A SYLLABLE BASED STATISTICAL TEXT TO SPEECH SYSTEM

Abhijit Pradhan, Aswin Shanmugam S, Anusha Prakash, Kamakoti Veezhinathan, Hema Murthy

IIT Madras, Chennai, India

ABSTRACT

A statistical parametric speech synthesis system uses tri-phones, penta-phones or full context phones to address the problem of co-articulation. In this paper, syllables are used as the basic units in the parametric synthesiser. Conventionally full context phones in a Hidden Markov Model (HMM) based speech synthesis framework are modeled with a fixed number of states. This is because each phoneme corresponds to a single indivisible sound. On the other hand a syllable is made up of a sequence of one or more sounds. To accommodate this variation, a variable number of states are used to model a syllable. Although a variable number of states are required to model syllables, a syllable captures co-articulation well since it is the smallest production unit. A syllable based speech synthesis system therefore does not require a well designed question set. The total number of syllables in a language is quite high and all of them cannot be modeled. To address this issue, a fallback unit is modeled instead. The quality of the proposed system is comparable to that of the phoneme based system in terms of DMOS and WER.

Index Terms— TTS, Syllable, HTS, Statistical TTS

1. INTRODUCTION

State of the art text-to-speech (TTS) synthesis systems are based on the Unit Selection Synthesis (USS) or Hidden Markov model based Statistical Parametric Speech Synthesis (SPSS). Speech produced by USS possesses natural quality but has prominent discontinuities at boundaries between segments [1]. In general, SPSS systems produce lower quality speech compared to that of USS systems. Speech produced by SPSS systems sometimes sounds muffled due to oversmoothing [2]. SPSS systems, however, have the advantage of being small footprint and they produce fairly continuous synthetic speech [2]. SPSS systems are well suited for small devices: like cellphones and tablets etc., where a small footprint is necessary. In conventional SPSS systems HMMs are built for phonemes of a language. These models are used in a generative framework to synthesise speech [3].

To address the issue of co-articulation mono phones are replaced by full context phones. The use of full context phones helps to preserve sufficient boundary information so that the transition from and to adjacent phones appears smooth. In this paper, syllables are proposed as an alternative to context dependent phones. The motivation behind choosing syllable as a unit is described in Section 2.

There are two problems associated with the modeling and use of full context phones: 1) Insufficient training data 2) Unseen full context model. Decision tree based state clustering solves both the problems by applying a Question Set on the seen phones [4]. A leaf node in a decision tree represents a cluster of similar states. All the states in a cluster are tied and trained together to overcome the problem of limited training data. The same decision-trees are used to make robust estimates of unseen full context models [4]. The quality of full context phone based SPSS depends on the choice of the Question Set used. These questions are relevant linguistic and phonetic classifications which influence the acoustic realisation of a phone. The Question Set preparation requires the knowledge of the acoustic phonetics of the sound units produced in a given language. For example, the short and long vowels for /a/, /i/ and /u/ are present in all the Indian languages, while Tamil, Malayalam, Kannada and Telugu also have long vowels for /e/ and /o/. There are some sounds that are unique to some languages. The retroflex /zh/ is present only in Tamil and Malayalam. Tamil does not have any aspirated consonants. The palatals /c/, /ch/, /j/, /jh/ are all aspirated consonants. The palatals /c/, /ch/, /j/, /jh/ are all affricates in Indian languages. Thus preparing the Question set for a new language is a nontrivial task. In the proposed system only context independent syllable models are built. The tree based context clustering is not performed.

In the absence of a decision tree, alternatives have to be provided for unseen syllables. To address this issue, we use a set of fallback units (see Subsection 4.1.3). In this work, the HTS based framework is used as the basic platform. HTS uses full-context (2 left + 2 right + 48 additional contexts) phones as units for synthesis. The parameter generation and synthesis are the same as that of full context phone based synthesis using HTS. The training process is slightly different in order to accommodate variable length syllable models.

This paper is organized as follows. In Section 2 we explain the motivation behind using a syllable as a unit of synthesis. In Section 3 we describe, in brief, a phone based HTS. In Section 4, we present the proposed syllable based HTS system, highlighting the differences between this system and the

\footnote{HTS [5] is an open source SPSS system/software.}
phone based system. In Section 5, we discuss the experiments and results. And finally, in Section 6 we summarise the fallouts of this proposed approach in the context of Indian languages.

2. IMPORTANCE OF A SYLLABLE AS A UNIT

The basic linguistic unit with a correspondence to a sound unit is the phoneme. Although the phoneme is a distinct acoustic unit, the basic production and cognition unit is a syllable [6, 7, 8, 9]. A syllable can be represented by a C*VC* where C is a consonant and V is a vowel. A syllable captures intra-syllable co-articulation between phonemes well. Although co-articulation is present across syllables, the co-articulation is less owing to the acoustic energy at the boundary between syllables being significantly lower [10] than that at the middle of a syllable.

Our brain learns and stores a dictionary of syllables. Whenever a new word is encountered, it is split into syllable-like units (syllable priming) based on some context [8, 9]. For example, /protect/ is syllabified as /pro/-/ect/, while /pro-tected/ is syllabified as /pro/-/ect/-/ed/. Clearly, the end /t/ in /ect/ is acoustically quite different from that of /t/ in /ed/. Experiments conducted in [11] indicate that the knowledge of the writing system is essential to determine and identify phones or sounds. A number of efforts exist in the literature where an attempt is made to use syllable as a basic unit. Especially in the context of Indian languages, it has been shown that syllable can be used for both synthesis and recognition [12]. Although the role of syllables in speech segmentation cannot be questioned [11], appropriate acoustic cues are required to segment the speech signal at the syllable level. In [13], it is shown that an utterance can be quite accurately segmented at the syllable level using acoustic cues. The body of work presented in this paper exploits this.

3. PHONE BASED HTS

HTS generates synthetic speech from HMMs by using a speech parameter generation algorithm [3] and a source-filter based speech production model. For each phone, two types of HMMs are created. Spectrum and pitch together are modeled as a multi-stream HMM. We henceforth refer to this as composite model. The duration is modeled using another HMM. We henceforth refer to this as duration model. The feature vector in a composite model has two streams: one for spectral parameter vector and another for pitch parameter vector [14]. Each duration model is represented by a multi-dimensional Gaussian distribution. The dimension of the duration density is $N$, where $N$ is equal to the number of states in the corresponding composite model and the $n$-th dimension corresponds to the $n$-th state in the composite model [15]. Both the HMM models are fullcontext phone models.

3.1. Training

A detailed phone training process can be found in [16]. A sequence of steps in the training process is:

(1) Monophone composite models are built.

(2) Fullcontext composite and duration models are built from monophone models.

(3) A context tree is built from these composite models using a Question set.

3.2. Synthesis

The training produces fullcontext composite and fullcontext duration models. The synthesis involves the following sequence of steps:

(1) Create fullcontext phone labels from input text.

(2) For an unseen model find the closest model from the context tree.

(3) From duration models get the state sequence.

(4) Form a sentence/utterance from the state sequence and full context composite models.

(5) Apply parameter generation algorithm [3] to get the speech observation vector that maximizes the output probability.

4. SYLLABLE BASED HTS

4.1. Training

We now discuss the development of a syllable based HTS. When building a syllable based HTS, three different issues must be addressed, namely, mapping from text to syllables, building syllable level HMMs and fallback when a syllable is not found in the database. Each of these tasks are described in this subsection.

4.1.1. Letter To Sound Rules (Parser)

In building a TTS system the first task is to establish a mapping between the written form and the spoken form. A parser is the component that performs the mapping. The conversion process is heavily language specific. For instance, in English grapheme to phoneme conversion cannot be performed using a set of rules. Instead the conversions are learnt from data or a dictionary. In Indian languages too, the conversion is language specific. A finite number of rules that are independent of the type of word are required [17, 18]. We briefly describe two rules one for each of the languages: Hindi and Tamil. In Hindi one such rule is schwa deletion. In this rule, for syllables ending with a unit of the CV type, the V part is replaced with a unit called halant. For example, the word चिनतन /chinatna/ without schwa deletion would be syllabified as चिन /china/ and तन /tan/, but with schwa deletion it is syllabified as चिन /chin/ and तन /tan/. In Tamil one such rule is voiced-unvoiced classification. The voiced and
unvoiced stop consonants have the same orthographic representation. The Tamil character ச can be pronounced as both /ka/ (UnVoiced) and /ga/ (Voiced). It is classified as UnVoiced when the character occurs at the beginning of a word or appears as a geminate; and classified as Voiced when it occurs in between the vowels or after the nasals (ந ம/ந ம ( N/a Na Na na na ma) or after the consonants ( ந ம/ந ம ( N/a Na Na na na ma) ) (l/la/ l/la/ l/la/ l/la/ /zh/). For example, in the word ந ம/ந ம (N/a Na Na na na ma), the character occurs as a geminate and thus is UnVoiced. In word ந ம/ந ம (an/gam), the character ச occurs after the nasal ம (N/a) and thus is Voiced.

4.1.2. Building Syllable Models

A semiautomatic labeling tool is used to obtain labels at the syllable level. Group delay segmentation [19] is used to find the syllable boundaries. Figure 1 illustrates the process to obtain syllable boundaries using an example TIMIT sentence (sa1): “She had your dark suit in greasy wash water all year”. The top panel shows the signal waveform, the middle panel shows the syllable transcription, and the bottom panel shows the group delay function. The peaks in the group delay function correspond to syllable boundaries. The boundaries are highlighted by the vertical lines in the figure. Observe that the group delay function detects most boundaries accurately, except for the insertion in /suwt/. The extra boundary is manually deleted to obtain accurate syllable boundaries.

In a phone based HTS, the phones are the basic units of synthesis. They are small discrete units of a speech stream. Even though a phone (e.g a vowel) can be longer than most other phones; it contains repetition of only a single sound. Therefore all phones can be modelled by an HMM that has fixed number of states. A 5 state HMM is generally used for phone models. A syllable contains from a single sound (e.g. a vowel) to a sequence of sounds. Thus a variable number of states are required to build syllable models. Given the transcription of a syllable into its constituent phonemes, the number of states for each syllable can be obtained. The number of states in the corresponding HMM is number of phones \* 5. Although syllables capture co-articulation well, the acoustic realisation of a syllable depends upon its position in a word. The prosody does differ significantly based on the syllable’s position in a word. Based on experiments with syllable-based USS for Indian languages positional context seems to improve USS performance [20]. Syllables are classified into three categories: begin, middle, and end based on their positions in a word. BEG,MID,END are added as suffixes to the text transcriptions and are modelled as separate HMMs. Once the positional context and the number of states for a syllable are known an appropriate HMM prototype is chosen to build the syllable models. Thus for a language having N syllables there can be at most 3N composite HMMs and 3N duration HMMs. Unseen syllables are modelled using fallback units. This is discussed in the next section.

4.1.3. Building Fallback Models

Since it is not practical to model the complete set of syllables in a language, during synthesis we encounter a syllable that is not in the train database. The unseen syllable is mapped to a sequence of \{(C*V) \cup \emptyset\} (where “C” stands for consonant and “V” for vowel) units. A fallback unit is a variable length unit like a syllable. Fallback unit models are built with the same positional context as syllables. An identical training process to that of syllables is adopted for fallback units.

![Fig. 1. Group Delay based Labeling](image-url)

<table>
<thead>
<tr>
<th>Syllable Label</th>
<th>Fallback Unit Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beg</td>
<td>End</td>
</tr>
<tr>
<td>0.000</td>
<td>0.196</td>
</tr>
<tr>
<td>0.196</td>
<td>0.464</td>
</tr>
<tr>
<td>0.271</td>
<td>0.464</td>
</tr>
<tr>
<td>0.464</td>
<td>0.700</td>
</tr>
<tr>
<td>0.549</td>
<td>0.700</td>
</tr>
<tr>
<td>0.700</td>
<td>1.152</td>
</tr>
<tr>
<td>1.025</td>
<td>1.152</td>
</tr>
</tbody>
</table>

Table 1. Fallback unit labels from syllable labels

The process of obtaining fallback unit boundaries is explained here. Table 1 shows a sequence of known syllable-boundaries and the boundaries of the basic fallback units that make up the syllable. Utterance waveforms are spliced into smaller waveforms such that a waveform consists of only a single syllable. Embedded re-estimation at the syllable-level is iteratively performed to build fallback unit models. Boundaries are generated for fallback units. As an example, for the syllable சbeg /pa/ (BeginTime=0.549 End-Time=0.700) the parser gives two fallback units: சbeg /pa/ and சend /hl/. Forced alignment gives boundaries of the fallback units. The fallback boundaries are obtained as: சbeg
In Figure 2, the top panel shows the fallback unit boundaries generated using flat start training and the bottom panel shows boundaries generated using syllable boundaries as described above. Observe that the syllable boundaries obtained using flat start are quite inaccurate, in that the end of the first syllable /in/ becomes part of the second syllable /caar/. Owing to the trill /rl/, the right edge of the boundary for /laa/ is not accurate in both systems, while the left edge for /ce/ is accurate in the syllable-based system. The right edge of /rl/ is correct in both systems. The boundary for a phone in the syllable-based system is contained within the syllable itself. This is primarily because the boundaries for syllable are obtained using event-based signal processing. From this example, it is clear that in the syllable-based approach the boundary of a phoneme is accurate when it occurs at the boundary of a syllable. It was therefore felt, that even for fallback units positional information in terms of begin, middle and end should be adequate.

4.2. Synthesis

The input text is parsed by a language-specific parser. The parser outputs syllable-level transcriptions of the text. The syllable is searched in the train syllable list. If a syllable is not found in the train syllable list then the syllable is split into fallback units. Since the complete set of fallback units are trained, the fallback unit models corresponding to the missing syllable are certain to be found. The text syllables and fallback unit sequences generated by the text parser are input to the parameter generation algorithm. The generation algorithm concatenates the syllable HMMs, fallback unit HMMs and builds a sentence HMM. The sentence HMM is used to generate the speech vectors.

5. EXPERIMENTS AND RESULTS

1643 Tamil sentences and 945 Hindi sentences are used for our experiments. All sentences are spoken by a single native female speaker of the respective languages. They are recorded in a studio environment at 16000Hz, 16bits per sample. After parsing these sentences and removing syllables with very less number of examples (syllables with less than 5 examples are removed), 984 and 1087 syllables (with positional contexts) for Tamil and Hindi are obtained, respectively. For the experiments in this paper HTK-3.4+HTS-2.1 is used and the Global Variance (GV) enhancement [21] is enabled. Degradation MOS [22] is performed (20 listeners for each language) to get a measure of the naturalness of the synthesised speech. For intelligibility, semantically unpredictable sentences are transcribed and Word error rate (WER) is calculated. Results are shown in Table 2. DMOS scores indicate that syllable based HTS systems are almost as good as phone-based HTS systems. However WER scores indicate that phone-based HTS systems are slightly better in overall quality than syllable based HTS systems. A set of Tamil and Hindi speech waveform files synthesized by the proposed system is available at https://sites.google.com/site/abhijitttshts/. A live synthesiser based on syllables can be found at http://www.iitm.ac.in/donlab/hts/.

6. CONCLUSION

In this paper we propose syllables as an alternative to full-context-phones for text-to-speech synthesis, in the HTS framework. One advantage of the syllable-based TTS is that the question set for decision-tree based context clustering is not required and thus it is easy to train the TTS for a new language. The intelligibility of synthetic speech produced by the syllable-based TTS is almost as good as the full-context-phone based TTS. Although the phone-based TTS system performs marginally better than the syllable-based HTS system, building the syllable-based system is quite simple. The results obtained in this paper do indicate that an SPSS for any new language can be built using the grapheme-to-phoneme (syllabification of text) from syllable to sound unit. In particular, such a system is relevant for syllable-timed languages.
7. REFERENCES


[5] “HMM-based speech synthesis system (HTS),” http://hts.sp.nitech.ac.jp/, [Last online access by authors; 01-Feb-2013].


[16] “HTS tutorial at interspeech, 2009,” http://www.sp.nitech.ac.jp/~tokuda/tokuda_interspeech09_tutorial.pdf, [Last online access by authors; 01-Feb-2013].


