DEFINING SYLLABLES AND THEIR STRESS LABELS IN MILE TAMIL TTS CORPUS

Laxmi Narayana M and A G Ramakrishnan

Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012.

ABSTRACT

We report our work on stress labeling of syllables in the speech corpus of Tamil Text to Speech Synthesis. Syllable is the minimum possible speech segment which can be spoken independent of the adjacent phones. Keeping this in mind, a new syllabification strategy, which preserves the coarticulation effects of the phones present in the identified syllables, is proposed for Tamil language. The syllables are stress labeled based on the combination of their Pitch, Energy and Duration (PED). The byproduct of the stress labeling of the corpus is the prosodic knowledge that the first syllable in Tamil is stressed (in most cases, except when there is some emphasis of a particular syllable in a word) and the second syllable is stressed if the first syllable is/has a short vowel. Based on this, a rudimentary prosody model is developed for Tamil TTS.

Index Terms— Syllable, Stress labeling, Duration, Coarticulation, Pitch, Energy, prosody model, speech synthesis.

1. INTRODUCTION

Text to Speech (TTS) synthesis is an automated encoding process which converts a sequence of symbols (text) conveying linguistic information, into an acoustic waveform (speech). The existing TTS systems for European languages like English, German and French have achieved a good level of intelligibility and naturalness. Though a number of research prototypes of Indian language TTS systems have been developed, none of these are of quality that can be compared to commercial grade TTS systems in languages mentioned above. Further, much work has not been carried out on prosody models or increasing naturalness in synthetic speech for Indian languages.

Speech synthesis based on syllables seems to be a good possibility to enhance the quality of synthesized speech compared to mono-phone or diphone-based synthesizers. This consideration is based both on the fact that more coarticulation aspects are included in syllable segments compared to diphone units and on the fact that the main prosodic parameters (pitch, duration, amplitude) are closely connected to syllables [3]. Section 2 describes the goal of the paper. Section 3 gives the definition of syllable and stress and the reason for associating the parameter stress (only) with syllable (only). Section 4 discusses the proposed structure of the Tamil syllable. Section 5 presents different cues of stress as mentioned in the literature, the attributes used to quantify stress in the current work and discusses the novelty of the work. Section 6 describes the proposed method for stress labeling of syllables and the corresponding results. Section 7 describes the prosodic knowledge acquired from stress labeling and the developed preliminary prosody model for Tamil TTS.

2. MOTIVATION FOR THE WORK

Predicting the characteristics of speech to be synthesized is known as prosody modeling. Prosody model needs to predict the characteristics like pitch, amplitude and duration of a speech segment to be concatenated. Although changing the characteristics of the available speech segments in the corpus is possible and in practice also, this is considered to be secondary to the availability of speech segments with characteristics matching the target context. One may not have a unit exactly matching the target context. However, if there is an interface which can clearly distinguish the characteristics of speech units in the corpus, it reduces to a great extent the complexity of unit selection process and exhaustive calculation of join cost for different combinations of units. This also reduces the burden of modifying the characteristics of the units during concatenation. So, it is useful to organize the database in such a way that the stressed and the unstressed syllables are distinguished. This is the motivation for our work. Deciding whether the syllable to be synthesized is stressed or not, is dictated by the prosody model (section 7). To our knowledge, there is no prior work done on this aspect of any Indian language so far.

3. WHY ONLY SYLLABLE? WHY ONLY STRESS?

As far as the production of speech is concerned, ‘syllable’ is the minimum possible speech segment, which can be spoken in isolation; i.e., without the help of its adjacent phones. A syllable contains only one vowel. The format of a syllable can be V, VC, CV, CVC (Ex: /a/, /ap/, /ra/, /ram/, respectively) etc. Section 4.1 elaborates on this. The

question is ‘why only stress’? Why not other expressions? It is good if the synthesizer can speak with different expressions: sadness, happiness, anger. But these expressions need not be considered while *reading* the text. Further, emotional expressions can be the next level of sophistication, once a basic quality of synthetic speech is achieved. But some syllables need to be emphasized (stressed) while reading. The immediate question, which follows is: Why ‘stress’ for ‘syllables’ only? Why not other units? Stress is the relative emphasis that may be given to certain *syllables* in a word. The parameter “stress” cannot be associated with a mono-phone, since a speaker cannot stress a phone without stressing the adjacent phones within a syllable. So, we hypothesize that syllable is the minimum possible unit with which the stress parameters can be associated.

4. SYLLABLE STRUCTURE

The problem of defining a syllable has always plagued the linguists. While it is true that native speakers can in most cases consistently say how many syllables a given utterance has, the format of a syllable cannot be globalized. Different languages have different formats of syllable in their grammar. The classical syllable structure of Tamil is defined for poetry and deals more with the written aspect of the language. As far as stress labeling of a spoken syllable is concerned, this definition of syllable is observed to be not so relevant.

4.1 Proposed Syllable structure for Tamil

The idea of syllabification lies in exactly defining the syllable segments (V, CV, VC, CVC, CCV or CVCC as the case may be) that are used for speech production and can be picked up from the database, so that good coarticulation of adjacent segments is preserved in the produced speech. These segments need not always be syllables in the traditional sense, but boundaries between them can well be determined from phonetic point of view, enabling automatic generation of the syllable database.

The following example shows a Tamil sentence (O), its phonetic transcription (P) and syllabic transcription (S).

O: நாயனக்காரன், மெல்ல, நாயனத்தை உதட்டில் வைத்து, பீ, பீ, என்று சத்தம் பார்த்தான்.

P: n A y a n a kk A r a n # m e ll a # n A y a n a tt ae # u d a TT i l # w ae tt u # p I # p I # e n R u # s a tt a m # p A r tt A n

S: nA ya nak kA ran mell lla nA ya natl tlae u daT Til waet tu pI pI en Ru sat tam pArt tAn

The syllables identified in the above example are of the format CV, CVC, V, VC and CVCC. Identifying V, CV, VC, and to some extent, CVC are trivial. But in cases where genitives are present, they are included in both the adjacent syllables. In the above example, in the first word ‘n A y a n a kk A r a n’, genitive /kk/ is included in both the syllables ‘nak’ and ‘kA’, to preserve coarticulation effects between /a/ & /k/ and /k/ & /A/. Similar is the case with other genitives.

Since a syllable contains only one vowel, the number of syllables identified in a word is equal to the number of vowels in it. In general, when a word starts with a vowel (say VCVCVC, Ex: u d a TT i l), the first vowel is identified as a syllable if the next to next phone is a vowel and CV syllables are identified unless there is a genitive in the sequence of phones, in which case the genitive manifests in both the adjacent syllables. In a word of the form VCCV, VC and CV syllables are identified (Ex: e n R u). If a word is of the form VCCVC, VC and CVC syllables are identified

5. ATTRIBUTES TO QUANTIFY NATURAL STRESS IN NORMATIVE SPEECH

Stress and its manifestation in the acoustic signal have been the subject of many studies. Researchers have attempted to determine the reliable indicators of stress by analyzing variables such as fundamental frequency (F0), amplitude, concentration of spectral energy, duration and others. Higher intensity, greater duration and higher fundamental frequency are believed to be the primary acoustic cues for stressed syllables, although how the three factors work together to make a syllable more prominent than the surrounding ones is still not very clear [2]. Stressed syllables are usually indicated by high sonorant energy, long syllable or vowel duration, and high and rising F0 [3]. The cues of stress mentioned above are found based on the studies carried out for stress detection of syllables in English and Dutch languages for the applications of speaker recognition or speech recognition. Further, the speech corpus analyzed, in most of the cases, was a biased one. For example, emotional speech was recorded (with happiness, anger, sadness) and used for stress analysis. It is not that only emotional speech has stress; normal speech too has some stress content in it, although it may not be very prominent. Otherwise, the pitch contour of speech signal spoken in neutral speaking style would have been flat. Much interest has not been shown in stress analysis of normal speech corpus as applied to TTS, that too in Indian languages. The present work deals with detecting such natural stress in normative speech.

For the present work, the attributes chosen to quantify stress are **Pitch, Energy and Duration**. Although another cue - amplitude was thought of, since energy is a function of amplitude, this attribute was not included. Hereafter, the combination of the above three parameters of a syllable will be called as **PED**. Syllables are stress labeled based on their PED values. “A syllable has a particular PED” means that it has P Hz of pitch, the amount of energy it has is E relative units and it exists for D seconds.

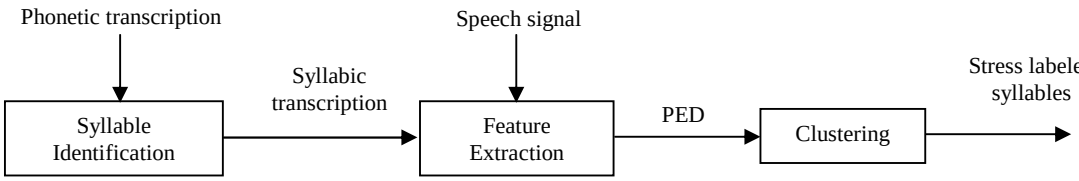

Figure 1: Outline of the Stress labeling system

6. STRESS LABELING METHOD PROPOSED

Figure 1 shows the outline of the stress determination system. Features that illuminate stress information (PEDs) are obtained for the syllables as follows. The duration (D) of each syllable is obtained from the boundaries established by the speech segmentation process. The pitch value for every 10 ms is calculated. 'P' is the maximum value among the pitch-values which fall within a syllable and 'E' is its RMS energy. From the segmented and labeled database, the number of occurrences, duration and the left and right phonetic contexts of each phone are known. Hence, phone clusters are formed (based on their duration and context). In addition, syllable clusters are also formed, which are further clustered such that the PEDs of syllables in each sub-cluster fall within a specified range. The two major clusters are:
1. with stress (relatively higher PED) 2. without stress (relatively lower PED).

6.1 Stress Determination

First, the phonetic transcription of the sentences is converted into syllabic transcription according to the syllabification strategy proposed. For each syllable, the number of occurrences in the database is found such that no syllable is a part of another syllable. For example, 'A', which is actually a phone, is a syllable in many words and 'nA' is another syllable. While searching for the instances of syllable 'A' in the database, the phone 'A' occurring as a part of the syllable 'nA' should not be included. This is taken care of by searching a syllable for its number of occurrences, in accordance with the syllabic transcription of the text corpus. For all the occurrences of a particular syllable, the PED statistics are computed and each feature (P, E and D) is normalized with respect to the variance of the feature values. For the determination of stress, each syllable is associated with a three dimensional vector, the components being its pitch, energy and duration. These vectors are plotted in three-dimensional space and clustered using k-means clustering algorithm. Figure 2 shows the clusters for the syllable 'nA', formed by k-means clustering, after variance normalization of features.

6.2 Specially recorded Stressed Corpus

The Tamil corpus being used for the TTS system is in a normal speaking style and there is not **much** emphasis given, unlike in an emotional speech. Nevertheless, some stress is naturally present in some syllables. To check the validity of the stress determination process, a new speech corpus is recorded specially from the same speaker, which can also be used for TTS along with the old corpus. The sentences in the new corpus are selected from a drama script and have lot of syllables that can be stressed. The drama is recorded in two different styles: 1. with full emotion, 2. in a neutral speaking style. Stress labeling is carried out for the syllables in both recordings. Figure 3 shows the syllable clusters formed by k-means clustering after variance normalization of features, for the syllable 'yA' in the drama corpus recorded with emotion.

6.3 Performance Evaluation – Perception experiments

To check the validity of the results of k-means clustering, listening tests are conducted. Four native Tamil people are asked to listen to the different instants of a syllable and rate them. The listeners are asked to rate the syllables with 3 levels (3: most stressed syllables, 1: non-stressed syllables and 2: medium ones). The results are compared with the results of k-means clustering. The clustering performed on the PED values normalized with respect to the variance showed better correlation than the case of normalizing PEDs with respect to the maximum value.

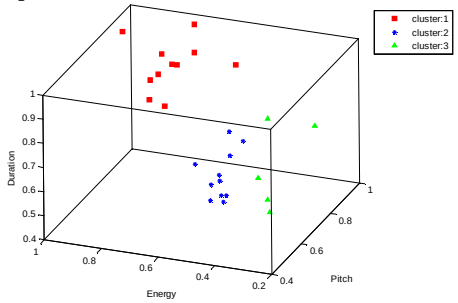

Figure 2: Result of k-means clustering on syllable 'nA'; Features – PED, Clusters: 1 - stressed, 2 - moderately stressed, 3 - unstressed.

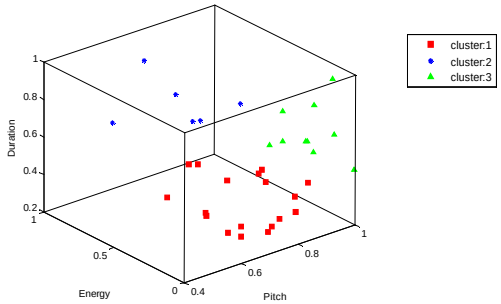

Figure 3: Result of k-means clustering on syllable 'yA' in the 'drama' corpus recorded with emotion; Features – PED, Cluster 1: unstressed, Cluster 2: moderately stressed, Cluster 3: stressed.

6.4 New labeling

It is found that many of the syllables with relatively higher values of PEDs are clustered to one class by the k-means clustering. So, a new type of clustering is performed to identify the stressed syllables. A reference vector is constructed from the maximum values of each feature across the syllables. The Euclidian distance from each syllable vector to the reference vector is computed. The syllables with distance below a threshold t_1 are designated as stressed (stress rating – 3), and the syllables with a distance above another threshold t_2 are marked as unstressed (stress rating – 1), with the remaining syllables marked as normal syllables (stress is not too high or absent; stress rating – 2).

7. PROSODY MODELING

After labeling the syllables with the stress assignment strategy proposed and comparing with the stress rating given by perception experiments, the location of a stressed syllable in the text corpus is found. Figures 4 and 5 show the syllabic transcription of some of the sentences in which the syllables 'nA' and 'rA' occur (only few words of the sentences are shown since the sentences are long, the symbol '#' separates two words). The stressed syllables, for which the rating is 3, are shown in bold black letters and the unstressed ones with the rating of 1, in bold grey color.

```
201 > nA ya nak kA ran # mel la #
220 > muk ku Ru Ni # wi nA ya har # kaR pa ha
# wi nA ya har # mAm ba zha # wi nA ya har #
mu da li ya # wi nA ya har # kO yil haL # nIng
ga lA ha # til lae # na ha ra meng gum # nart
ta na # wi nA ya har # kO wil # na ra mu ha #
wi nA ya gar #
236 > siT Tuk ku ru wi haL # nA rA #
273 > ti ruk ku RaL # nA Da hat tin #
276 > # nA ra dan # kan dar wa ku lat taec
cErn da # ku mA ra nA ha #
284 > me ylo li haL # toN Dae # nA win <
```

Figure 4: Location of stressed and unstressed 'nA' syllables in words. **nA: Stressed (rating – 3)** *nA: Unstressed (rating–1)*

It is observed that the stressed syllables found by the labeling method described above and confirmed by the listening tests occurred as the first syllable in most of the cases. In the examples shown, all the stressed syllables ('nA's) except one occurred as the first syllable. All the unstressed syllables occurred in the middle of the word. Some of the remaining occurrences of 'nA' in the sentences have a stress rating of 2 and some are part of another syllable. In the case of syllable 'rA', its frequency of occurrence as the first syllable of a word is less. As observed from Figure 5, the frequency of the second syllable of the word being stressed increases, when the first syllable is/has a short vowel (hrasva).

So, the preliminary prosody model is as follows: "The first syllable in Tamil is stressed and if the first syllable is/has a short vowel (*hrasva*), the second syllable is stressed". As we analyze the recorded speech, this may not be found always, because the speaker was specifically asked to constrain his speech to normative style, with as minimal explicit stress as possible. We believe that the strategy of selecting syllables as the basic units, selecting them according to the prosody model mentioned above and then modifying their characteristics if required might give better synthesis with computational savings in unit selection. This hypothesis needs to be verified experimentally.

```
202> sol la ga rA di # a ha ra mu da li #
219> wI Dum # aL ip pa wa rA hi ya #
220> kO wil # i rA sa # wi nA ya har #
221> i rA sa sa bae # na Da rA sap pe ru mAn #
234> pu zhak kat tiR ku wa rA da daR kA na
236> paL Lik kuc cen Ra # nA rA # 258> sA laeg
rA mam # a ra wa mang ga lam
269> i wae # gi rA mi ya mA na wae #
279> in da # rA gam # ka ru nI la #
280> i da nae # ae da rA bAd #
285> kan na Da # nAT Til # rA ma #
289> bi rA ma Nar ka Luk ku <
293> pEr # bi rA ma Nar haL <
```

Figure 5: Location of stressed and unstressed 'rA' syllables in words. **rA: Stressed (rating–3)** *rA: Unstressed (rating–1)*

8. CONCLUSION

A new method for dynamically choosing the syllable structure, is proposed for Tamil language. The proposed syllabification strategy preserves the coarticulation effects of all the phones present in the identified syllables. A method of Stress labeling of syllables based on their PEDs is proposed. Stress labeling of syllables in the Tamil corpus is performed. Based on the location of stressed syllables in a word, a rudimentary prosody model has been developed for Tamil TTS. The prosody model says that "first syllable in Tamil is stressed in general and the second syllable is stressed if the first syllable is/has a short vowel. This confirms the earlier studies of some linguists. The fact has also been verified acoustically.

9. REFERENCES

[1] Min Lai, Yining Chen, Min Chu, Yong Zhao, Fangyu Hu, "A hierarchical approach to automatic stress detection in English sentences", Proc. *ICASSP* 2006, pp 753-756.

[2] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. on Speech and Audio Processing*, 2(4), pp. 469-481, 1994.

[3] Chao Wang and Stephanie Seneff, "Lexical Stress Modeling for Improved Speech Recognition of Spontaneous Telephone Speech in the JUPITER Domain1", *EUROSPEECH 2001*, Sept 2–7, 2001, Aarlborg, Denmark.

[4] Kopecek, I., Pala, K. "Prosody Modelling for Syllable- Based Speech Synthesis", *Proc. IASTED Conf. AI and Soft Computing*, 1998, pp. 134-137.

[5] Cairns, Douglas A.; Hansen, John H. L, "Nonlinear analysis and classification of speech under stressed conditions", *JASA*, Vol 96 (6), pp.3392-3400, Dec 1994.