Building Probabilistic Corpus-based Speech Synthesis Systems from the Blizzard Challenge 2006 Speech Databases

Shinsuke Sakai

Academic Center for Computing and Media Studies
Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
sakai@ar.media.kyoto-u.ac.jp

Abstract

In this paper, we describe the development of probabilistic corpus-based concatenative speech synthesis systems with the Blizzard Challenge 2006 speech databases. In the current probabilistic approach, 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. Two kinds of voices for speech synthesis were developed using 1,052 and 4,273 utterances of the Blizzard Challenge 2006 speech databases.

1. Introduction

Corpus-based concatenative approaches 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 speech from the 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, as well as more symbolic features to prefer the database units that occurred in contexts similar to the context in the output sentence.

In our corpus-based speech synthesis framework [4], we adopt a probabilistic approach to unit selection for 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 subclass within the larger class of general cost-based approach.

In the following section, we review 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 finally describe the voice-building process with American English speech databases provided from ATR for Blizzard Challenge 2006, 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, $s = s_1, ..., s_N$. A major job 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 is 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 attempt to find the 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 input 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 on the right side of (3) is assumed to be decomposable into a product of the probabilities specific to various features such as the duration fea-

ture $d(u_i)$, F_0 feature $f(u_i)$, spectral feature $o(u_i)$, near-boundary spectral features at the left and right side of the unit-concatenation boundary, $t(u_{i-1})$ and $h(u_i)$,

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

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*.

2.1. 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. The spectral target of a unit is represented by mean spectral features of m evenly divided regions of the unit. The overall spectral target probability is the product of the probabilities associated with m regions,

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

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:

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

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. In the current implementation, we use 14 MFCC coefficients, with dimensionality reduced to 8 by principal component analysis.

2.2. Duration target models

The duration models characterize tendencies of phone duration lengths 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.

2.3. F_0 target models

The F_0 model is based on a three-layered statistical additive F_0 model [5, 6, 7]. 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.

2.4. Spectral concatenation models

The likeliness of the occurrence of the spectral shape of a unit after another unit is given by the spectral concatenation models. We currently assume that it is good enough to look at the regions near the concatenation boundary of the units being connected. The region near the end (or *tail*) of the unit u, on the left side of the concatenation boundary, is denoted by $h(u)$. The initial region (or *head*) of the unit v, on the right side of the concatenation boundary, is denoted by $h(v)$. In the current implementation, head and tail are averages of the MFCC-derived spectral features of the 10ms intervals at the both ends 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,

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

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 model parameters B_s,b_s, and Σ_s are trained with a maximum likelihood estimation from the training data using a decision tree-based parameter tying [8].

2.5. 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. As mentioned before, $m = 2$ in the current implementation.

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.

2.6. 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].

3. Voice development with Arctic corpora

We developed two voices from the speech corpus for the Blizzard Challenge 2006 provided by ATR, spoken by a male speaker of American English. The sampling frequency is 16,000 Hz. We developed a voice using the 4,264 out of 4,273 utterances consisting of 1,052 news utterances, 2,130 travel conversation utterances, and 1,082 Arctic utterances. We discarded four travel utterances and five news utterances that included sounds not easy to transcribe in our system due to foreign words, filler words, and dysfluent pronunciation of the names of unfamiliar chemical substances. No utterances were discarded from the Arctic subset. We refer to this corpus as well as the voice built with it as `atr`, hereafter. We also developed a voice with the subset of 1,082 Arctic utterances, which we call `atr-arctic`.

3.1. Corpus transcription

For the development using the `atr` corpus, we needed to add a little more than 400 new words. We noted that there were many Japanese words for which a systematic and consistent assingment of pronunciations using English phone set may not be easy. Furthermore, some instances of those words sounded like influenced by the original pronunciations of Japanese language. Due to the limited time for development, we decided not to try to get rid of undesirable instances of words by listening and just used all of the 4,264 utterances.

The corpus was transcribed at the phonetic level with possible different allophonic variations derived from applying phonological rules [10] to phonemic baseform dictionary, using acoustic models adapted to the corpus speaker, as described in [4].

3.2. Prosodic annotation

We utilize prosodic information such as boundary tones and pitch accent types as well as syllable and word labels for training F_0 models and duration models. Since assigning these labels by hand is expensive and time-consuming, we generated Festival "Utterance"' structures [11] and extracted labels that bear boundary tone and pitch accent information from them. Since the labels were generated top-down from the prompt texts using a Festival command, without any reference to the speech data, they are not guaranteed to be the same as the way the prompts were actually spoken. We are required to perform additional process of matching syllable labels that bear the prosodic information coming from Festival Utterance to the syllable labels in our own framework. Since the way of grouping of phones to syllables as well as the set of phones themselves are different between Festival and ours, this matching was not a trivial task. Due to the limited amount of time, we adopted a rough approximation scheme in which we linearly warped the label times to obtain the match between Festival Utterance labels and the labels in the phonetic transcriptions of the `atr` corpus.

3.3. Training of target and concatenation models

The three layer additive F_0 models were trained using the syllable and word labels as well as intonational phrase and pitch accent labels generated using the method described in the last subsection. Pitch trains used for training the models were extracted every 10ms from the corpus using the Snack Sound Toolkit with "esps" method [12]. The duration models were trained using the phone, syllable, word, pitch accent, and intonational phrase labels. No hand corrections were performed on the labels for training these prosodic models. Spectral target and concatenation models for phone-sized units were trained using the phone labels mentioned above.

3.4. 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.

4. Speech Synthesis from test sentences

To perform a whole text-to-speech conversion process, we need 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. When we joined the Blizzard Challenge 2005 evaluation, we did not have our own front-end module yet. Therefore, we chose to use the front-end module in the Festival system [11] and developed an interface module that takes the Festival "Utterance" structure and convert it into the format for input to our speech synthesis system. We used the same interface in the Blizzard Challenge 2006.

5. Discussion

The mean opinion scores for the both of `atr` and `atr-arctic` were rather poor (e.g. 2.00 and 1.84, respectively for the listener category S) and approximately four times larger corpus size for `atr` did not help much to make a difference. These scores are comparable to our previous MOS score for the speaker `bdl` (1.80) which was considerably worth than that for `slt` (2.49) for the same listener category. We are aware that perceived intonation contour is often strange and some phones have irrelevant durations. We have recently noticed that the labels in the Festival "Utterance" structures provided with `slt` database, which we used for training our prosodic models, are time-aligned to the waveforms, but it was not the case for other databases. As we described in a previous section, we applied a linear warping of label times to match the labels extracted from Festival "Utterances" against the labels obtained by transcribing the data. This may have generated poor-quality training labels for F0 and duration models. In fact, whereas correlation coefficient of the model-generated F_0 curve to the training data $F_0$was 0.467 for `slt`, that for `atr-arctic` was 0.379, which is considerably worth. We would like to develop a better matching tool for labels and see if it contributes to improve the synthesis quality.

For the Blizzard Challenge 06, we have been developing a prosody modification module based on TD-PSOLA. Unfortunately, we did not have time to incorporate the module to synthesizer. Therefore, the system is basically the same as the one we used for the Blizzard Challenge 05. We plan to integrate the prosody modification module to our synthesizer in a couple of months and see it contributes to better synthesis quality.

6. Conclusion

In this paper, we described the development of a probabilistic concatenative speech synthesizer with Blizzard Challenge 06 speech databases. Although the result was not satisfactory, we believe the experience of the evaluation this time will help us make a progress in our speech synthesis research.

7. 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] S. Sakai and H. Shu, "A probabilistic approach to unit selection for corpus-based speech synthesis," in *Proc. Interspeech 2005*, Lisbon, Portugal, Sept. 2005, pp. 81–84.

[5] 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.

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

[7] 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.

[8] S. Sakai and T. Kawahara, "Decision tree-based training of probabilistic concatenation models for corpus-based speech synthesis.," in *Proc. Interspeech 2006*.

[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] 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.

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

[12] Kåre Sjölander, "The snack sound toolkit," `http://www.speech.kth.se/snack/`.