Continuity Metric for Unit Selection based Text-to-Speech Synthesis

Vikram Ramesh Lakkavalli, Arulmozhi P and A G Ramakrishnan
Medical Intelligence and Language Engineering (MILE) Laboratory
Department of Electrical Engineering
Indian Institute of Science, Bangalore, 560012, INDIA
Email: {vikram.ckm, p.arulmozhi}@gmail.com, ramkiag@ee.iisc.ernet.in

***Abstract*—A new method based on unit continuity metric (UCM) is proposed for optimal unit selection in text-to-speech (TTS) synthesis. UCM employs two features, namely, pitch continuity metric and spectral continuity metric. The methods have been implemented and tested on our test bed called MILE-TTS and it is available as web demo. After verification by a self selection test, the algorithms are evaluated on 8 paragraphs each for Kannada and Tamil by native users of the languages. Mean-opinion-score (MOS) shows that naturalness and comprehension are better with UCM based algorithm than the non-UCM based ones. The naturalness of the TTS output is further enhanced by a new rule based algorithm for pause prediction for Tamil language. The pauses between the words are predicted based on parts-of-speech information obtained from the input text.**

***Index Terms*—unit selection, MFCC, unit continuity metric, pitch continuity metric, spectral continuity metric, MILE-TTS, part-of-speech, pause model, Tamil, Kannada**

I. Introduction

Text-to-speech (TTS) synthesis transforms written information into spoken information for easier and efficient access of data. In a multilingual country like India, TTS helps the group of people who know many languages orally but are not aware of the script. Also, TTS is a boon to the blind and people with visual disorders for reading of text from the computer screen or printed books.

TTS synthesis can be broadly classified into i) parametric and ii) non-parametric methods. Parametric systems have the issue of naturalness of output speech whereas the non-parametric systems have challenges in the choice of basic speech unit, corpora selection and concatenation. Non-parametric TTS can further be categorized into demisyllable based, diphone based [3], syllable based [1] and polyphone based [10],[11] systems. Appropriate selection of the basic speech-unit for concatenation not only has the potential to produce better quality of synthesized speech but also drives the corpora size and the challenges in signal processing.

The quality of a TTS system is determined by the intelligibility of synthesized speech and its naturalness, a quality that indicates closeness to a human voice. Natural speech has good prosody, where prosody is defined as the collection of the dynamic features of speech such as pitch, duration, pause and stress. Prosody prediction from text would help to make the synthesis natural, however such a model needs to be developed for Indian languages.

For Indian languages, there hasn't been much progress in TTS technology. Some of the reasons being:

1) Lack of good prosody model for Indian languages.
2) Lack of concerted efforts to build good annotated speech corpora.
3) Absence of research and study in computational linguistics.

In our lab, we have a developed a TTS framework called MILE-TTS [11] and it can be used to develop concatenative speech synthesis system for any language. At present MILE-TTS supports speech synthesis for Kannada and Tamil languages.

II. Description of MILE-TTS

MILE-TTS is a concatenation based TTS synthesis system and it employs variable length polyphonic unit as the basic-unit for concatenation. The length of polyphonic unit is selected depending on the word and sequence of phonetic units. For speech synthesis, the required polyphone speech segments are selected from the manually segmented, annotated speech database and concatenated. The Kannada database has 8 hours of speech data with 1110 phonetically rich sentences recorded by a professional Kannada male speaker and stored at 16 kHz sampling frequency. Tamil database contains 5 hours of speech data with 1027 phonetically rich sentences stored at the same rate. The database is segmented and labeled at phoneme level. Database contains 64841 basic polyphonic units in Kannada and 42012 in Tamil.

A. Details of MILE-TTS

MILE-TTS engine accepts Kannada or Tamil text. The text is processed by the natural language processing (NLP) module to perform text-normalization and grapheme-to-phoneme (G2P) conversion. Unit-selection module employs a decision tree approach to search for units with best left and right phonetic context match for the target-unit. The synthesis database is rich and an average of 70 to 75% of the polyphonic-units selected for input text have all the phonetic contexts with multiple occurrences. If the target unit with context match is not found then the required polyphone is searched for within the available choices. If the polyphonic-unit is not found in the database, a stripped down version of the polyphone is searched and if no match is found, the target unit is dropped in the

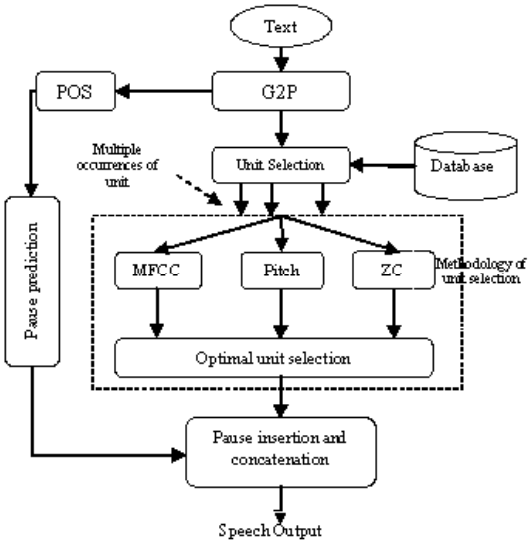

Fig. 1. Block diagram of MILE TTS

current version of MILE-TTS. The missing polyphone can be generated by a simple concatenation of the constituent phones.

Speech synthesis is carried out on a word by word basis for the input text. From the candidate units chosen by NLP module, the optimal units are selected by Viterbi search considering the lowest total join cost for a word. Selected units may be concatenated by spectral smoothing algorithm such as LSF interpolation for further enhancement of the output speech quality. A detailed block diagram of MILE-TTS is shown in Fig.1.

In the proposed method, emphasis is laid on selecting a best unit for concatenation based on prosody match. Prosody is an important aspect of speech, which includes rhythm, stress, and intonation. In quantifiable terms, prosody is defined as the collection of the features of speech such as pitch, intonation, timbre, duration and pause. Prosody information is crucial to a natural sounding TTS [7], and it is also a challenge to incorporate prosody into TTS without compromising on the output speech quality.

MILE-TTS incorporates an online parts-of-speech (POS) tagging module shown in Fig.1. From the sequence of POS output, pause between the words is decided. This work is carried out for Tamil and may be extended to all Dravidian languages taking into account the nuances of the language structure. The complete description of pause modeling and POS tagging is explained for Tamil language in the next section.

III. Pause modeling with POS information

One of the important characteristic of a natural sounding synthesized speech is the right amount of pauses at appropriate places in a sentence. The stress and pause can be modeled using the POS and other syntactic information. A wrong pause inserted between two words may make the synthesized speech unnatural. This is illustrated in Fig. 2 with an example each from English and Tamil. Here, the notation <np> denotes

1. The<np>book<np>is<p> on<np>the<np>table.(sounds natural)
2. The<np>book<np>is<np> on<p>the<np>table.(sounds unnatural)
3. சுவற்றின் <np> மேல் <p> சித்திரங்கள் <np> உள்ளன.(sounds natural)
4. சுவற்றின் <p> மேல் <np> சித்திரங்கள் <np> உள்ளன. (sounds unnatural)

Fig. 2. Effect of pause between words on naturalness of speech

"no pause" between the words, whereas <p> refers to the "required pause". Speech synthesized as per the tags for sentences 1 and 3 will be perceptually natural, whereas for the sentences 2 and 4, a wrong pause will make the speech sound unnatural. Hence POS information and pause are important in the context of TTS. In this section, we focus on the pause insertion between successive words by predicting pause from the estimated POS tags.

A. Parts-of-Speech

POS indicates the classification of words into different classes such as noun and verb, which decide the prosody of a sentence. An automated POS tagger is developed for this purpose, which uses lexical and context sensitive rules [8] for finding the POS of each word in a sentence. POS tagging is language specific and a recent system reported to be the comprehensive by the authors provides low accuracy while being computationally intensive [6]. Nevertheless even an approximate POS tagger helps to improve the naturalness of TTS synthesis. Hence an approximate, computationally efficient POS tagger is implemented for Tamil in MILE-TTS. Twelve main-tags and 27 sub-tags have been defined for Tamil, some of which are given below:

Main tags:

NN-noun, VB-verb, ADJ-adjective, AJP–adjectival participle, Q–quantifier, PP–post position

Sub-tags:

prs-present tense, pst–past tense, fut–future tense, 1,2,3–first, second, and third person respectively, s–singular pl–plural, neg–negative, acc–accusative case, dat–dative case

<PW>–specifies the default pause between any two words.

B. Pause Estimation using POS

Basic syntactic information of POS of words in a sentence is considered for forming rules for pause insertion. The prosodic structure of a sentence can be represented by different levels of break markers in the model [7]. For natural pause between words in a sentence, nine different levels of pause are considered as shown in Table I. Pause duration is highest at the sentence end.

Fig.3 shows a sample Tamil sentence in the first line, the corresponding meaning of the respective words in English in the second line and predicted POS information in the third line. Pause level estimation between the words is indicated within < > in the fourth line. Some of the pause estimation rules employed for the above sentence are explained below.

1) Rule1: Considerable pause duration is needed after semi-colon, <P6>, colon <P6> or a comma <P4>. As per

TABLE I
DIFFERENT LEVELS OF PAUSE BETWEEN WORDS

Pause Label	Pause duration
P0	No Pause
P1	10 to 15 ms
P2	100 to 150 ms
P3	200 to 250 ms
P4	400 ms
P5	400 to 500 ms
P6	600 to 700 ms
P7	1000 ms
PW	15 to 25 ms

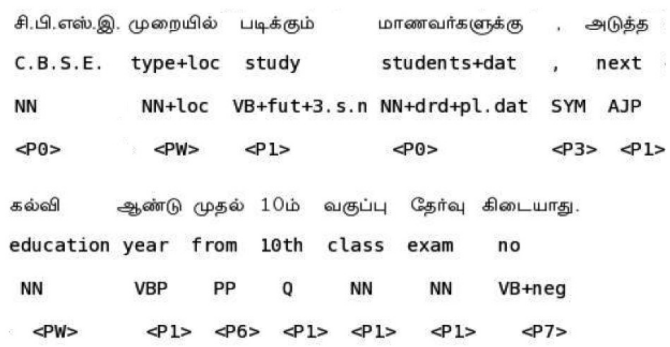

Fig. 3. Estimated POS tags and pause levels for a sample Tamil sentence. The pause level indicated below any word corresponds to the pause duration between the current and the following words.

this rule, <P4> is inserted wherever a comma occurs in the sentence.

2) Rule2: There should be a pause level <P6> before any quantifier.
3) Rule3: Combine the words with POS tag AJP and NN (any noun) occurring together, with a pause duration <P1> between them

There are 15 such rules which identify the pause level according to the word and its parts-of-speech.

The pause model has been incorporated into MILE-TTS. The performance is evaluated by group of natives for naturalness of the output speech on sentences synthesized with and without pause rules.

IV. UNIT SELECTION-REVIEW

This section describes the traditional techniques [9] employed for unit selection. The selection of possible candidate units for matching the specifications of target unit. The cost incurred in selecting the candidate units is called target cost and the cost of concatenating any pair of units is known as join cost. Total cost incurred is the sum of target and join costs [2].

$$C_{total} = C_{sel} + C_{join} \tag{1}$$

where, C_{total} is the total cost, C_{sel} is target cost or selection cost and C_{join} is the join cost between the two consecutive units. Target and join costs, in turn are a weighted combination of sub-costs [2]:

$$C_{sel} = \sum_{n=1}^{N} w_{sel}^{n} C_{sel}(n) \tag{2}$$

$$C_{join} = \sum_{n=1}^{Q} w_{join}^{n} C_{join}(n) \tag{3}$$

where N and Q are the number of sub-costs for selection and concatenation costs respectively.

In order to minimize the total cost, the individual costs C_{sel} and C_{join} need to be minimized. Some of the recent methods have focussed on minimizing the target and join costs separately. In [12], statistical prosody models are used in unit selection to minimize target and join cost. This work employs separate probabilistic models for pitch, duration and energy. In [13], signal dependent transformations are used to obtain the discontinuity metric for each candidate unit instead of considering features such as pitch and spectrum separately to compute the total cost. Lambert [14] talks about unit selection employing the Mel frequency cepstral coefficients (MFCC) as features at the concatenation boundaries with phonetic context. Another method of unit selection is to choose complete words or sentences from the database [4].

V. UNIT CONTINUITY METRIC

We propose a new algorithm to minimize the total cost C_{total} by employing unit-continuity-metric (UCM) to minimize C_{sel} in (1). The sub-costs of C_{sel} can be broadly grouped under spectral and pitch based features. It must be noted that the sub-costs of C_{sel} in (1) are considered only at the concatenation boundary. However the proposed continuity metric looks for continuity of features up-to several frames on both sides of the concatenation boundary. In the present work, two continuity metrics are proposed namely, the pitch continuity metric (PCM), and the spectral continuity metric (SCM). PCM employs pitch track and SCM employs the sequence of spectral centroids of the signal about the concatenation boundary as features. Hence C_{sel} in (1) can be written as,

$$C_{sel} = w_p C_{sel}^{p} + w_s C_{sel}^{s} \tag{4}$$

where, C_{sel}^{p} and C_{sel}^{s} are the costs incurred for pitch based features and the spectral features respectively. Optimal values for the weights w_s and w_p can be obtained by training on a reference database as suggested in [2] with the condition $w_s + w_p = 1$. Obtaining optimal values for w_p and w_s will be involved as it requires a reference database and objective evaluation of all the output sentences with different values for each variable.

Let us consider the boundary of concatenation between the units i and $i+1$ in a word as shown in Fig. 4. The individual unit-continuity-metrics at the i^{th} concatenation boundary in a word can be defined as $\bar{C_{sel}}(i)$ consisting of PCM and SCM

contributions, $C^{p\hat{}}_{sel}(i)$ and $C^{s\hat{}}_{sel}(i)$ respectively. The above discussion can be symbolically stated as,

$$C^{-}_{sel}(i) = w_p C^{p^{-}}_{sel}(i) + w_s C^{s^{-}}_{sel}(i) \quad (5)$$

Each of these sub-costs are computed as,

$$C^{p^{-}}_{sel}(i) = \sqrt{\sum_{k=-K}^{K} |p_i(k) - p_{i+1}(k)|^2} \quad (6)$$

and,

$$C^{s^{-}}_{sel}(i) = \sqrt{\sum_{k=-K}^{K} |s_i(k) - s_{i+1}(k)|^2} \quad (7)$$

where, $p_i(k)$ is the average pitch value and $s_i(k)$ is the spectral centroid of the k^{th} frame from the i^{th} unit concatenation boundary shown in Fig.4 and K is the number of frames employed on either side of the concatenation boundary. The value $K = 0$ represents the matching based only on the frames at the concatenation boundary.

From (6) and (7), we may observe that larger value of K would lead to better unit selection. However value of K is limited by the duration of the polyphonic unit and it is experimentally found that $K = 4$ is sufficient for the application.

A. Algorithm

In Fig. 4, the unfilled boxes indicate the units to be concatenated and the continuity metric features are extracted from either side of the concatenation boundary. In this paper, SCM and PCM are considered separately for evaluation. SCM features are obtained by low pass filtering the signal at 2000 Hz and calculating the spectral centroid by zero crossing rate.

1) Given a word input for TTS, the candidate units for concatenation are obtained from NLP module.
2) PCM features across the concatenation boundary are computed. PCM features are the pitch values $p_i(k)$ in (6). Typically 4 frames on either side of the boundary with 20 msec frame size without overlap is considered in experiments.
3) Pitch tracks of the consecutive units for a word are compared by Euclidian or absolute distance measure. This new distance measure is termed as pitch continuity measure (6).
4) Viterbi algorithm is employed to find the best path to minimize the continuity metric accumulated over the word.
5) Similarly SCM features are also employed to synthesize speech.

Perceptual experiments have shown that pitch continuity is more relevant in selecting units rather than units which are spectrally similar and the argument is supported in [5]. If the pitch continuity is ignored across the concatenation boundary, the output speech sounds unnatural. The spectral continuity is also an important factor in prosody, and it is shown through perceptual experiments that spectral continuity is not so important as the pitch continuity.

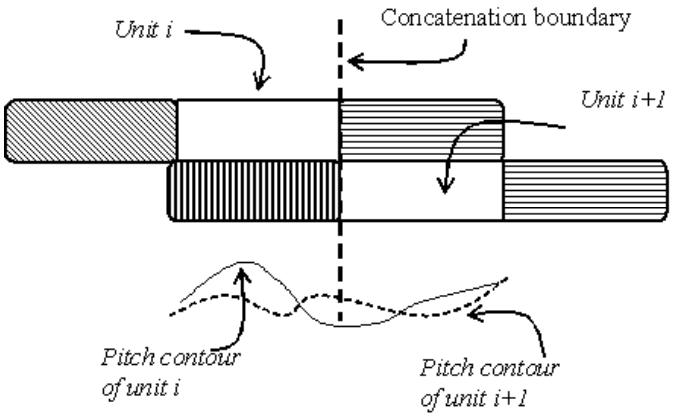

Fig. 4. Proposed unit selection method using PCM features.

TABLE II
RATING GUIDELINES FOR SUBJECTIVE EVALUATION OUTPUT SPEECH

Score	Subjective Perception
1	Poor speech, with discontinuities and very low comprehension
2	Poor speech with discontinuities, but comprehendible
3	Good speech quality with less discontinuity, and comprehendible
4	Very good speech quality, with less naturalness
5	As good as natural speech

B. Test Setup

To validate the proposed method, sentences corresponding to the synthesis database are given as input to the MILE-TTS. Also, speech is synthesized for input sentences from outside the database for subjective evaluation by MOS. Table II shows the basis for evaluation of synthesized sentences. Subjects are requested to rate the output speech on a scale of 1 to 5, based on intelligibility, speech continuity and naturalness.

Three methods are compared for evaluation, (i) MFCC based unit selection, which considers single frames at the boundary (ii) SCM and (iii) PCM based unit selection. MOS score is given by 10 native people on a set of eight paragraphs each for Tamil and Kannada.

VI. EVALUATION

For the validation of our claim for better unit selection using the proposed method, PCM and SCM features are tested independently on 10 test sentences from the database as input for synthesis. PCM and SCM features for unit-selection returned an accuracy of 100%, which means all the polyphonic units are selected from the corresponding polyphones in synthesis database. The same set of 10 test sentences are employed to synthesize speech by non-continuity metric for unit-selection. For such a case MFCC feature is employed, and the method returned an average self-selection of 83% for the same database sentences in Kannada and Tamil.

To compare the performance of different features for unit-selection, test input that is not present in the synthesis corpus

TABLE III
PERFORMANCE COMPARISON OF UCM AND NON-UCM BASED UNIT SELECTION. MOS SHOWN ARE MEANS OVER 8 PARAGRAPHS OF KANNADA AND TAMIL EVALUATED BY 10 PEOPLE EACH

Language	MFCC	PCM	SCM
Kannada	2.5	3.1	2.7
Tamil	3.0	3.6	2.8

TABLE IV
MOS EVALUATION OF A PARAGRAPH SYNTHESIZED WITH AND WITHOUT PAUSE MODEL

	Pause Model	No pause Model
MOS	3.2	2.6

is used. A set of eight input paragraphs are synthesized using PCM, SCM and MFCC features for Kannada and Tamil. The MOS score is evaluated for synthesized speech by a group of ten people native to each language. The MOS score for the proposed PCM method is 3.1 on a scale of 1 to 5 for Kannada and 3.6 for Tamil and the MOS scores for the other features are shown in Table III.

To test the effectiveness of pause prediction, ten Tamil sentences were synthesized using MILE-TTS and MOS scores are evaluated for naturalness as compared to the fixed pause insertion. Table IV shows that by employing pause prediction model in TTS the output sounds natural.

VII. CONCLUSION AND DISCUSSION

In this paper, we have proposed a new approach to evaluate and minimize the join cost of speech units. We have found from experiments that the pitch continuity measure is key to achieve better naturalness of output speech when compared to spectral continuity. With the new pause estimation module, the naturalness of TTS output is improved. The Kannada and Tamil TTS with PCM features for unit-selection can be tested at http://mile.ee.iisc.ernet.in/tts. In the future work, we propose to consider the combination of the spectral and pitch continuity metrics for a unified unit selection.

ACKNOWLEDGEMENT

The authors would like to thank Ministry of Social Justice and Empowerment, Government of India for funding the TTS project, Sri.K. Suchendra Prasad and Sri. Jayam Kondan for the Kannada and Tamil TTS voices, respectively, Dr G Sita for giving valuable suggestions and feedback, and K. Parthasarathy, Shiva Kumar HR, Arun Sriraman and Abhinava Shivakumar for the TTS code and creating the web demo, and Shanthi Devaraja and Sri Lakshmi for segmenting and annotating the synthesis databases.

REFERENCES

[1] Samuel Thomas, M. Nageshwara Rao, Hema A. Murthy, C.S. Ramalingam, "Natural sounding TTS based on syllable -like units", *EUSIPCO*, 2006.

[2] 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] M. Beumagel, A. Conkie, A. Syrdal, "Diphone synthesis using unit selection", *Proceedings. third ESCHCOCOSDA Workshop Speech Synthesis*, pp. 185–190, 1998.

[4] Esther Klabbers, Jan P. H. van Santen and Alexander Kain, "The Contribution of various sources of spectral mismatch to audible discontinuities in a diphone database", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 3,pp. 949–956, 2007.

[5] Kun Liu, Zhiwei Shuang, Yong Qin, Jianping Zhang, Yonghong Yan, "Mandarin accent analysis based on formant frequencies", *ICASSP*, Vol. 4, pp. 637–640, 2007.

[6] Dhanalakshmi V, Anand Kumar, Shivapratap G, Soman KP and Rajendran S, "Tamil POS tagging using linear programming", *International Journal of Recent Trends in Engineering*, Vol. 1, No. 2, pp. 166–169, 2009.

[7] Fu-chiang Chou, Chiu-yu Tseng, Keh-jiann Chen and Lin-shan Lee, "A Chinese Text-To-Speech System based on Part-of-Speech analysis, prosodic modeling and non-uniform units", *IEEE International Conf on Acoustics, Speech, and Signal Processing*, Vol 2, pp. 923–926, 1997.

[8] Arulmozhi. P, Sobha. L, Kumara Shanmugam. B. "Parts of speech tagger for Tamil", *Proc. Symposium on Indian Morphology, Phonology & Language Engineering*, pp. 55-57, 2004.

[9] Paul Taylor, "Text to speech synthesis", Cambridge Press, First edn, 2009.

[10] G L Jayavardhana Rama, A G Ramakrishnan, R Muralishankar and P Prathibha, "A complete text-to-speech synthesis system in Tamil", *Proc. IEEE 2002 Workshop Speech Synthesis*, pp. 191–194, 2002.

[11] K Parthasarathy, AG Ramakrishnan, "A research bed for unit selection based text to speech synthesis", *Proc. II IEEE Spoken Language Technology (SLT) workshop*, pp. 229–232, 2008.

[12] Wei Zhang, Liang Gu, Yuqing Gao, "Recent improvements of probability based prosody models for unit selection in concatenative text-speech", *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pp. 3777–3780, 2009.

[13] Jerome R. Bellegarda, "A Global Boundary-Centric Framework for Unit Selection Text-to-Speech Synthesis", *IEEE Trans. on Audio Speech and Language Processing*, Vol. 14, pp. 990–997, 2009.

[14] T. Lambert, Andrew P. Breen, Barry Eggleton, Stephen J. Cox, Ben P. Milner, "Unit selection in concatenative TTS synthesis systems based on mel filter bank amplitudes and phonetic context", *Proc. 8th European Conf. on Speech Commn. and Tech, EUROSPEECH*, pp. 273–276, 2003.