Annotating Speech Corpus for Prosody Modeling in Indian Language Text to Speech Systems

Kiruthiga S[1] and Krishnamoorthy K[2]

[1] Faculty of CSE, Anna University Coimbatore
Coimbatore, Tamil Nadu, India

[2] Department of CSE, Sudharsan Engineering College
Pudukkottai, Tamil Nadu, India

Abstract

A spoken language system, it may either be a speech synthesis or a speech recognition system, starts with building a speech corpora. We give a detailed survey of issues and a methodology that selects the appropriate speech unit in building a speech corpus for Indian language Text to Speech systems. The paper ultimately aims to improve the intelligibility of the synthesized speech in Text to Speech synthesis systems. To begin with, an appropriate text file should be selected for building the speech corpus. Then a corresponding speech file is generated and stored. This speech file is the phonetic representation of the selected text file. The speech file is processed in different levels viz., paragraphs, sentences, phrases, words, syllables and phones. These are called the speech units of the file. Researches have been done taking these units as the basic unit for processing. This paper analyses the researches done using phones, diphones, triphones, syllables and polysyllables as their basic unit for speech synthesis. The paper also provides a recommended set of combinations for polysyllables. Concatenative speech synthesis involves the concatenation of these basic units to synthesize an intelligent, natural sounding speech. The speech units are annotated with relevant prosodic information about each unit, manually or automatically, based on an algorithm. The database consisting of the units along with their annotated information is called as the annotated speech corpus. A Clustering technique is used in the annotated speech corpus that provides way to select the appropriate unit for concatenation, based on the lowest total join cost of the speech unit**.**

Keywords: *Speech corpus, Indian languages, Concatenative Speech synthesis, Syllables, Prosody*

1. Introduction

The area of speech processing has gone to the extent of synthesizing natural speech that highly resembles a human voice. The idea behind is so simple. The system is fed with a human voice, processed and stored as smaller units. When the target text is given for converting into speech, these smaller units are concatenated to build a continuous speech. This technique is called as Concatenative Speech synthesis. The processing of human voice into smaller units is called the Speech Corpus development. The corpus can either be general purpose or application specific, based on the type of Speech Processing system. To build a corpus, a text file is first selected, which has phonetically rich and prosodically stable sentences. For some application specific Text to Speech synthesis systems, prosody rich text file may also be selected. The text file may be a News bulletin, forum interviews, every-day conversations in an organization, conversation in road traffic, etc. Then, with the help of a native speaker, who may be a news reader, this text file is made to be read and recorded. This recording is the corresponding speech file. A speech file is an audio file whose size generally ranges from several minutes to hours. This speech file is the phonetic representation of the selected text file. The speech file is processed in different levels viz., paragraphs, sentences, phrases, words, syllables and phones, which are called the speech units of the file. Many researches have been carried out using these units as the basic units of processing. Research in the area of speech synthesis have an importance in many applications, including information retrieval services over telephone such as banking services, public announcements at places like train stations and reading out manuscripts for collation, reading emails, faxes and web pages over telephone and voice output in automatic translation systems.

2. Speech Corpus for Indian Languages

The speech synthesis systems are developed for the Indian languages such as Hindi, Tamil, Telugu, Kannada, Bengali and Marathi. These Indian languages share phonological features. For Indian languages, building a speech corpus is a different task than that of the English speech corpus. Since prosodic processing such as duration and intonation prediction has to be done in the corpus-development stage itself, some more information has to be

annotated with the basic units after storing them in the corpus. Methodologies which include clustering are incorporated to improve the corpus, thus improving the entire system. Issues such as detection of mispronunciation, detection of pronunciation variants, un-transcribed speech are to be noted and addressed.

Phonetizers or the Grapheme to Phoneme converters are tools that convert the text corpus into its phonetic equivalent [7]. The phonetic nature of Indian scripts reduces the effort to building mere mapping tables and rules for the lexical representation. These rules and the mapping tables together comprise the Grapheme to Phoneme converters.

3. Concatenative Speech Synthesis

Concatenative speech synthesis uses phones, diphones, syllables, words or sentences as basic speech units. Speech is synthesized by selecting appropriate units from a speech database, called as a speech corpus, and concatenating them. Many researches have been made, selecting each separate unit as the basic unit. The size of the database differs, as the number and size of individual units differ. This also affects the quality of synthesized speech. If large units such as phrases or sentences are stored and used, the quality of synthesized speech is good, although the domain of synthesis is not unrestricted text. When small units such as phones are used, a wide range of words or sentences can be synthesized but with poor speech quality.

3.1 Phones as the basic unit

When phones are considered as basic units, the size of the database will be less than 50 units for Indian languages. Because, including allophones, the number of phone units of any Indian language is less than 50. The database is small, but phones provide very less co-articulation information across adjacent units, thus failing to model the dynamics of speech sounds with their large variability depending on context. Thus phones are found to be inefficient for speech synthesis.

3.2 Diphones and triphones as the basic unit

A diphone is made of two connected half phones and captures the transition between two phones by starting in the middle of the first phone and ending in the middle of the second one. It captures the co-articulation effects and minimizes the discontinuities at the concatenation points. Diphones are relatively bigger units than phones. There are about 1000 to 2000 diphones found in Indian languages. Unlike phones, they do not exhibit allophonic variations. i.e., each diphone has only one instance of pronunciation. Though diphone concatenation can produce a reasonable quality speech, a single example of each diphone is not enough to produce good quality speech. Moreover, diphone-based synthesizers need elaborate prosody rules to produce natural speech. In newer Speech Processing systems, one other unit was introduced, called triphones. A triphone provides more co-articulation information than diphones, since they involve in blending with both previous and next phones in the phrase. Compared to diphones, triphones are larger in number thus leading to a larger size of database. But some triphones rarely occur in the language model, thus making the database unnecessarily large. An attempt has been made for word and triphone based speech synthesis [4], where a full dictionary, fast dictionary and a Simple Breadth First Search Manager were experimented with. In scenarios where the vocabulary is limited, repeatability of sentences is more and speakers are limited, the word-based recognizer with trained voice on trained sentences is highly suitable. In the same scenario, when the vocabulary is high and speakers are limited, triphone based model is suitable.

3.3 Syllables as the basic unit

Indian languages are syllable centered, where pronunciations are mainly based on syllables. A Syllable can be the best unit for Indian language Speech synthesis systems. Intelligible speech synthesis is possible for Indian languages with syllable as the basic unit. Syllable units being larger in comparison to phones or diphones, can capture co-articulation better than phones. The number of concatenation points decreases when syllable is used as the basic unit. Syllable boundaries are characterized by regions of low energy, providing more prosodic information. A grapheme in Indian languages is close to a syllable. The general format of an Indian language syllable is C*VC*, where C is a consonant, V is a vowel and C* indicates the presence of 0 or more consonants. There are about 35 consonants and 18 vowels in Indian languages [12]. There are defined set of syllabification rules formed by researchers, to produce computationally reasonable syllables. A rule based grapheme to syllable converter was used for syllabification. Some of the rules used to perform grapheme to syllable conversion [9] are:

- Nucleus can be Vowel(V) or Consonant (C)
- If onset is C then nucleus is V to yield a syllable of type CV
- Coda can be empty or C
- If characters after CV pattern are of type CV then the syllables are split as CV and CV.
- If the CV pattern if followed by CCV then syllables are split as CVC and CV.

- If the CV pattern is followed by CCCV then the syllables are split as CVCC and CV
- If the VC pattern is followed by V then the syllables are split as V and CV.
- If the VC pattern is followed by CVC then the syllables are split as VC and CVC

Researches were conducted to find which order of syllables is best acceptable for synthesis [8]. The following are the recommended combinations:

a) Monosyllables at the beginning of a word and bisyllables at the end.

b) Bisyllables at the beginning of a word and monosyllables at the end.

c) Monosyllables at the beginning and trisyllables at the end of a word.

d) Trisyllables at the beginning and monosyllables at the end of a word.

3.4 Polysyllables as the basic unit

An attempt exploring a Speech synthesizing system using polysyllabic units has been made [5]. Since polysyllable units are formed using the monosyllable units already present in the database, the synthesis quality can be improved without augmenting any new set of units. The system uses a large database, which consists of syllables, bisyllables and trisyllables. While synthesizing, the first matching trisyllable is selected followed by the bisyllable and monosyllable units, as needed. Picking up the largest possible unit in the database improves the quality of speech, since the number of co-articulation points greatly reduce.

Selection of an appropriate candidate unit set is carried out using search algorithms. Units which incur lowest total join cost for a word are preferred. Total join cost is the sum of selection cost of the candidate unit and the join cost of the selected candidate units [3].

$$C_{total} = C_{Sel} + C_{Join} \quad (1)$$

where, C_{total} is the total join cost, C_{Sel} is the selection cost and CJ_{oin} is the cost of join.

4. Clustering the syllables

The need for selecting the phonetically and prosodically best units for synthesis needs clustering the units in the database. An acoustic distance measure is defined to measure the distance between two units of the same phone type. Factors concerning prosodic and phonetic context are evaluated to form cluster units within a unit type. A decision tree is built based on questions concerning the phonetic and prosodic aspects of the grapheme. Ultimately the leaves of the decision tree are the list of database units that best suit the required aspects. At the time of synthesis, for each target, the appropriate decision tree is used to find the best cluster of candidate units. A search is then made to find the best path through the candidate units.

Pruning is performed to remove spurious atypical units which may have been caused by mislabeling or poor articulation in the original recording. It also removes those units which are so common that there is no significant distinction between candidates. Databases were tested with this clustering method in hand. The method [3] produced both extreme high quality examples and extremely low quality ones. Minimizing these bad examples was the important target.

A clustering technique is suggested in [2], which includes pre-clustering works that tags the syllables as begin, middle and end, depending on the occurrence of the syllable in the word. Further clustering is performed by tagging the syllables based on type of the syllable (v, c*v, vc*, c*vc*) and nature of the constituent vowels and consonants. Syllables of the same type were clustered using features like word length of the phrase, relative position of the syllable in the phrase, relative position of the parent phrase and the features of the preceding and the following syllables in the phrase. The acoustic distance measure is calculated using the mel frequency cepstral coefficients (MFCC). Using the feature set and the acoustic distance measure, the decision tree was built for each of the unique syllable in the database. Questions were used at the nodes to find the best set of candidate syllables. Morpheme tags are used for phrase prediction. This technique has improved the quality of the synthesized speech to a greater extent.

4.1 Prosody

The Speech synthesis system exhibits its naturalness when the corpus is annotated with prosodic information. Prosody modeling is subdivided into modeling the following constituents of prosody - phrasing, duration, intonation and intensity. Two major approaches for prosody modeling are the rule based approach and the corpus based approach. In the rule based approach, linguistic experts derive a complicated set of rules to model prosodic variations by observing natural speech. In the corpus based approach, a well-designed speech corpus, annotated with various levels of prosodic information is used. The corpus is analyzed automatically to create a prosodic model which is then made to synthesize a training data set, following which the test dataset is evaluated. Based on the performance on test data, the models are then improved. The syllables have sufficient duration information as it improves the quality of synthetic speech when used as a duration model. Thus syllables are identified as the best-suited processing units for Indian language Speech synthesis.

5. Proposed Work

A prosodically stable set of sentences, such as a dictated full length text, is selected. The text is broken down to appropriate segmental or suprasegmental structure of utterances such as phones, syllables, or words. These structures are called speech units. A repository of speech units is created and each unit is annotated with prosodic features. i.e., each entry in the repository is a set which consists of a speech unit and its prosodic information. These speech units are henceforth being called as candidate units. During synthesis a candidate unit is selected, considering the lowest total join cost for a word. One of the important prosodic aspects is the prediction of phrase boundary. Some morpheme tags are identified as phrase boundary indicators. The work aims to automatically detect these morpheme tags in order to predict the phrase boundary. These phrase boundary details are annotated with the candidate units, as prosody information. The synthesized speech shall have to cater the issues of pitch normalization across speakers.

6. Conclusions

In this paper, the importance of annotating every individual candidate unit in the speech corpus with necessary prosody information is analyzed. The use of various units as basic units for developing a speech corpus for Indian language Text to Speech synthesis system is also sorted. The analysis significantly shows the need for Prosody modeling to obtain high quality intelligent speech. The appropriate speech unit for Indian languages is the syllable, which is the good constituent of prosodic features of the Indian languages. Selection of the appropriate candidate unit annotated with necessary prosodic information play a vital role in the development of Speech corpus, which obtains naturalness in the synthesized speech. A Clustering Technique is also proposed which selects the best unit when more than one candidate unit shows the lowest total join cost.

References

[1] M. Nageshwara Rao, Samuel Thomas, T. Nagarajan and Hema A. Murthy, “Text-to-speech synthesis using syllable-like units,” in National Conference on Communication, Kharagpur, India, Jan 2005, pp 277-280.

[2] G L Jayavardhana Rama, A G Ramakrishnan, R Muralishankar and Vijay Venkatesh. ”Thirukkural - A text to speech synthesis system”. Proc. Tamil Internet 2001, Kuala Lumpur 2001,92-97.

Hunt, A.J. and Black, A.W., “Unit selection in a concatenative speech synthesis system using a large speech database”, Proc. ICASSP, vol. 1,pp. 373–376, 1996.

[3] Alan W. Black and P. Taylor, “Automatically clustering similar units for unit selection in speech synthesis”, Proc. EUROSPEECH 97, Rhodes, Greece, 1997, Vol. 2, pp. 601-604. K. Elissa, “Title of paper if known,” unpublished.

[4] R. Thangarajan, A.M. Natarajan and M. Selvam, “Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language”, in WSEAS Transactions on Signal Processing, Issue 3, Volume 4, March 2008.

[5] Vinodh M Vishwanath, Ashwin Bellur, Badri Narayan K, Deepali M Thakare, Anila Susan, Suthakar N M and Hema A Murthy,“Using Polysyllabic units for Text to Speech Synthesis in Indian languages,” Proceedings of National Conference on Communication (NCC),pp.1-5, 29-31 Jan. 2010

[6] Kishore Prahallad, Arthur R Toth, Alan W Black, “Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases”, in Proceedings of Interspeech, Antwerp, Belgium 2007.

[7] Anumanchipalli Gopalakrishna, Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinder Singh, R.N.V Sitaram andS.P. Kishore, "Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems", Proceedings of International Conference on Speech and Computer (SPECOM), Patras, Greece, Oct 2005.

[8] T.Jayasankar, R.Thangarajan, J.Arputha Vijaya Selvi, “Automatic Continuous Speech Segmentation to Improve Tamil Text-to-Speech Synthesis”, in International Journal of Computer Applications (0975 – 8887), Volume 25– No.1, July 2011.

[9] S. Saraswathi and T.V. Geetha, “Design of language models at various phases of Tamil speech recognition system”, International Journal of Engineering, Science and Technology Vol. 2, No. 5, 2010, pp. 244-257.

[10] Kiruthiga S, Krishnamoorthy K, “Design Issues in Developing Speech Corpus for Indian Languages”, 2nd International Conference on Computer Communication and Informatics, Jan 2012, Vol. 2, 978-1-4577-1581-5.

[11] Samuel Thomas, M. Nageshwara Rao, Hema A. Murthy and C.S. Ramalingam, “Natural sounding TTS based on syllable-like units,” in the Proceedings of the 14th European Signal Processing Conference, Florence, Italy, Sep 2006.

[12] G. L. Jayavardhana Rama, A. G. Ramakrishnan, R. Muralishankar and R Prathibha, “A Complete Text-To-Speech Synthesis System In Tamil”, in 0-7803-7395-2/02, IEEE proceedings 2002.

[13] Ashwin Bellur, K Badri Narayan, Raghava Krishnan K, Hema A Murthy, “Prosody Modeling for Syllable-Based Concatenative Speech Synthesis of Hindi and Tamil”, DOI: 10.1109/NCC.2011.5734737 , IEEE proceedings 2011.

[14] Grażyna Demenko, Stefan Grocholewski, Agnieszka Wagner, Marcin Szymanski, “Prosody Annotation for Corpus based Speech Synthesis”, Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9.

Kiruthiga S is a research scholar in Anna University Coimbatore. She obtained her Master of Engineering Degree in 2006 under Anna University Chennai. She completed her Bachelor of Engineering Degree under Bharathiyar University in the year 2003. She has worked as Lecturer in Sona College of Technology Salem for 4 years. Her research interest includes Text to Speech Synthesis system and other Natural Language Processing applications.

Krishnamoorthy K is a Professor in Sudharsan Engineering College, Pudukkottai. He is one of the recognized research

supervisor in Anna University Coimbatore. He obtained his Ph. D degree in 2007 and Master of Engineering Degree in 2003 both under Dayananda Sagar University, Bangalore. He has an experience as an academician for a span of a decade. His research interest includes Natural Language Processing and Digital Image Processing.