Journal Of Harmonized Research (JOHR)

Journal Of Harmonized Research in Engineering
2(1), 2014, 162-168

ISSN 2347 – 7393

Original Research Article

TEXT TO SPEECH SYNTHESIS OF GALO AND ADI LANGUAGES USING POLYSYLLABIC UNITS

Laba Kr. Thakuria, Prof. P.H. Talukdar

Department of Instrumentation & USIC, Gauhati University
Guwahati, India

Abstract
This research paper deals with the design and development of Galo and Adi language Text-To-Speech (TTS) synthesis systems, using polysyllabic units. First we have create a phone based Text To Speech and then a syllable cluster unit Text To Speech. We have observed that the performance of the synthesized sentences can improve using polysyllable units, since the effects of co-articulation will be preserved in such a case and hence, we create Galo and Adi Text To Speech with polysyllabic units that contains cluster units of more than one type (monosyllable, bisyllable and trisyllable). The system selects the best set of units during the unit selection process, so as to minimize the join and concatenation costs. In preliminary listening tests indicate the polysyllable Text To Speech has better performance.

Key Word: - galo , adi, tts, polysyllabic units .

Introduction

The Concatenative speech synthesis systems combine the sound units from the database to generate the utterance. The main advantage of using unit selection based concatenative synthesis is that there may not be a need for separate prosody modeling for the availability of many units under varied contexts. The sound units could be a phoneme, diphone, syllable or word. To build Galo and Adi language speech synthesis systems, its more appropriate to use syllables as the the basic unit. For example a syllable can be defined as taking the form D*VD*, where 'D' denotes a consonant and 'V' denotes a vowel. The experiment of Kishore and Black suggests the usage of syllables as the basic unit for any Indian languages. The basic advantages of using syllables as basic units is that they have fairly long duration when compared to phonemes or diphones and hence, the task of segmentation becomes relatively easier. Also, since the boundaries of most of the syllables are low energy regions the concatenations will result in reduced

For Correspondence:
thakurialabaATgmail.com
Received on: January 2014
Accepted after revision: February 2014

perceivable distortions. The basic case of making use of poly-syllables is that, as the number of concatenation points will be less, the synthesized speech quality is expected to be higher performance. It so happen that the Festival speech synthesis system chooses two units from the different parts, even when both the units occur in sequence in the database. This will indicate that the criteria used by Festival for unit selection may be inappropriate, which can be avoided by making use of polysyllabic units because polysyllable units are formed using the monosyllable units which is already present in the database and the synthesis quality can be improved without augmenting any new set of units or realizations. The method that we have discussed in the paper is different in the sense that the cluster units are not just built for one type of unit, but for different types of units say mono, bi and tri syllables and then we combine the results of three different phases i.e. for mono, bi and tri syllables of training appropriately to produce polysyllable TTS. The usage of pronunciation dictionary would help in choosing the appropriate sound unit, based on context, in the case of Galo and Adi languages. The research paper is organized as - **Section I** describes the design considerations involved to build a polysyllable TTS and the steps pertaining to corpus collection, syllable coverage, building pronunciation dictionary and voice talent/artist selection process. **Section II** describe the implementation aspects of building a polysyllable TTS using the Festival framework and elaborates on the labeling process (using both Ergodic Hidden Markov Model [EHMM] **[**8**]** and group delay **[**9**]** based segmentation algorithms) and on the multiple stages of TTS training/building, in order to produce polysyllabic speech synthesis. **Section III** describe the possible research directions and finally, **Section IV** describe conclusions.

(i) Galo

The Galo language(Agom) is inherited from the Tibeto-Burman family of languages. This language is mostly spoken by the Galo people. of Arunachal Pradesh. The Galo is one of the major tribes of Arunachal Pradesh. Around 95% of Galo people learn Galo as a first language, although most are also bilingual and borrow frequently from Assamese, Hindi and English(The major languages of Indian subcontinent). In the Arunachal Pradesh there are total 25 major tribes and almost 110 sub- tribes. There is high degree of mutual intelligibility among the different languages of Arunachal Pradesh like language spoken by the Adis, Apatanis, Galos, the Hill Miris, the Nyishis and the Tagins. Moreover they share many characteristic features in their cultural code and trace their ancestry from a common forefather, namely Abotani. Hence, the language spoken by them can rightly be given generic name –Tani language[37]. The languages in Arunachal Pradesh can broadly be classified into two groups: namely Abotani group and Non-Abotani(Buddhism). The major Galo dialects are Pugo, spoken around the district capital(Itanagar), Aalo and Lare spoken in the south of Aalo, and Subdialects are numerous, and often correspond to regional or clan groupings. The Galo have decided to adopt the Roman script with certain additions of two owels and two consonants whose phonetics are common in international phonetic alphabets (IPA). The most common additions are the Roman symbol v,w,q and x, which are not otherwise use for writing Galo. These symbols now represent IPA ə, ɨ, ŋ and ɲ/ñ respectively. With the assignment of Galo sounds to the Roman letters as given below, the Galo Script has devised as accurate and as practical as possible.

Table 1: Script Comparison of Galo and Adi Vowels

Galo	**A A**	**I I**	**U u**	**E E**	**O O**	**V V**	**W W**
Adii	**A A**	**I I**	**U U**	**EY EY**	**O O**	**E E**	**UI UI**
Phonetic	**A**	**I**	**U**	**E**	**O**	**Ə**	**ɨ**

Table 2 Script Comparison of Galo and Adi Consonant

Galo	**K k**	**G G**	**Q Q**	**C C**	**J J**	**X X**	**T T**	**D d**	**N n**	**P P**	**B B**	**M m**	**Y Y**	**R R**	**L L**	**S S**	**H H**
Adii	**K k**	**G G**	**NG ng**	**CH Ch**	**J J**	**NY Ny**	**T T**	**D d**	**N n**	**P P**	**B B**	**M m**	**Y Y**	**R R**	**L L**	**S S**	**H H**
Phonetic	**K**	**G**	**ŋ**	**tɕ**	**dʑ**	**ɲ**	**T**	**d**	**n**	**P**	**B**	**m**	**J**	**R**	**L**	**s/ɕ**	**H**

ii. Adi

The Adis have decided to adopt the Roman script with certain additions of two vowels (gaayo merey) and two consonants (merey) whose phonetics are not very common in international phonetic alphabets (IPA). The four new alphabets coined for use in Adi language (agom) are dual-letter vowels - EY & UI and consonants - NG & NY. With this restructured script, AAK, the apex literary body of Adi community, will develop study materials for use in the schools of Adi areas as third language. All books, poetries and archive writings, already in printed form, will be re-written with this script in due course of time. Describing as "historic" the reform brought in towards enriching the traditional cultural heritage of Adi community, the ABK exudes hope that this will usher in a new era of linguistic and literary development of Adi.

Table 3 Adi Vowel named as Gayo Merey

Vowels (Gaayo Merey)				
Short Vowels	Long Vowels	Adii Words	English Alike	Description of Phonetics (Tongue Position)
A	AA	ALAK/AAYI	PALM/EARN	Tip at near front lower jaw & back open(Lips/jaw normal open)
E	EE	EKUM/EEJO	-NA-	Tip floating at middle(lips normal open, jaw half normal open)
EY or E'	EEY	EYSUL/ EEYRUNG	PEN/PAIN	Tip processing lower front jaw(lips in smile position at half normal open)
I	II	IYYI/IIPANG	PIN/PEAK	Tip pressing lower front jaw(lips in smile position at jaw near close)
UI or I'	UUI	UIRBUNG/U UIK	-NA-	Tips floating middle (lips in smiling positions at jaw near close)
O	OO	OMRUI/ OOSONG	OPEN/OAK	Tip floating at middle(lips rounded at near open)
U	UU	UROM/UUT	PUT/YOUR	Tip floating at middle

There are 24 phonemes in both the languages. The phoneme consists of seventeen (17) consonant and seven (7) vowels. The structure of Sino-Tibetan family of language is as given below.

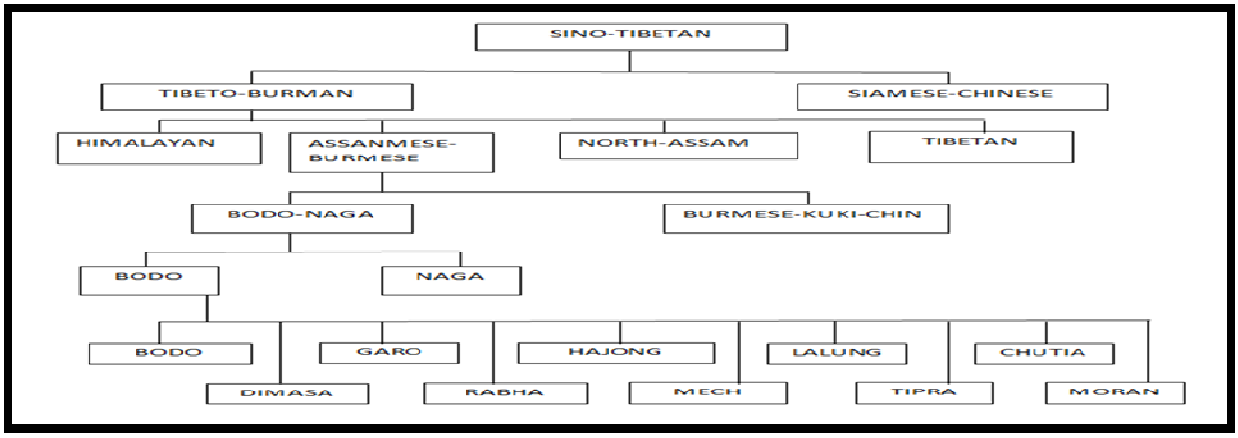

Figure 1: Sino-Tibetan family of Language

(iii) Text To Speech(TTS)

TTS requires the conversion of text into semi-continuous sounds. The most obvious approach seems to be to store digitized sound sequences representing complete words or phrases, and then simply select and output these sounds. For those who travel by train, a good example of this approach is the automated announcement system used at many UK railways stations, including New Street Station. This approach has the advantage of simplicity and can also produces very natural-sounding speech; however it also has several serious drawbacks. In a limited domain (e.g. speaking weather forecasts or train announcements), it is feasible to construct a complete dictionary giving the digitized pronunciations of all relevant words or phrases. However, this is impossible for general text. Languages contain a very large number of words, many of which occur only rarely, so that a huge dictionary would be required. Furthermore, natural languages are constantly changing, with new words being introduced (especially in the names of commercial products). Such words could not be handled by a system relying on pre-stored sounds.

I. **Design**:

A. Syllable Coverage

For syllable coverage we have consider GU_Golo_Adi corpus. The syllable set that we have considered includes V, VC, CV, CVC and CCVC and then the processed sentences are segmented into syllables and their frequency and uniqueness of occurrence are identified. The syllable occurs is taken into account, for the uniqueness of a syllable. For example, the same syllable /alo/ in Galo, occurring in three different contexts say begin, middle and end, are treated as three unique syllables. The sentences which contribute towards coverage of syllables are selected and are used for recording by the voice talent.

B. Pronunciation Dictionary

Since Galo and Adi languages are syllable-centric, so there is a need of pronunciation definition for all possible contexts. Since one-to-one correspondence between the spoken and written form does not always hold good. In the case of Galo and Adi language, the same letter is used for producing different sounds under different contexts. Thus the pronunciation dictionary for Adi and Galo language should be capable of providing multiple pronunciations for the same word, occurring in different contexts The pronunciation dictionary (Lexicon) module has many advantages over the Letter-To-Sound

(LTS) rules. But Pronunciation dictionary provides appropriate pronunciations, based on context and binary search is done in case of pronunciation dictionary, where as LTS uses linear search. The pronunciation dictionary is created using unique list of words generated from the text corpus and appropriate pronunciations are provided for the same, based on their context. Another advantage of using pronunciation dictionary is to perform a Viterbi search, so that concatenation cost is minimized.

C. Speaker selection

The speech, are recorded by multiple speakers is subjected to variations in pitch, tempo and amplitude. The level of variations done is limited such that the quality of the voice does not change. The speaker, whose voice did not loose its quality upon applying these prosodic variations, is chosen for further recording.

D. Voice Recording

The collected data is recorded in a professional setting (anechoic chamber), with the following technical specifications - 16KHz sampling rate, 16 bits, single channel recording. The recorded files, are converted to NIST sphere file format. The main advantage of sphere file format is that it can hold all the audio information as part of its header, alleviating the need for storing it separately.

IV. Implementation**:**

A. Data Collection

The GU_Galo_Adi corpus that is used for create TTS, initially for 2 hour, consisted of 1000 sentences but 30% of these sentences are used as hold-out sentences that are used later for quality evaluation purposes (sample synthesized files can be found at **[**15**]**). The held-out sentences are chosen based on syllable coverage criterion i.e. the sentences which do not add a new unique syllable, in terms of coverage, are chosen to be part of held-out sentences. The held-out sentences hence will not affect the syllable coverage of the speech synthe sis system.

III. Labeling Tool*:*

The process of manual labeling is a time consuming and daunting task and so It is not trivial to label waveforms manually, at the syllable level.

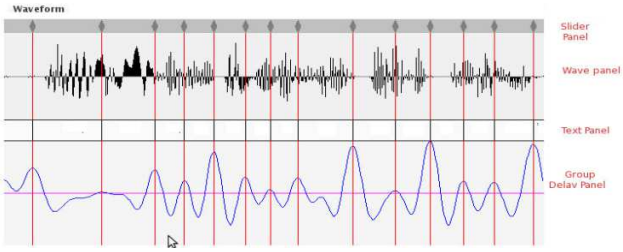

Fig. 1. Screenshot of Labeling Tool with EHMM

The tool makes use of group delay based segmentation to provide the segment boundaries. The size of the segment labels generated can vary from monosyllables to poly-syllables, as the Window Scale Factor (WSF) parameter is varied from small to large values. The tool contains 4 panels : • Slider panel – which is used to change/adjust the labels. • Wave panel – which shows the waveform in segmented format• Text panel - Shows the segmented text with syllable as the basic units and Group Delay panel - Displays the group delay plot. Our labeling process use of both Ergodic HMM (EHMM) labeling procedure provided by Festival and the group delay based algorithm provided by the labeling tool. This is achieved by enhancing the Labeling tool to display a new panel, which would show the segment boundaries as estimated by the EHMM process. The screenshot of the labeling tool, with a missed boundary, i.e. the boundary which is not indicated by group delay, but by EHMM, is as shown in Fig.2. and the screen shot of the labeling tool, with the boundary corrected by finding the peak which lies below the threshold is shown in Fig.3. The highlighted section of Fig.3 shows the group delay peak which is missed and also the corresponding included boundary.

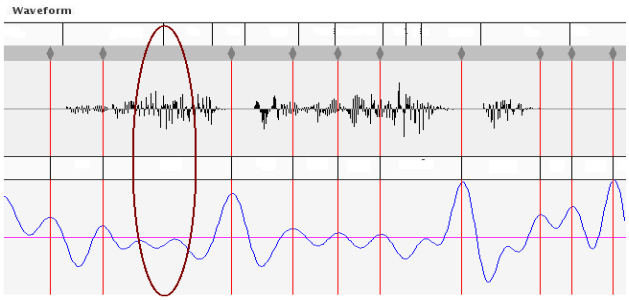

Fig. 2. Screenshot of Labeling Tool showing missed boundary

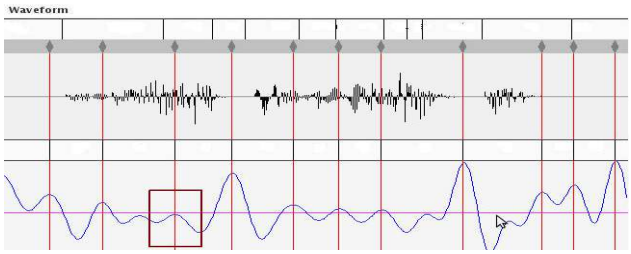

Fig. 3. Screenshot of Labeling Tool showing corrected boundary

syllable• In Phase 2: The text is broken down into bi-syllables at the word level and mono syllables if bi-syllables is not possible. In Phase 3 : The Text is first broken down into tri-syllables at the word level followed by bi and mono syllables. After completing the processes, all the cluster unit trees generated by the three phases are merged into a single tree and also the catalogue files are merged. This is done to provide a large number of instances of units with varying sizes during synthesis. For polysyllabic speech synthesis, the LTS rules are such that it first attempts to break down the text into the largest possible unit trained, at the word level.

Conclusion

We have discussed the design and implementation steps involved in building a polysyllabic speech synthesis system for Galo and Adi languages using the Festival framework. The polysyllable Text To Speech takes the largest possible unit from the database, which improve the quality, since the number of concatenation points is greatly reduced. The prosodic variations across the smaller units which make up the polysyllabic units will remain intact.

VI. Acknowledgment

Firstly, I would like to express my sincere gratitude and heartfelt thanks to my guide Prof. Pran Hari Talukdar, Professor, Department of Instrumentation and USIC, Gauhati University. I would not have been able to complete the research work and shape it in the form of the research paper without his consistent advice, and never ending enthusiasm, positivity, encouragement, support and understanding. I am very fortunate for having an opportunity to work with him from which I benefited enormously.

References

[1] T. Dutoit, *An introduction to text-to-speech synthesis*, Kulwer Academic Publishers, 1997.

[2] S.P. Kishore and A.W. Black, *Unit size in unit selection speech synthesis*, proceedings of EUROSPEECH, pp. 1317-1320, 2003.

[3] M. Nageshwara Rao,S. Thomas, T. Nagarajan and Hema A. Murthy *Text-to-speech synthesis using syllable like units*, proceedings of National Conference on Communication (NCC) 2005, pp. 227-280, IIT
Kharagpur, India, Jan 2005.

[4] Samuel Thomas, M. Nageshwara Rao, Hema A. Murthy and C.S. Ramalingam,*Natural Sounding TTS based on Syllable-like Units*, proceedings of 14th European Signal Processing Conference, Florence, Italy, Sep 2006.

[5] Venugopalakrishna.Y.R. et.al., *Design and Development of a Text-To-Speech Synthesizer for Indian Languages*, pp. 259-262, proceedings of National Conference on Communication (NCC) 2008, IIT-Bombay, February 2008.

[6] Venugopalakrishna.Y.R.,Vinodh.M.V., Hema A. Murthy and C.S. Ramalingam, *Methods for Improving the Quality of Syllable based Speech Synthesis* , proceedings of

Spoken Language Technology (SLT) 2008 workshop, pp. 29 -32, Goa, December 2008.

[7] Sreekanth.M and A.G.Ramakrishnan, *Festival based maiden TTS system for Tamil language*, Proceedings of 3rd Language and Technology Conference, pp. 187-191, Poznan, Poland, Oct 5-7, 2007.

[8] Alan W. Black, John Kominek, *Optimizing segment label boundaries for statistical speech synthesis* icassp, pp.3785-3788, 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009.

[9] T. Nagarajan and Hema A. Murthy, *Groupd Delay based Segmentation of Spontaneous speech into syllable-like units*, EURASIP Journal of Applied Signal Processing, Vol. 17, pp. 2614-2625, 2004.

[10] E. Veera Raghavendra, B. Yegnanarayana and Kishore Prahallad, *Speech Synthesis using approximate matching of syllables*, proceedings of Spoken Language Technology (SLT) 2008 workshop, Goa, December 2008.

[11] P. G. Deivapalan, Mukund Jha, Rakesh Guttikonda and Hema A. Murthy, *DONLabel: An Automatic Labeling Tool for Indian Languages*, pp. 263-266, National Conference on Communication (NCC)
2008, IIT-Bombay, February 2008.

[12] ITU-T Recommendations P.800, *Methods for subjective determination of transmission quality (formerly Rec. P.80*, 1996.

[13] ITU-T Recommendations P.862, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs* ,02/2001.

[14] H. Sakoe, S. Chiba, *Dynamic programming algorithm optimization for spoken word recognition*, IEEE Transactions on Acoustics, Speech and Signal Processing In Acoustics, Speech and Signal Processing, IEEE Transactions on, Vol. 26, No. 1., pp. 43-49, 1978.

[15] Sample synthesized sentences, http://www.lantana.tenet.res.in/ website files/research/Speech/TTS/contents/main.html.