A Probabilistic Approach to Unit Selection for Corpus-based Speech Synthesis

Shinsuke Sakai and Han Shu

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{sakai,hshu}@csail.mit.edu

Abstract

In this paper, we present a novel statistical approach to corpus-based speech synthesis. Unit selection is directed by probabilistic models for F_0 contour, duration, and spectral characteristics of the synthesis units. The F_0 targets for units are modeled by statistical additive models, and duration targets are modeled by regression trees. Spectral targets for a unit is modeled by Gaussian mixtures on MFCC-based features. Goodness of concatenation of two units is modeled by conditional Gaussian models on MFCC-based features. Although the system is in its early stage of development, we implemented an English speech synthesizer with CMU Arctic corpora and confirmed the effectiveness of this new framework.

1. Introduction

Corpus-based concatenative approach to speech synthesis has been widely explored in the research community in recent years [1, 2, 3]. In this approach, a best sequence of phone or subphone-sized units are chosen from a large inventory of possible units to synthesize input text, by minimizing the overall cost function. The overall cost is often modeled as the weighted sum of target costs and concatenation costs on the various features such as spectral, intonational and duration features.

In the MIT Envoice system [4], an information-theoretic approach was proposed, in which Kullback-Leibler divergence that captures a distance between unit classes was utilized for defining the substitution cost. The concatenation costs were defined based on a mutual information metric representing the dependency between unit classes across the concatenation boundary. In the new corpus-based speech synthesis framework that we present in this paper, we go further and propose a probabilistic approach to unit selection in concatenative speech synthesis. We are pursuing this approach in the hope that a probabilistic approach will make it easy to establish a method that is mathematically manageable, needs fewer tuning parameters, and is easy to train, by taking advantage of statistical properties emerging from the data. It can be regarded as a more constrained sublcass within the larger class of general cost-based approach.

In the following section, we introduce our probabilistic framework for unit selection. It is followed by the descriptions of the target and concatenation models in our probabilistic approach. We then briefly describe the unit search mechanism after that. We then describe the way we generate the target phone sequence from input using a wFST. We finally describe the implementation with CMU Arctic corpora [5] for Blizzard Challenge evaluation, followed by discussions.

2. Probabilistic approach to unit selection

In a speech synthesis framework where units are selected from the corpus, we are given some input specification such as specifications for phone-sized or even finer subphone units, $s = s_1, ..., s_N$. The major work of the synthesizer is to find a best sequence of units $u = u_1, ..., u_N$ for this input specification. A specification for a unit s_i can be a collection of target features, $s_i = (f_i(1), ...f_i(p))$. These features may include such things as a phone label, a duration target, and an F_0 target for the i-th unit.

In a probabilistic framework, we would like to find a best sequence of units that maximizes the probability, $P(u|s)$, i.e.

$$u^* = \arg\max_u P(u|s) \tag{1}$$

In general, the probability of generating a unit u_i can be dependent on the specification s (hopefully a small neighborhood of s_i), and the units preceding u_i,

$$P(u|s) = \prod_{i=1}^{N} P(u_i|u_1, ..., u_{i-1}, s, i). \tag{2}$$

If we assume that the choice of unit is dependent only on one unit before that, it reduces to the simpler form,

$$P(u|s) = \prod_{i=1}^{N} P(u_i|u_{i-1}, s, i). \tag{3}$$

The conditional probability of generating a unit is assumed to be a product of the probabilities to have particular values for various feature values of the unit, such as the value of duration feature $d(u_i)$, F_0 feature $f(u_i)$, spectral feature $o(u_i)$, near-boundary spectral features at the left (head) and right (tail) ends of the units $h(u_i)$ and $t(u_i)$,

$$\begin{aligned} & P(u_i|u_{i-1}, s, i) \\ &= P(d(u_i), f(u_i), o(u_i), h(u_i)|u_{i-1}, s, i) \\ &= P(d(u_i)|s_i)\ P(f(u_i)|s_i)\ P(o(u_i)|s, i) \\ &\quad P(h(u_i)|t(u_{i-1}), s, i). \end{aligned} \tag{4}$$

The conditional probability $P(h(u_i)|t(u_{i-1}), s, i)$ of having a left boundary feature after a right boundary feature of the previous unit corresponds to what is often referred to as *concatenation cost* in the context of corpus-based speech synthesis. The rest of the component probabilities corresponds to so-called *target costs* or *substitution costs*.

3. Spectral target models

The purpose of the spectral target model is to measure the appropriateness of the spectral shape of the unit for the phone context specified by the input. We model the spectral target through m mean spectral features for each unit,

$$\begin{aligned} P(o(u_i)|s) &= P(o_{i,1}, ..., o_{i,m}|s) \\ &= P(o_{i,1}|s) \cdots P(o_{i,m}|s). \end{aligned} \quad (5)$$

In the current implementation adopting phone-sized units, m is set to be 2. Therefore, spectral target models accounts for the average spectral shape of the first half and the second half of the unit. The probability of each part is assumed to be conditioned on the triphone context:

$$P(o_{i,j}|s) = P(o_{i,j}|l_i, c_i, r_i), \quad j = 1 \dots m, \quad (6)$$

where l_i,c_i, and r_i represents left phone, center phone, and right phone for the unit u_i. Each of these densities are to be tied by phonetic decision-tree based clustering for robust estimation and to handle unseen contexts in the runtime. It could also be worth considering a quinphone context. In the current implementation, we use 14 MFCC coefficients, with dimensionality reduced to 8 by principal component analysis.

4. Duration target models

The duration models characterize tendencies of phone durations based on the surrounding phonological, lexical, and phrasal context. A duration model for each phone class is represented as a scalar Gaussian model and it is clustered using a regression tree. The features used for tree building are the number of syllables in word, the position of the syllable containing the unit in word, the position of the syllable containing the unit in intonational phrase, lexical stress of the syllable, pitch accent of the syllable, function word identity if the unit occurs in a function word, phone position in syllable, and the left and right phone identities.

5. F_0 target models

The F_0 model is based on a three-layered statistical additive F_0 model [6, 7, 8]. The first layer is an intonational phrase-level component determined by the intonational phrase type and its syllable length. The second layer is the word-level component identified by the lexical stress positions and the number of syllables in the word. The third layer accounts for the effect of pitch accent at the syllable granularity. The output from the additive F_0 model is the sum of these three layers and a constant and gives a prediction of the F_0 contour. We regard this predicted contour as the mean of a constant variance Gaussian model. The variance is computed based on the overall error of the model against the original F_0 data in the corpus during training. Although we currently assume a constant variance, it would be interesting to consider a way to estimate different variances for subclasses of intonational phrases or accentual phrases in some way from the training data.

6. Spectral concatenation models

The likelihood of the occurrence of the spectral shape of a unit after another unit is given by the spectral concatenation models. Here we make an assumption that this is best done by using spectral features at the both ends of the unit, namely the initial portion (or *head*) $h(u_i)$ and the portion at the end (or *tail*) $t(u_i)$ of the unit u_i. In the current implementation, head and tail are average spectrum of the 10ms interval at the both end of the unit. The concatenation probability is modeled as a linear conditional Gaussian density of observing the head of a unit given the tail of the preceding unit,

$$P(h(u_i)|t(u_{i-1})) = \mathcal{N}(h(u_i)|B_s\, t(u_{i-1}) + b_s,\ \Sigma_s), \quad (7)$$

where $h(u_i)$ and $t(u_{i-1})$ are d-dimensional vectors, B_s is a $d \times d$ matrix with the j-th row representing a regression coefficients for the j-th component of $h(u_i)$, and b_s is a d-dimensional vector of intercepts, and Σ_s is a $d \times d$ covariance matrix. B_s, b_s, and Σ_s are determined by the diphone context, i.e. a phone symbol pair (p_{i-1}, p_i), for the units u_{i-1} and u_i.

The maximum likelihood (ML) estimation of the parameters, B and b for a training data $\mathcal{D} = \{(t_1, h_1), ..., (t_N, h_N)\}$, where (t_i, h_i) is a pair of tail and head meeting at a boundary of a pair of consecutive units in the corpus, are given by solving a linear regression problem. By defining a $d \times (d+1)$ matrix A and a $(d+1)$-vector s_i such that,

$$A = \left[\ b_s \ \middle|\ B_s\ \right], \quad and \quad s_i = \left[\begin{array}{c} 1 \\ t_i \end{array}\right], \quad (8)$$

we see a relationship $Bt_i + b = As_i$, and we can obtain the estimates of B and b from the estimate of A. By differentiating the log likelihood of the training data with respect to A and setting it to zero, we obtain the ML estimate of A,

$$\hat{A} = (\sum_{i=1}^{N} h_i s_i^T)(\sum_{i=1}^{N} s_i s_i^T)^{-1}. \quad (9)$$

The covariance matrix Σ_s will be estimated from the sample covariances around the means given by $\hat{B}_s$ and $\hat{b}_s$.

These densities are also to be tied by a decision-tree clustering for robust training and handling unseen contexts.

7. Unit search

The unit database is organized in the shape of decision trees. We utilize the phonetic decision trees constructed in the training of spectral target models for this purpose. A set of units are associated with each node of a tree, in which the nodes closer to the root represent broader classes of units and the nodes closer to leaves represent more specific classes of units. In the synthesis time, we walk down each of m trees from the root to the most specific node with enough number of units associated with it. This is controlled by the prespecified threshold value for the minimal number of units for a node. The union of the sets of units coming from m trees makes the whole candidate unit set for a phone target.

The runtime search module performs a Viterbi beam search through the space formed as a sequence of sets of units preselected from the trees mentioned above for the best sequence of units for the input.

8. Output rendering

To achieve a smooth sound quality around concatenation points, unit concatenation is done using a simple overlap-and-add smoothing technique which is a simplified version of a technique previously proposed for error concealment of packet-based speech transmission through the Internet [9].

9. Phone sequence generation

A set of phonological rules is used to model allophinic variations. For example, one rule for flappable t is expressed by:

$$\{\text{VOWEL}\}\ t\ \{\text{VOWEL}\} \Rightarrow dx \mid tcl\ t.$$

This rule only applies to intervocalic ts. In this case, $/t/$ can be mapped to a flap $[dx]$ or t closure, $[tcl]$, followed by a t release. We currently use about 120 hand-crafted phonological rules that map 60 phonemic input symbols to 55 phonetic output symbols.

The set of phonological rules can be represented by a weighted finite-state transducer (wFST) [10]. The weights associated with the example rule would model how often an intervocalic t is flapped. The weights in the wFST were trained using a generic wFST EM-training algorithm [11] with reference word sequences and forced aligned phone sequences from a training corpus.

At runtime, with the learned weights on the wFST, we compose it with an input phoneme sequence to obtain the corresponding set of weighted sequences of phones. The single best phone sequence is the result of Viterbi search on the set.

There can be another approach which treat this whole network (from word sequences to phone sequences) as a probabilistic version of the target specification, but for simplicity reason, we did not explore it at this time.

10. Voice development with Arctic corpora

We developed two voices for the new speech synthesizer using Arctic SLT (female) and BDL (male) corpora, each of which consists of some eleven hundred utterances.

10.1. Corpus transcription

The corpus was transcribed at the phonetic level with possible different allophonic pronunciations derived from applying phonological rules [12] to phonemic baseform dictionary. The baseform dictionary also has multiple pronunciations for some words in the vocabulary. For example, the preposition "to" has two pronunciations, /t uw/ and /t ax/. We also generated transcriptions at word and syllable levels for use in the training of prosodic models.

We bootstrapped the transcription process using a speaker-independent acoustic models trained on lecture data [13]. We then adapted acoustic models to the corpus speaker and transcribed the whole corpus again using the adapted models.

10.2. Prosodic annotation

When we first investigated the effectiveness of statistical additive F_0 models on English intonation modeling [7], we made use of Boston University Radio News Corpus [14], in which prosodic labels such as break indices, boundary tones, and pitch accent markers are assigned by hand. In the current implementation using Arctic corpora, which do not (yet) have hand-labeled prosodic annotations, we used the ToBI labels for a reduced set of boundary tone and pitch accent types available from Festival "Utterance" structure which is apparently generated automatically using the Festival system [15]. We utilized those labels as they were to generate intonational phrase labels and pitch accent labels without checking or correcting by hand, due to the limited resource and time.

10.3. Training of target and concatenation models

The three layer additive F_0 models were trained using the syllable, word, and intonational phrase labels generated from the corpora in the automatic way, as described in the previos subsection, with no hand corrections. The duration models, which are regression trees modeling phone durations, were trained using the phone, syllable, word, pitch accent, and intonational phrase labels, again with no hand corrections.

Spectral target models and concatenation models for phone-sized units were trained using the phone labels mentioned above.

10.4. Construction of synthesis unit databases

A waveform unit database populated with phone-sized units was constructed using the whole waveform data of each corpus. Other kinds of information such as F_0 fragments, mean spectral features and edge spectral features for phone-sized units were also stored associated with units. A tree-shaped access mechanism of the unit database was constructed using the phonetic decision tree constructed during the training of spectral target models.

10.5. Pronunciation training

To reflect possible phonological differences among speakers, two seperate wFSTs representing the phonological rules were trained, one using the Arctic SLT corpus and another using the Arctic BDL corpus.

11. A text-to-speech prototype for Blizzard Challenge

To perform a whole text-to-speech conversion process, we needed a front-end, or a text analysis module that places phrase boundaries and pitch accents as well as choosing a proper reading based on the grammatical and discourse knowledge when needed. Since we do not have our own front-end module yet, we chose to use the front-end module in the Festival system [15] and developed an interface module that takes the Festival "Utterance" structure and convert it to the format for input to our synthesizer.

The whole computation time from text input to waveform generation is roughly 30 times the length of the output waveform file. This rather slow speed is due to various factors including the rapid prototyping making use of script language and file interface, loading of large database files every time synthesizer command is invoked, use of full covariance Gaussians in concatenation models, and rather loose beam width (2000 candidates survive in every pruning). Therefore we are optimistic about the future improvement in processing speed.

Figure 1 shows a fragment of the spectrogram of synthesized speech from the input, *"I would like to fly from Boston to San Francisco."* In the figure, we can see a smooth trajectories of formants in splicing points such as $[ay]$ to $[w]$, $[w]$ to $[uh]$, $[l]$ to $[ay]$, despite the fact that many of them are taken from a different context of surrounding phones. For example, the $[w]$ unit chosen before $[uh]$ was originally followed by $[ah]$ in the corpus, and the first $[l]$ unit was surrounded by $[p]$ and $[eh]$ in the corpus, although it is used between $[f]$ and $[ay]$.

12. Discussion

Although we have not analyzed the evaluation results enough, we saw a general trend that it was more difficult to have a good synthesis quality for BDL (male) corpus than SLT (female) corpus. Mean opinion scores were better with conversation and novel sentences as compared to news sentences. From a informal listening of the synthesized speech, we feel it is important

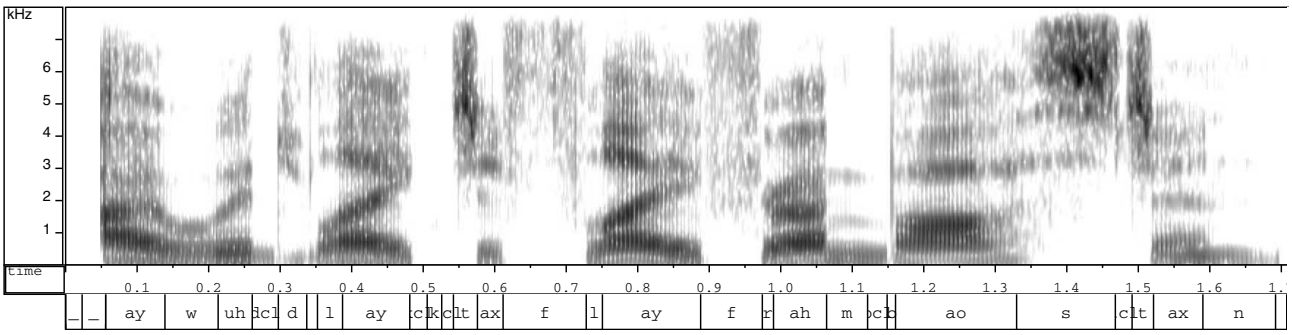

Figure 1: Initial part of the synthesized speech for the input text, "I would like to fly from Boston to San Francisco."

to look at the consistency of the boundaries of the units for the same phone symbol across various different surrounding phone contexts, which was determined by the forced transcription.

Since our current framework does not involve prosodic modification of the units by signal processing, we feel it would work best with a rather large-sized corpora, e.g. 7-10 hours of speech. It would be necessary to consider prosodic modification of units if we were to decide to optimize our approach with the corpus of the size of one hour or so. The current unit granularity is at the phone level and it may not be fine enough to control the smooth movement and splicing of the spectral and prosodic features. We are interested in using finer units such as half phones in the future improvements.

13. Conclusion

In this paper, we proposed a probabilistic approach to unit selection in concatenative speech synthesis, where all the "costs" are formulated in a probabilistic framework. We also proposed a novel probabilistic modeling scheme to account for the goodness of concatenation based on the conditional Gaussian models. The system is still in its infancy and we plan to improve on various aspects of the system.

14. Acknowledgments

The authors would like to thank T. J. Hazen for developing phonological rules and syllable-based recognizers for transcription of the corpora, Lee Hetherington for his help in implementing the pronunciation module, and Jim Glass for the helpful advice for this research. We are also grateful to the people in the Spoken Language System group at MIT and other people who participated in the evaluation. Finally, we would like to thank the organizers of the Blizzard Challenge for their patience with our rather late participation and development. This research was supported in part by the SLS Affiliate Program.

15. References

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP '96*, 1996, pp. 373–376.

[2] E. Eide et al., "Recent improvements to the ibm trainable speech synthesis system," in *Proc. ICASSP 2003*, 2003, pp. I–708–I–711.

[3] M. Chu, H. Peng, Y. Zhao, Z. Niu, and E. Chang, "Microsoft Mulan – a bilingual TTS system," in *Proc. ICASSP 2003*, 2003, pp. I–264–I–267.

[4] J. Yi and J. Glass, "Information-theoretic criteria for unit selection synthesis," in *Proc. ICSLP 2002*, Denver, 2002, pp. 2617–2620.

[5] J. Kominek and A. Black, "The cmu arctic speech databases for speech synthesis research," Tech. Rep. Tech. Rep. CMULTI-03-177, Language Technologies Institute, CMU, 2003.

[6] S. Sakai and J. Glass, "Fundamental frequency modeling for corpus-based speech synthesis based on a statistical learning technique," in *Proc. ASRU 2003*, 2003, pp. 712–717.

[7] S. Sakai, "Additive modeling of english f0 contour for speech synthesis," in *Proc. ICASSP 2005*, Philadelphia, PA, Mar. 2005, pp. I–277–I–280.

[8] S. Sakai, "Fundamental frequency modeling for speech synthesis based on a statistical learning technique," *IEICE Transactions on Information and Systems*, vol. E88-D, no. 3, pp. 489–495, 2005.

[9] A. Stenger, K. Ben Younes, R. Reng, and B. Girod, "A new error concealment technique for audio transmission with packet loss," in *Proc. EUSIPCO 96*, Trieste, Italy, Sept. 1996, pp. 1965–1968.

[10] I. Lee Hetherington, "An efficient implementation of phonological rules using finite-state transducers," in *Proc. Eurospeech 2001*, Aalborg, Sept. 2001, pp. 1599–1602.

[11] H. Shu and I. Lee Hetherington, "Em training of finite-state transducers and its application to pronunciation modeling," in *Proc. ICSLP 2002*, Denver, Colorado, Sept. 2002, pp. 1293–1296.

[12] T. J. Hazen, I. Lee Hetherington, H. Shu, and K. Livescu, "Pronunciation modeling using a finite-state transducer representation," in *Proc. of the ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation*, Estes Park, Colorado, Sept. 2002, pp. 99–104.

[13] Alex Park, Timothy J. Hazen, and James R. Glass, "Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling," in *Proc. ICASSP'05*, Philadelphia, Mar. 2005, pp. I–497–I–500.

[14] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," Tech. Rep. ECS-95-001, Boston University, Mar. 1995.

[15] A. Black and K. Lenzo, "Building voices in the festival speech synthesis system," `http://festvox.org/bsv`, 2000.