Improved Tamil Text to Speech Synthesis

M.Vinu Krithiga * and T.V.Geetha **
* Research Scholar <vinu_krithiga@yahoo.com>;
** Professor <rctamil@annauniv.edu>
Department of Computer Science and Engineering,
Anna University, Chennai-25, India.

Abstract.
This paper describes the improvement of the quality of Tamil text to speech using LPC based diphone database and the modification of syllable pitch through time scale modification. Speech is generated by concatenative speech synthesizer. Syllable units need to be concatenated such that spectral discontinuities are lowered at unit boundaries without degrading their quality. Smoothing is done by inserting suitable diphone at the concatenation boundary and changing the syllable pitch by performing time scale modification. The suitable diphone is chosen based on LPC coefficient files and their corresponding residuals.

1 Introduction

In this paper, the aim is to improve the quality of Tamil text to speech system. One of the important issues in Text-to-Speech systems is the quality of smoothing. This paper describes two different methods to improve joint and individual smoothness of speech units. Smoothing when speech is synthesized using concatenation method has been dealt with many ways. Among the important methods are Frequency-Domain Pitch-Synchronous Overlap-Add algorithm (FD-PSOLA), Time-Domain Pitch-Synchronous Overlap-Add algorithm (TD-PSOLA), Multi-Band Re-synthesis Pitch-Synchronous Overlap-Add model (MBR-PSOLA), Multi-Band Re-synthesis Overlap-Add (MBROLA) [1], [4], [6], [10]. All the PSOLA methods can be applied only for voiced sounds and when applied to unvoiced signal parts it generates a tonal noise. Text-to-Speech using MBROLA technique gives better quality when compared to PSOLA. MBROLA technique is preferred for Tamil TTS. MBROLA, a speech synthesizer based on the concatenation of diphones. It takes a list of phonemes as input, together with prosodic information (duration of phonemes and a piecewise linear description of pitch), and produces speech samples with bit depth 16 bits (linear), at the sampling frequency of the diphone database used, However MBROLA does not accept raw text as input. While PSOLA used syllables as speech unit and MBROLA used only diphones as the speech unit.

The aim of this work is to further improve smoothness compared to MBROLA method and to accommodate raw text as input. In the system described in this work speech output is obtained by concatenation of syllables. Syllable is an intermediate unit which is the intermediate form between the phones and the word level.They need to be concatenated such that spectral discontinuities are lowered at unit boundaries without degrading their quality. The corresponding diphone is inserted between syllable-syllable concatenations to remove the discontinuity at the concatenation point. The diphone is chosen and the end segments of the diphone are smoothened by the LPC coefficient and residue value. Syllables are used as speech unit and diphones are inserted between the syllables to smooth the output. Thus in this work smooth-

ing across phoneme boundaries is performed by appropriate addition of diphone based on LPC coefficients and residues. To further improve the quality of speech, intra syllable smoothing through pitch modification required for adjusting duration is performed in this work using time scale modification.

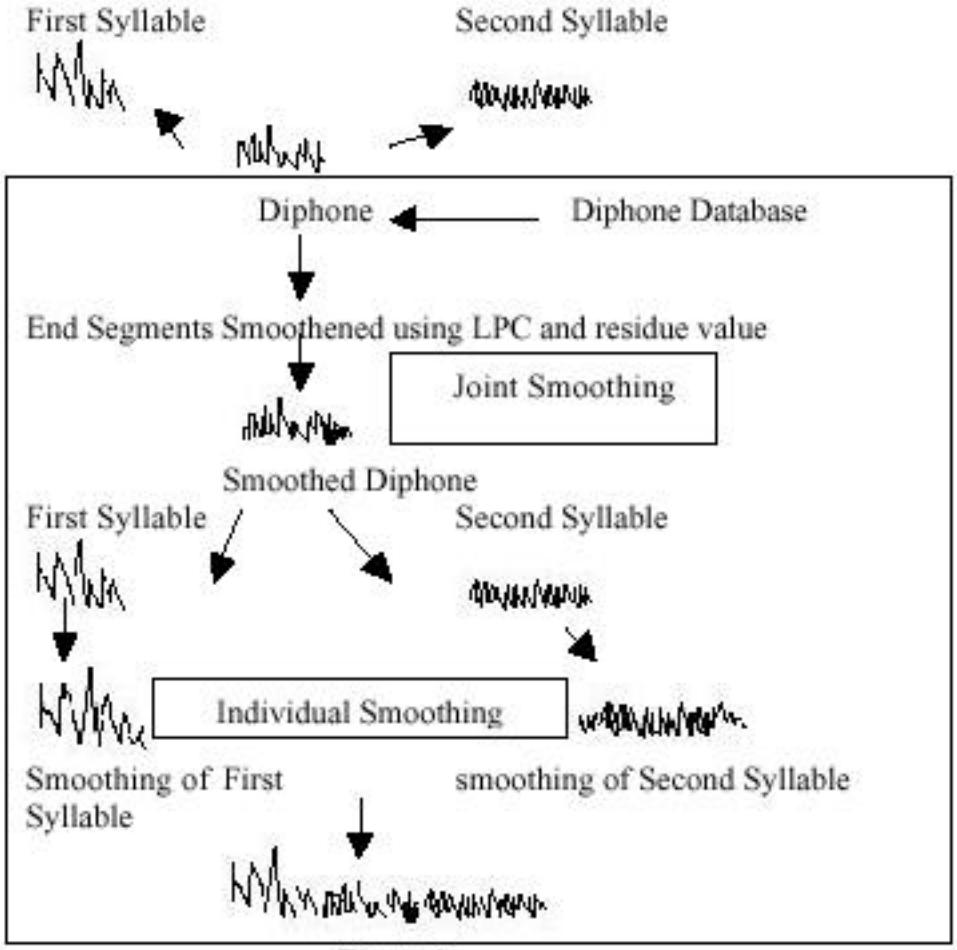


Figure 1

Figure 1 explains the steps followed to smooth the speech output of Tamil Text-to-Speech.

2. LINEAR PREDICTIVE CODING (LPC)

As already described smoothing at concatenation joints is performed using LPC. In general, LPC is used for representing the spectral envelope of a digital signal of speech using the information of a linear predictive model. It is one of the most powerful method for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters [12].

LPC starts with the assumption that a speech signal is produced by a buzzer at the end of a tube. The glottis (the space between the vocal cords) produces the buzz, which is characterized by its intensity (loudness) and frequency (pitch). The vocal tract (the throat and mouth) forms the tube, which is characterized by its resonances, which are called formants. LPC analyzes the speech signal by estimating the formants and the intensity and

frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue. The numbers which describe the formants and the residue can be stored. Because speech signals vary with time, this process is done on short chunks of the speech signal, which are called frames [12].

LPC is a very successful model used for smoothing and is mathematically efficient (IIR filters), remarkably accurate for voice (fits source-filter distinction) and satisfying physical interpretation (resonance). This model outputs a linear function of prior outputs and hence is called Linear Prediction [11]. Synthesis model splits the speech signal into LPC coefficients and residual signal. Since the LPC analysis is pitch synchronous, waveform interpolation could be used for the residual, but we found that the residual method is better due to the frame-to-frame variability of natural speech which would be lost if a series of residual frames were strictly interpolated [11]. Hence only the amplitude of the residual signal is scaled in order to smooth the energy of the output signal.

2.1. LPC COEFFICIENT REPRESENTATIONS

As mentioned earlier, LPC is used for transmitting spectral envelope information. Transmission of the filter coefficients directly is undesirable, since a very small error can distort the whole spectrum. Hence LPC values has to be represented as coefficients. More advanced representations of LPC values are log area ratios (LAR), line sprectrum pairs (LSP) decomposition and reflection coefficients. Of these, especially LSP decomposition has gained popularity, since it ensures stability of the predictor and localization of spectral errors for small coefficient deviations [12]. Hence in this work, LPC coefficients are represented using Line Spectrum Pairs (LSP) decomposition.

2.2. LPC IN TAMIL TEXT TO SPEECH

This section describes how LPC is used to smooth the end segments of diphone in Tamil Text to Speech engine. Initially LPC coefficient and residue value of the end segments for each diphone corresponding to CV-CV combination are calculated from the recorded voice and maintained as an LPC database. The LPC and residue value is calculated using Matlab program. When input text is given in Tamil Text-to-Speech, Tamil text is split into syllables and diphone is chosen corresponding to Syllable-Syllable combination from the diphone database. The diphone is extracted depending on the first and the second syllable. In Tamil, if the first syllable is “ka” and the second syllable is “sa”, the diphone “ka_s” is inserted between the concatenation point and the LPC value of the “ka_s” combination is calculated and the value is compared with the value of the already stored for “ka-sa” combination. The “ka_s” diphone is common if any one of the syllable “ka,kaa,ki,kee,ku,koo,ke,keq,kai,ko,koa,kov” comes first and the second syllable is “sa”. Thus the diphones are grouped to reduce the database size. Tamil diphone database is built by storing the wave files corresponding to the syllable-syllable concatenation. A small end portion of the first syllable and a small start portion of the second syllable are extracted for each

syllable-syllable combination and the database is built. The diphone has been extracted for CV-CV combination, where C stands for consonant and V for vowel. A female speaker voice is recorded to develop diphone database. The speaker reads a set of carefully designed Tamil words, which have been constructed to elicit particular phonetic effects. At present almost 1000 diphones have been created but around 3000 will probably be the final number of diphones.

LPC coefficient and residue value is calculated for end segments of the chosen diphone. This value is compared with the LPC value of already stored database. All coefficient values are compared to find the exact spectral features. If the LPC value of the syllable-syllable combination does not coincide with the LPC value of the stored segments, then the value is changed between the concatenation depending on the "CV-CV" combination to smooth the output.

The following Table 1 shows a sample of LPC coefficient values of the start segment for the diphones "ka_d", "kaa_r" and "ka_s".

Diphone	1	2	3	4	5
Ka_d	1.000000	-1.038448	-1.401614	-1.211756	-1.410790…
Kaa_r	1.000000	0.080596	-1.314152	-1.530058	-1.494284…
Ka_s	1.000000	-0.048233	-0.414213	0.008168	-0.088925…

Table 1 : LPC coefficient values of the start segment for the diphones "ka_d", "kaa_r" and "ka_s".

The following Table 2 shows a sample of list of first syllable, second syllable and the corresponding diphone to be inserted between the syllable-syllable combination.

First Syllable	Second Syllable	Diphone
Ka	Da	Ka_d
Ka	Sa	Ka_s
Ma	Za	Ma_z
Na	La	Na_l
Ta	Ma	Ta_m

Table 2 : Shows the list of first syllable, second syllable and the corresponding diphone to be inserted between the syllable-syllable combination.

3. TIME SCALE MODIFICATION IN TAMIL TEXT-TO-SPEECH

As mentioned earlier, to introduce individual smoothness for syllable in Tamil Text-to-Speech, time scale modification is carried out for each syllable. The pitch value for tamil syllable is changed by performing time scale modification. Duration value is calculated for each syllable using Praat software [10] and the value is maintained as database. A sample list is shown below. Syllable duration is calculated in milli seconds. The duration of a syllable affects the quality of sound produced. Duration, as a supra segmental feature, is a dependent feature as an element of intonation [3]. This feature operates in association with pitch pattern and accent. A pitch pattern has its own duration. Duration is one of the dimensions of intonation. It is counted at various segment levels, viz., syllable, word, utterance and pause [3]. It is also an effective factor as it exerts certain constraints over rhythm and tempo. The duration of sounds, syllables, words or phrases will have their share in the prosodic system of a language.

As per phase vocoder algorithm [13], pitch change is done by the following procedure. Let y(n) be the segment (syllable) whose pitch has to be changed according to some defined pitch profile information. Let x(n) be each sub-segment and x(n) has N number of sampling point and its period is T. Let the required period is T1. When T1 is greater than T, i.e., target pitch is lower than the original pitch, and then glottal period is extended by making a intermediate signal such that $x_{int}(n)$ is the concatenation of signals x(n) and _x(n), where 0< _ < 0.25. Thus we will get a signal whose period is 2T having pitch half of the original and contains 2N numbers of sampling point. Now a window W(n) is defined whose length is equal to the desired pitch period on the intermediate signal. Now concatenating those changed pitch periods generate the required segment. This procedure is adapted for Tamil syllables. One of the issues that were tackled was the fact that Tamil syllables in general were of variable duration.

Matlab program has been written to change the duration value. The following examples shows the duration value for some sample words.

Examples

Word 1	:	mozhi
Syllable		Duration
Mo	174	
Zhi	458	
Word 2	:	oli
Syllable		Duration
O_l	267	
li	482	
Word 3	:	inimai
Syllable		Duration
I_n	290	
Ni	268	
Mai	557	

Word 4	:	namadhu
Syllable		Duration
Na	232	
Ma	209	
Dhu	435	

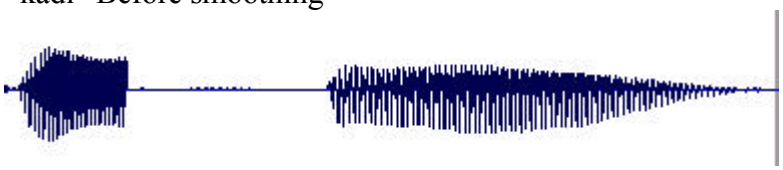
4. EXPERIMENTAL RESULTS

Speech output of Tamil text to speech based on simple concatenation technique is compared with the speech output of Tamil Text to Speech using residual excited LPC based synthesis. The quality is improved. The required diphone is inserted between CV-CV combinations and thus the spectral discontinuities are lowered at unit boundaries.

Example:

An example comparison is done for the word “kadi”

“kadi” Before smoothing



“kadi” After smoothing

Figure 2: Experimental result for word “kadi”

It is seen from figure 2 that the preceding “ka” waveform takes on the characteristics of the succeeding “di” waveform.

5. CONCLUSION

The quality and smoothness of Tamil text to speech output has been improved. At present, 1000 diphones have been created. LPC based diphone selection improves the quality of Text to speech synthesis than simple concatenation of syllables. Efforts were taken to develop the complete diphone database and to create a table, which includes the duration value for all the syllable-syllable combination. The diphone database was developed for CV-CV combination and further improvement can be done by developing diphone database for CV, VC combination.

References

1. R. Muralishankar and A G Ramakrishnan (2000), "*Robust Pitch detection using DCT based Spectral Autocorrelation*", Conference on Multimedia Processing, Chennai, Aug. 13-15, pp. 129-132.
2. R Muralishankar and A G Ramakrishnan (2001), "*Human Touch to Tamil Speech Synthesizer* ", Tamilnet 2001, Kuala Lumpur, Malaysia, pp. 103 - 109, 2001.
3. Soumen Chowdhury, A.K Datta, B.B.Chaudhuri (2001), "*Study of Intonation Patterns for text reading in standard colloquial Bengali*", Proceedings of IWSMSP-2001.
4. A.Bandyopadhyay (2002), "*Some Important aspects of Bengali Speech Synthesis System*", IEMCT JUNE ,Tata McGraw-Hill.
5. Aniruddha Sen (2001), "*Speech Synthesis in Indian Languages*", Pre-Workshop Tutorial on Speech and Music Signal Processing IWSMSP-2001.
6. A. G. Ramakrishnan (2001), "*Issues in standardization for Text to Speech in Tamil*", Tamilnet 2001, Kuala Lumpur, Malaysia.
7. Douglas O'Shaughnessy (2000), "*Speech Communication - Human and Machine*", Second Edition, IEEE press.
8. G. L. Jayavardhana Rama, A. G. Ramakrishnan, V. Vijay Venkatesh, and R. Muralishankar, (2001) "*Thirukkural: a text-to-speech synthesis system*", Proc. Tamil Internet 2001, Kuala Lumpur, pp. 92-97, August 26-28.
9. Min Tang, Chao Wang, Stephanie Seneff (2001), "*Voice Transformations: From Speech Synthesis to Mamalian Vocalizations*", Conference on Speech Communication and Technology, Denmark.
10. R. Muralishankar, A. G. Ramakrishnan and Prathibha P (2002), "*Dynamic Pitch changes for concatenative synthesis*", SPPRA, Greece, 2002.
11. S. Varho and P. Alku (1997), *A Linear Predictive Method Using Extrapolated Samples for Modelling of Voiced Speech*, Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, Session IV, pp. 13-16.
12. Susanna Varho (2001), "*New Linear Predictive Methods for Digital Speech Processing*".
13. John Garas and Piet C.W.Sommen (1980), "*Time/Pitch Scaling Using The Constant-Q Phase Vocoder",* Eindhoven University of Technology
14. "Issues in high quality {LPC} analysis and synthesis" Hunt, M. and Zwierynski D. and Carr R Eurospeech 89, Vol 2 pp 348-351, Paris, France