Decision Tree-based Training of Probabilistic Concatenation Models for Corpus-based Speech Synthesis

Shinsuke Sakai and Tatsuya Kawahara

Academic Center for Computing and Media Studies
Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

sakai@ar.media.kyoto-u.ac.jp kawahara@i.kyoto-u.ac.jp

Abstract

The measure of the goodness, or cost, of concatenating synthesis units plays an important role in concatenative speech synthesis. In this paper, we present a probabilistic approach to concatenation modeling in which the goodness of concatenation is represented as the conditional probability of observing the spectral shape of a unit given the previous unit and the current phonetic context. This conditional probability is modeled by a conditional Gaussian density whose mean vector has a form of linear transform of the past spectral shape. A phonetic decision-tree based parameter tying is performed to achieve a robust training that balances between model complexity and the amount of training data available. The concatenation models are implemented in a corpus-based speech synthesizer trained with a CMU Arctic database and the effectiveness of the proposed method was confirmed by a subjective listening test.

Index Terms: speech synthesis, unit selection, join costs.

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, the best sequence of phone or subphone-sized synthesis units are chosen from a large inventory of units to synthesize speech from the input text through the minimization of the overall cost. The overall cost is often modeled as the weighted sum of target costs and concatenation (or join) costs defined on various features of synthesis units such as spectral shape, intonation contour, and segmental duration. Establishing a good model of concatenation cost is one of the most important aspects that influence the quality of concatenative speech synthesis, and there has been a number of research efforts to find a good measure of concatenation cost [4, 5, 6, 7], in which various spectral feature parameters and distance measures are investigated. There is also a research effort to find optimal mapping functions from distance measures to costs based on perceptual evaluation [8].

In our probabilistic framework for concatenative speech synthesis [9], we depart from the traditional view of cost based on "distance" and attempt to take a probabilistic view of concatenation cost where concatenation modeling is done with a probabilistic model that captures how likely it is to observe the spectral shape of the current unit given the spectral shape of the previous unit. For the modeling of this conditional probability, we make use of *conditional Gaussian* models. The mean vector of a conditional Gaussian density has a form of linear transform of some other vector, which is useful for representing the correlation between two random variables. An example of the use of conditional Gaussian in speech processing is found in *autoregressive HMMs* [10], where the observation vector from a state is conditioned not only on the identity of the current state, but also on the observation from the previous state.

In this paper, we present a roubst and efficient training method and an experimental evaluation of the probabilistic concatenation models. Section 2 gives an overview of the model. A robust and efficient training method for the models based on phonetic decision tree-based context tying is described in section 3. Experimental results are presented in section 4 where we examine how linear transforms for conditional means of the models are trained from the corpus. Subjective evaluation results are also reported. The last section presents our conclusion.

2. Probabilistic concatenation models

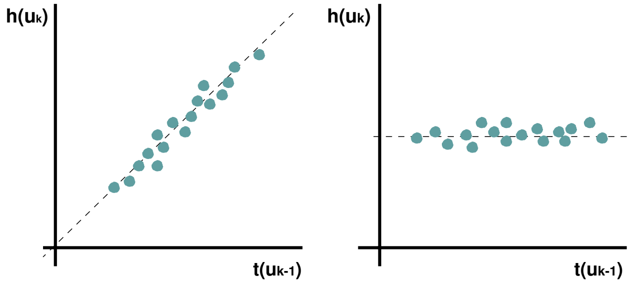
We model the goodness of concatenation of the spectral shapes of the synthesis units in terms of the conditional probability of observing the spectral shape $o(u_k)$ of the unit u_k given that of the previous unit u_{k-1} and the phonetic context c_k for the k-th unit. We currently assume that it is enough to consider the spectral shapes near the concatenation boundary, so that

$$P(o(u_k)|o(u_{k-1}), c_k) \approx P(h(u_k)|t(u_{k-1}), c_k),$$

where $h(u_k)$ represents some initial portion (or *head*) of the spectral shape of the unit u_k, and $t(u_{k-1})$ represents some portion at the end (or *tail*) of the spectral shape of the unit u_{k-1}. In the current implementation, head and tail are spectral feature vectors averaged over a 10 ms interval (two 5-ms frames) at the both end of the unit. As a spectral feature vector, we use 14 MFCC coefficients with dimensionality reduced to 8 by principal component analysis. This concatenation probability is modeled by a conditional Gaussian density,

$$P(h(u_k)|t(u_{k-1}), c_k) = \mathcal{N}(h(u_k)|B_{c_k}\, t(u_{k-1}) + b_{c_k},\ \Sigma_{c_k}), \tag{1}$$

where $h(u_k)$ and $t(u_{k-1})$ are d-dimensional vectors, B_{c_k} is a $d \times d$ regression matrix with the j-th row representing a regression coefficients for the j-th component of $h(u_k)$, b_{c_k} is a d-dimensional vector of intercepts, and Σ_{c_k} is a $d \times d$ covariance matrix. In the current implementation using phone-sized units, we adopt the phone identities of the units u_k and u_{k-1} as the context c_k that



(a)

h(uk) is very similar to t(uk-1).

(b)

h(uk) almost independent of t(uk-1).

Figure 1: Schematic diagram representing the relationship between $h(u_k)$ and $t(u_{k-1})$ in two extreme cases.

identifies the model parameters $\{B_{c_k}, b_{c_k}, \Sigma_{c_k}\}$. We drop the suffix c_k for simplicity of notation hereafter.

Fig. 1 shows conceptual graphs using hypothetical one-dimensional features that describe the relationships between the tail and the head of two consecutive units for two extreme cases. Fig. 1 (a) corresponds to a case where spectral shapes are very similar across the unit boundary, e.g. a vowel followed by the same vowel. In this kind of situation, the regression matrix B is considered to be close to identity matrix and the constant vector b is close to zero. On the other hand, if there is a case like Fig. 1 (b), where the head of the current unit is almost independent of the tail of the previous unit, the regression matrix B is considered to be close to zero matrix and b will be the significant contributor to the mean vector. In general cases in between two extremes, B and b are considered to have some meaningful values that represent u_k's characteristics that is dependent on u_{k-1} in some aspects and independent of it in some other aspects.

2.1. ML estimation of conditional Gaussian model parameters

The maximum likelihood (ML) estimate of the model parameters, B and b from the training data is derived as a solution to a simple convex optimization problem, like ML estimation of a multivariate Gaussian. The training data $\mathcal{D} = \{(t_1, h_1), ..., (t_N, h_N)\}$ for a conditional Gaussian model for a given phonetic context consists of all the pairs (t_i, h_i) of tail and head spectral feature vectors available from the corpus for that context.

By defining a $d \times (d+1)$ matrix A and a $(d+1)$-vector s_i, where d is the dimensionality of t_i and h_i, such that,

$$A = \left[\begin{array}{c|c} b & B \end{array}\right], \quad and \quad s_i = \left[\begin{array}{c} 1 \\ t_i \end{array}\right], \tag{2}$$

it holds that $B\,t_i + b = As_i$. Thus, we obtain the estimates of B and b from the estimate of A. Then the conditional Gaussian density function can be written as

$$\begin{aligned} &\mathcal{N}(h|B\,t+b,\,\Sigma) = \mathcal{N}(h|A\,s,\,\Sigma) \\ &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\{-\frac{1}{2}(h-As)^T\Sigma^{-1}(h-As)\}.(3) \end{aligned}$$

The log likelihood $\mathcal{L}$ with the training data $\mathcal{D}$ is, therefore,

$$\begin{aligned} \mathcal{L}(A,\Sigma;\mathcal{D}) &\triangleq \log\prod_{i=1}^{N}\mathcal{N}(h_i|As_i,\Sigma) \\ &= -\frac{dN}{2}\log 2\pi - \frac{N}{2}\log|\Sigma| \\ &\quad -\frac{1}{2}\sum_{i=1}^{N}(h_i - As_i)\Sigma^{-1}(h_i - As_i). \end{aligned} \tag{4}$$

Taking the partial derivative of $\mathcal{L}$ with regard to A, and utilizing the formula (see, e.g., [11]),

$$\frac{\partial\{(Xa+b)^T C(Xa+b)\}}{\partial X} = (C + C^T)(Xa+b)a^T,$$

we have

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial A} &= -\frac{1}{2}\Sigma_{i=1}^{N}\{-(\Sigma^{-1} + \Sigma^{-1T})(h_i - As_i)s_i^T\} \\ &= \Sigma^{-1}\sum_{i=1}^{N}(h_i - As_i)s_i^T. \end{aligned} \tag{5}$$

Setting the partial derivative to zero, we obtain the ML estimate of A as

$$\hat{A} = (\sum h_i s_i^T)(\sum s_i s_i^T)^{-1}. \tag{6}$$

The covariance matrix Σ can be estimated as the sample covariance around the conditional mean $\hat{A}\,s_i$, and it reduces to

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} h_i {h_i}^T - \hat{A}\frac{1}{N}\sum_{i=1}^{N} s_i {h_i}^T. \tag{7}$$

3. Robust training with decision-tree clustering

The number of the types of contexts that determines the specific conditional Gaussian (CG) model to use for measuring the goodness of concatenation can be very large and we may have rather few training data points (or, even worse, no data points at all) available from the corpus for some types of phonetic contexts. In the current implementation where the context is simply determined by the phone identities of the current unit and the preceding unit, the number of possible combination is already close to 3000. In order to achieve robust training of the conditional Gaussian concatenation models, we tie the model parameters using phonetic decision-tree clustering. Contexts for the models are clustered according to the questions about the phone symbol of the preceding units (tail phones). The process of parameter tying is performed by the following steps.

1. Initial CG model parameters are trained for all the distinct contexts, i.e. the combinations of tail and head phones available in the training data.
2. For each head phone, CG models with this same head phone are clustered using phonetic decision tree:
   (a) All the CG models with this head phone is tied and associated with the root node of the decision tree.

(b) Each terminal node of the tree is examined and recursively split into two child nodes based on the phonetic question that yields the maximum increase of the likelihood.

The node is not split if the likelihood gain is below the prespecified threshold or the number of training data points after split is smaller than the prespecified minimum number of elements in the node.

Suppose we have a subset of the (augmented) training data $\mathcal{S} = \{(s_1, h_1), ..., (s_n, h_n)\}$ associated with a node, where s_i is a $(d+1)$-dimensional augmented tail vector like in the equations (2). Let $\mathcal{L_S}$ be the log likelihood with regard to $\mathcal{S}$ of the model trained with $\mathcal{S}$ itself. Noting the relationship,

$$\sum_{i=1}^{n}(h_i - A_{\mathcal{S}}s_i)^T\Sigma_{\mathcal{S}}^{-1}(h_i - A_{\mathcal{S}}s_i) = trace(\Sigma_{\mathcal{S}}^{-1} \cdot n\,\Sigma_{\mathcal{S}}) = n \cdot d,$$

where $A_{\mathcal{S}}$ and $\Sigma_{\mathcal{S}}$ are augmented regression matrix and covariance matrix trained with $\mathcal{S}$, we can reduce $\mathcal{L_S}$ into

$$\begin{aligned} \mathcal{L_S} &= \log \prod_{i=1}^{n} \mathcal{N}(h_i | A_{\mathcal{S}} \cdot s_i, \Sigma_{\mathcal{S}}) \\ &= -\frac{n}{2}(d \log(2\pi) + \log|\Sigma_{\mathcal{S}}| + d). \end{aligned} \quad (8)$$

Therefore, we see that the log likelihood with $\mathcal{S}$ depends only on the covariance matrix $\Sigma_{\mathcal{S}}$ and the number of data points n. When $\mathcal{S}$ is divided into the subsets $\mathcal{A}$ with a data points and $\mathcal{B}$ with b $(= n - a)$ data points by a phonetic context question, the increase in the log likelihood $\mathcal{G}$ becomes

$$\begin{aligned} \mathcal{G} &= \mathcal{L_A} + \mathcal{L_B} - \mathcal{L_S} \\ &= \frac{1}{2}\{(a+b)\log|\Sigma_{\mathcal{S}}| - a\log|\Sigma_{\mathcal{A}}| - b\log|\Sigma_{\mathcal{B}}|\}. \end{aligned} \quad (9)$$

$\mathcal{G}$ can be computed efficiently utilizing the sufficient statistics $\sum_i h_i s_i^T$, $\sum_i s_i s_i^T$, $\sum_i h_i h_i^T$, and $\sum_i s_i h_i^T$ and the formulas (6) and (7). We compute these sufficient statistics for all the untied models in the stage 1 of the decision tree-based clustering process described earlier. The likelihood at any node can be computed reusing these sufficient statistics without direct reference to the training data points.

Figure 2 shows part of the decision tree grown for clustering the context for the head phone `[aa]`, obtained through the training of CG models in the experiment described in the next section.

4. Experiments

We trained the conditional Gaussian concatenation models using the speaker SLT of the CMU Arctic speech databases [12]. It is spoken by a female speaker of American English and consists of 1132 utterances. The total duration is roughly 50 minutes. The phone inventory we used consists of 53 detailed phones. For the decision tree-based clustering of the phonetic context, the likelihood gain threshold was set to 1.0 and the minimum number of data point per node was set to 17. As a result, the whole 2809 (= 53^2) combinations of the tail and head phones were clustered into 677 clusters. Figure 3 depicts three examples of the augmented regression matrices of the conditional Gaussians. From the left matrix, we see that the constant vector part b is dominant in the linear transform $Bt + b$ for the phonetic context `[s]` for `[ah]`, whereas we also note a slight diagonal pattern in the regression matrix. On the other hand, the diagonal components of the regression matrix B appears to be very dominant in the transition from `[axr]` or `[r]` to `[ax]` (Figure 3 (b)), suggesting that the spectral shape is very similar on the both sides of the boundary. In Figure 3 (c), we notice significant contributions from both of the constant vector b and the regression matrix B for the boundary of a nasal consonant (`[en]`, `[n]`, or `[ng]`) and the vowel `[ow]`.

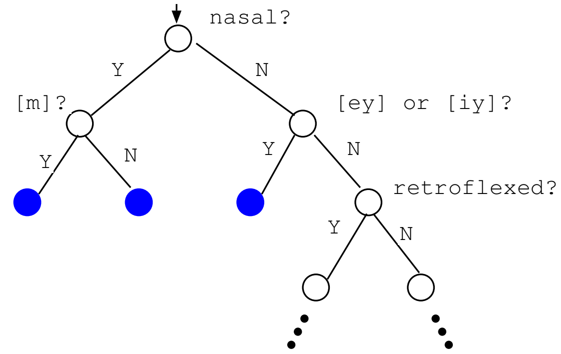


Figure 2: A phonetic decision tree for clustering the context for the head phone `[aa]`. Open circles represent nonterminal nodes and filled circles represent terminal nodes. Nodes are split by phonetic questions on the preceding preceding unit.

In order to investigate the effectiveness of the proposed approach to concatenation cost, we performed a subjective listening test, using Euclidean distance as the baseline for comparison, which has been reported to be a good predictor of perceived discontinuity when measured on Mel-cepstral feature parameters [13]. For synthesizing the utterances, we made use of the speech synthesizer reported in [9], trained also with the Arctic SLT database. In this synthesizer, the total cost C is the sum of three kinds of target costs (c_d^t for duration, c_f^t for F_0, and c_s^t for spectrum) and the spectral concatenation costs c_s^c,

$$C = \sum_{k=1}^{N}\{c_d^t(u_k) + c_f^t(u_k) + c_s^t(u_k)\} + \sum_{k=2}^{N} c_s^c(u_{k-1}, u_k), \quad (10)$$

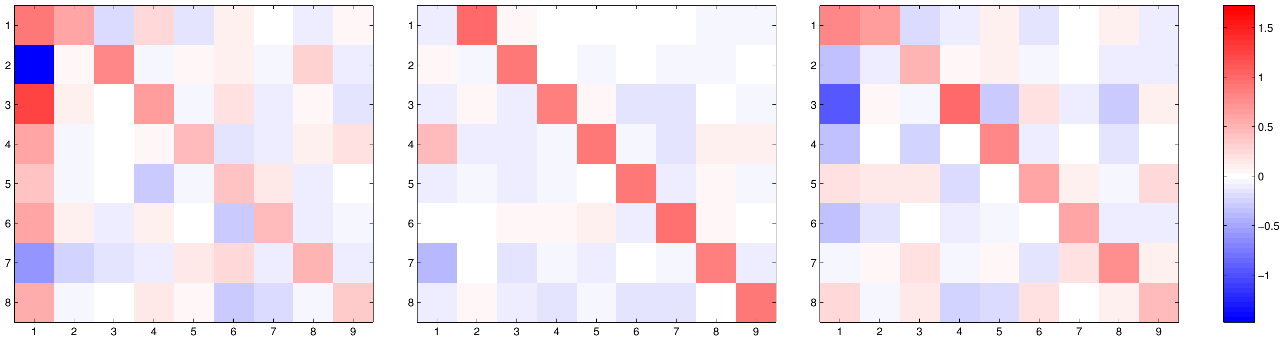
where the concatenation cost c_s^c with the proposed models is defined as

$$c_s^c(u_{k-1}, u_k) = -w \cdot logP(h(u_k)|t(u_{k-1}), c_k). \quad (11)$$

When the Euclidean distance is used, it is defined to be

$$c_s^c(u_{k-1}, u_k) = w \cdot \|h(u_k) - t(u_{k-1})\|. \quad (12)$$

In order to determine the weight for the Euclidean distance, we preliminarily synthesized ten utterances with varying values of the weight w and picked the one that yielded the best sounding synthetic speech by informal listening. A set of twenty sentences were extracted for the listening test from the sentences used for Blizzard Challenge 2005 [14]. Ten sentences were taken from "novels" part and ten other sentences were taken from the "conversation" part. Eight subjects listened to the speech synthesis output from two synthesizers, one of which adopting Euclidean distance and



(a) `s`→ `ah` (b) `axr,r`→ `ax` (c) `en,n,ng`→ `ow`

Figure 3: Graphical representations of the $8 \times (1 + 8)$ augmented regression matrices $A = [b|B]$ trained using Arctic SLT corpus for concatenation boundaries of (a) from `[s]` to `[ah]`, (b) from `[axr]/[r]` to `[ax]`, and (c) from `[en]/[ng]/[n]` to `[ow]`. Small squares represent matrix elements and the color bar on the right shows the mapping from the element's value to its color. Darker squares have larger absolute values. Red means positive and blue means negative if full color is available.

the other with the proposed conditional Gaussian (CG) models for concatenation cost. They were asked to give scores of 1 to 5 to each utterance. The results of the listening test is summarized in Table 1. The mean opinion score with the proposed method turned out to be significantly higher than the baseline at the 1% level by the paired t-test, with a p-value of 5.178×10^{-11}.

Table 1: 5-level mean opinion scores for the two synthesizers.

Euclidean	CG
2.44	2.97

5. CONCLUSION

In this paper, we presented a novel probabilistic approach to concatenation modeling using conditional Gaussian models. We described a maximum likelihood estimation formula for the models and a robust and efficient training scheme using decision-tree based context clustering. We implemented the proposed method with a CMU Arctic speech database and confirmed the effectiveness of the proposed method by a subjective listening test. In the current work, we only look at spectral features to measure the goodness of concatenation. It would further help improving the synthesized speech quality if we also consider a prosodic feature such as F_0.

6. Acknowledgments

The authors would like to thank people at MIT Computer Science and Artificial Intelligence Laboratory who participated in the listening test.

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] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 1, pp. 39–51, 2001.

[5] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. ICASSP 2001*, Salt Lake City, USA, 2001.

[6] R.E. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesisers," in *Proc. 4th ESCA Tutorial and Research Workshop on Speech Synthesis*, Scotland, Sept. 2001.

[7] J. Vepa and S. King, "Join cost for unit selection speech synthesis," in *Text to Speech Synthesis*, A. Alwan and S. Narayanan, Eds. Prentice Hall, 2004.

[8] T. Toda, H. Kawai, and M. Tsuzaki, "Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis," in *Proc. ICASSP 2004*, Montreal, Canada, May 2004, pp. 657–660.

[9] 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.

[10] J. Bilmes, "Graphical models and automatic speech recognition," in *Mathematical Foundations of Speech and Language Processing*. Springer-Verlag, 2003.

[11] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Feb. 2006, `http://2302.dk/uni/matrixcookbook.html`.

[12] 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.

[13] J. Wouters and M. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis," in *Proc. ICSLP 98*, Sydney, Australia, 1998, pp. 2747–2750.

[14] A. Black and K. Tokuda, "Blizzard challenge – 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc. Interspeech 2005*, Lisbon, Portugal, 2005, pp. 77–80.