NATURAL SOUNDING TTS BASED ON SYLLABLE-LIKE UNITS

Samuel Thomas[1], *M. Nageshwara Rao*[1], *Hema A. Murthy*[1], *C.S. Ramalingam*[2]

[1]Department of Computer Science & Engineering, [2]Department of Electrical Engineering,
Indian Institute of Technology Madras,
Chennai - 600 036, India
email:{samuel,mnrao,hema}@lantana.tenet.res.in, ramli@ee.iitm.ac.in

ABSTRACT

In this work we describe a new "syllable-like" speech unit that is suitable for concatenative speech synthesis. These units are automatically generated using a group delay based segmentation algorithm and acoustically correspond to the form C*VC* (C: consonant, V: vowel). The effectiveness of the unit is demonstrated by synthesizing natural-sounding speech in Tamil, a regional Indian language. Significant quality improvement is obtained if bisyllable units are also used, rather than just monosyllables, with results far superior to the traditional diphone-based approach. An important advantage of this approach is the elimination of prosody rules. Since f_0 is part of the target cost, the unit selection procedure chooses the best unit from among the many candidates. The naturalness of the synthesized speech demonstrates the effectiveness of this approach.

1. INTRODUCTION

The extent of naturalness of synthetic speech produced by state-of-the-art speech synthesizers is mainly attributed to the use of concatenative synthesis [1]. This synthesis method uses basic speech units that produce the sounds of the particular language, along with the coarticulation, prosody, and transitions of the language [2]. These basic units are selected from a repository of stored waveforms. The quality of the synthetic speech is thus a direct function of the available units, making unit selection very important. For good quality synthesis, all the units of the language should be present. Moreover, the units should also be generic so that they can be used for unrestricted synthesis, which means that they should have minimum prosodic variations [3]. The commonly used basic units are phonemes, diphones, syllables, words or sentences. Of these units, phones are found to be inefficient for speech synthesis because they fail to model the dynamics of speech sounds with their large variability depending on context [4]. Syllables on the other hand, are inherently of longer duration and it has been observed that the relative duration of syllables is less dependent on speaking rate variations than that of phonemes [5]. The human auditory system integrates time spans of 200 msecs of speech, which roughly corresponds to the duration of syllables [6]. Syllables also capture the co-articulation between sounds better than the phonemes.

The scripts of Indian languages are characters that are orthographic representations of speech sounds. A character in Indian languages is close to a syllable and can be typically of form C*VC* (C:consonant, V: vowel). There are about 35 consonants and 18 vowels in Indian languages [7]. For Indian languages, syllable units are a much better choice than units like diphone, phone, and half-phone [8]. Our experience also confirms this [9]. We demonstrate that we don't actually need units that are syllables in the linguistic sense, but merely that "syllable-like" units suffice. A "syllable-like" unit is one that has a vowel nucleus. We have automated the extraction of these units from the data by using the group delay based segmentation algorithm [10, 11]. This leads to synthesized speech that has a very high degree of naturalness. Diphone-based synthesizers need elaborate prosody rules to produce natural speech; lot of care and effort are required for the rules to be put in place. Our approach does not need prosody rules; instead, we rely on the f_0 based component of the target cost for selecting the best unit required. The effectiveness of this approach is demonstrated through examples of synthesized speech for the Indian language Tamil, which are available at the website [12].

In Section 2, we discuss the identification of the syllable-like units using a multilevel segmentation scheme on a database of Tamil news bulletin data. In Section 3, we use the speech units with the FestVox voice building framework [13] to build a cluster unit selection synthesizer for Tamil. In Section 4, the syllable-like unit is evaluated and in Section 5, results from the evaluation of the synthetic speech are presented and discussed.

2. AUTOMATIC GENERATION OF SYLLABLE-LIKE UNITS

To generate natural sounding synthesized speech by concatenating prototype waveforms, a speech database with large number of units having adequate variations in prosody and spectral characteristics is needed. Before such a database can be used for synthesis, it has to be segmented into basic units and also labeled. If manually done, this task becomes tedious, time consuming, and error prone. Moreover, due to variability in human perception, large inconsistencies have been observed. In [10, 11] we have proposed an automatic segmentation and identification of syllable-like speech units, based on the group delay function. Syllable boundaries correspond to minimum energy regions. Tracking these boundaries from the signal energy is error prone because of the localized fluctuations present, giving rise to missed and extraneous boundaries. Instead, identifying them using group delay is not only consistent but also produces results that are close enough to manual segmentation [14, 15].

We now review the group delay based segmentation algorithm. Given a speech signal $x[n]$, we first compute its short-term energy $E[k]$. This short-term energy function is viewed as the positive-frequency part of the magnitude spectrum of a minimum-phase signal, and we examine the corresponding minimum-phase signal's group delay. We observe that the peaks and valleys in the group delay correspond to the peaks and valleys in the short-term energy function [16, 17]. The syllable boundaries are obtained as the location of the valleys of the group delay function. In practice, we work with $(1/E[k])^{\gamma}$ instead of $E[k]$. This causes the syllable boundaries to be associated with the *peaks* of the group delay function, rather than its valleys (see [14, 15]).

The steps are given below:

- Let $x[n]$ be the digitized speech signal. Compute its short-term energy (STE) function $E[k]$, $k = 0, 1, \ldots, M-1$, using overlapped windows. Denote the minimum value of the STE function by $E_{\min}$.
- For $2M$ let N be its nearest power of 2. Append $E_{\min}$ to $E[k]$ for values of k beyond M up to $N/2$.
- Let $E'[k] = 1/(E[k])^{\gamma}$, where $\gamma = 0.001$. This step reduces the dynamic range and thus prevents large peak excursions.
- Make $E'[k]$ symmetric by reflecting it around the y-axis. View the symmetrized sequence as a magnitude spectrum (having N points between $-\pi$ and π).
- Compute IDFT of the symmetrized sequence. The causal part of the resulting sequence is a minimum-phase signal $\hat{e}[n]$.
- Compute the group delay function of $\hat{e}[n]\,w[n]$, where $w[n]$ is a cepstral lifter window of length N_c.
- Locations of the positive peaks in the group delay function give approximate syllable-like boundaries.

Segmenting the speech signal at the minimum energy points gives rise to units that have $\mathrm{C}^*\mathrm{V}\mathrm{C}^*$ structure (C: consonant, V: vowel). Note that we can have polysyllables as basic units in the concatenative synthesis approach. We have explored the use of units up to trisyllables. The group delay based algorithm has in it the ability to locate polysyllables by adjusting the so-called "window scale factor (WSF)" (N_c is inversely proportional to WSF; see [14, 15] for more details).

As an example, consider segmenting the Tamil phrase "*ennudaya dAimozhi*". The signal, its energy function, and the group delay are shown in in Fig. 1. The positive peaks in the group delay function give the syllable boundaries, and are identified by the vertical lines. The window scale factor has been chosen such that the segmentation yields monosyllable-like units. VCs appear only at the onset, CVs only at the coda, and CVCs anywhere. For this example, the derived boundaries are */en/, /nud/, /day/, /ya/, /dAim/, /moz/, /zhi/*, where (i)*/en/* is a VC unit, (ii) */nud/, /moz/* are CVC units and (iii) */ya/, /zhi/* are CV units. These units are automically generated and assigned labels after listening to them.

For the identification and analysis of speech units an existing news bulletin database, called DBIL [18], was used. A prompt-list created from the the analysis stage was used to generate the basic units needed for synthesis. About 45 minutes of speech data was read by a native Tamil speaker in an anechoic chamber. We refer to this as the voice-talent (VT) database.

To get the polysyllable boundaries, we observed that it is better to first segment the DBIL sentences into word-like units using the group delay algorithm, and then seek the syllables within each word-like unit. To obtain word boundaries, the window scale factor was set to a high value, whereas various low values of WSF yielded monosyllables, bisyllables, and trisyllables.

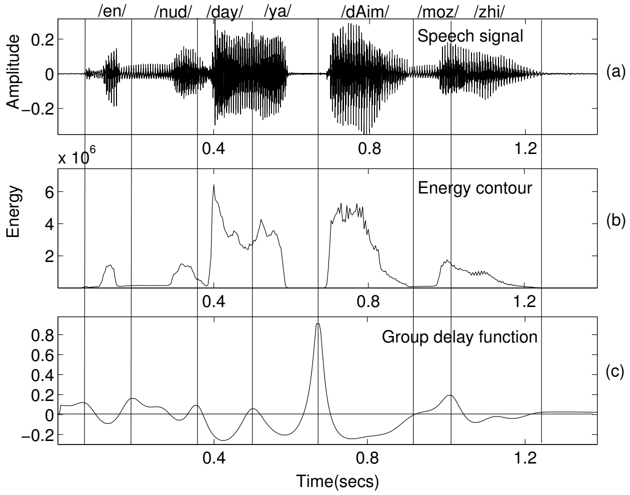

Figure 1: (a) Speech signal, (b) short-term energy function, and, (c) corresponding minimum-phase signal's group delay for the Tamil phrase "*ennudaya dAimozhi*".

The WSF values were determined over a set of 30 sentences (average of six words per sentence), taken from the above-mentioned Tamil news bulletin. We varied WSF from 15 to 25; the corresponding word boundary accuracy varied from 48.2% to 54.1%, with the maximum accuracy of 55.9% occuring for WSF = 22.

The word-level segmentation is now made finer by varying WSF from 4 to 10. When WSF is low, monosyllables dominate, whereas for high WSF, bisyllables and trisyllables dominate. The number of mono-, bi-, and trisyllables for 100 words taken from the 30-sentence subset are given in Table 1 for various WSF values. Fig. 2 shows the word-level and finer segmentation for "*vanakkam*".

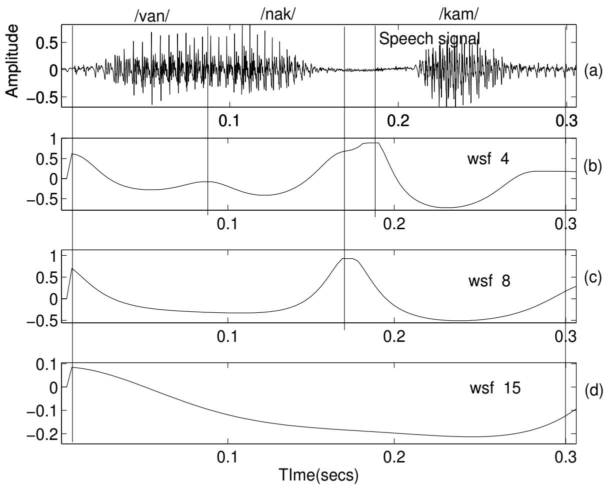

Figure 2: (a) Speech signal for the Tamil word "*vanakkam*". (b) For WSF = 4, we get the monosyllables */van/, /nak/, /kam/*. (c) When WSF = 8, we get one bisyllable */vanak/* and one monosyllable */kam/*. (d) When WSF = 15, we get the trisyllable */vanakkam/*.

The complete segmentation exercise on the DBIL database resulted in 1,325 unique monosyllable-like units,

Window scale factor	Monosyllables	Bisyllables	Trisyllables
4	273	62	3
6	168	70	10
8	96	94	18
10	52	101	28

Table 1: Number of *syllable-like* units for different window scale factors WSF.

4,882 unique bisyllable-like units, and 4,795 unique trisyllable-like units.

3. BUILDING A CLUSTER UNIT SELECTION SYNTHESIZER FOR TAMIL

The above syllable-like speech units were used in the FestVox cluster-unit-selection based speech synthesizer [13]. An algorithm that clusters units based on their phonetic and prosodic context was used [3]. The selection criteria has two costs, viz., (a) target cost: evaluates how close the speech unit's features are to the desired actual phonetic and prosodic features, and (b) concatenation cost: measures how well the speech units match and join with each other when concatenated [3]. The unit that minimizes both costs is selected.

Synthesis Voice Acoustic Units: We identified the phoneset for the VT database containing the needed monosyllables, bisyllables and trisyllables and prepared a promptlist that covers their occurrence in various contexts. The VT database consisted of 6,421 sentences containing 59,478 syllable units (including repetitions). This database was labeled automatically at different syllable levels, and the boundary errors were manually corrected (which is a much simpler task than labeling the entire database manually). There were 749 monosyllable (35,126 realizations), 2,649 bisyllables (17,670 realizations), and 2,342 trisyllables (6,682 realizations) in the above speech database.

Text-to-phoneset: A comprehensive set of letter-to-sound rules were created to syllabify the input text into the syllable-like units. These rules are framed in such a way that each word is split into its largest constituent syllable units: trisyllables first, and then bisyllables only if trisyllable combinations are not present, and finally monosyllables if bisyllable combinations are also not present.

4. EVALUATION OF THE SYLLABLE-LIKE UNIT

In order to test the improvement of synthesized speech quality using syllable-like units, a perceptual evaluation of 20 sets of synthesized Tamil sentences was conducted using 20 native Tamil subjects. Each set had 4 different sentences synthesized using different methods: the first in each set was synthesized using a multilingual diphone synthesizer [19]; the second was synthesized using monosyllables only; the third used both monosyllables and bisyllables units, with the monosyllables being used only when bisyllables are not present; the final sentence was created with trisyllables, bisyllables, and monosyllables. For the last case, each word used the largest possible syllable unit. As an illustration, the phrase "*inda nikazchchiyil*", contained following units for the 4 cases:

- Diphones: */i-n/ /n-d/ /d-a/ /n-i/ /i-k/ /k-a/ /a-z/ /z-ch/ /ch-ch/ /ch-iy/ /iy-l/*
- Monosyllables: */in/ /da/ /ni/ /kaz/ /chchi/ /yil/*
- Mono- and bisyllables: */in/ /da/ /nikaz/ /chchiyil/*
- Mono-, bi-, and trisyllables: */in/ /da/ /nikazchchi/ /yil/*

The subjects were asked to score the naturalness of each waveform on a scale from 1 to 5 (1=Bad, 2=Poor, 3=Fair, 4=Good 5=Excellent). The Mean Opinion Score(MOS) for each of the synthesis units is given in Table 2.

Diphone	Monosyllable	Bi- and monosyllable	Tri-, bi-, and monosyllable
1.34	1.47	3.74	3.97

Table 2: MOS score for Tamil sentences synthesized using different speech units.

A second perceptual test was conducted to find which order of syllable-like units was best acceptable for synthesis. The combinations used were:

- Monosyllables at the beginning of a word and bisyllables at the end. E.g., "*palkalaikkazaha*" is synthesized as */pal/ /kal/ /a/ /ik/ /kaz/ /a/ /ha/.*
- Bisyllables at the beginning of a word and monosyllables at the end. E.g., */palka/ /laikka/ /zaha/.*
- Monosyllables at the beginning and trisyllables at the end of a word. E.g., */pal/ /kal/ /a/ /ik/ /kazaha/.*
- Trisyllables at the beginning and monosyllables at the end of a word. E.g., */palkalaik/ / kaza/ /ha/.*

20 common Tamil words were synthesized using these 4 combinations and the 20 Tamil subjects were asked to score the naturalness of each waveform with a score as in the first test. The MOS scores for each of the waveforms is given in Table 3.

Mono- and bisyllable	Bi- and monosyllable	Mono- and trisyllable	Tri- and monosyllable
3.1	3.9	3.8	4.1

Table 3: MOS for Tamil words using different syllable combinations.

5. RESULTS OF EVALUATION OF THE SYLLABLE-LIKE UNIT

The results of the first MOS test show that speech synthesis with syllable-like units is much better than diphone based synthesis. Using only monosyllables is only slightly better than the diphone synthesis. Of the three different types of syllable-like units, trisyllables as the primary concatenation unit is preferred to the other two. The disadvantage is, of course, the number of such units needed. However it is also evident from the scores that bisyllable units can also be used in place of trisyllables to achieve very good results, but without an explosion in the number of units needed.

The results of the second test show that it is better to use large units (trisyllables or bisyllables) in the beginning, and monosyllables only at the end. A further examination of our sentence repository showed the presence of certain morphemes that appear at word boundaries. These morphemes are essentially monosyllables, and since they appear at word boundaries, they help in the prediction of phrase boundaries [20]. A set of 54 such tags have been identified. It is also observed that more than 50% of the words in our sentence repository have one of these tags.

The appropriateness of the syllable-like units is evident from the naturalness of the synthesized speech. It is important to observe that no duration, intonation or energy modelling have been applied. No modifications have been done on the default weights of various parameters used in calculating the target and concatenation costs in the FestVox framework. This is an important difference between diphone-based synthesis, which relies heavily on prosody modelling for good quality speech.

The primary reason for the good quality of the synthesis using syllable-like units is that they have more prosodic and acoustic information and less discontinuities when compared to other synthesis techniques using phones, diphones, or half-phones. As described earlier, the boundaries of the syllable-like units correspond to low energy regions of the short-term energy function. These low energy regions correspond to minimum coarticulation points and are hence preferable for concatenative waveform synthesis. The speech waveform and spectrogram plot for the Tamil phrase "*inda saNtAIkku mukkiya kAraNam*" is shown in Fig. 3. The spectrogram shows that the formant changes are not abrupt at the concatenation points. Moreover, spectral changes are uniform across the syllable boundaries and hence reinforce the idea that the syllable-like unit is indeed a good candidate for concatenative speech synthesis. The number of concatenation points for such units is also very less, which is 4 for this example, instead of 26 present in the diphone synthesis.

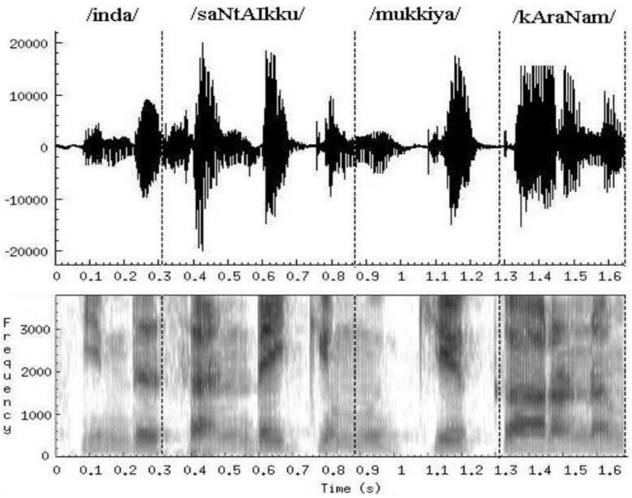

Figure 3: Speech signal and spectrogram for the synthesized Tamil phrase "*inda saNtAIkku mukkiya kAraNam*"

The observation from perceptual experiments that sentences synthesized using bisyllable units as the primary units perform almost as well as those using trisyllables, makes them an ideal choice for unrestricted speech synthesis since they strike a balance between quality and number of units needed. Our results also indicate that bisyllables, along with a sufficient number of the monosyllables (especially those that appear as morphemes) in various prosodic variations, are sufficient to produce very good quality synthetic speech. With these units in place the unit selection algorithm performs well and is able to pick the best units based on its target and concatenation cost estimates.

6. CONCLUSIONS

We have demonstrated natural sounding synthesized speech for Tamil using syllable-like units. Using units up to bisyllables strikes a balance between quality and number of units needed. The generation of these units has been automated using our group delay based algorithm. This approach is general in that it is suitable for other languages as well.

REFERENCES

[1] A.W. Black and K.A. Lenzo, "Building synthetic voice," `http://festvox.org/bsv/`, 2003.

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

[3] A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1996, vol. 1, pp. 373–376.

[4] Z. Hu, J. Schalkwky, E. Baranrd and R. Cole, "Speech recognition using syllable-like units," in *Proceedings of Int. Conf. Spoken Language Processing*, Philadelphia, PA, USA, 1996, vol. 2, pp. 1117–1120.

[5] S. Greenberg, "Understanding speech undestanding: Towards a unified theory of speech perception," in *In Proc. ESCA Workshop on The Auditory Basis of Speech Perception*, Keele Univ., UK, 1996.

[6] A. Hasuenstein, "Using syllables in a hybrid HMM-ANN recognition system," in *Proceedings of EUROSPEECH*, Rhodes, Greece, 1997, vol. 3, pp. 1203–1206.

[7] S.P. Kishore, R. Sangal and M. Srinivas, "Building Hindi and Telugu voices using Festvox," in *ICON*, Mumbai, India, Dec 2002.

[8] S.P. Kishore and A.W. Black, "Unit size in unit selection speech synthesis," in *Proceedings of EUROSPEECH*, 2003, pp. 1317–1320.

[9] M. Nageshwara Rao, S. Thomas, T. Nagarajan and H.A. Murthy, "Text-to-speech synthesis using syllable-like units," in *National Conference on Communication*, IIT Kharagpur, India, Jan 2005, pp. 227–280.

[10] T. Nagarajan and H.A. Murthy, "Group delay based segmentation of spontaneous speech into syllable-like units," *EURASIP Journal of Applied Signal Processing*, vol. 17, pp. 2614–2625, 2004.

[11] T. Nagarajan, H.A. Murthy, and, R.M. Hegde, "Segmentation of speech into syllable-like units," in *Proceedings of EUROSPEECH*, 2003, pp. 2893–2896.

[12] *`http://www.lantana.tenet.res.in/Research/Speech/TTS/demo.html`*.

[13] A.W. Black, P. Taylor and R. Caley, "The festival speech synthesis system," `http://festvox.org/festival/`, 1998.

[14] V.K. Prasad, *Segmentation and recognition of continuous speech*, Phd dissertation, Indian Institute of Tech-

nology, Department of Computer Science and Engg., Madras, 2002.

[15] V.K. Prasad T. Nagarajan and H.A. Murthy, “Minimum phase signal derived from the root cepstrum,” *IEEE Electronics Letters*, vol. 39, no. 12, pp. 941–942, June 2003.

[16] H.A. Murthy and B. Yegnanarayana, “Formant extraction from minimum phase group delay functions,” *Speech Communication*, vol. 1, pp. 209–221, 1991.

[17] H.A. Murthy, “The real root cepstrum and its applications to speech processing,” in *National Conference on Communication*, IITM Chennai, India, 1997, pp. 180–183.

[18] *Database for Indian languages*, Speech and Vision Lab, IIT Madras, Chennai, India, 2001.

[19] N.Sridhar Krishna and H.A. Murthy, “Duration modeling of Indian languages Hindi and Telugu,” in *Proceedings of 5th ISCA Speech Synthesis Workshop*, CMU, Pittsburgh, 2004, pp. 197–202.

[20] N.Sridhar Krishna and H.A. Murthy, “A new prosodic phrasing model for Indian language Telugu,” in *Proceedings of INTERSPEECH*, 2004, pp. 793–796.