceptual test. The improved perceptual quality of the duration and mean frequency modified phonemes proved to be a promising result for perception as demonstrated in the subjective evaluation test with resynthesis stimuli. The main application leads to develop a multimodel data-based assistance system for self learning activities for old aged people affected with brain related diseases e.g., Parkinson's disease.
Index Terms: prosody, phoneme, duration, rhythm, intonation and intensity

1. Introduction
The methods of computer-assisted language learning (CALL) and so-called intelligent language tutoring systems (ILTS) play an increasing role in the second language education. The project supports teaching and private studies of languages in neighbour countries, e. g. using e-learning infrastructure and computer-based language courses. The baseline platform AzAR (German acronym for 'automat for accent reduction') was developed in preceding projects [1][2][3]. The core function is based on different phonetic-phonologic and prosodic measures, involving typical cross-lingual influences from a native source language on the target language. It leads to the marking of mispronounced phones within a spoken utterance using a coloured scale from red ("very bad") to green ("very good"). Rhythmic based prosodic analysis play a tremendous role for the algorithmic development in speech synthesis and recognition. Unlike English in which the speech rhythm is mainly characterized by stress, German rhythmic phonemes are marked by a varying syllable length depending on germination of consonants and by certain phrasing features as well as glottalization as a consequence of rhythmic structure [1]. Duration modification is the process of modifying the speech rate according to desired modification factors without affecting the pitch, spectral and speaker characteristics of the original speech [2]. The main objective in duration modification is to modify the speech rate according to desired modification factors with minimum perceptual distortion thereby improving the intelligibility and naturalness. The main motivation for this kind of analysis is that prosodic features carry a substantial part of the language identity that may be sufficient for humans to perceptually identify some languages [3]. Rhythmic unit corresponding to the phoneme combined with an optional stress pattern plays an important role as an intermediate level of perception between the acoustic signal and the word level [4]. Stress is considered as a prominent factor in this analysis because it is the basis of rhythm in all languages and it was well proposed a rhythmic continuum that does not stretch from phoneme to stress timing [5]. In this analysis we considered only 43 German phonemes which includes 15 Vowels, 3 Diphthongs, 2 unstressed vowels and 22 consonants as shown in Table 1. This table presents limited german phonemes.

Vowels	a:,a,e:,E,i:,I,o:,O,u:,U,y:,Y,E:,2:,9
Diphthongs	aI, aU, OY
Unstressed Vowels	6 (turned A), ə(schwa)
Consonants	S,Z,C,N,Q,b,d,f,g,h,j,k,l,m,n,p,r,s,t,v,x,z

Table 1: Limited German phonemes in prosodic corpus

This paper is organized as follows: In section 2, we present examination methods which includes the speech data collection, test corpus and rhythmic analysis, resynthesis, modification factors and perceptual investigation. Section 3 presents results and discussion which includes significance of stress and intonation, phoneme duration, intensity variation, relation among duration, f0, intensity (Int) and perception evaluation. Section 4 concludes this analysis.

2. Data and Investigation Methods

2.1. Speech data collection

The data collection originally targeted for development of multimodel data-based assistance system. It involves 50 male and 60 female speakers aged from 55 to 75 years. Each speaker read 105 German sentences collected for different projects. The data was recorded in quiet office environment and 5 speakers were recorded in IAS recording studio at 16 kHz, 16 bit PCM. The recordings were semi-automatically segmented, pitch marked in a multilayered Praat TextGrid format [10]. The database includes 30 minutes of read speech. Table 2 shows an overview of the prosodic test corpus used.

2.2. Test Corpus and Rhythmic Analysis

For Rhythmic based prosodic analysis, we randomly selected 105 sentences from 5 male and 1 female speaker. The phoneme segmentation is done accordingly duration and intensity analysis relied on ASR alignment [4]. According to audio perception,

Parameter	**Neutral**	**Stressed(+ +)**	**Reduced(- -)**
dpho(m)	2071	2115	2173
dpho(f)	515	536	526
f0(m)	53024	24921	11947
f0(f)	10690	4057	1776

Table 2: Number of phonemes in prosodic test corpus

the first author, non-native German speaker annotated three stress levels: unstressed or "neutral", "stressed (++)", "reduced (- -)" phoneme. The duration of the phoneme d_{pho} was measured successively at phoneme boundaries considering signal pauses. The f0 was extracted using the ESPS algorithm from Wave surfer (ver: 1.8.8). Segments containing voiced parts with f0 below 60 Hz (caused by glottalization effects or algorithmic restriction) were excluded from intonation analysis reducing the number of phonemes to limited data phonemes. For each phoneme segment, we observed minimum, mean and maximum f0 and $f0_{range}$ (difference of maximum and minimum f0). Finally, the values of d_{pho}, f0 were averaged over all utterances and speakers within the same gender.

2.3. Resynthesis, Modification factors and Perceptual Investigation

To test the perceptual relevance, we manipulated six sentences(2 to 3 words). In all the six sentences we modified two to three words as per the modification factors stated here. We have taken the combination of duration and pitch modifications: d_{pho}±35\%, +65%, +100% and f0±10%, ±25%, ±50% as modification factors. In three sentences, we increased duration and mean f0 in a single phoneme to form a synthetic accent. In the other three sentences, duration and mean f0 of a stressed phoneme were decreased aiming at deaccentuation. All modifications were performed using [10]. We presented 6 natural sentences and 80 test sentences with modifications. During testing, we asked listeners to mark for emphasized word instead of phoneme position. Listeners evaluated the naturalness according (Table 7)

3. Results and Discussion

3.1. Significance of Stress and Intonation

Table 3 summarizes the phoneme-based mean f0 values for male utterances. The percentage in parentheses specifies the deviation from the reference given by unstressed ("neutral") segments. Stress is associated with higher, and reduction with lower f0 values as known from other languages [3]. The increased *f0min* in reduced phonemes might be a result of co-articulation or might be due to rare number of occurrences. Female speakers (Table 4) feature a similar relative f0 variation in stressed and reduced phonemes. For stressed phonemes, the average deviation of *f0mean*, *f0min* and *f0max* shows greater significance in both genders. Figures 1(a) and 1(b) compare the *f0* variations for both genders.

Parameter	**Neutral**	**Stressed(+ +)**	**Reduced(- -)**
f0mean	896 Hz	777 Hz	731 Hz
f0min	64 Hz	56 Hz	61 Hz
f0max	101 Hz	100 Hz	100 Hz
f0range	37 Hz	43 Hz	39 Hz

Table 3.Stress and mean f0 variation in phoneme (m)

3.2. Significance of phoneme Duration and Intensity (Int)

Table 5 summarizes the stress-related variation of the phoneme duration in both genders. The average deviation of stressed phonemes confirms the assumption with regard to the importance of duration in previous studies [6]. Figure 2 compares the duration modification for both genders. The mean intensity (Int) modification in stressed phonemes in Table 6 suggests a lower importance of intensity parameters in German prosody generation and perception [7].

Parameter	**Neutral**	**Stressed(+ +)**	**Reduced(- -)**
f0mean	1195 Hz	1180 Hz	1163 Hz
f0min	109 Hz	92 Hz	114 Hz
f0max	184 Hz	179 Hz	178 Hz
f0range	75 Hz	87 Hz	63 Hz

Table 4. Stress and mean f0 variation in phoneme (f)

Parameter	**Neutral**	**Stressed(+ +)**	**Reduced(- -)**
dpho(m)	437 µsec	3.5 µsec (+99%)	9.5 µsec (+99%)
dpho(f)	0.5 µsec	3.4 µsec(+84%)	1.6 µsec(+66%)

Table 5.Stress-related modification of mean duration

Parameter	**Neutral**	**Stressed(+ +)**	**Reduced(- -)**
Int(m)	21 dB	19 dB (-7%)	30 dB (+30%)
Int(f)	23 dB	39 dB (+41%)	11 dB (+100%)

Table 6.Stress-related modification of mean Intensity (Int)

3.3. Significance of relation among Duration, f0, Intensity

To verify the significance among the prosodic parameters, we have chosen 16 phonemes with lowest and highest Δ*dpho* (stressed Vs neutral) assuring a minimum frequency of occurrence of 10 neutral and 5 stressed segments. Figure 3 shows the mean prosodic parameters for these phonemes, sorted by their Δ*dpho*. This significance and selection of phonemes have been well studied in [8]. Compared with their neutral phonemes, stressed phonemes in Figure 3a show higher (apparently uncorrelated) deviations in mean f0 and

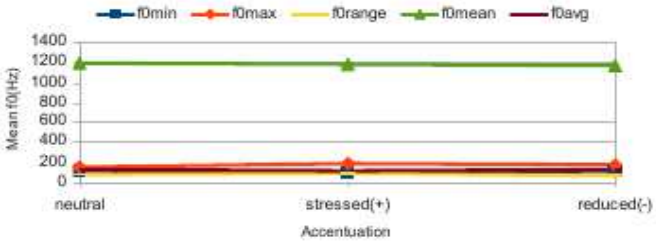

Fig.1(a) f0 variation in phonemes for female speaker (f)

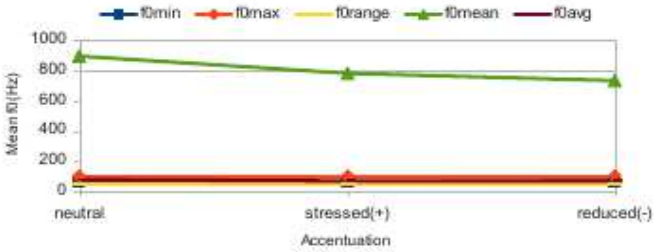

Fig.1(b) f0 variation in phonemes for male speaker (m)

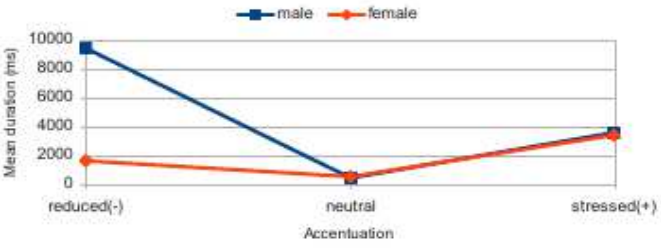

Fig.2. Stress-related modification of mean duration

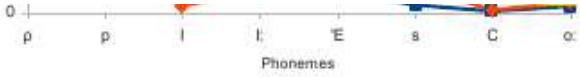

Fig.3(b) Phonemes with highest Δduration

Justification	Speech Quality	Rating
De-accented	Bad	1
Slightly De-accented	Poor	2
Perceptible	Fair	3
Just perceptible	Good	4
Imperceptible	Excellent	5

Table 7: Ranking used for judging the quality of speech

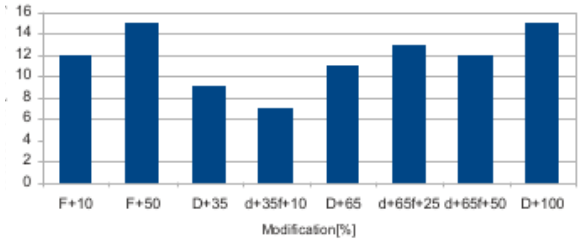

Fig.4(a) identification of accent position

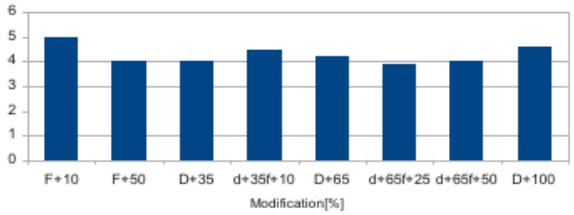

Fig.4(b) Naturalness of the modified utterance

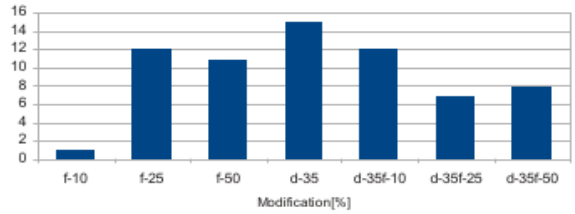

Fig.5(a) Identification of accent deletion

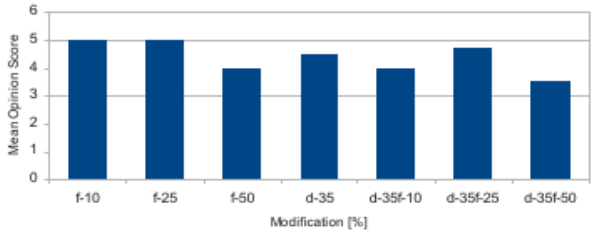

Fig.5(b) Naturalness of modified utterance

intensity than the phonemes in Figure 3(b). The results indicate different strategies in stress forming, e.g. duration vs. f0 and intensity. The dominating vowels {*a:,u:,o:,e:* } refer to phonological aspects of duration [6]. We observe from the perception investigation that perceived rhythm is not just a product of durational structuring and modeling, but also the strong accent position in the rhythm-index hierarchy. Change in f0 also effect strongly to the impression of rhythm, where stronger f0 differences with weaker durational differences (between stressed and unstressed phonemes) sounded more rhythmical than more monotone line with stronger durational differences.

3.4. Perception Investigation and Evaluation

20 evaluators (10 of them are experts in speech technology and 10 are non-experts), native German speakers participated in the experiment. Evaluators listened to 80 modified sentences at different durations and frequencies in randomized manner and decided which resynthesized signal has a more preferable quality based on rating. The description of each of the scores are given in Table 7. There were a total of 86 speech files (80 modified + 6 original files). The mean of the scores obtained for all the files for a given duration modification factor is calculated as mean opinion score. Figure 4(a) shows the perceptual identification of the synthetic accents for combined manipulations of phoneme duration and f0.In all tables the notation (m) and (f) indicates male and female speaker respectively.

4. Conclusion

We conclude that assuming a baseline approach with three stress levels, we have studied phoneme-based prosodic attributes in German read speech. We have confirmed the important role of duration modification suggested by previous studies on phoneme level. In addition, we have examined stress-related modifications in f0 and intensity. To test the perceptual relevance of observed variations in duration and f0,we manipulated single phonemes aiming at synthetic accentuation or deaccentuation and presented resynthesized sentences to native speakers of German who recognized a significant word of intended accent positions and deletions. The analysis results can be used in the development of rhythmic based prosody models and also a multimodel data-based assistance system for old aged people affected with brain related diseases e.g., Parkinson's disease. This work is a part of ongoing project titled *RehaVox.* Project Nr. 1209036999/ 42853 (2012-2014) funded by Federal Ministry of Economics and Technology, Germany.

References

[1] Malisz Z, Żygis M, Marschall B.P., "Glottalisation as a consequence of rhythmic structure? A study of different speech styles in Polish and German" *Proceedings of LabPhon,13th Conference on Laboratory Phonology*. Stuttgart 2012

[2] K.S.Rao and B.Yegnanarayana, "Duration modification using glottal closure instants and vowel onset points, *Speech Communication*., vol.40, no.3, pp.1263-1269, Dec.2009

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

[4] Neijt A.,Schreuder R.,"Rhythm versus Analogy:Prosodic Form Variation in Dutch Compounds " *Language and Speech,*Page(s):533 – 566, 2007

[5] Felicitas Kleber, Nadine Klipphahn "An acoustic investigation of secondary stress in German" *Institute of Phonetics and Digital Speech Processing,* C. A. University,Germany,2011

[6] Bernd Mobius ,Jan van Santen "Modeling segmental duration in German Text-to-Speech Synthesis" *Spoken Language ICSLP Proceedings,*PP: 2395 - 2398 vol.4 , 1996

[7] Jokisch, Oliver Ding, Hongwei ; Kruschke, Hans "Towards a multilingual prosody model for text-to-speech " *IEEE Proceedings* Volume: 1 Page(s): I-421 - I-424 *ICASSP,* 2002

[8] Jokisch, O.,Birhanu, Y., Hoffmann, R. "Syllable-based prosodic analysis of Amharic read speech " *Spoken Language Technology Workshop (SLT)*, Page(s): 252 - 257 , I*EEE* 2012

[9] D. Talkin, In W. Kleijn and K. Paliwal “A robust algorithm for pitch tracking (RAPT).” *Speech Coding and Synthesis* pp. 495–518. *Elsevier,* 1995

[10] P.Boersma,D.Weenink, Praat:doing phonetics by computer (ver-5.3.05) *http://www.praat.org* , 2012