SYLLABLE-BASED PROSODIC ANALYSIS OF AMHARIC READ SPEECH

Oliver Jokisch

Leipzig University of Telecommunication
Dept. of Communications Engineering
jokisch@hft-leipzig.de

Yitagessu Birhanu, Rüdiger Hoffmann

Dresden University of Technology
Chair System Theory & Speech Technology
yitagesu.gebremedhin@ias.et.tu-dresden.de

ABSTRACT

Amharic is the official language of Ethiopia and belongs to the under-resourced languages. Analyzing a new corpus of Amharic read speech, this contribution surveys syllable-based prosodic variations in f_0, duration and intensity to develop suitable prosody models for speech synthesis and recognition. The article starts with a brief description of the Amharic script, the prosodic analysis methods, an accentuation experiment using resynthesis and a perceptual test. The main part summarizes stress-related analysis results for f_0, duration and intensity and their interrelations. The quantitative variations of Amharic are comparable with the range in well-examined languages. The observed modifications in syllable duration and mean f_0 proved to be relevant for stress perception as demonstrated in the perceptual test with resynthesis stimuli.

Index Terms— Amharic language, prosody, syllable, duration, intonation

1. INTRODUCTION

Amharic (አማርኛ) – the official language of Ethiopia – is the world's second widely spoken Semitic language with at least 27 million native speakers. In contrast, Amharic is an under-resourced language with very limited research on acoustic features and spoken language technologies. The Amharic language is based on a syllabic alphabet and in a simplified way, 33 initial consonants are combined with 7 final vowels, resulting in 231 CV syllables with 196 different pronunciations. The Amharic script (Fidel) is coded using the Ge'ez system as shown in Table 1. For example, the syllable transcription **ho** (see also Figure 3a) is equivalent to the Fidel character [ሆ] in the table.

Beside the creation of speech corpora, prosodic analysis and generation play a critical role for the algorithmic development in speech synthesis and recognition. Unlike English in which the speech rhythm is mainly characterized by stress, Amharic rhythm is marked by a varying syllable length depending on gemination of consonants and by certain phrasing features [1, 2]. The gemination allows for the word distinction and for the proper pronunciation in terms of naturalness. First approaches to Amharic speech synthesis were described in [3, 4]. Anberbir and Takara [4] manipulated syllable stress. In [5], Grimay proposed a qualitative prosody model of Amharic following the ToBI approach [6]. He analyzed intona-

Table 1. Syllabic alphabet of Amharic – initials and finals.

consonants (initial)	vowels (final)						
	a	**u**	**i**	**A**	**E**	**e**	**o**
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
h	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
s	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
S	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
c	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ
h	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
N	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ
H	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
K	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ
w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
H	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
Z	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ
y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
D	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
T	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
C	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ
P	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
x	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
x	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

tion and duration in some phrases – without systematically surveying a speech corpus. We developed a prosodic corpus, extracted from a training database for Amharic speech recognition, and analyzed the prosodic variations in f_0, duration and intensity with regard to a syllable-based stress annotation. To test the relevance of the observed deviations in stressed syllables, we manipulated f_0 and duration in natural voices, aiming at accentuation or deaccentuation in resynthesized phrases. In a listening experiment, we evaluated the stress localization and overall naturalness.

2. EXAMINATION METHODS

2.1. Speech data collection

The data collection originally targeted the training of an Amharic speech recognizer [7] and involved 73 male and 60 female native speakers. Each speaker read 60...150 phonetically balanced sentences from different text sources (Internet, news etc.). The majority of data (120 speakers) was captured in quiet office environment, and 13 speakers were recorded in a studio – both groups at 16 kHz, 16 bit PCM. Some signals were amplified according to voice strength and to achieve a proper signal-to-noise ratio. In total, the database includes 30 hours read speech.

2.2. Test corpus and prosodic analysis

For prosodic analysis, we randomly selected 200 sentences from 10 male and 10 female speakers. The syllable segmentation – accordingly duration and intensity analysis – relied on forced ASR alignment [7]. According to his audio perception, a native speaker of Amharic annotated three stress levels: unstressed (in the following "neutral"), stressed or reduced syllable. The duration d_{syl} was measured at successive syllable boundaries without considering signal pauses or fillers. The intensity was calculated as effective value of a segment with N samples:

$$RMS = \sqrt{\frac{1}{N}\sum_{k=i}^{i+N} x^2(k)} \qquad (1)$$

and then logarithmized ($log.RMS$). The f_0 was extracted using the ESPS algorithm [8] from WaveSurfer software (version 1.8.8). Segments containing voiced parts with an f_0 below 60 Hz (caused by glottalization effects or algorithmic restriction) were excluded from intonation analysis – reducing the number of syllables by 20...30 %. Table 2 shows the prosodic test corpus in overview.

For each syllable segment, we observed minimum, mean and maximum f_0 and the f_0 range (difference of maximum and minimum). Finally, the values of d_{syl}, $log.RMS$ and f_0 were averaged over all utterances and speakers within the same gender.

Table 2. Number of syllables in prosodic corpus.

analyzed parameters	neutral	stressed	reduced
d_{syl} (♂), $log.RMS$ (♂)	4,371	456	27
d_{syl} (♀), $log.RMS$ (♀)	3,367	210	30
f_0 (♂)	3,029	273	7
f_0 (♀)	2,693	195	6

2.3. Resynthesis and perceptual evaluation

To test the perceptual relevance of observed prosodic variations, we manipulated six phrases (4...9 words) of a male speaker. In three phrases, we increased duration and mean f_0 in a single syllable to form a synthetic accent. In the other three phrases, duration and mean f_0 of a stressed syllable were decreased – aiming at deaccentuation. All modifications were performed with the "overlap-add" resynthesis algorithm from Praat software [9]. The combination of duration and pitch modifications: $d_{syl} \pm 35$ %, $+65$ %, $+100$ % and $f_0 \pm 10$ %, ± 25 %, ± 50 % resulted in 45 accentuation and 21 deaccentuation stimuli. Completed by six natural stimuli, we presented 72 test phrases to six native speakers of Amharic (male, 31.5 ± 3.7 years). We tested the perceived stress positions. To avoid misunderstanding among naives, we asked for emphasized words instead of syllable positions. In addition, the listeners evaluated the naturalness of stimuli in absolute category rating from 1 "bad" to 5 "excellent" (MOS scale).

3. RESULTS AND DISCUSSION

3.1. Stress and intonation

Table 3 summarizes the syllable-based mean f_0 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 f_0 values as known from other languages. The increased f_{0min} in reduced syllables might be a result of co-articulation or a mistake due to rare occurrence.

Table 3. Stress level and mean f_0 variation in syllable (♂).

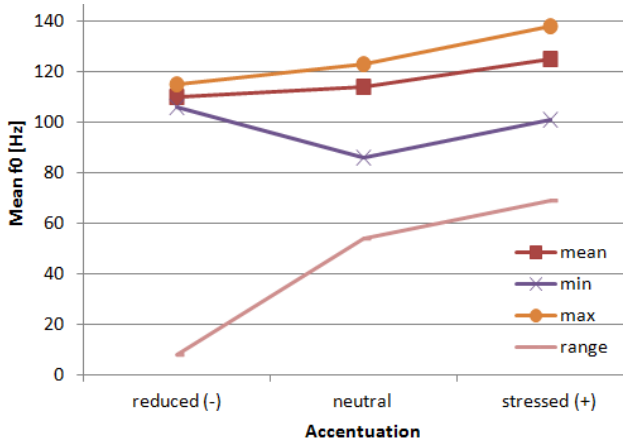
	neutral	stressed (+)	reduced (−)
f_{0mean}	114 Hz	125 Hz (+10%)	110 Hz (−4%)
f_{0min}	86 Hz	101 Hz (+17%)	106 Hz (+5%)
f_{0max}	123 Hz	138 Hz (+12%)	115 Hz (−7%)
f_{0range}	54 Hz	69 Hz (+28%)	8 Hz (−85%)

Female speakers (Table 4) feature a similar relative f_0 variation in stressed and reduced syllables.

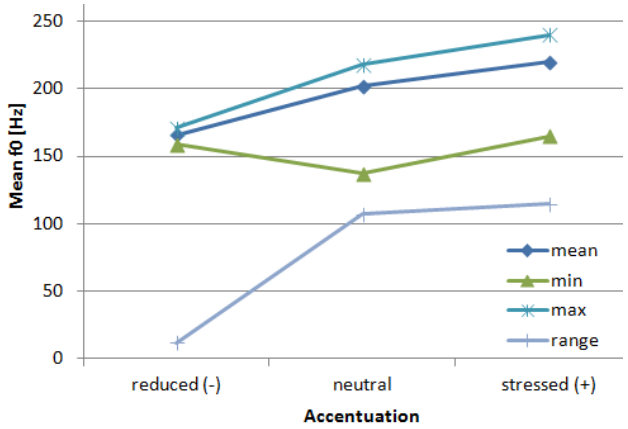
For stressed syllables, the average deviation of f_{0mean}, f_{0min} and f_{0max} amounts to 13.1 % in both genders. The Figures 1a and 1b compare the f_0 variations for both genders.

Table 4. Stress level and mean f_0 variation in syllable (♀).

	neutral	stressed (+)	reduced (−)
f_{0mean}	202 Hz	220 Hz (+9%)	166 Hz (−18%)
f_{0min}	137 Hz	165 Hz (+20%)	159 Hz (+16%)
f_{0max}	218 Hz	240 Hz (+10%)	171 Hz (−22%)
f_{0range}	107 Hz	115 Hz (+7%)	12 Hz (−89%)



(a) Male speakers



(b) Female speakers

Fig. 1. Gender-specific f_0 variation in syllable (averaged).

3.2. Syllable duration

Table 5 summarizes the stress-related variation of the syllable duration in both genders. The average deviation of stressed syllables ($\Delta d_{syl} = 41.5$ ms) is equivalent to a difference of 32.7 % and confirms the assumption with regard to the importance of duration in previous studies [2, 4]. Figure 2 compares the average duration modification for both genders.

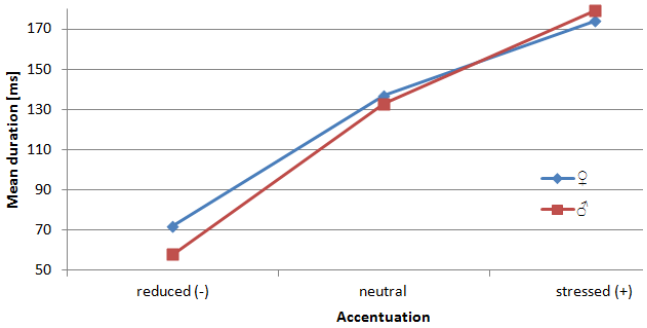


Fig. 2. Stress-related modification of mean duration.

3.3. Intensity variation

The mean intensity modification in stressed syllables of 8.5 % in Table 6 suggests a lower importance of intensity parameters in Amharic prosody generation and perception.

Table 6. Stress-related modification of mean intensity.

	neutral	stressed (+)	reduced (−)
$log.RMS$ ♂	38 dB	40 dB (+5%)	31 dB (−18%)
$log.RMS$ ♀	43 dB	48 dB (+12%)	32 dB (−26%)

3.4. Interrelation between duration, f_0 and intensity

To survey interrelations among the prosodic parameters, we chose the 20 syllables with lowest and highest Δd_{syl} (stressed vs. neutral) – assuring a minimum frequency of occurrence of 15 neutral and 5 stressed segments. Figure 3 shows the mean prosodic parameters for these syllables, sorted by their Δd_{syl}.

Compared with their neutral reference, stressed syllables in Figure 3a show higher (apparently uncorrelated) deviations in mean f_0 and intensity than the syllables in Figure 3b. The results indicate different strategies in stress forming, e. g. duration vs. f_0 / intensity-based.

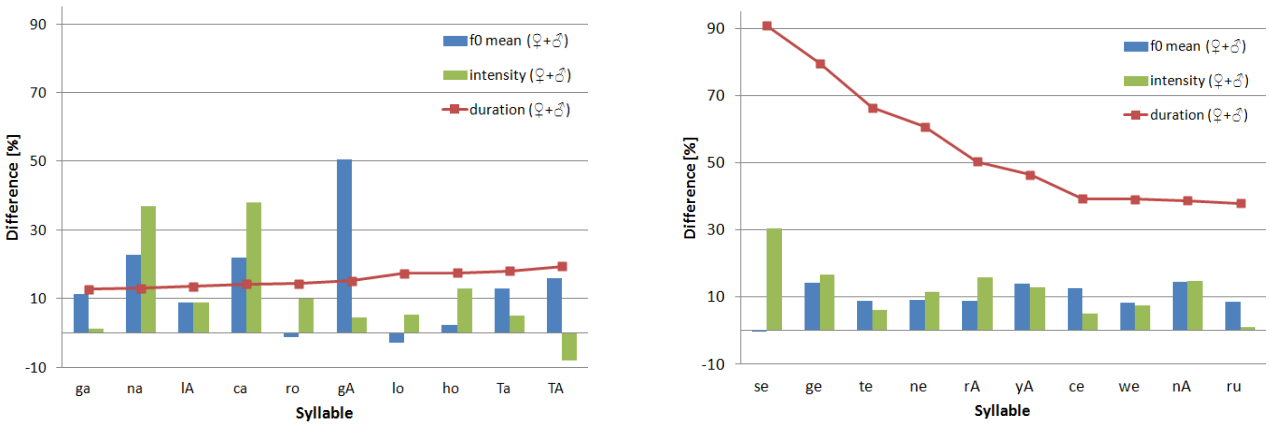
The dominating vowels /a/, /A/, /o/, and /e/ refer to phonological aspects of duration. Table 7 compares the correlations between duration (d_{syl}), mean f_0 and intensity ($log.RMS$) and confirms the assumptions for our limited data sets.

Table 5. Stress-related modification of mean duration.

	neutral	stressed (+)	reduced (−)
d_{syl} (♂)	133 ms	179 ms (+35%)	58 ms (−56%)
d_{syl} (♀)	137 ms	174 ms (+27%)	72 ms (−47%)

Table 7. Pearson's correlation ρ among prosodic parameters.

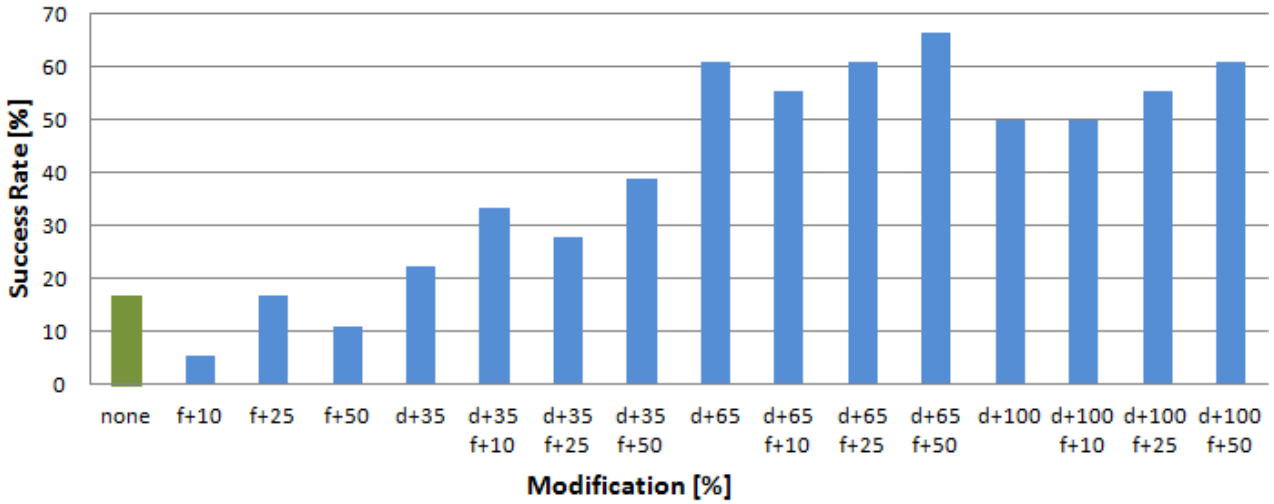
data set	dur. vs. $f0$	dur. vs. int.	$f0$ vs. int.
lowest Δd_{syl}	−0.18	−0.52	0.14
highest Δd_{syl}	−0.51	0.72	−0.46



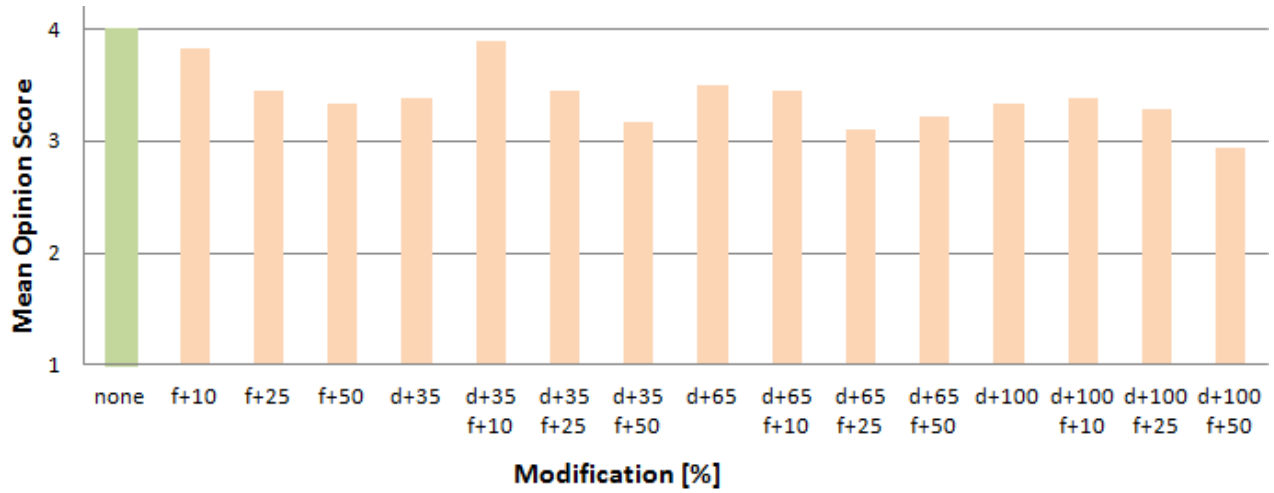
(a) Syllables with the lowest Δ in duration

(b) Syllables with the highest Δ in duration

Fig. 3. Relative parameter modification for both genders in selected stressed syllables (ordered by Δ in duration).



(a) Identification of the intended accent position

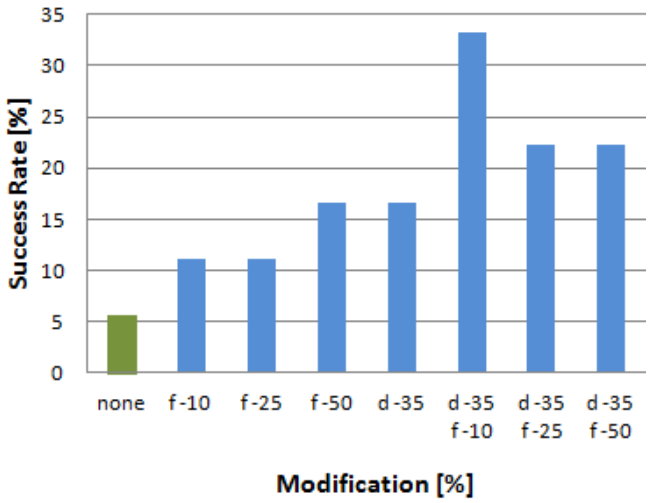


(b) Naturalness of the manipulated utterance

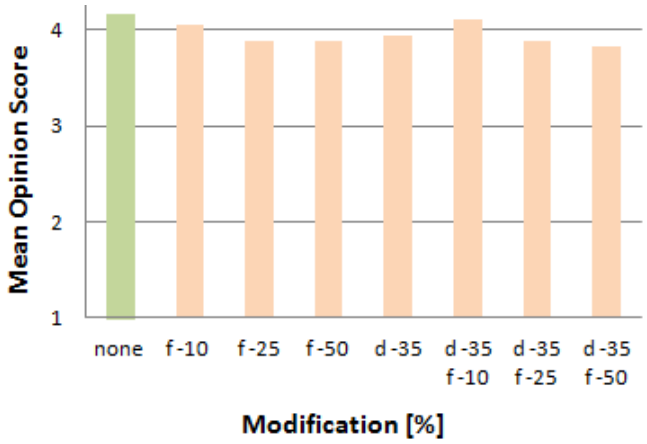
Fig. 4. Perceptual evaluation of synthetic syllable **accentuation** – depending on strength of manipulation.

3.5. Perception experiment

Figure 4a shows the perceptual identification of the synthetic accents for combined manipulations of syllable duration and f_0. The category "none" characterizes non-manipulated (natural) reference phrases. Higher modifications of duration and f_0 are not necessarily leading to a better detection of synthetic accents – the best success rate of 66.7 % was achieved for the combination of $d_{syl} + 65$ % and $f_0 + 50$ %. By contrast, the combination of $d_{syl} + 35$ % and $f_0 + 10$ %, which simulates the mean natural deviations (cf. sections 3.1 and 3.2), received the highest MOS of 3.89 in naturalness – comparable with the assessment of the natural voices (MOS of 4.00) as illustrated in Figure 4b. The success rate in synthetic deaccentuation (Figure 5a) was lower but the naturalness scores achieved the same level as in the accentuation test, and the near-to-natural simulation with $d_{syl} - 35$ % and $f_0 - 10$ % performed best (Figure 5a and Figure 5b).



(a) Identification of the intended accent deletion



(b) Naturalness of the manipulated utterance

Fig. 5. Perceptual evaluation of synthetic **deaccentuation**.

4. CONCLUSION

Amharic is one of the least researched languages in the world. Assuming a baseline approach with three stress levels, we have studied syllable-based prosodic attributes in read speech. We have confirmed the important role of duration suggested by previous studies on phoneme level.

In addition, we examined stress-related modifications in f_0 and intensity. They show similar variations and interrelations as known from e. g. Indo-European languages. To test the perceptual relevance of observed variations in duration and f_0, we manipulated single syllables – aiming at synthetic accentuation or deaccentuation – and presented resynthesized phrases to native probands who recognized a significant part of intended accent positions.

The analysis results can be used in the development of prosody models for speech synthesis or recognition. Another application domain is related to intelligent language tutoring systems.

5. REFERENCES

[1] M.L. Bender, J.D. Bowen, R.L. Cooper, C.A. Ferguson. Language in Ethiopia, Oxford University Press, 1976.

[2] T. Anberbir, T. Takara, and D.Y. Kim, "Modeling of geminate duration in an amharic text-to-speech synthesis system." Proc. 2^{nd} Workshop on SLT for Under-Resourced Languages, Penang, pp. 122–129, 2010.

[3] S.H. Mariam, S.P. Kishore, A.W. Black, R. Kumar, and R. Sangal, "Unit selection voice for amharic using festvox." Proc. 5^{th} ISCA Speech Synthesis Workshop, Pittsburgh, pp. 103–107, 2004.

[4] T. Anberbir and T. Takara, "Amharic speech synthesis using cepstral method with stress generation rule." Proc. Interspeech (ICSLP) Pittsburgh, pp. 1340–1343, 2006.

[5] G. Girmay, "Prosodic Modeling for Amharic." Master thesis, Addis Abeba University, 2008.

[6] M.E. Beckman and G.A. Elam, "Guidelines for ToBI labeling." Manuscript, Ohio State University, 1994.

[7] Y. Birhanu and R. Hoffmann, "Development of automatic amharic speech recognizer.", Proc. ESSV conference, pp. 118–122, Aachen, 2011.

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

[9] P. Boersma and D. Weenink, Praat: doing phonetics by computer (version 5.3.05). Retrieved February 24, 2012 from http://www.praat.org