Amharic Speech Synthesis Using Cepstral Method with Stress Generation Rule

Tadesse Anberbir, Tomio TAKARA

Faculty of Engineering, University of the Ryukyus, Okinawa, Japan
takara@ie.u-ryukyu.ac.jp

Abstract

Amharic is the official language of Ethiopia. In this paper, we present our study on Amharic stress. Stress (Gemination of consonants) in Amharic language is very important for proper pronunciation of words. It is also one of the most distinctive characteristics of the rhythm of the speech. We discuss a method employed for generating stressed syllables from unstressed syllables, and its application to our speech synthesizer. First, we analyzed waveforms of minimal pair words concerned with stressed and unstressed syllables into the time patterns of pitch, power and spectrum. Then, by combining or exchanging these patterns, speech sounds were synthesized. Using the synthesized sounds, listening tests were performed to examine the acoustic correlates of stress among pitch, spectrum, power and duration. We found that consonant's duration is the most important factor. A further listening test was performed to determine the threshold of duration of consonants between unstressed and stressed syllables, and we observed that 50ms is the average threshold duration for voiced consonants and 70ms is for unvoiced consonants.

Index Terms: Amharic, stress, speech synthesis, cepstrum

1. Introduction

The general objective of the present research is to develop a speech synthesizer for Amharic language. Amharic is the official language of Ethiopia and it is one of the most widely spoken languages in the country. It has its own script called "Fidel". The orthographic representation of the language is syllabic and it is organized into orders. It has 32 consonants and 7 vowels. Each of the 32 consonants has seven derivative syllables.

Amharic words are mainly characterized frequent presence of gemination. Gemination in Amharic is one of the most distinctive characteristics of the cadence of the speech, and also caries a very heavy semantic and syntactic functional weight [1]. As it is in many other languages such as Italian, gemination in Amharic is phonemically meaningful. For example, /gena/ (yet) and /genna/ (Christmas) are written as "ገና", but gives different meaning by geminating the consonant "ና" /n/. Gemination in Amharic is often done with emphasis (stress) [2].

Stress is the relative emphasis given to a certain syllables in a word. The feature of stress in speech stream is highly language dependent. Stress in Amharic is related with gemination. Getahun [3] defines stress as a phonological phenomenon related with longer production of consonants and louder perception. In our study, we consider gemination as a physical phenomenon related with duration (production) and stress as a psychological phenomenon related with prominence (perception).

Stress in Amharic language is very important for proper pronunciation of words and plays a key role in speech synthesis because it provides information that increases the intelligibility and naturalness of synthesized speech. However, the acoustic correlate of Amharic stress is not studied. As far as our knowledge, there is no research conducted on the acoustic analysis of Amharic stress particularly in relation to speech synthesis.

In this paper, we present a syllabic based Amharic speech synthesizer with stress generation rule. Section 2 discusses the overview of Amharic speech synthesis system. Section 3 explains about the acoustic characteristics of stress. Section 4 and 5 discusses the experimental results performed to determine, the acoustic correlates of the stress and the threshold duration of consonants respectively. Discussion about stress generation rule and conclusion are given in section 6 and 7 respectively.

2. Amharic speech synthesis system

Speech synthesis is a process that artificially produces speech for various applications. Many speech synthesis systems are available for different languages (like English and Japanese). However, researches on Amharic speech synthesis [4] are very limited.

Amharic speech synthesis system is a parametric and rule based system designed based on the general speech synthesis system [5]. Figure 1 shows the scheme of Amharic speech synthesis system. The system has two main components, a text analysis subsystem and a speech synthesis subsystem. First, the input is analyzed by a text analysis subsystem. Then, synthetic speech output is produced by the synthesis subsystem using linguistic and prosodic information (like stress) that are extracted from the text analysis subsystem. The database contains syllable parameters consisting of voiced/unvoiced decision parameters, pitch period and cepestral coefficients. The system rule section contains syllable connection rules and stress generation rule.

In the synthesis part, the synthetic sound is produced using Log Magnitude Approximation (LMA)[6] filter as the system filter, for which cepstral coefficients are used to characterize the speech sound. The LMA filter is controlled by cepstrum parameters as vocal tract parameters, and driven by fundamental period impulse series for voiced sounds and by white noise for unvoiced sounds. The gain of the filter or the power of synthesized speech is set by the 0th order cepstral coefficient, c[0]. The fundamental frequency (F0) of the speech is controlled by the impulse series of the fundamental period.

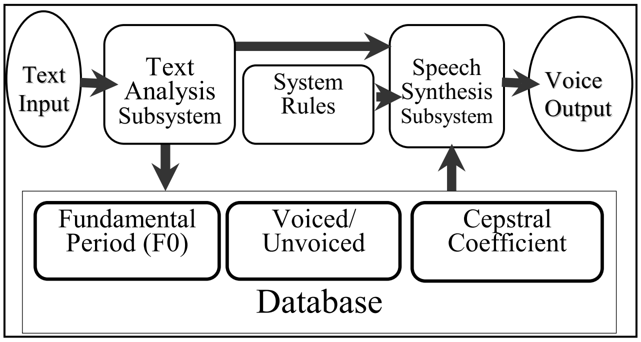


Figure 1: Amharic Speech Synthesis System

3. Acoustic characteristics of Amharic stress

In order to study the acoustic characteristics of Amharic stress, first, we extracted all the possible minimal pair of contrastive words which differ only by presence or absence of geminated consonants. Then, we estimated the time-varying patterns of their power and pitch, and observed the difference. The number of extracted minimal pairs was 15. Among these, we selected 6 pair of words (as shown in Table 1) which can be used as a representative for continuant (voiced and unvoiced) consonants. However, for non-continuant consonants, we could not find such comparable words. The pairs of contrastive words were selected from Amharic-English dictionary and they are written with the same character but have different meaning due to the geminated consonants. In our study, we used an apostrophe mark (’) as a stress marker preceding the geminated consonants. All words were uttered by a male speaker. Then, sampled at 10 kHz and quantized into 16 bit.

Table 1: Minimal pair of words with stressed and unstressed syllable

WORD	AM	MEANING	WORD	AM	MEANING
/gena/ /ge’na/	ገና	(still/yet) (christmas)	/kefa/ /ke’fa/	ከፋ	(region name) (worse)
/lega/ /le’ga/	ለጋ	(fresh) (hit)	/sefii/ /se’fii/	ሰፊ	(tailor) (wide)
/wana/ /wa’na/	ዋና	(swimming) (main/core)	/sxixfixta/ /sxix’fixta/	ሽፍ:ታ	(rebel) (rash)

AM=Amharic script

We mainly observed that there is a large duration difference between the singletons (much shorter) and geminates (much longer), and vowels preceding the singletons tend to have shorter length than vowels preceding geminates. We also observed that stressed words have longer duration, higher pitch and greater intensity than unstressed words. However, intensity and pitch are significantly affected by the position. For example a C_1V_1 syllable with geminated consonant has higher pitch and greater intensity only when it precedes the same C_1V_1 syllable with singleton but not follows it. This shows that the higher pitch and the greater intensity may not be necessarily caused by the stress. Therefore, we assumed that duration is the main factor for stress than pitch and intensity.

4. Acoustic correlates of the stress

Stress is a phonological phenomenon and it is related to perception of prominence. This prominence can be realized acoustically as any of the combination of greater intensity, longer duration or higher pitch. Stress is usually considered to be the vowel feature, however in our study, we presupposed that consonants to be the main factor for Amharic stress.

In order to confirm our supposition, first, we analyzed speech sounds of minimal pair words (Table.1) into time patterns of pitch, power and spectrum. Then, by combining or exchanging these patterns with the time axis of stressed words, we synthesized sixteen kinds of speech data and performed a perceptual listening test.

4.1. Stimuli

The minimal pairs of words used in this experiment were /kefa/, /ke’fa/; /sefii/, /se’fii/; /wana/, /wa’na/. The words were uttered by male speaker . The filter used for synthesizing the speech was the Log Magnitude Approximation filter (LMA filter) as explained in section 2. The speech sound was sampled at 10 kHz and quantized into 16 bit. Pitch, power and spectrum were estimated at a frame length of 25.6ms and frame shift of 10ms.

The number of synthesized speech words used for the listening test was 16, among which, 2 were analysis/synthesis sounds and 14 were synthesized by exchanging or combining the F0 patterns, power patterns, spectrum patterns and the time-axis (duration) patterns between unstressed and stressed words. Two original speech words were also added for comparison purpose. In this study, we use the term spectrum to mean the log power spectrum of speech because the magnitude of sound is perceived in log scale. The spectrum can be exchanged by exchanging the cepstral coefficients because the log magnitude spectrum is set by the cepstral coefficients. Exchanging conditions of the synthesized sounds are shown as DATA(c)–DATA(p) at the bottom of fig.3. The numbers “0” and “1” shows which parameter is exchanged. “1” is for parameters come from stressed words and “0” is for parameters come from unstressed words. For example, in the case of word /kefa/, DATA(c), “0001” shows the synthesized speech in which the spectrum of unstressed word /kefa/ is exchanged, i.e., the spectrum from stressed word /ke’fa/ is used and the time-axis (duration), power and F0 from unstressed word /kefa/ are used for synthesis. DATA(a) and DATA(r) are original speech and DATA(b) and DATA(q) are analysis-synthesis speech.

4.2. Matching of stressed and unstressed words

Although the duration of words changes every time with non-linear expansion and contraction depending on the speaking-rate, stressed words have longer duration than unstressed words mainly due to the long consonant of stressed syllables. Because of this duration mismatch between stressed and unstressed words, we cannot exchange the parameters directly. Therefore, in order to exchange the parameters appropriately, we employed a “Dynamic Programming (DP) Matching Method” and expanded unstressed words with the time-axis of stressed words using the spectral distance.

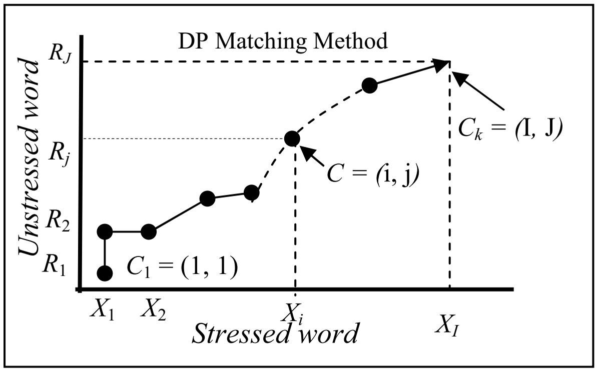


Figure 2: The most appropriate matching sequence.

Xi is defined as the i^{th} vector of the spectral pattern of stressed word; *Rj* is defined as the j^{th} vector of the spectral pattern of unstressed words. In equation (1), $d(i, j)$ stands for the difference between spectral parameters.

$$d(i,j) = |Xi-Rj| \qquad (1)$$

Furthermore, the distance of the matching sequence $d(i, j)$ between each frame is calculated with equation(3).

$$g\,(1, 1) = 2d(1, 1) \qquad (2)$$

$$g\,(i, j) = min \begin{pmatrix} g(i,j\text{-}1)+ d(i,j) \\ g(i\text{-}1, j\text{-}1)+ 2d(i,j) \\ g(i\text{-}1, j)+ d(i, j) \end{pmatrix} \qquad (3)$$

The initial condition is set in equation (2). Then, equation (3) is calculated, increasing *i* until *i*=*I*. The procedure is then repeated until *j* =*J*. As shown in Fig.2, the values from equation (3), which select minimum value, are stored, and then we require the most appropriate matching sequence by tracing the path in reverse to get the corresponding frame.

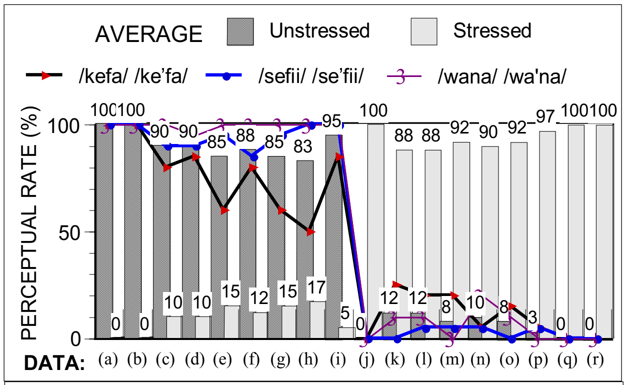
4.3. Procedure of the listening test

The listening test was performed in a soundproof room using a headphone by three native speakers of the language. All listeners have normal hearing ability and participated in the word intelligibility test we performed in our previous study. Each sound was played to each listener randomly and twice in two-second interval. The listener listens to the sound and selects what he/she perceived among the list of six words (three pairs of words). Each listener performed the listening test ten times. Totally each word data was presented 3x10=30 times.

4.4. Results and discussion

Results of the listening tests are shown in Fig.3. The vertical axis of the figures shows the perceptional rate (%) and the horizontal axis DATA(a) -DATA(r) shows the data type. The bar graph shows the average of three words. In the bar graph, the number indicated at the shaded part is the percentage of responses that the words perceived as unstressed and that of the white part is for the words perceived as stressed. The line graph shows the result for each pair of words.

In the figure, we can see that all unstressed words lengthened with the time axis of stressed words, i.e., DATA(j) (1000), perceived as stressed and most of stressed words shortened with the time axis of unstressed words, i.e., DATA(i) (0111), perceived as unstressed. However, power, F0 and spectrum exchanged words shows few changes. This shows that duration is the most important acoustic feature correlate with stress than power, F0, and spectrum.



	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)	(o)	(p)	(q)	(r)
Duration:	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Power:	0	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	1
F0:	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	1
Spectrum:	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1

Figure 3: Result of the listening test concerning the contribution of duration, power, F0, and spectrum.

5. Threshold duration of consonants between stressed and unstressed syllables

In order to determine the threshold duration of consonants between unstressed and stressed syllables, we performed two listening tests by grouping the above six pair of words (shown in Table 1) into two groups.

5.1. Stimuli

For both listening tests, we prepared 16 types of data for each word, among which, twelve were synthesized by repeating the parameters of the consonant part of unstressed syllable. The conditions of repetition are shown as DATA(c) -DATA(n) at the bottom of Fig.4 and Fig.5. And two were analysis-synthesis words with unstressed and stressed syllables shown as DATA(b) and DATA(o), respectively. Two original speech words were also added for comparison purpose and shown as DATA(a) and DATA(p). In the synthesized words, the parameters of the consonant part of unstressed syllable (DATA(b)) are repeated one frame per data. That means the duration of each data from DATA(c) - DATA(n) has increased by an interval of 10ms. These are: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, and 120 [ms] plus the duration of consonant part of unstressed syllable.

5.2. Procedure of the listening test

The procedure for the listening test was the same as the previous experiment except that the sound played only once.

5.3. Results and discussion

Results of the listening tests are shown in Fig.4 and Fig.5. The vertical axis of the figures shows the perceptional rate (%). In both figures, the bar graph shows the average of words. In the bar graph, the number indicated at the shaded part is a percentage of responses that the words perceived as unstressed and that of the white part is for the words perceived as stressed. The line graph shows the result for each pair of words. In both figures, we can see that the more the duration increase to the right, the more the words perceived as stressed. The listener perceived the difference, but, they did not perceive the feature as long consonant duration rather they perceived as stress.

Fig.4 shows that the average threshold duration of voiced consonants is 50ms, and fig.5 shows the average threshold duration of unvoiced consonants is 70ms. Note that, in Fig.5, we did not include the result of word /sxixfixta/ which is with 6th ordered syllable because its result is different comparing with the other two words. Amharic 6th order syllables are exceptional and have special property comparing with other orders. We observed that the unstressed 6th order syllable /fix/ in a word /sxixfixta/ is without vowel but when it becomes stressed it is associated with epenthetic vowel /ix/. In the listening test, all 12 synthesized speeches for word /sxixfixta/ were perceived as unstressed. This shows that vowels are needed for syllables to be stressed.

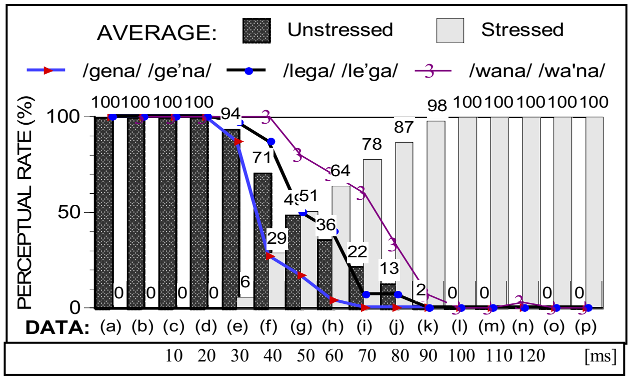


Figure 4: Result of the listening test for pair of words with voiced consonants.

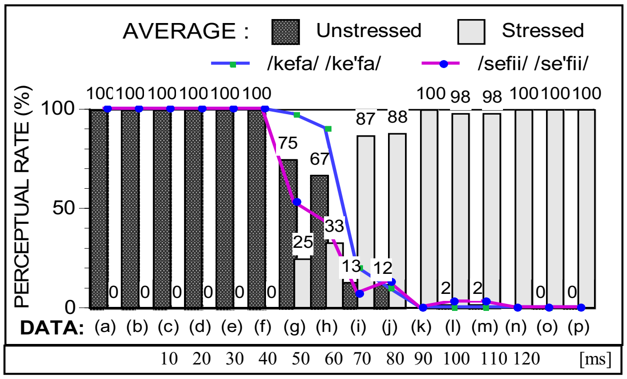


Figure 5: Result of the listening test for pair of words with unvoiced consonants.

6. Stress generation rule

In our previous study, we had evaluated the intelligibility of our system and we found 100% correct-rate. However, although, the intelligibility of synthesized words was very high, it lacks naturalness mainly because we did not considered words with stressed syllables. Among the 200 words we used, 20% of the words were with one or more stressed syllables. In this study, we prepared a stress generation rule and employed in our previous system. The stress generation rule is programmed in the system and it generates stressed words by lengthening the duration of the consonants. For syllables with continuant (voiced and unvoiced) consonants, the parameter (F0, power and spectrum) of the first three frames of the consonant part is repeated for 90ms. For non-continuant consonants, a 100ms silence is added preceding the stressed syllables. For example, if the user inputs the word "ge'na" where apostrophe (') is the stress mark, the duration of the consonant /n/ will be lengthened by 90ms. We found 90ms to be long enough to make the syllables stressed. Our stress generation rule is very simple but it is very effective and made a significant better natural quality.

7. Conclusions

In this paper, we examined the acoustic correlate of Amharic stress, determined the threshold duration of consonants, and prepared a stress generation rule in Amharic speech synthesizer. Though stress is usually considered to be a vowel feature, in our study we found consonants to be the main factor for Amharic stress. Our stress-rule achieves 100% correct-rate to generate stressed words except for words with 6th order syllables and significantly improves the naturalness of synthesized stressed words.

However, our system does not have a mechanism to identify the stressed syllables from the input text, unless the user inputs the stress mark. As a future work, we have a plan to prepare a lookup-dictionary as a reference to locate the stressed syllables. It is also important to study the effect of stress on the neighboring syllables especially on vowels preceding geminates. Finally, we have a plan to extend our work at sentence level.

References

[1] M. Lionel Bender, Hailu Fulass, "Amharic Verb Morphology: A Generative Approach", Carbondale, 1978.

[2] C.H DAWKINS, "The Fundamentals of Amharic", Bible Based Books, SIM Publishing, Addis Ababa, Ethiopia, pp.5-7,1969.

[3] Getahun Amare, "Modern Amharic Grammar in a Simple approach", 96 (in Amaharic)

[4] Sebsibe H/Mariam, S P Kishore, Alan W Black, Rohit Kumar, and Rajeev Sangal,"Unit Selection Voice for Amharic Using Festvox", 5th ISCA Speech Synthesis Workshop, Pittsburgh, 2004.

[5] T. Takara and T. Kochi, "General speech synthesis system for Japanese Ryukyu dialect", Proc. of the 7th WestPRAC, pp. 173-176 ,2000.

[6] S. Imai, "Log Magnitude Approximation (LMA) filter",Trans. of IECE Japan, J63-A, 12, PP. 886-893 ,1980. (in Japanese)