Syllable based models for prosody modeling in HMM based speech synthesis

Srikanth Ronanki[1], *Oliver Watts*[2], *Simon King*[2], *Rob Clark*[2]

[1]Speech and Vision Lab, IIIT Hyderabad, India
[2]Center for Speech Technology Research, University of Edinburgh, Edinburgh
srikanth.ronanki@research.iiit.ac.in, robert@cstr.ed.ac.uk

Abstract

Simple4All is a speech synthesis research project that aims to ease the production of synthetic voices in new languages by means of unsupervised modeling techniques. In this work, we introduce syllable based models for prosody modeling in Hidden Markov Model based Text-to-Speech system (HTS). As a part of investigating the potential for building speech synthesis systems in new languages with little data, we are investigating alternate formulations for the pitch and duration models within HMM based speech synthesis frame-work, specifically investigating models that explicitly model prosody for named syllabic contexts. A comparison between phone and syllable methods demonstrating the differences in spectral and prosody features was carried out. In the end, a hybrid system with spectral features from phoneme models and prosody features from syllable models has been designed to synthesize speech with robust quality and naturalness.

Index Terms: Speech synthesis, Prosody Modeling, Hidden Markov Models.

1. Introduction

In early days of Text-to-Speech (TTS) research [1], researchers mainly focussed on parametric synthesis techniques, where the parameters are determined using rules designed by experts. Most of these approaches exploits one of the three basic technologies: articulatory, formant based phonemic synthesis and Linear Predcition Coefficient (LPC) based concatenative synthesis [2, 3, 4]. These rule-based parametric synthesis techniques, though smooth sounding, lack naturalness. Hence data-driven systems which provide comparatively more natural sounding speech have received research focus in the last decade. Data-driven synthesis, also referred to as corpus-based synthesis, makes it possible to dramatically improve the naturalness of speech over rule based methods. There are two types of prominently used data-driven synthesis techniques; the statistical parametric synthesis technique [5, 6] and the unit selection synthesis technique [7].

Unit selection synthesis has shown itself to be capable of producing high quality natural sounding synthetic speech when constructed from large databases of well-recorded, well-labeled speech. However, the cost in time and expertise of building such voices is still too expensive and specialized to be able to build individual voices for everyone. As an alternative, statistical parametric speech synthesis with HMMs has been widely used and is particularly well known as HMM-based speech synthesis(HTS) [8, 10]. In the HMM-based speech synthesis, the speech parameters of a speech unit such as the spectrum, fundamental frequency (F0), and duration are statistically modeled and generated by using HMMs based on maximum likelihood criterion [9, 10, 11].

The use of syllables rather than phonemes for TTS navigates the output towards more natural in terms of quality. Since using longer units provides prosodic and acoustic variability found in natural speech, the synthesized speech quality is enhanced. However, unlike phonetic symbols, syllables don't have a widely accepted symbol. The naming of syllables and labeling it is quite different in various languages. Majority of the algorithms based on syllable modeling in literature concatenates phones in the form of C*VC* to form a syllable unit. The number of such unique syllables may vary from few hundreds to thousands depending on the training database, language and its corresponding phoneset. Also, synthesis of missing units [12] which are not covered in the training data has to be handled carefully. In general, a simple rule based back-off technique is used in such cases and it varies from one language to another.

Hence, in our work, we propose an unsupervised algorithm to form named segments from syllables using the linguistic and syntactic features of the corresponding syllable and thereby modeling the prosody explicitly using syllable HMMs. For missing units, the back-off strategy is designed to find the nearest optimal unit based on Levenshtein distance. The use of syllable HMMs to estimate prosodic features and phone HMMs for spectral features has motivated us to design a hybrid system combining both by dynamic time-warping.

The rest of the paper is organized as follows: Section 2 gives an overview of HMM based speech synthesis using baseline system. Section 3 forms the basis for generating speech features using syllable HMMs and back-off strategy. The experiments along with comparative results are evaluated in section 4. A new kind of hybrid algorithm using dynamic time warping is presented in section 5 and finally conclusions are presented in section 6.

2. Baseline system for HMM based TTS

2.1. Database

We experimented with English and two major Indian languages. EMMA by Jane Austen, an audio book from librivox [13] is considered for building voice in English. For Indian languages, Indic databases [14] have been used for synthesis experiments. For training HMM models, one hour of data is considered from each of the languages.

2.2. HMM based synthesis using phoneme as basic unit

HMM-based speech synthesis consists of a training and synthesis phase. In the training phase, spectral parameters, namely, Mel generalized cepstral coefficients (mgc) and their dynamic features, the excitation parameters, namely, the log fundamental

Table 1: Formation of segments

Words	Syllables	Feature representation	Segments
All	(aol)	(020 011101)	Seg413
librivox	(lax, brax, vaaks)	(102 000100, 102 000000, 112 001011)	Seg708, Seg704, Seg907
recordings	(rax, kaor, daxngz)	(102 000100, 121 000011, 122 101010)	Seg708, Seg1027, Seg1130
are	(aar)	(220 011101)	Seg1565
in	(ihn)	(320 001100)	Seg2124
the	(dhax)	(002 001100)	Seg140
public	(pah, blaxk)	(101 000111, 112 001000)	Seg647, Seg904
domain	(dow, meyn)	(102 000100, 122 101011)	Seg708, Seg1131

frequency (lf0) and its dynamic features, are extracted from the speech data. Using these features and the time-aligned phonetic transcriptions, context independent monophone HMMs are trained. By default, the basic subword unit considered for the HMM-based system is the context-dependent pentaphone for most of the languages. During building, the context-dependent models are initialized with a set of context independent monophone HMMs. A sequence of steps based on the question set, is used for state-tying.

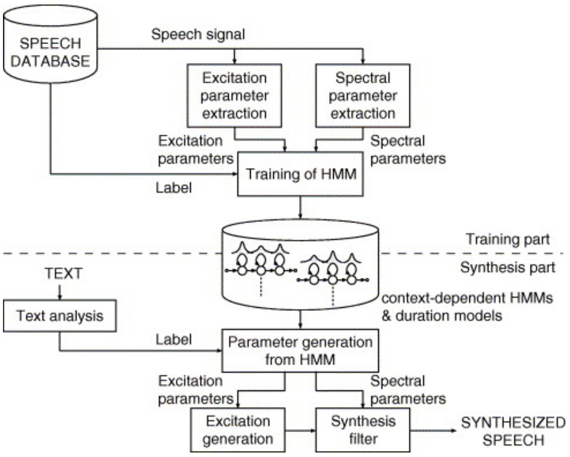

Figure 1: Overview of HMM based Speech Synthesis System

Then, the decision-tree-based context clustering technique [15], [16] is applied separately to the spectral and logF0 parts of the context-dependent phoneme HMMs. In the clustering technique, a decision tree is automatically constructed based on the MDL criterion. We then perform re-estimation processes of the clustered context-dependent phoneme HMMs using the BaumWelch (EM) algorithm. Finally, state durations are modeled by a multivariate Gaussian distribution [17], and the same state clustering technique is applied to the state duration models

In the synthesis phase, first, an arbitrarily given text is transformed into a sequence of context-dependent phoneme labels. Based on the label sequence, a sentence HMM is constructed by concatenating context-dependent HMMs. From the sentence HMM, spectral and F0 parameter sequences are obtained based on the ML criterion [9] in which durations are determined using state duration distributions. Finally, by using an Mel Log Spectral Approximation (MLSA) filter [18, 19] or STRAIGHT vocoder [20], speech is synthesized from the generated mel-cepstral and F0 parameter sequences. For Indian languages, except for the text and speech data, which are language-dependent, the rest of the modules can be made language-independent by preparing a common phone set and common question set. [24, 14] The whole process is illustrated in Fig. 1.

3. HMM based synthesis using syllable as basic unit

3.1. Formation of named segments from syllables

The full-contextual phoneme label file contains the linguistic and syntactic features which are derived from text using Festival [21]. Using the key information such as position of the phone in the syllable and position of the syllable in the word, we concatenate the phones in the form of C*VC* to form a syllable unit [25]. Later, we clustered these syllables into named segments as shown in table 1 using the feature notation from table 2. Feature representation in table 1 lists the primary and secondary features from table 2, with a space in-between, from right to left.

Table 2: *Features and their representation*

Primary features		Secondary features	
Features	Notation	Features	Notation
stress	0/1	onset	0/1/2
accent	0/1	coda	0/1/2
word initial	0/1	parts of speech	0-8
word final	0/1		
phrase initial	0/1		
phrase final	0/1		

As part of initial experiments, the syllables are clustered into named segments using the primary features as shown in table 2. The reason for such a clustering is to derive the models which can do better prosody modeling when compared to phoneme based system. But, from table-1, it is observed that words “in” and “the” both having same set of primary features and hence can be categorized into one segment. However, both words differ a lot in duration and their pitch contours. Hence, we considered secondary features such as onset and coda. We have also added Parts of Speech (POS) as well for better clustering. For onset (0/1/2), zero represents non-existent, one represents voiced and two represents un-voiced and same in the case of coda as well. Nine different kinds of POS are tagged for English.

3.2. Training

Syllable HMMs are trained using the same baseline framework with few changes in label file and question file. The label file contains named segments with same number of contextual features as phoneme. For context clustering, questions based on previous to previous of the current segment, next to the next of current segment are ignored. Since, the total number of unique segments are more in number compared to total number of phones, the top 200 segments in the sorting order of their frequency in the training data are considered for clustering with current segment followed by any other. For rest of the segments, questions which are relevant to syllables are used for context clustering.

3.3. Back-off technique

Since, the syllables are huge in number and can't be covered in the training data, an appropriate back-off strategy is essential for synthesis. Since, the named segments are independent of language, a common strategy can be used to find the nearest optimal segment. The Levenshtein distance is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (insertion, deletion, substitution) required to change one word into the other. Here, in our case, we used the Levenshtein distance [22] to measure the distance between two digital strings to find the nearest optimal segment. However, word initial and word final features are given heighest weighted preference to find a similar segment with rest of the features same.

4. Experimentation and Results

4.1. Evaluation of spectral features

Mel Cepstral Distortion (MCD) is an objective error measure, which is known to have correlation with the subjective test results [23]. Thus, MCD is used to measure the deviation of estimated Mel-generalised cepstral features from original. MCD is essentially a weighted Euclidean distance defined as:

$$MCD = (\frac{10}{\log 10}) * \sqrt{2 * \sum_{i=1}^{40} (mc_i^e - mc_i^o)^2} \qquad (1)$$

where mc_i^o and mc_i^e denote the original and the estimated Mel-generalised cepstral features, respectively.

Table 3: *MCD measure for both the systems*

Language	Original Vs Phoneme	Original Vs Syllable
English	5.57	7.57
Telugu	5.85	6.6
Kannada	5.45	6.31

4.2. Evaluation of prosodic features

Root Mean Square Error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed and is defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i^2 - \hat{y_i}^2)}{n}} \qquad (2)$$

where y_i and $\hat{y_i}$ denotes original and predicted f0 contours respectively.

Linear Correlation Coefficient, measures the strength and the direction of a linear relationship between two variables.

$$CORR = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} \qquad (3)$$

where x and y denotes original and predicted f0 contours respectively.

If x and y have a strong positive linear correlation, CORR is close to +1. A CORR value of exactly +1 indicates a perfect positive fit. If x and y have a strong negative linear correlation, CORR is close to -1. A CORR value of exactly -1 indicates a perfect negative fit.

Table 4: *F0 measures for both the systems*

Voices	RMSE		CORR	
	Phoneme	Syllable	Phoneme	Syllable
English	24.82	23.06	0.78	0.87
Telugu	19.04	15.84	0.81	0.88
Kannada	18.56	14.72	0.80	0.88

Table 4 objectively compares the contours generated by phone model and the proposed syllable model in terms of mean error and correlation. From both the metrics, it has been observed that the estimation of prosodic features is performed to be good from syllable HMMs compared to phone HMMs. It demonstrates that inclusion of syntactic and linguistic features add prosody to the speech synthesis. From MCD scores in table 3, it has been observed that syllable HMMs are not that good at estimating spectral features, since the number of unique units are more in number and hence less number of training examples for each syllable. However, there is something pecular about syllable based system which underlines the RMSE scores of f0 and motivated us to build a hybrid system making use of both the models.

5. Hybrid synthesis using both phone and syllable models

As described in the above sections, we have built two Text-To-Speech systems with one system having models trained on sub-word unit phoneme where as the other system trained on sub-word unit as syllable. From Table 3, the phoneme based system is performed to be good at estimating spectral features since the number of training examples for each sub-word unit is more when compared to syllable. Therefore, we propose a hybrid system which makes use of phoneme based HMM for estimating spectral features and syllable based HMM for estimating prosodic features using dynamic time warping (DTW).

5.1. Dynamic Programming

Let $X = \{\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(M)\}$ and $Y = \{\boldsymbol{y}(1), \boldsymbol{y}(2), \ldots, \boldsymbol{y}(T)\}$ be two observed feature vectors. The dynamic programming algorithm aligns the feature vector Y with feature vector X. The result is the stretched or shrunk signal $X' = \{\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(T)\}$. The algorithm to compute X' is as explained below (This is explained in the probability-like domain, as apposed to tradition Euclidean distance domain) [26, 27].

Let $1 \leq j \leq M$, $1 \leq i \leq M$, and $1 \leq t \leq T$. Let us define $\alpha_t(j)$ as a cost or score incurred to align j^{th} feature of X with t^{th} feature vector of Y.

The $\alpha_t(j)$ could be computed frame-by-frame using the recursive Viterbi equation

$$\alpha_t(j) = \max_i \{\alpha_{t-1}(i) a_{i,j}\} P(\boldsymbol{y}(t), \boldsymbol{x}(j)), \tag{4}$$

where $P(\boldsymbol{y}(t), \boldsymbol{x}(j)) = exp(\|\boldsymbol{y}(t) - \boldsymbol{x}(j)\|^2)$, and $\|.\|^2$ represents the Euclidean distance between two feature vectors. Here $i = \{j, j-1, j-2\}$. The value of $a_{i,j} = 1$, thus making all paths (including non-diagonal) leading from $(i, t-1)$ to (j, t) are given uniform weightage.

The value $P(\boldsymbol{y}(t), \boldsymbol{x}(j))$ is typically less than 1. For large values of t, $\alpha_t(.)$ tends exponentially to zero, and its computation exceeds the precision range of any machine. Hence $\alpha_t(.)$ is scaled with term $\frac{1}{\max_i\{\alpha_t(i)\}}$, at every time instant t. This normalization ensures that values of $\alpha_t(.)$ are between 0 and 1 at time t.

Given $\alpha_t(.)$, a backtracking algorithm is used to find the best alignment path. In order to backtrack, an addition variable ϕ is used to store the path as follows.

$$\phi_t(j) = arg \max_i \{\alpha_{t-1}(i) a_{i,j}\} \tag{5}$$

where $\phi_t(j)$ denotes the frame number (index of the feature vector) at time $(t-1)$ which provides an optimal path to reach state j at time t.

5.2. Best path

Given values of $\phi_t(.)$, a typical backtracking done to obtain the best path is as follows:

$$y(T) = N \tag{6}$$

$$y(t) = \phi_{t+1}(y(t+1)), \; t = T-1, T-2, \ldots, 1. \tag{7}$$

5.3. Synthesis

To synthesize speech from text using such a hybrid system requires Dynamic Time Warping (DTW). The algorithm aligns the spectral features from source (phoneme) to that of target (syllable). Such an alignment gives the trajectory of phoneme aligned against syllable with respect to time. The prosodic features such as f0 and duration are preserved from syllable models. The F0 contour estimated from syllable based HMM along with spectral features obtained from the DTW alignment are given as input to either MLSA filter or STRAIGHT vocoder to synthesize speech.

5.4. Subjective Evaluation

We need to compare the baseline phoneme based HTS system with hybrid system to determine if there has been any improvement. The two important qualities that are considered to be important in synthesized speech are naturalness, intelligibility. We evaluated both subjective tests on hybrid Text-To-Speech system. Subjective tests involve humans listen to examples of speech and rate them.

We conducted a Mean Opinion Scoring (MOS) test to evaluate the performance of the each system against the original. A total of 10 subjects were asked to participate in the two experiments. Each subject was asked to listen to 5 utterances corresponding to one of the experiments. In the MOS test, listeners evaluated speech quality of the converted voices using a 5-point scale. (5: excellent, 4: good, 3:fair, 2: poor, 1: bad)

Table 5: *MOS measure for both the systems*

Language	Phone based HTS	Hybrid HTS
English	4.2	4.3
Telugu	4.2	4.35
Kannada	4.1	4.3

6. Conclusions

It has been shown that the HMMs trained on syllable units performed to be good at estimating prosodic features. The formation of segments from syllables ease the synthesis of missing units and made the algorithm unsupervised. Our future work concentrates on building more robust TTS systems with cross-lingual acoustic models and speaker adaptation.

7. Acknowledgements

The research leading to these results was partly funded from the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement Simple4All.

8. References

[1] D. H. Klatt, “Review of text-to-speech conversion for English”, Journal of the Acoustical Society of America, vol. 82, pp. 737-793, 1987.

[2] A.R. Greenwood, “Articulatory Speech Synthesis Using Diphone Units”, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1635-1638, 1997.

[3] D. Klatt, “Software for a cascade/parallel formant synthesizer”, Journal of the Acoustical Society of America, vol. 67, no. 3, pp. 971-995, 1980.

[4] R. Carlson and B. Granstrom, “A text-to-speech system based entirely on rules”, in Proceedings of ICASSP, vol. 1, Apr. 1976, pp. 686 - 688.

[5] Al.W. Black and H. Zen and K. Tokuda, “Statistical parametric speech synthesis”. In Proc. ICASSP. pp. 1229-1232, 2007.

[6] H. Zen and T. Toda, “An overview of Nitech HMM based speech synthesis system for Blizzard Challenge 2005”. In Proc. Blizzard Workshop, 2005.

[7] Hunt, A.J., Black, A.W., “Unit selection in a concatenative speech synthesis system using a large speech database,” Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on , vol.1, no., pp.373-376 vol.1, 1996.

[8] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. “Simultaneous modeling of spectrum, pitch and duration in HMMbased speech synthesis”. In Proc. EUROSPEECH-99, pages 2374-2350,September 1999.

[9] K. Tokuda, T. Kobayashi, and S. Imai. “Speech parameter generation from HMM using dynamic features”. InProc. ICASSP-95, pages 660-663,May 1995.

[10] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai. “Speech synthesis using HMMs with dynamic features”. In Proc. ICASSP-96, pages 389-392, May 1996.

[11] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. “Speech parameter generation algorigthms for HMM-based speech synthesis”. In Proc. ICASSP 2000, pages 1315-1318, June 2000.

[12] Vijayaditya Peddinti. “Synthesis of missing units in a Telugu text-to-speech system”, MS thesis, May 2011.

[13] http://librivox.org

[14] Kishore Prahallad, E. Naresh Kumar, Venkatesh Keri, S. Rajendran and Alan W Black “The IIIT-H Indic Speech Databases”, in Proceedings of Interspeech 2012, Portland, Oregon, USA.

[15] K. Shinoda and T. Watanabe. "MDL-based context-dependent subword modeling for speech recognition". J. Acoust. Soc. Japan (E), 21:79-86,March 2000.

[16] S.J. Young, J.J. Odell, and P.C. Woodland. "Tree-based state tying for high accuracy acoustic modeling". In Proc. ARPA Human Language Technology Workshop, pages 307-312, March 1994.

[17] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. "Duration modeling for HMM-based speech synthesis". In Proc. ICSLP-98, pages 29-32, December 1998.

[18] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," IECE Trans. A, vol.J66-A, no.2, pp.122-129, Feb. 1983 (in Japanese).

[19] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai. "An adaptive algorithm for mel-cepstral analysis of speech". In Proc. ICASSP-92, pages 137-140, March 1992.

[20] Hideki Kawahara, Ikuyo Masuda-Katsuse and Alain de Cheveign. "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds". Speech Communication, pages 187 - 207, 1999.

[21] A.W. Black, P. Taylor, R. Caley, "The Festival Speech Synthesis System", University of Edinburgh, 1999.

[22] Yujian, L.; Bo, L. "A normalized levenshtein distance metric". IEEE Trans. Pattern Anal. Mach. Intell. 2007, 29, 1091-1095

[23] Tomoki Toda, Alan W Black, and Keiichi Tokuda, " Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in 5th ISCA Speech Synthesis Workshop, 2004, pp. 31-36.

[24] B. Ramani et al., "A Common Attributebased Unified HTS framework for Speech Synthesis in Indian Languages," 8th ISCA Speech Synthesis Workshop (SSW8), pp. 291-296, Aug 31st - Sept 2nd 2013, Barcelona, Spain.

[25] K. Prahallad, R. Kumar, and R. Sangal, "A data-driven synthesis approach for Indian languages using syllable as basic unit", in Proceedings of International Conference on Natural Language Processing (ICON), Mumbai, India, 2002, pp. 311-316.

[26] K. Prahallad and A. W. Black, "Segmentation of monologues in audio books for building synthetic voices", IEEE Transactions on Audio, Speech and Language Processing, 2010.

[27] L. Prahallad, R. Mamidi, and K. Prahallad, "A template matching approach for detecting pronunciation mismatch", in Proceedings of Workshop on Speech and Language Processing Tools in Education (SLP-TED) in 24th International Conference on Computational Linguistics (COLING), Mumbai, India, 2012, pp. 79-84.