Text-to-Speech Synthesis using syllable-like units

M. Nageshwara Rao, Samuel Thomas, T. Nagarajan, and Hema A. Murthy

Department of Computer Science and Engineering
Indian Institute of Technology, Madras
{mnrao,samuel,raju,hema}@lantana.cs.iitm.ernet.in

Abstract

In this work we describe the design of a syllable based concatenative waveform synthesizer for Indian languages. A syllable-like unit is chosen primarily because Indian languages are syllable centred. Further, concatenation at syllable boundaries can lead to smaller error owing to the spectrum being similar across different syllable boundaries. But the number of syllables in a language can be large. This issue is addressed using a data driven approach to syllable identification. A group delay based algorithm automatically segments continuous speech at syllable-like boundaries. The unique syllable-like units are then added to a syllable repository. We present our preliminary results on speech synthesis with a speech repository obtained from a small continuous speech database.

1. Introduction

The high quality of synthetic speech produced by state-of-the-art speech synthesizers in terms of naturalness is mainly attributed to the use of concatenative waveform synthesis [1]. Concatenative waveform synthesis maintains a waveform repository of basic speech units that encapsulate the sounds in the particular language with the co-articulation, prosody and transitions exhibited by the system [2]. Since the synthesis technique derives its strength from the speech units in its repository it is of utmost importance to choose the best speech units that represent the language being modeled.

The basic units that can be used to make the speech repository can be phonemes, diphones, syllables, words, sentences or any such unit. The evaluation criteria for a speech unit used for the repository is defined in terms of two costs [3] namely,

- **Target cost**, which evaluates how close the speech unit's features are to the desired actual speech waveform features.
- **Concatenation cost**, which measures how well the speech units match and join with each other when concatenated.

An ideal speech unit for the speech repository would be the unit that minimizes both costs.

The organization of the paper is as follows. In Section 2, we discuss a syllable-like speech unit. In Section 3, the algorithm used to generate the syllable-like unit and speech repository is discussed. In Section 4, the procedure used to concatenate the appropriate syllable-like units is described. Finally, in Section 5, we address other issues namely,the modeling of prosody and syllable based speech synthesis.

2. A syllable-like speech unit

Speech in Indian language is based on basic sound units which are inherently syllable units made from C, CV, CCV, VC and CVC combinations, where C is a consonant and V is a vowel . From perceptual results, it is observed that from four different choices of speech units: syllable, diphone, phone and half phone, the syllable unit performs better than all the rest and is a better representation for Indian languages [4].

In this work, we describe a new speech unit namely, a "syllable-like" unit as the basic unit to be used in a concatenative waveform synthesizer. We create a speech repository and demonstrate that the syllable-like unit is indeed a good choice.

The syllable-like unit is defined in terms of the Short-Term Energy (STE) function. In terms of the STE function the syllable-like unit can be viewed as an energy peak in the nucleus (vowel) region. It tapers off at both the ends of the nucleus where a consonant may be present. In the STE function, the presence of the consonants at both the ends of a vowel result in local energy fluctuations. If these local energy fluctuations are smoothed out, then the valleys at both the ends of the nucleus can be considered as the boundaries of a syllable-like unit. To create the syllable-like unit repository, a group delay based segmentation algorithm [5] is used, which is explained in the following section.

3. Segmentation of speech into syllable-like units

We have earlier proposed a method [5] for segmenting the acoustic signal into syllable-like units, in which we derive a minimum phase signal from the short-term energy function as if it were a magnitude spectrum. We have found that the group delay function of this minimum phase signal is a better representative of the short-term energy function to perform segmentation.

For example, a speech signal corresponds to the phrase *ennudaya dAimozhi* is considered and segmented as described above. The original speech signal, and the group delay function are plotted in Figure 1. The positive peaks in the group delay function (see Figure 1(b)) correspond the boundaries of the syllable-like units. According to the segmentation algorithm, the boundaries derived are, /en/, /nud/, /day/, /ya/, /dAim/, /moz/, and /zhi/. We can note that, the resultant speech segments (units) correspond to

- VC or CVC, if it at the beginning of a word (e.g., /en/, /dAim/)
- CVC, if it is at the middle of a word (e.g., /nud/, /moz/)
- CV (CVC is also possible), if it at the end of the word (e.g., /ya/, /zhi/)

The unit that are derived as above constitute the speech repository used.

For the present task, continuous speech is collected from a female speaker in the lab environment. This speech data corresponds to Tamil text. The total speech data of 20 minutes is split into speech files each of duration 10 secs. These speech files are then segmented into syllable-like units using the above algorithm. The text is also split into syllable-like units and aligned to the speech segments. Each of the syllable-like units is tagged as one of the three categories, namely, (i) initial unit, (ii) medial unit, and (iii) final unit. Our speech repository currently contains only one unique candidate for each syllable-like unit.This resulted in total of 1242 syllable-like units.

4. Syllable-like unit concatenation

Naturalness in synthetic speech generated by concatenative speech synthesis increases when the number of concatenation points required to create a waveform is minimal. Concatenation performed using the syllable-like units described in the preceeding section, results in minimal number of junctures and discontinuity effects across the waveform being generated.

Since Indian languages already have a well defined syllable structure, choosing such a unit increases the quality of synthesis. A coarse classification of syllable in Indian language is a monosyllable (one syllable), a bisyllable (two syllables) or a polysyllable (two or more syllables). In synthesis, junctures are usually between the stable parts of the same vowel or between consonants that belongs to two different syllables resulting in high degree of acoustic alignment [6]. In our case the concatenation performed uses the latter.

The primary criteria for choosing a specific syllable-like unit for concatenation from the speech repository depends upon its appropriateness in the given context [7]. This requires the syllable to be properly tagged prior to its being used with positional identity. The tagging is done at the time of repository gneration itself i.e when a syllable-like unit is identified. The tagging information we use indicates whether the unit is picked from isolated utterances or from the initial, medial or final positions, of its origin word. Based on this, syllable-like unit are categorized into any one of the following groups:

- a monosyllable (also a word)
- a onset-syllable (occurs at the initial of a word)
- a medial-syllable (occurs at the middle of a word)
- a coda-syllable (occurs at the final of the word)

Use of such tagging information prevents a stressed syllable from appearing in an unstressed position and vice-versa. This categorization of certain syllables based on tags is also helpful in ensuring that these syllables do appear at their appropriate positions in the synthetic waveform also.

Further tagging ensures that the pause at word boundaries is automatically incorporated as each unit itself possesses a silence boundary. A simple chaining strategy of units in the time-domain, i.e concatenation of speech unit is infact sufficient because the concatenation of any two contiguous units is performed at a juncture between the stable parts of the same consonants of two different syllables like units. To maintain the acoustic stability at the junctures the concatenation of units is governed by the following heuristic rules:

- a VC is an initial unit.
- a VC is also an isolated unit.
- a CV occurs at the end unit.
- a CVC can be an initial, medial or final unit.
- a VC - CVC combination appears only at the initial part of the synthetic waveform.
- a CVC - CV combination appears only at the end of the synthetic waveform.

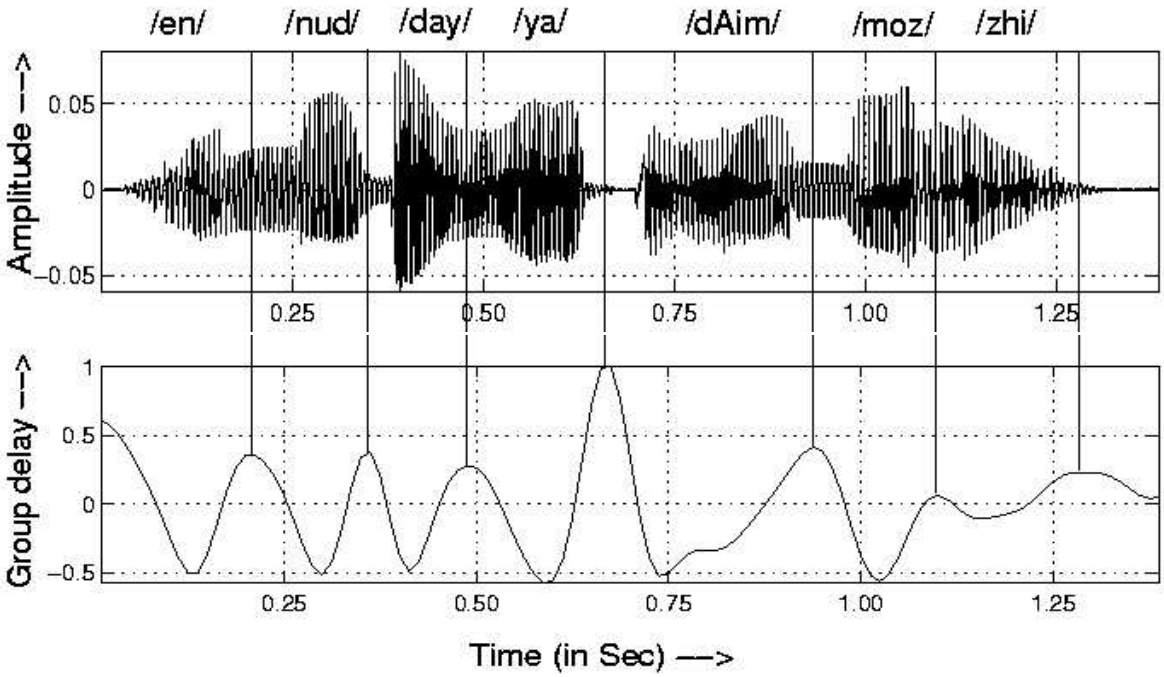

Figure 1: (a) Speech signal (b) Group delay function derived from the short-term energy function of the signal.

- a CVC - CVC combination appears in all other parts but with the syllables chosen based on tagging information and rules as described above.

Illustration of concatenation procedure for Tamil phrase *"mudal vanakam kurriukirar"*

/mud/+/dal/+/van/+/nak/+/kam/
+/kur/+/riuk/+/kirar/

The concatenated synthetic speech waveform and spectrogram for the above are plotted in Figure 2. From the spectrogram plots, it is obvious that formant changes are not abrupt at the concatenation points. Clearly the spectral changes are unifrom across syllable boundaries which suggests that the syllable-like unit is a good candidate for speech synthesis. Informal listening tests show that abrupt spectral changes which annoy the ear in diphone-based synthesis are absent in syllable based synthesis.If the phrase had been synthesized using diphones it would have resulted in 26 concatention points as opposed to 8 points for syllable base synthesis. Further, the spectral changes would have been abrupt, leading to significant error at the boundaries.

5. Criticism

Our current work uses a single unique syllable-like unit from the repository for synthesis. Optimum cluster unit selection algorithms need to be experimented. Spectral aspects of synthetic speech signal are to be carefully studied. Significant prosodic analysis in terms of duration, pitch and stress across a syllable in a given context need to be performed to improve the quality of speech synthesis. Extensive analysis on large number of synthesized sentences need to be performed estimate the average concatenation and target costs.

6. Conclusion

We have presented a new syllable like unit for concatenative waveform synthesizers in Indian languages. We show how the automatic segmentation algorithm has indeed created a useful speech unit that has low target and concatenation costs. Using a positional tagging scheme in association with a heuristic algorithm we produce good quality synthetic speech with a lab quality speech corpus. Issues related to prosody and intonation are yet to be addressed and are likely to improve the quality of synthesis.

7. References

[1] Alan W Black and Kevin A. Lenzo, Building Synthetic Voices, January 2003.

[2] Thierry Dutoit, An Introduction to Text-to-Speech Synthesis, 1st edition, The Netherlands, Kluwer Academic Publishers.

[3] Daniel Jurafsky and James H. Martin, Speech and Language Processing, 1st edition, Delhi, Pearson Education.

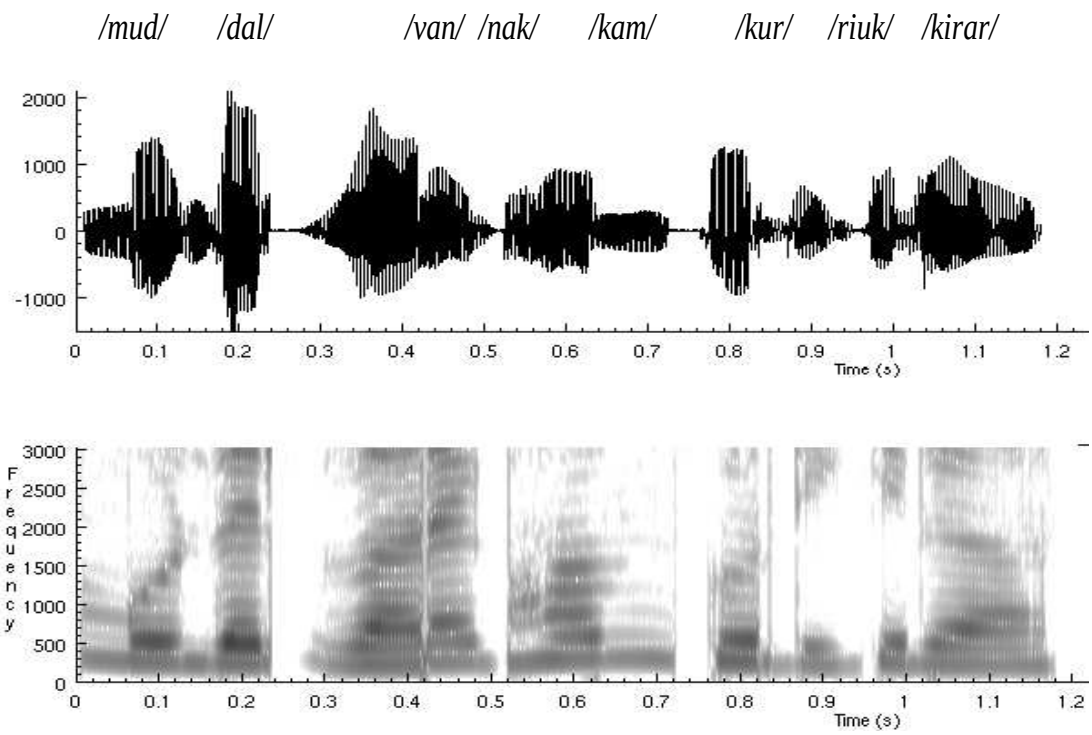

Figure 2: (a) Concatenated synthetic speech signal for Tamil phrase *"mudal vanakam kurriukirar"* (b) Corressponding spectrogram.

[4] Kishore, S., Black, A., Kumar, R., and Sangal R., "Experiments with Unit Selection Speech Databases for Indian Languages", National seminar on Language Technology Tools: Implementation of Telugu, October 2003, Hyderabad, India.

[5] V.Kamakshi Prasad, T.Nagarajan and Hema A.Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions, *Speech Communications*, Elsevier publications, Vol. 42, pp. 429-446, 2004.

[6] El-Imam,Y.A.;Banat,k; "Text-to-Speech Conversion on a Personal computer", Micro, IEEE, Volume:10, Issue:4, Aug.1990, Pages:62-74.

[7] Eric Lewis and Mark Tatham,"Word and syllable concatenation in text-to-speech synthesis".In:Sixth European Conference on Speech Communications and Technology, pages 615–618.ESCA, September 1999.