Development of Syllable Based Unit Selection Text-To-Speech Synthesis System for Tamil Using Three Level Fall Back Technique

[1] V. Tamilselvi, [2]Dr.P.Visu
[1]Student of CSE Department, [2] Professor and Head of the CSE Department
Vel Tech University, Avadi, Chennai, Tamil Nadu
[1]tamilselvi17690@gmail.com, pvisu@veltechuniv.edu.in

***Abstract*—A text-to-speech synthesis system is one that is capable of producing intelligible and natural speech corresponding to any given text. A popular approach to speech synthesis is unit selection synthesis (USS). The current work focuses on developing a USS system for Tamil. Literature suggests that syllable is a suitable unit for Indian languages. Creating a database that covers all the syllables of Tamil is tedious and expensive, and the footprint size of the system would be in the order of GBs. Therefore, to reach a compromise between the quality and the footprint size, the current work proposes to use a database containing all the phonemes, consonant-vowel (CV) units, and the most frequently occurring syllables of Tamil. This way, given a text, it is first decomposed into syllables. If a particular syllable is not available in the database, it is broken down to CV units and phonemes. The appropriate speech units are then chosen from the database and concatenated to produce a speech utterance. The performance of the system will be evaluated subjectively by the mean opinion score.**

***Keywords*— Speech Synthesis, Unit Selection, Text-to-Speech, Mean Opinion Score**

I. INTRODUCTION

Speech Synthesis is the artificial generation of speech signal from text. Speech synthesis system has mainly two parts; first part converts speech to linguistic specification (grapheme-to-phoneme conversion). The second part generates the speech waveform. An unrestricted text-to-speech system is expected to produce a speech signal, which is corresponding to the given text. A popular approach to speech synthesis are Unit Selection Synthesis (USS), Hidden Markov model-based speech synthesis. A successful concatenative speech synthesis technique is the unit selection synthesis.

Unit selection speech synthesis [1] is used to synthesize speech for the given input text. Festival framework is required for synthesizing the voice for unit selection speech synthesis approach. It involves the concatenation of appropriate pre-recorded speech units, for the given text, based on the target and concatenation costs. The target cost identifies the units in the database that best match the required specification and the concatenation cost identifies the units that join smoothly. The speech units can be words or sub-word units such as phonemes, diphones, syllables, etc. The quality of speech synthesized varies based on the size of the unit. If the units are longer, naturalness is better-preserved and the number of concatenation point are less. However, the amount of data required to train the synthesis system increases, by increasing the unit-size, thereby increasing the footprint size of the system.

Earlier phoneme based system and CV based systems are developed, and the result shows that phoneme system performs better since the concatenation points are less. For smaller units, phoneme is selected as the best unit, but there are more sonic glitches in the phoneme based system. For building a system using syllable as a unit, more amount of training speech data are required, creating a huge database is tedious and not a time consuming process and also more memory space is required for building such a system. If less amount of data are used then some syllables [6] may not be present in the database. Hence there are some issues in maintaining the good quality of the synthesized Speech and naturalness has to be considered to a greater extent for a better synthesized system.

In the current work, the unit selection speech synthesis systems for Tamil are developed with less amount of speech data with syllable as the major unit. The given text is first decomposed to syllables. If the particular syllable in the given input is not present in the internal database, then the syllable unit is broken down into CV units and phonemes. If the particular CV is not present in the database, then it picks the phoneme unit from the database for developing the synthesis system. Finally the appropriate speech units are chosen from the entire database and the units are concatenated which are used to produce a speech utterance. The given text is synthesized by combining the pre-recorded units from the database.

The performance of the synthesized voice is analyzed by obtaining the mean opinion score. The creation of the database is tedious process, so in this approach with less amount of training data, the system generates a speech which is highly intelligible and natural [9] and quality is also maintained.

The paper is organized as given below: Section II describes the speech corpus for building the synthesis system. Section III describes the Unit Selection systems developed for Tamil in detail. Section IV describes the performance analysis for the developed systems. Section V describes the conclusion of the given work.

II. SPEECH CORPUS

The speech corpus consists of one hour of recorded Tamil speech data for training the system. The data are recorded from a native Tamil speaker. The recorded speech has the sampling rate of 16kHZ. Recording was done in a noise free environment at laboratory using a unidirectional carbon microphone.

An M-Audio Mixer was used to suppress the noise and was set in mono mode to capture only the speaker's voice and the sampling rate was set to be 16 KHz. For recording, Audacity software was used. Precautions were taken to maintain a constant energy in the recorded speech. Festival supports different forms of audio files such as ulaw, snd, aiff (audio interchange file format) and riff (resource interchange file format chunks) format.

Segmentation

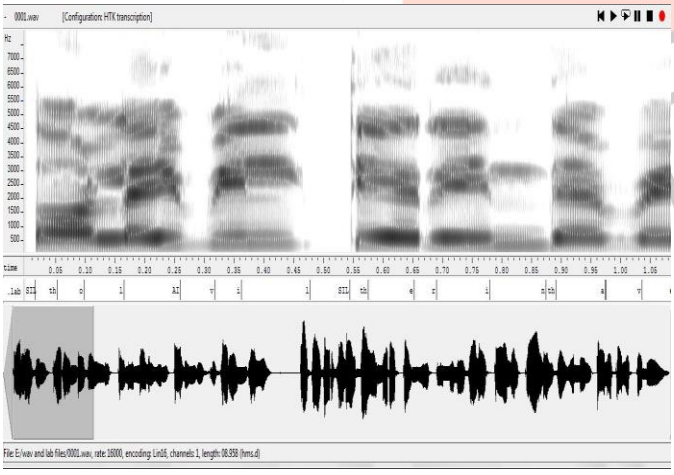
Speech segmentation is the process of identifying the boundaries between words, syllables, or phonemes in spoken natural languages. The lowest level of speech segmentation is the breakup and classification of the sound signal into a string of phones.

The lab files are required for the corresponding wave files. The lab files are generated by segmenting the data. The datas are segmented by using forced-viterbi algorithm. Initially five minutes of data are manually segmented, which contains of 50 sentences. The manual segementation is done at phoneme level for the representations waveforms and the spectrograms, HTK transcriptions, TIMIT transcriptions, etc. Models are generated for all the phonemes and the lab files are generated. Forced-viterbi alignment are performed to segment the rest of the data. The following steps are used to segment the data :

1) By using 5 minutes of data and the corresponding time aligned phonetic transcriptions, context-independent phoneme models are trained.
2) Using these models and the phonetic transcriptions, the speech data are segmented using forced-Viterbi alignment procedure.
3) Using the obtained phonetic transcription (phone-level label files), new context-independent phoneme models are trained.
4) Steps 2 and 3 are repeated for *N* times.
5) After *N* iterations, the resultant HMMs are used to segment the entire speech data, again. These boundaries are considered as final boundaries.

Finally the lab files are obtained by itearatively performing forced-viterbi alignment for the speech data.

Figure 1 Segmented wave file



From the phoneme level lab file obtained, the CV and syllable lab files are obtained by using the script which gives the phonemes as CV units and syllable units. Finally phoneme, CV and syllable lab files has been created and the segementation is checked manually.

III. UNIT SELECTION SPEECH SYNTHESIS

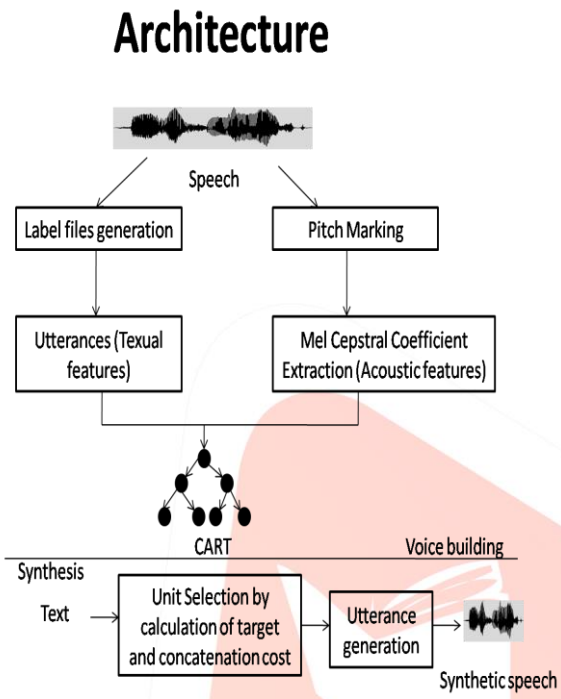
In the current work, the phoneme based system, CV based system, syllable based system and syllable with fall back systems were developed using Unit Selection Speech Synthesis to compare the performance of all the systems.

Concatenative synthesis generates speech by connecting natural, prerecorded speech units. These units can be words, syllables, half-syllables, phonemes, diphones or triphones [3]. The unit length affects the quality of the synthesized speech. With longer units, the naturalness increases, less concatenation points are needed, but more memory is needed and the number of units

stored in the database becomes very numerous. With shorter units, less memory is needed, but the sample collecting and labeling techniques become more complex.

The unit selection is based on two cost functions.
– Target cost, $C_{t\,(}u_i, t_{i\,)}$ is an estimate of the difference between a database unit, u_i and the target, t_i which it is supposed to represent.
– Concatenation cost, $C_{c\,(}u_{i-1}, u_{i)}$, is an estimate of the quality of a join between consecutive units, u_{i-1} and u_i .

The architecture of unit selection system is shown in the below figure:



The following systems are developed using unit selection speech synthesis:

Phoneme based System:

The phoneme based system is developed using the phoneme level lab files. In the word “ammA”. It is transcribed as /a/ /m/ /m/ /A/. The utterances are generated for each of the labels and the corresponding text. Phoneme is the smaller unit in the speech corpus, hence the concatenation points are more in the phoneme based system. Finally the system is built with one hour of speech data. The phoneset features have been defined for all the phonemes in the Tamil language. Using the festival framework, the system is developed and the performance of the system is also analysed.

CV-Based System:

The CV-based system was developed using CV level lab files. The label files for the CV-based system are obtained from the phoneme-level label files by combining two successive phonemes if a vowel follows a consonant. The CV-based system contains vowels (V), consonants(C), and consonant-vowel (CV) units. For example, the word “ammA” is split as /a/ /m/ /m/ /A/.

Syllable based system:

The syllable based unit selection speech synthesis systems [4] are developed. More amount of training data are required to develop a synthesis system with syllable as a unit. In this system, there are less number of examples so some of the syllable units are not synthesized. The syllable level lab files are generated from the CV lab files, whereas the CV lab files are obtained from the phoneme lab files by concateneting the consonants followed by a vowel. Finally these lab files with corresponding text and wave files are required for building the synthesis system.

Syllable with three level fall back:

The syllable based system with fallback is similar to syllable based system but in addition the CV and phoneme lab files are also considered. The systems are built using all the subword units such as phoneme, Consonant vowel (CV) and syllable units. The syllables which occur frequently and contains more number of examples are sorted out, and for these syllables alone the syllable units are picked. In case if particular syllable is not present in the database the syllables are decomposed into CV units. If the particular CV is not present in the database finally it is fall back to phoneme units and the corresponding word has been synthesized and the speech

was generated. In this system, if the syllable units are not present in the database, then also the text is synthesized by fall back to CV units and phonemes.

Hence the syllable with fall-back requires only less amount of data but can synthesize all the syllable units and generate the corresponding speech voice. Therefore we introduce a system which contains all the units and the text is synthesized. Therefore we infer that this systems outperforms the phoneme, CV and syllable based systems.

The steps for building voice in festival framework are as follows:

1. The features for each of the units has to be mentioned and is used for the pronunciation of the particular unit, the features are vowel/ consonant/ nasal/ fricative etc.
2. The utterance are created for all the sentences used in the entire database.
3. Letter-to-sound rules are used to break a sentence into required subword units, to generate the initial utterances.
4. Extracting the pitchmarks and building LPC coefficients.
5. Post processing is done to tune the pitch marks. Pitch marking plays a vital role in the extraction of Mel cepstral coefficients, because synchronous framing was used for Festival.[7]
6. The units in the database are clustered using Classification and Regression Trees (CART) [6]. The number of questions has been defined to classify the units into clusters. If the number of questions are greater then the number of units in a cluster are less and tree becomes deeper.
7. Testing of the voice for out domain sentences.

To synthesize a new sentence or paragraph, the text is given as the input, the system splits the text into the required subword units and identifies the most suitable unit to be concatenated based on the target cost and the concatenation cost. The utterance are created for the corresponding text with the units selected and the final speech is synthesized from the utterance.

IV. PERFORMANCE ANALYSIS

The synthesis systems developed are evaluated using the conventional mean opinion score (MOS). MOS is a five-point grading scale, which represents score from 5 to 1. The score 5 corresponds to excellent and good intelligibility and quality and 1 corresponds to highly unintelligible and annoying. The MOS scores were collected from 14 native listeners. 25 wave files were synthesized and the Mean Opinion Score was obtained for each of the wave files.

Table 1: MOS Obtained for all the Systems

Syllable with Fall back System	Phoneme based System	CV based System	Syllable based System
3.5	3.0	2.7	2.5

The testing is done by synthesizing sentences from out domain, which is not present in the training data. The corresponding wave files are synthesized form the festival system for the sentences or paragraphs. From the MOS score obtained, the syllable with fall back system has the highest score and intelligibility.

V. CONCLUSION

Thus from the systems developed and the observations made, we conclude that the syllable with fall back system outperforms the other systems. The synthesized speech is highly intelligible and the quality is improved to a greater extent. Although the syllable based system is intelligible, more amount of training data are required to synthesize all the out domain sentences.

The voices are tested and analyzed and found from the analysis made, it is observed that many of the syllables are missing in syllable based system due to less amount of training data in the corpus, if data increased these can be neglected, which results in high memory and more time for creating a large speech database. So we conclude that the systems developed using fall back can be considered as the best synthesis technique to building speech voices.

VI. REFERENCES

[1] Alimilious Chalamandaris, Sotiris Karabetes, Syros Raptis, "*A Unit Selection Text-to-Speech synthesis system optimized for use with screen readers*", IEEE Transactions on Consumer Electronics, Vol. 56, No. 3, August 2010.
[2] Y. Tabet and M. Boughazi, "*Speech Synthesis Techniques: A Survey*", in Proc. Of IEEE/WOSSPA, 2011, 99. 67-70.
[3] A. J. Hunt and A. W. Black, "*Unit Selection in a Concatenative Speech Synthesis Using a Large Speech Database*", in ICASSP, 1996, pp. 373-376.
[4] M. Nageshwara Rao, Samuel 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.
[5] Aby Louw, "*A Short Guide to Pitch-Marking in the Festival Speech Synthesis System and Recommendations for Improvements*", CSIR, Pretoria, 2004
[6] A. W. Black and P. Taylor, "*Automatically Clustering Similar Units for Unit Selection in Speech Synthesis*" in Proc. of Eurospeech, 1997, pp. 601-604.
[7] A. Bellur, K. B. Narayan, K. Raghava and H. A. Murthy, "*Prosody Modelling for Syllable-Based Concatenative Speech Synthesis of Hindi and Tamil*", in NCC, 2011, pp. 1-5.
[8] K. Tokuda, H. Zen and A. W. Black, "*An HMM-Based Speech Synthesis System Applied to English*", in Proc. of IEEE Workshop on Speech Synthesis, 2002, pp. 227-230.
[9] Young S., G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, "*HTK Book*", ver 3.2.1 Cambridge University Engineering department, 2002.
[10] Black A. W, Zen H. and Tokuda K. (2007), '*Statistical parametric speech Synthesis*', in Proc. of ICASSP 07, vol.4, pp.1229 –1232.
[11] Black A.W and Lenzo K. (2004), '*Multilingual text to speech synthesis*', in Proc. of ICASSP 04, vol.3, pp. 761–764.
[12] Black A.W. and Lenzo K. (2003), 'Building synthetic voices', available at http://festvox.org/bsv/.
[13] Douglas O'Shaughnessy (1999), '*Speech Communications: Human and Machine*', Universities press, second edition, pp. 337-365.